UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Análisis Estadístico y Geométrico de los Modelos de Mezcla Locales y una propuesta de Pruebas de Bondad de Ajuste para datos censurados

PARA OBTENER EL TÍTULO DE:

DOCTOR EN CIENCIAS (MATEMÁTICAS)

Karim Alejandro Anaya Izquierdo

Director de Tesis:
Dr. Federico Jorge O'Reilly Togno

Universidad Nacional Autónoma de México
Posgrado en Ciencias Matemáticas
Facultad de Ciencias

México, D.F. Junio, 2006

Agradecimientos

A la UNAM a la cual debo mi completa formación académica, gracias a la promoción del desarrollo científico y humano.

A mi director de Tesis Dr. Federico O'Reilly Togno por su incondicional apoyo tanto académico como el de un gran amigo.

I am indebted to Paul Marriott who believed in my ideas and kindly supported me during my academic visits to where he was.

A mis otros sinodales Dra. Lilia del Riego, Dr. Raul Rueda, Dr. Eduardo Gutierrez Peña, Dr. Eugenio Garnica y Professor Frank Critchley por la cuidadosa revisión de la tesis.

A mis amigos y maestros del departamento de Probabilidad y Estadística del IIMAS: Silvia Ruiz-Velasco, Leticia Gracia, Patricia Romero, Rebeca Aguirre, Eduardo Gutierrez Peña, Raul Rueda, Alberto Contreras, Ramsés Mena, Jose María Gonzalez Barrios, Ignacio Mendez, Carlos Diaz, Juan Gonzalez, Mogens Bladt y por supuesto a Elida Estrada y Hernando Ortega por su apoyo y amistad.

A mis amigos y maestros del ITAM: Manuel Mendoza, Rubén Hernández y Victor Aguirre por haberme inducido en el camino de la Estadística.

A la coordinadora del Posgrado en Matemáticas Dra. Begoña Fernández por su apoyo.

A las secretarias del Posgrado Laurita, Socorro Alexia y Tere por soportar toda la carga que implica que un estudiante se titule desde el extranjero.

A Consejo Nacional de Ciencia y Tecnlogía y a la Dirección General de Estudios de Posgrado por el apoyo económico.

Al programa de Movilidad Internacional de Estudiantes de Posgrado de la UNAM por el apoyo económico.

A mis compañeros y amigos estudiantes en el IIMAS: Edilberto, Gonzalo, Salvador, Hugo, Arturo, Humberto, Carlos Ernesto y Yesica, Jesica, Juan, Osvaldo, Yazmin, Jaime, Monica y Jose Luis por su apoyo y amistad.

To my new friends now at the the Open University: Frank Critchley, Chris Jones, Kevin McConway, Paddy Farrington, Paul Garthwaite, Alvaro Faria, Nickolay Trendafilov, Heather Whitaker, Karen Vines and Casper Albers for their advice.

Last but not least I also want to express my gratitude to the friends I made during my visit to the ISDS, Duke University in North Carolina. My warmest thanks to Max and Kelly, Michele, Ananda, Fabio, Anna, Natesh, Margaret, Dawn, Abel and Gavino. To them all I have to say is: WE MADE IT!!!!

Y también gracias a toda la gente que de alguna u otra forma me ayudó y por desgracia (o más bien mala memoria) omití.

Dedicatorias

A mi amada, cariñosa, paciente, incansable ... esposita Hadid.

A mi querida madre Lourdes y a mi abuelita Juanita que me cuida desde donte está.

A mi querido padre Jorge.

A mis hermanos Juan Pablo, Marcelo, Yazmin, José y Jorge.

A mis queridas sobrinas Fernanda, Ximena y la recientemente llegada Natalia.

A mis amigos del alma: Jose Antonio (Compadrito), Nashyeli (Comadrita), Luis Adolfo, Raúl y Manolo.

A mi ahijado Emiliano que espero orientar de la mejor manera.

A mis amigos del alma: Miguel, Toño, Alberto, Leonor y Mitsuo.

A mis queridos amigos de la Santa: Alexis, Jorge, José, Daniel y Omar.

A mis queridos amigos del ITAM: Mauricio, Rita, Lizette y Mauricio.

A mis queridos suegros (en serio !!!) Juana Cruz, Roberto Vera y mi cuñado Roberto.

Statistical and Geometrical Analysis of Local Mixture Models and a proposal of some new Tests of Fit with censored data

Contents

Resumen

Esta tesis está dividida en dos apartados, debido a que se estudiaron dos temas diferentes.

En la parte 1, se estudió el comportamiento local de los Modelos de Mezclas. Primero, haciendo algunas analogías con la geometría de las Familias Exponenciales Naturales, se desarrolló una teoría acerca de la geometría de los modelos de mezcla donde los Espacios Afín jugaron un papel primordial.

Después, se derivaron expansiones asintóticas que revelan la naturaleza afín de los modelos de mezcla cuando se asume el supuesto de que la distribución de mezcla es un modelo de dispersión propio. La clase de los modelos de mezclas locales se redefinieron de forma ligeramente distinta a la literatura, en la nueva definición se hace explicita la dependencia de los parámetros de mezcla con respecto a las distribuciones base.

La motivación principal para la construcción de estos modelos, es que poseen propiedades geométricas simples que son explotadas para llevar a cabo inferencias estadísticas. Se analizaron las propiedades estadísticas y geométricas de estos modelos y los resultados fueron aplicados para estudiar mezclas de la distribución exponencial negativa.

La conclusión principal es que los modelos de mezclas locales constituyen un conjunto de modelos paramétricos que son flexibles, identificables e interpretables; además, generalizan a las familias exponenciales naturales con función de varianza cuadrática, mediante la introducción de parámetros adicionales diseñados para capturar información acerca de una posible estructura de mezcla. La estructura de fronteras, hace de estos modelos una clase innovadora con respecto a la literatura sobre modelos de mezclas.

En la parte 2, se estudió el problema de bondad de ajuste en las distribuciones Gaussiana Inversa y Gamma cuando existe censura y los parámetros son desconocidos. Se derivaron estadísticas de prueba del tipo Cramér-von Mises para verificar el ajuste de dichas distribuciones. Se estudiaron las distribuciones asintóticas asociadas y se obtuvieron fórmulas para evaluar las funciones de covarianza de los procesos empíricos subyacentes. Algunos porcentiles asintóticos fueron tabulados para mostrar las conexiones existentes con otros casos ya estudiados, como la distribución normal, exponencial negativa y de Levy.

Se sugirió un procedimiento basado en valores de significancia, en lugar de la construcción y uso de tablas. Se llevó a cabo un estudio de Monte Carlo para mostrar las propiedades de este procedimiento con muestras pequeñas.

Abstract

This thesis is divided in two parts as two different problems have been studied.

In part 1, we studied the local behavior of mixture models. First, by making some analogues with the well known geometry of Natural Exponential Families, we developed a comprehensive theory of the geometry of Mixture models where Affine Spaces played a key role. After that, we derived a set of asymptotic expansions which revealed the affine nature that mixture models have, under the assumption that the mixing distribution follows a proper dispersion model.

The class of Local Mixture Models is then defined in a slightly different way as was previously done in the literature. The new definition makes explicit the dependence of the mixing parameters with respect to the baseline distribution family. The main motivation for the construction of these models is that they have simple geometric properties, which are exploited to make statistical inference. The statistical and geometrical properties of these models were analyzed, and the results applied to study mixtures of the negative exponential distribution.

The main conclusion is that local mixture models constitute a set of flexible, identifiable and interpretable parametric statistical models which generalize Natural Exponential Families with Quadratic Variance Function by adding extra parameters, which are intended to capture possible mixing structure. The boundary structure makes these models an innovative class with respect to the literature in mixture models.

In part 2, we studied the problem of testing the fit of the Inverse Gaussian and Gamma distributions under censoring and when the parameters are unknown. We derived tests statistics of the Cramer-von Mises' type for testing the fit of such distributions. The theory for the asymptotic distributions of the test statistics was studied and formulae were obtained to evaluate the covariance functions of the underlying empirical processes. Some asymptotic percentiles are tabulated to show connections with other known cases such as the normal, the negative exponential and the Levy distribution.

A procedure was suggested to compute p-values instead of constructing and using tables. A small Monte Carlo study is carried out to show the small properties of this procedure.

Preface

I only realized the hard work required to complete a PhD degree until I actually did it. I spent almost five years trying to solve my own scientific enquiries until I finally discovered that research is all about finding partial answers and that new questions arise all the time. This thesis is just the beginning of a hopefully long scientific journey. Its structure, divided in two parts, appears to be quite unusual and therefore I will start giving a brief explanation of such structure.

During the first year of my PhD, I worked as Research Assistant of my supervisor Dr. Federico O'Reilly. Two publications emerged from that period of work. The first, was merely completing a note by Dr. O'Reilly on some simple asymptotic based inferences under censoring for location-scale families. During that time, we came across with the problem of testing the fit of parametric families under censoring. Soon after, I decided to start working on some simple new procedures for such purpose in the special case of the Gamma and Inverse Gaussian distributions. The final result is **?**) and appear as Part II of this thesis.

While writing the second paper my scientific inquisitiveness started giving me trouble. Why families like the Gamma and Inverse Gaussian have so many nice properties? Digging into the immense statistical literature, I discovered the Geometrical-Statistical literature which motivates Part I of this thesis. The motivating example was a simple one. A largely used distribution in the analysis of positive data, the negative exponential distribution, when sampled with censoring becomes a Curved Exponential Family!!! If censoring means loss of information, why the Exponential Family structure is retained in some way after censoring? Why the geometrical concept of Curvature comes into play? What is the Statistical meaning of this curvature? So many questions, very few answers.

I needed a special guidance to start in an ascending direction. My first choice was asking Bradley Efron, the author of the seminal paper in statistical curvature. After receiving a very short but encouraging reply, I continued my search for guidance abroad which ended up with Paul Marriott which kindly agreed to collaborate with me in a common research area, the statistical and geometrical analysis of mixtures of positive distributions. The first approach was a 6 month visit to the Institute of Statistics and Decision Sciences, Duke University, U.S. in 2004 and then later a 1 month visit to the Department of Statistics and Actuarial Sciences, University of Waterloo, Canada in 2005. Both visits were always supported by UNAM. Part I of this thesis is the result of this collaboration as well as two forthcoming publications.

After participating in the most important conference in the area, Information Geometry and its Applications (Tokyo, December 2005) I finished to understand two main issues: first, how fertile the area is at the moment and second, how crucial is an efficient interaction between Geometers and Statisticians. Many relatively simple tools still underused by both communities. In that respect, writing this thesis was a challenge. In an almost philosophical way, I can justify the use of Information Geometry by the following argument.

We all know that Statistical Theory is solidly constructed upon the foundations of Probability using Measure Theory. Any additional mathematical structure which help in improving the understanding of any statistical issue is always welcome. Group Theory is a good example, Transformation models (such as Location-scale models) have a substantial geometrical content which has been successfully exploited in Statistics over many years. Bayesian Decision Theory is another clear example. It contains Bayesian Statistical Inference as a subset and, in fact, enriches it to better understand the behavior of people. We can continue giving examples with the only purpose of convincing ourselves independently of the statistical paradigm we follow.

Karim Anaya

Milton Keynes, U.K., June 2006

Part I

Local Mixture Models

Chapter 1

Introduction

1.1 Some History

There are two main areas within the interplay between Geometry and Statistics. One of them studies statistical models using *Group Invariance.* See Giri (1996) for an excellent review on this area. The other area, which is the one we follow in this work, studies statistical models using Riemannian Geometry and began with the seminal papers of Rao (1945) and Jeffreys (1946). The geometric structure implicit on those seminal papers did not attract too much attention to the statisticians of that time and thirty years later, Efron (1975) retook the subject to emphasize the importance that geometrical insight has in the asymptotic theory of parametric inference. But Efron did not exploit the Geometry in its full multivariate splendor. It was Amari (1985) who first really exploited the whole multivariate machinery of Geometry to better understand parametric asymptotic inference in Exponential Families.

Amari first elucidated the Geometry of Exponential Families and exploited it to develop a powerful, as well as insightful theory that clearly related Geometrical and Statistical concepts in a natural way from both points of view. After that account, there had been many contributions to the area, now called *Information Geometry* (a name coined by Amari), but they still compose an insignificant proportion compared to the now vast statistical literature. Information Geometry is a fascinating and fertile area which is fastly developing and with a promising future (see the Proceedings of the Second International Symposium on Information Geometry and its Applications (2005)).

One of the major challenges in Information Geometry is to make the tools developed accessible to both Statistics and Geometry audiences. Part I of this thesis is a contribution to tackle that challenge. It gives a second step to the work inaugurated by Marriott (2002) and his subsequent articles: Marriott (2003), Critchley and Marriott (2004), Marriott and Vos (2004) and Marriott (2005) with the aim of a better understanding of the Geometry inherent to mixture models. As in Amari's work, we use very simple Geometrical and Statistical ideas and put them to interact together naturally, making our results readable to both audiences.

1.2 Local Mixtures

It is well known that Exponential Families are one the most important models used in Statistical Inference. Two of their main virtues are that they offer powerful statistical properties and at the same time are tractable analytically, see for example Barndorff-Nielsen (1978), Brown (1986) and Letac (1992). What it is not so well known is that those virtues are mainly due to the fact that they can be interpreted as *Affine Spaces* (see Appendix B).

Affine Geometry is simple. It is just the "usual" Geometry but without the notions of distance and angles. Roughly, an affine space can be thought of as a set which becomes a vector space by simply selecting a point to be the origin. We use Affine Spaces to explain the local nature that mixture models have under natural statistical assumptions. Consider the space of functions

$$\mathcal{D}_\nu^m := \left\{ g(x) \, : \, \int g(x) \, \nu(dx) = 1 \right\} ,$$

where is ν is a known measure. Note that the elements of $\mathcal{D}_\nu^m$ are not necessarily positive everywhere, so the set of all positive densities with respect to ν is only a convex subset. Any regular parametric family of densities $\mathcal{F} = \{f(x;\theta) \,:\, \theta \in \Theta\}$ with respect to ν, can be embedded into $\mathcal{D}_\nu^m$ by using the map $\theta \mapsto f(x;\theta)$. Now consider the vector space of functions

$$\mathcal{V}_\nu^0 := \left\{ s(x) \,:\, \int s(x)\,\nu(dx) = 0 \right\}.$$

An affine structure is basically defined by the following fact:

$$g(x) + s(x) \in \mathcal{D}_\nu^m$$

for any $g \in \mathcal{D}_\nu^m$ and $s \in \mathcal{V}_\nu^0$, where $+$ is the usual sum operator between real valued functions. The structure $(\mathcal{D}_\nu^m, \mathcal{V}_\nu^0, +)$ is then an Affine Space. In general this affine space can be infinite dimensional. We consider mixtures of $\mathcal{F}$ of the form

$$g(x;Q) := \int_\Theta f(x;\theta)\,dQ(\theta)\,,$$

for some unknown distribution Q defined over Θ. For the set of allowed Q's, we consider the family of *Proper Dispersion Models*

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

with densities

$$dQ(\theta;\vartheta,\epsilon) = a(\epsilon) V^{-1/2}(\theta) \exp\left\{ -\frac{1}{2\epsilon} d(\theta;\vartheta) \right\} d\theta.$$

These models generalize the idea of *localizing mixing distribution* of Marriott (2002). We can therefore embed the generated family of densities in $\mathcal{D}_\nu^m$ via the mapping

$$(\epsilon,\vartheta) \mapsto g(x; Q(\cdot\,;\vartheta,\epsilon))$$

and think of it as a family of curves in the space $\mathcal{D}_\nu^m$ approaching the family $\mathcal{F}$ at each point $f(x;\vartheta)$, when ϵ goes to zero. See Figure 1.1.

We will show the following asymptotic result valid when ϵ is small:

$$g(x; Q(\cdot;\vartheta,\epsilon)) \sim f(x; M_1(\vartheta,\epsilon)) + \sum_{i=2}^{d} M_i(\vartheta,\epsilon)\, f^{(i)}(x; M_1(\vartheta,\epsilon)) \in \mathcal{D}_\nu^m \quad (1.1)$$

at each $\vartheta \in \Theta$ and for some functions $M_1(\vartheta,\epsilon), \ldots, M_d(\vartheta,\epsilon)$.

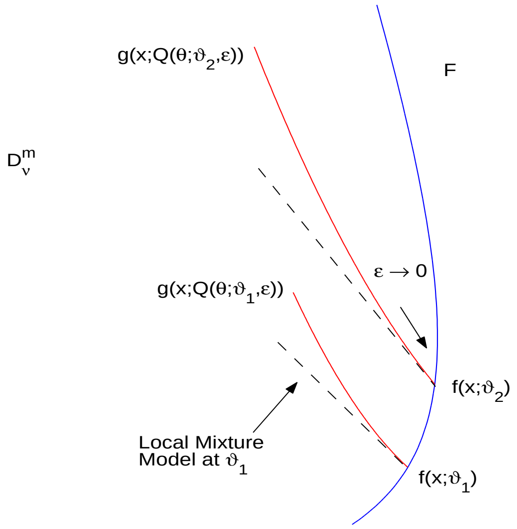
$g(x;Q(\theta;\vartheta_2,\varepsilon))$
F
D^m_ν
$\varepsilon \to 0$
$g(x;Q(\theta;\vartheta_1,\varepsilon))$
$f(x;\vartheta_2)$
Local Mixture Model at ϑ_1
$f(x;\vartheta_1)$

Figure 1.1: Local mixtures approaching $\mathcal{F}$ (red) and local mixture models with $d = 2$ (dashed)

Motivated from this geometrical structure we will propose a new class of statistical models, called Local Mixture Models, of the form

$$\left\{ g(x;\theta,\lambda_2,\ldots,\lambda_d) = f(x;\theta) + \sum_{i=2}^{d} \lambda_i(\theta)\, f^{(i)}(x;\theta) \ : \ (\theta,\lambda_2,\ldots,\lambda_d) \in \Upsilon \right\} \tag{1.2}$$

where $\Upsilon \subset \mathbb{R}^d$ is a set to be defined later. For each fixed $\theta \in \Theta$ these models are subsets of a finite dimensional affine subspace of $(\mathcal{D}^m_\nu, \mathcal{V}^0_\nu, +)$. Also, from (1.1), these models mimic the behavior of genuine mixtures when the parameter ϵ is small. See Figure 1.1.

Exponential families constitute other important example of statistical families which are subsets of finite dimensional affine subspaces. The nice and powerful statistical properties of exponential families are well known and have already been studied from this geometric point of view (see Amari and Nagaoka (2000) and the references therein). We do the mixture counterpart to show that Local Mixture Families have also nice and powerful statistical properties under very natural statistical assumptions and which can be exploited to make inference. When the underlying family $\mathcal{F}$ is *Natural exponential family with quadratic variance function*, we show that Local mixture models constitute a set of models which identifiable, flexible and interpretable. In particular, we study in some detail the application of Local mixture models to the analysis of mixtures of the negative exponential distribution.

As in exponential families, only a subset of the affine subspace can be considered as a proper density. That is, the natural parameter space of a k-dimensional exponential family is not necessarily the whole $\mathbb{R}^k$. This induces a boundary structure which has been largely ignored in the statistical literature about exponential families. We recognize an analogue boundary structure in our models as they are also subsets of affine spaces. In our case, the boundaries are given by the fact that the sum of a function which integrates to one and a function which integrates to zero is not necessarily positive everywhere. Our families have that structure, so we have to bound the parameter space to always ensure positivity.

We study, in some detail, the effect that this boundary structure has on the statistical inference of this type of models. These positivity boundaries are difficult to dealt with as they represent singularities and, for that reason, we call them *Hard Boundaries*. But, ensuring positivity is not enough to make local mixture models capture mixture structure. We have to restrict a bit more the parameter space using new boundaries, which are very simple to manage and have a meaningful interpretation. For that reason, we call those *Soft Boundaries*. A simple example arises when we need to further restrict the parameter space to ensure positivity of a certain variance function.

1.3 Outline

The outline for Part I of the thesis is the following. In chapter two, we described the relevant geometric properties of exponential families when they are viewed as affine spaces. For our own convenience, we emphasize the use of general exponential families instead of natural exponential families. Then we show that mixture models also follow a natural affine geometry. We define general mixture families basically by analogy from the exponential family structure. At the end of the chapter we briefly described the construction of statistical bundles with fibers which are either general exponential or mixtures families.

In chapter three, we begin motivating local mixture models by describing asymptotic expansions of mixtures. Taylor and Laplace based expansions are described in detail. To construct Laplace expansions we made the important assumption that the mixing distribution is a proper dispersion model. Motivated on the affine structure on those expansions we formally defined local mixture models in a slightly different way as has been done in the literature. We made explicit the dependence on the mixing parameters on the baseline distribution family. Then true local mixture models are defined as an attempt to improve the mixture behavior of our models. Local mixture models for exponential families are described in detail and some statistical issues like identification are discussed. A new class of local mixture models called local scale and local location mixture models are then defined. The general estimation problem is analyzed and some important statistical properties of local mixture models are proved.

In chapter four, the results found are applied to the estimation problem of mixtures of the negative exponential distribution. Some particular models are studied. The use of boundaries in particular situations is described in detail. Some connections are made with the literature about testing for overdispersion. We described some interesting relationships with other models which, apparently, are not related to local mixture models. A simulation study is performed at the end of the chapter to show many practical issues that arise when trying to fit local mixture models to a particular set of data.

Finally, we included three appendixes at the end of the thesis. They correspond to background definitions and results on: Regularity Conditions, Affine Spaces and Dispersion Models. Such material is only used in Part I of this thesis.

Chapter 2

Geometry of Mixture Models

In this chapter, the geometry of exponential and mixtures families is described using Affine Spaces. The affine geometry of exponential families is described in detail first and then summarized in Theorem 3 which states that exponential families are convex subsets of subspaces of a specific affine space. Immediately after, a geometric interpretation of the log-likelihood in exponential families is given. Motivated on the exponential case, General Mixture Families are defined as convex subsets of subspaces of a specific affine space different than the exponential one. The boundary structure and the visualization of the families is then described. Finally a brief and simple description of Affine Bundles and an Euclidean Structure is given because of its relevance to the next chapters.

The following material is a mixture of ideas taken basically from Barndorff-Nielsen (1978), Brown (1986), Amari (1990), Letac (1992), Murray and Rice (1993), Pistone and Rogantin (1999), De Sanctis (2002) and Grasselli (2005). The contribution of this chapter is the form in which we state the results and the connections made between them.

2.1 Geometry of Exponential Families

2.1.1 Relevant properties of GEFs

Let ν be a σ-finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $\boldsymbol{S} : \mathbb{R} \to \mathbb{R}^d$ a Borel measurable function. The *Laplace Transform of* $\boldsymbol{S}$ with respect to ν is the function $L_{\nu,\boldsymbol{S}} : \mathbb{R}^d \to (0, \infty]$ defined by

$$L_{\nu,\boldsymbol{S}}(\boldsymbol{\lambda}) := \int_{\mathbb{R}} \exp\langle\, \boldsymbol{\lambda}, \boldsymbol{S}(x)\rangle\, \nu(dx) = \int_{\mathbb{R}^d} \exp\langle\, \boldsymbol{\lambda}, \boldsymbol{s}\rangle\, \nu_{\boldsymbol{S}}(d\boldsymbol{s})\,,$$

where $\nu_{\boldsymbol{S}} = \nu \circ \boldsymbol{S}^{-1}$ is the image measure of ν under $\boldsymbol{S}$. Define

$$D^e(\nu_{\boldsymbol{S}}) := \left\{\boldsymbol{\lambda} \in \mathbb{R}^d \;:\; L_{\nu,\boldsymbol{S}}(\boldsymbol{\lambda}) < \infty\right\}$$

and

$$K_{\nu_{\boldsymbol{S}}}(\boldsymbol{\lambda}) := \log L_{\nu,\boldsymbol{S}}(\boldsymbol{\lambda}).$$

Using Hölder's inequality, see Letac (1992), it is possible to show

Theorem 1 *$D^e(\nu_{\boldsymbol{S}})$ and $K_{\nu_{\boldsymbol{S}}}(\boldsymbol{\lambda})$ have the following properties:*

1. *$D^e(\nu_{\boldsymbol{S}})$ is a convex set in $\mathbb{R}^d$,*
2. *$K_{\nu_{\boldsymbol{S}}}(\boldsymbol{\lambda})$ is a convex function on $D^e(\nu_{\boldsymbol{S}})$ and*
3. *$K_{\nu_{\boldsymbol{S}}}(\boldsymbol{\lambda})$ is strictly convex if and only if $\nu_{\boldsymbol{S}}$ is not concentrated on a proper affine subspace of $\mathbb{R}^d$.*

Denote by $\mathcal{R}(\mathbb{R}^d)$ the set of σ-finite measures η on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that

1. $\Lambda^e(\eta) := \text{interior}(D^e(\eta)) \neq \emptyset$
2. η is not concentrated on a proper affine subspace of $\mathbb{R}^d$.

Definition 1 Let ν be a σ-finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $\boldsymbol{S} : \mathbb{R} \to \mathbb{R}^d$ a Borel measurable function such that $\nu_{\boldsymbol{S}} \in \mathcal{R}(\mathbb{R}^d)$. The family $\mathcal{GE}(\nu, \boldsymbol{S}) = \{P_{\boldsymbol{\lambda}} : \boldsymbol{\lambda} \in \Lambda^e(\nu_{\boldsymbol{S}})\}$ of probability measures with Radon-Nikodym derivatives

$$f_\nu(x; \boldsymbol{\lambda}) := \frac{dP_{\boldsymbol{\lambda}}}{d\nu}(x) = \exp\left[\langle\, \boldsymbol{\lambda}, \boldsymbol{S}(x)\rangle - K_{\nu_{\boldsymbol{S}}}(\boldsymbol{\lambda})\right] \tag{2.1}$$

is called the *General Exponential Family* (GEF) generated by ν and $\boldsymbol{S}$. $\boldsymbol{\lambda}$ and $\Lambda^e(\nu_{\boldsymbol{S}})$ are called the *natural parameter* and *natural parameter space* associated with $\mathcal{GE}(\nu, \boldsymbol{S})$, respectively. The boundary of $\Lambda^e(\nu_{\boldsymbol{S}})$ will be called the *Hard Exponential Boundary* of the family. The function $K_{\nu_{\boldsymbol{S}}}(\boldsymbol{\lambda})$ is called the *cumulant function* and the number d is the dimension of the family.

Remarks:

R.1 All the measures in a GEF are mutually absolutely continuous. In particular, this implies that they share the same support.

R.2 If ν and ν' are such that $\nu_{\boldsymbol{S}}, \nu'_{\boldsymbol{S}} \in \mathcal{R}(\mathbb{R}^d)$ then $\mathcal{GE}(\nu, \boldsymbol{S}) = \mathcal{GE}(\nu', \boldsymbol{S})$ if and only if there exist $\boldsymbol{a}_0 \in \mathbb{R}^d$ and $c_0 \in \mathbb{R}$ such that

$$\nu'(A) = \int_A \exp\left\{\langle\, \boldsymbol{a}_0, \boldsymbol{S}(x)\rangle + c_0\right\} \nu(dx)\,, \qquad A \in \mathcal{B}(\mathbb{R}).$$

The natural parameter space is not the same though. In fact,

$$D^e(\nu'_{\boldsymbol{S}}) = D^e(\nu_{\boldsymbol{S}}) - \boldsymbol{a}_0 := \{\boldsymbol{\kappa} : \boldsymbol{\kappa} + \boldsymbol{a}_0 \in D^e(\nu_{\boldsymbol{S}})\}\ .$$

In this sense, the natural parameter space is defined up to translations. More generally, the natural parameter space is defined up to affine transformations, see Brown (1986) for details. In particular, note that a GEF is generated by any of its elements, that is

$$\mathcal{GE}(\nu, \boldsymbol{S}) = \mathcal{GE}(P_{\boldsymbol{\lambda}}, \boldsymbol{S}) \qquad \forall\, \boldsymbol{\lambda} \in \Lambda^e(\nu_{\boldsymbol{S}}).$$

This implies that, by choosing any member of the GEF as starting measure, we can always make the natural parameter space to include the origin.

R.3 It is possible to "center" the distribution of $\boldsymbol{S}$ anywhere we like without changing the generated GEF. That is,

$$\mathcal{GE}(\nu, \boldsymbol{S}) = \mathcal{GE}(\nu, \boldsymbol{S} + \boldsymbol{\tau}) \,,$$

for any $\boldsymbol{\tau} \in \mathrm{I\!R}^d$.

R.4 The measure

$$\int \exp\left[\langle \boldsymbol{\lambda}, \boldsymbol{S}(x) \rangle - K_{\nu_{\boldsymbol{S}}}(\boldsymbol{\lambda})\right] \, \nu(dx)$$

is called an *exponential tilting* of ν with $\boldsymbol{S}$. Hence the members of a GEF are exponential tiltings of each other.

R.5 The conditions on the definition of the set of measures $\mathcal{R}(\mathrm{I\!R}^d)$ are important because of the following. If $\nu_{\boldsymbol{S}}$ is supported on a proper affine subspace of $\mathrm{I\!R}^d$ then it is possible to reduce the dimension of the GEF to a number less than d. On the other hand, if $D^e(\nu_{\boldsymbol{S}})$ is contained in a proper affine subspace of $\mathrm{I\!R}^d$ (and therefore $\Lambda^e(\nu_{\boldsymbol{S}})$ is empty) then it is also possible to reduce the dimension of the GEF to a number less than d. In both cases, that number is called the *order* of the family. Then, the restriction to measures in $\mathcal{R}(\mathrm{I\!R}^d)$ ensure that GEF's always have order equal to the dimension of the support of $\nu_{\boldsymbol{S}}$. See Brown (1986) for details.

R.6 The following Corollary of Theorem 1, states a result of important statistical consequences.

Corollary 1 *The logarithm of the density function $f_\nu(x; \boldsymbol{\lambda})$ of a General Exponential Family is a strictly concave function of $\boldsymbol{\lambda}$ for each fixed x. In particular, the log-likelihood for $\boldsymbol{\lambda}$ will be a a strictly concave function.*

2.1.2 Relevant properties of NEFs

The family of probability distributions of $\boldsymbol{S}$ in $\mathcal{GE}(\nu, \boldsymbol{S})$ is called the *Natural Exponential Family* (NEF) generated by $\nu_{\boldsymbol{S}}$. We will denote such family by $\mathcal{NE}(\nu_{\boldsymbol{S}})$. The associated family of Radon-Nikodym derivatives is given by

$$f_{\nu_{\boldsymbol{S}}}(\boldsymbol{s}; \boldsymbol{\lambda}) := \frac{dP_{\boldsymbol{\lambda}}}{d\nu_{\boldsymbol{S}}}(\boldsymbol{s}) = \exp\left\{\langle \boldsymbol{\lambda}, \boldsymbol{s} \rangle - K_{\nu_{\boldsymbol{S}}}(\boldsymbol{\lambda})\right\}. \tag{2.2}$$

We now review some relevant properties of NEFs for the case $d = 1$ which will be used later in this work. The *mean and variance* are given by

$$\begin{aligned}\frac{\partial}{\partial\lambda}K_{\nu_s}(\lambda) &= \int_{\mathbb{R}} s\, f_{\nu_s}(s;\lambda)\,\nu_s(ds) =: E_\lambda[s\,] \\ \frac{\partial^2}{\partial\lambda^2}K_{\nu_s}(\lambda) &= \int_{\mathbb{R}} (s - E_\lambda[s])^2\, f_{\nu_s}(s;\lambda)\,\nu_s(ds) =: Var_\lambda[s].\end{aligned}$$

If we define

$$K'_{\nu_s}(\lambda) := \frac{\partial}{\partial\lambda}K_{\nu_s}(\lambda)\,,$$

then the set

$$M(\nu_s) := K'_{\nu_s}(\Lambda^e(\nu_s)) \subset \mathbb{R}$$

is called *the domain of the means* of $\mathcal{NE}(\nu_s)$. It is possible to use $M(\nu_s)$ to parametrise $\mathcal{NE}(\nu_s)$ as follows. If $\nu_s \in \mathcal{R}(\mathbb{R})$ then $K'_{\nu_s} : \Lambda^e(\nu_s) \to M(\nu_s)$ is a diffeomorphism. We denote by $\psi_{\nu_s} : M(\nu_s) \to \Lambda^e(\nu_s)$ the inverse of K'_{ν_s}, then the map

$$P_{\psi_{\nu_s}(m)} \mapsto m$$

is a new parametrization of $\mathcal{NE}(\nu_s)$ by its domain of the means. We can thus define $\tilde{f}_{\nu_s}(s;m) := f_{\nu_s}(s;\psi_{\nu_s}(m))$. For simplicity in notation, we will omit the tilde over the density f in the mean parametrization.

Denote by $C^0(\nu_s)$ the interior of the convex hull of the support of $\mathcal{NE}(\nu_s)$. Then it is clear that $M(\nu_s) \subseteq C^0(\nu_s)$. In most practical cases both sets are equal but not always. When $M(\nu_s) = C^0(\nu_s)$ the NEF is said to be *steep*. It is possible to show that if $D^e(\nu_s)$ is open then $\mathcal{NE}(\nu_s)$ is steep (see Barndorff-Nielsen 1978). We shall only consider steep NEF's throughout this thesis.

Definition 2 Let $\mathcal{NE}(\nu_s)$ be a univariate NEF. Then the map $V_{\nu_s} : M(\nu_s) \to \mathbb{R}^+$ defined by

$$V_{\nu_s}(m) := Var_{\psi_{\nu_s}(m)}[s]$$

is called the *variance function* of $\mathcal{NE}(\nu_s)$.

The variance function, together with its domain of means, characterizes the NEF.

Theorem 2 *Let $\mathcal{NE}(\nu_s)$ be a univariate NEF. Then*

$$V_{\nu_s}(m) = K''_{\nu_s}(\psi_{\nu_s}(m)) = [\psi'_{\nu_s}(m)]^{-1}\,.$$

Furthermore, if $\mathcal{NE}(\nu_s)$ and $\mathcal{NE}(\nu_t)$ are two NEF's such that V_{ν_s} and V_{ν_t} coincide on a non-empty subset of $M(\nu_s) \cap M(\nu_t)$ then $\mathcal{NE}(\nu_s) = \mathcal{NE}(\nu_t)$.

V_{ν_s} is always positive since $\mathcal{NE}(\nu_s)$ is not concentrated on a proper affine subspace of $\mathbb{R}$, that is, since $\mathcal{NE}(\nu_s)$ is not degenerate. See Letac (1992) for details about Natural Exponential Families.

2.1.3 Affine Geometry of GEFs

In this section we show that General Exponential Families have an Affine Geometry. Following Murray and Rice (1993) and De Sanctis (2002), when considering measures up to scale, any GEF is the interior of a convex subset of an affine subspace of measures. The key ideas are:

A. Embed any GEF in a common space using the fact that measures in a GEF are mutually absolutely continuous,

B. extend the GEF by adding measures which are possibly non-finite or does not necessarily integrate to one,

C. treat measures which are proportional to each other as equal and

D. construct an appropriate vector space of random variables such that any extended GEF corresponds to an affine subspace of a common bigger affine space.

We now develop these ideas in detail.

A. Let ν be a σ-finite measure defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $\mathcal{M}_\nu$ be the set of all σ-finite measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ which are equivalent to the measure ν in the following sense. Two σ-finite measures ν and η are considered equivalent if and only if ν is absolutely continuous with respect to η and viceversa. We will denote this equivalence relation by $\nu \equiv \eta$. Obviously $\mathcal{M}_\nu$ is just the equivalence class that contains ν. Note that for any σ-finite measure ν we have

$$\mathcal{GE}(\nu, \boldsymbol{S}) \subset \mathcal{M}_\nu$$

for all measurable $\boldsymbol{S}$. Recall remark R.1. Also note that any a.e.$[\nu]$ assertion is equivalent to the same a.e.$[\eta]$ assertion for any $\eta \in \mathcal{M}_\nu$.

B. Consider the following extension of any $\mathcal{GE}(\nu, \boldsymbol{S})$:

$$\mathcal{EE}(\nu, \boldsymbol{S}) := \left\{ \eta \,:\, \eta = \int \exp\left[\langle \boldsymbol{\lambda}, \boldsymbol{S}(x)\rangle + c\right] \nu(dx),\ \boldsymbol{\lambda} \in \mathrm{I\!R}^d,\ c \in \mathrm{I\!R} \right\}$$

which clearly satisfies

$$\mathcal{GE}(\nu, \boldsymbol{S}) \subset \mathcal{EE}(\nu, \boldsymbol{S}) \subset \mathcal{M}_\nu.$$

$\mathcal{EE}(\nu, \boldsymbol{S})$ includes non-finite measures (those with $\boldsymbol{\lambda} \notin \Lambda^e(\nu_{\boldsymbol{S}})$) and all the multiples of measures in $\mathcal{GE}(\nu, \boldsymbol{S})$ which do not integrate to one.

C. Consider now the following equivalence relation $\cong$ defined in $\mathcal{M}_\nu$. We define $\eta_1 \cong \eta_2$ if and only if $\eta_2 = \exp(c)\,\eta_1$ for some $c \in \mathrm{I\!R}$ and denote by $\overline{\eta}$ the equivalence class of η. The use of this equivalence relation means that we are identifying measures up to scale. In other words, treating as equal, measures in $\mathcal{M}_\nu$ which are proportional to each other. For any $\mathcal{H} \subseteq \mathcal{M}_\nu$ define $\overline{\mathcal{H}}$ to be the quotient space $\mathcal{H}/\cong$. That is,

$$\overline{\mathcal{H}} = \{\bar{\eta} \,:\, \eta \in \mathcal{H}\}.$$

Then we have

$$\overline{\mathcal{GE}}(\nu, \boldsymbol{S}) \subset \overline{\mathcal{EE}}(\nu, \boldsymbol{S}) \subset \overline{\mathcal{M}}_\nu.$$

We declare $\bar{\eta} \in \overline{\mathcal{EE}}(\nu, \boldsymbol{S})$ finite if the representative η is a finite measure. Then $\overline{\mathcal{GE}}(\nu, \boldsymbol{S})$ is the interior of the set of finite elements of $\overline{\mathcal{EE}}(\nu, \boldsymbol{S})$.

D. Now, we are going to construct a vector space $\overline{\mathcal{V}}_\nu$ and a translation operator $\boxplus^{\mathrm{E}}$ such that $(\overline{\mathcal{M}}_\nu, \overline{\mathcal{V}}_\nu, \boxplus^{\mathrm{E}})$ is an affine space and any extended GEF $\overline{\mathcal{EE}}(\nu, \boldsymbol{S})$ is an affine subspace of $(\overline{\mathcal{M}}_\nu, \overline{\mathcal{V}}_\nu, \boxplus^{\mathrm{E}})$ of dimension d.

Consider the vector space $L^0(\nu)$ of all measurable functions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ quotiented by the subspace of functions which are 0 a.e.$[\nu]$. The elements of $L^0(\nu)$ are equivalence classes $\tilde{s}$ of the form

$$\tilde{s} = \left\{ t \in \mathcal{L}^0(\mathbb{R}, \mathcal{B}(\mathbb{R})) \,:\, t = s \;\text{ a.e. } [\nu] \right\}.$$

For simplicity in the notation, we will write s for $\tilde{s}$, always keeping in mind is an equivalence class of functions. Define the following translation operator on $\mathcal{M}_\nu$

$$\eta \oplus^{\mathrm{E}} s := \int_{\mathbb{R}} e^{s(x)}\, \eta(dx) \tag{2.3}$$

where $\eta \in \mathcal{M}_\nu$, $s \in L^0(\nu)$. Note that this operator is well defined due to the properties of the integral. It is easy to check that $(\mathcal{M}_\nu, L^0(\nu), \oplus^{\mathrm{E}})$ is an affine space[1].

Denote by $\mathbf{1}$ the constant function equal to one. To construct a convenient vector space, consider the subspace $\mathcal{K}$ of $L^0(\nu)$ generated by $\mathbf{1}$ and denote by s^* the coset

$$\{s + r \,:\, r \in \mathcal{K}\} = \{s + c\,\mathbf{1} \,:\, c \in \mathbb{R}\}.$$

The quotient vector space

$$L^0(\nu)/\mathcal{K} = \left\{ s^* \,:\, s \in L^0(\nu) \right\}$$

has the appropriate structure to make $\overline{\mathcal{M}}_\nu$ an affine space and means that we are identifying vectors up to the addition of a constant. We will denote this space by $\overline{\mathcal{V}}_\nu$. Recall remark R.3.

For any $\overline{\eta} \in \overline{\mathcal{M}}_\nu$ and $s^* \in \overline{\mathcal{V}}_\nu$ define the following translation operator

$$\overline{\eta} \boxplus^{\mathrm{E}} s^* := \overline{\rho \oplus^{\mathrm{E}} r}$$

for any $\rho \in \overline{\eta}$ and $r \in L^0(\nu)$ such that $r - s \in \mathcal{K}$. The operator $\boxplus^{\mathrm{E}}$ is well defined in the sense that does not depend on the representatives ρ and r so we can take them as η and s. Then we have the following result,

[1] Denote by $\mathcal{P}_\nu$ the set of probability measures equivalent to ν. The geometry of $\mathcal{P}_\nu$ has been studied by some authors (Pistone and Sempi 1995, Gibilisco and Pistone 1998, Pistone and Rogantin 1999, Cena 2003 and Grasselli 2005). The set $\mathcal{P}_\nu$ is endowed with the structure of a $\mathcal{C}^\infty$-Banach manifold using the Orlicz space of an exponentially growing function. With this added structure, they call $\mathcal{P}_\nu$ the *exponential statistical manifold.*

Theorem 3 *$(\overline{\mathcal{M}}_\nu, \overline{\mathcal{V}}_\nu, \boxplus^E)$ is an affine space and any $\overline{\mathcal{E}\mathcal{E}}(\nu, \boldsymbol{S})$ is an affine subspace of $(\overline{\mathcal{M}}_\nu, \overline{\mathcal{V}}_\nu, \boxplus^E)$ of dimension d.*

Proof: First, $\overline{\eta} \boxplus^{\mathrm{E}} 0^* = \overline{\eta \oplus^{\mathrm{E}} 0} = \overline{\eta}$ for all $\overline{\eta} \in \overline{\mathcal{M}}_\nu$. Also

$$\begin{aligned}(\overline{\eta} \boxplus^{\mathrm{E}} s_1^*) \boxplus^{\mathrm{E}} s_2^* &= \overline{(\eta \oplus^{\mathrm{E}} s_1) \oplus^{\mathrm{E}} s_2} \\ &= \overline{\eta \oplus^{\mathrm{E}} (s_1 + s_2)} \\ &= \overline{\eta} \boxplus^{\mathrm{E}} (s_1^* + s_2^*).\end{aligned}$$

Finally, s^* with $s = \log(d\eta_2/d\eta_1)$ is the only vector such that $\bar{\eta}_1 \boxplus^{\mathrm{E}} s^* = \bar{\eta}_2$. This proves that $(\overline{\mathcal{M}}_\nu, \overline{\mathcal{V}}_\nu, \boxplus^{\mathrm{E}})$ is an Affine Space. Now, we can write

$$\overline{\mathcal{E}\mathcal{E}}(\nu, \boldsymbol{S}) = \left\{\bar{\eta}(\boldsymbol{\lambda}) := \bar{\nu} \boxplus^{\mathrm{E}} s^*_{\boldsymbol{\lambda}}(x) \;:\; s_{\boldsymbol{\lambda}}(x) = \langle \boldsymbol{\lambda}, \boldsymbol{S}(x)\rangle \,,\, \boldsymbol{\lambda} \in \mathrm{I\!R}^d\right\}$$

or, more clearly

$$\overline{\mathcal{E}\mathcal{E}}(\nu, \boldsymbol{S}) = \bar{\nu} \boxplus^{\mathrm{E}} \mathrm{span}\left\{s_1^*, \ldots, s_d^*\right\}. \tag{2.4}$$

To show that $\overline{\mathcal{E}\mathcal{E}}(\nu, \boldsymbol{S})$ is an affine subspace of $(\overline{\mathcal{M}}_\nu, \overline{\mathcal{V}}_\nu, \boxplus^{\mathrm{E}})$ we need to show that it is closed under affine combinations (see Appendix B), that is, for any finite set $\{\boldsymbol{\lambda}_i\}_{i\in I} \subset \mathrm{I\!R}^d$ and for any $\{\theta_i\}_{i\in I}$ such that $\sum_{i\in I} \theta_i = 1$ the affine combination

$$\sum_{i\in I} \theta_i \bar{\eta}(\boldsymbol{\lambda}_i) \in \overline{\mathcal{E}\mathcal{E}}(\nu, \boldsymbol{S}).$$

Indeed, for any $\boldsymbol{\lambda}_0 \in \mathrm{I\!R}^d$

$$\begin{aligned}\sum_{i\in I} \theta_i \bar{\eta}(\boldsymbol{\lambda}_i) &= \bar{\eta}(\boldsymbol{\lambda}_0) \boxplus^{\mathrm{E}} \sum_{i\in I} \theta_i \left[\log\left(\frac{d\bar{\eta}(\boldsymbol{\lambda}_i)}{d\bar{\eta}(\boldsymbol{\lambda}_0)}\right)\right]^* \\ &= \overline{\int \exp\left[\sum_{i\in I} \theta_i \log\left(\frac{d\bar{\eta}(\boldsymbol{\lambda}_i)}{d\bar{\eta}(\boldsymbol{\lambda}_0)}\right) + \langle \boldsymbol{\lambda}_0, \boldsymbol{S}(x)\rangle\right] d\nu} \\ &= \overline{\int \exp\left[\sum_{i\in I} \theta_i \langle \boldsymbol{\lambda}_i - \boldsymbol{\lambda}_0, \boldsymbol{S}(x)\rangle + \langle \boldsymbol{\lambda}_0, \boldsymbol{S}(x)\rangle\right] d\nu} \\ &= \overline{\int \exp\left[\left\langle \sum_{i\in I} \theta_i \boldsymbol{\lambda}_i, \boldsymbol{S}(x)\right\rangle\right] d\nu} \in \overline{\mathcal{E}\mathcal{E}}(\nu, \boldsymbol{S}).\end{aligned}$$

By using Lemma 2 in Appendix B, the dimension of the affine subspace is the same as the dimension of the vector space

$$\overline{\mathcal{V}}_\nu(\boldsymbol{S}) = \left\{ \left[\log\left(\frac{d\bar{\eta}(\boldsymbol{\lambda})}{d\bar{\eta}(\boldsymbol{\lambda}_0)} \right) \right]^* \,:\, \boldsymbol{\lambda} \in \mathrm{I\!R}^d \right\} = \left\{ [\langle \boldsymbol{\lambda} - \boldsymbol{\lambda}_0, \boldsymbol{S}(x) \rangle]^* \,:\, \boldsymbol{\lambda} \in \mathrm{I\!R}^d \right\}.$$

The assumption that $\nu_{\boldsymbol{S}}$ is not concentrated on a proper affine subspace of $\mathrm{I\!R}^d$ implies that a.e.$[\nu]$

$$a_0 + \sum_{i=1}^{d} a_i\, s_i(x) = 0 \quad \Rightarrow \quad a_0 = a_1 = \ldots = a_d = 0$$

and this implies the vector space $\overline{\mathcal{V}}_\nu(\boldsymbol{S})$ above has dimension d. ■

To clarify the ideas presented, consider the following example.

Example 1 Consider the measure

$$\nu(A) := \int_{A \cap (0,\infty)} \exp(-x)\, dx\,, \qquad A \in \mathcal{B}(\mathrm{I\!R})$$

and the function $\boldsymbol{S}(x) = (s_1(x), s_2(x))^t = (x, \log x)$ which is measurable over $(0, \infty)$. Then we have

$$\overline{\mathcal{E}\mathcal{E}}(\nu, \boldsymbol{S}) = \left\{ \bar{\eta}(\lambda_1, \lambda_2) = \overline{\int \exp\left[\lambda_1 x + \lambda_2 \log x - x\right] dx} \,:\, \lambda_1, \lambda_2 \in \mathrm{I\!R} \right\}$$

and $D^e(\nu_{\boldsymbol{S}}) = (-\infty, 1) \times (-1, \infty)$. The GEF generated by $(\nu, \boldsymbol{S})$ has densities

$$f_\nu(x; \lambda_1, \lambda_2) = (1 - \lambda_1)^{\lambda_2 + 1} x^{\lambda_2} \exp\left\{\lambda_1\, x\right\} / \Gamma(\lambda_2 + 1).$$

This family is known as the Gamma Family with two parameters. The subfamily generated by the restriction $\lambda_2 = 0$ is known as the Negative Exponential Family. This subfamily is clearly also a GEF.

Consider now the family of measures with densities with respect to ν,

$$h_\nu(x; \beta) = \beta x^{\beta - 1} \exp\left\{x - x^\beta\right\}\,, \quad \beta > 0.$$

This family is known as the Weibull family with shape parameter β. This is not a General exponential family because its extended family

$$\left\{ \overline{\int x^{\beta-1} \exp\left\{x - x^{\beta}\right\} \nu(dx)} \;:\; \beta > 0 \right\}$$

is not closed under affine combinations. Figure 2.1 shows graphically the geometry of this example. We use the notation of an overline for any family name as we are considering elements in $\overline{\mathcal{M}}_\nu$, that is, measures up to scale. The Weibull family appears as a curve because it is not an affine subspace of $\overline{\mathcal{M}}_\nu$ (see section 2.2.4).

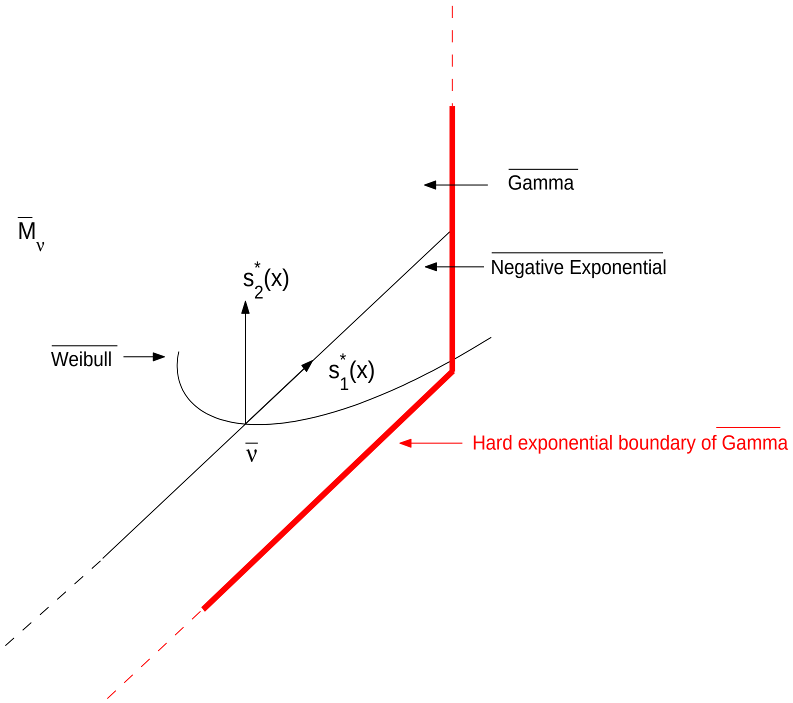
$\overline{\text{Gamma}}$
$\overline{M}_\nu$
$\overline{\text{Negative Exponential}}$
$s^*_2(x)$
$\overline{\text{Weibull}}$
$s^*_1(x)$
$\overline{\nu}$
Hard exponential boundary of $\overline{\text{Gamma}}$

Figure 2.1: Affine Geometry of GEFs

From a geometrical point of view, the construction of a GEF is quite clear. Assume we have a σ-finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and measurable functions

$s_i : \mathbb{R} \to \mathbb{R}$ for $i = 1, \dots, d$. The most familiar examples of finite dimensional affine subspaces are hyperplanes of dimension d in $\mathbb{R}^n$. For a given point $p \in \mathbb{R}^n$ and a set of d vectors (directions) $v_1, \dots, v_d$ also in $\mathbb{R}^n$, we can construct any such such hyperplane by

$$\left\{ p + \sum_{i=1}^{d} \lambda_i v_i \,:\, (\lambda_1, \dots, \lambda_d) \in \mathbb{R}^d \right\}.$$

Here the translation operator is the sum between vectors in $\mathbb{R}^n$. If we change the translation operator to $\oplus^{\mathrm{E}}$, substitute $\mathbb{R}^n$ by $\mathcal{M}_\nu$, take the point p as ν and the directions v_i as s_i, we obtain (apart from multiplicative constants) a General Exponential Family by performing the same procedure. In this respect, the representation of the exponential family as a plane in Figure 2.1, makes sense.

As a final note, if the starting measure ν has a finite number of atoms (say k) then the dimension of $\overline{\mathcal{V}}_\nu$ is $k-1$ and therefore $\overline{\mathcal{M}}_\nu$ is a finite dimensional affine space.

2.1.4 Curved Exponential Families

We can attempt to generalize GEF's by simply replacing $\boldsymbol{\lambda}$ by a function h of another parameter, say $\boldsymbol{\xi}$, valued in $\mathbb{R}^q$. If $q = d$ and h is a diffeomorphism then it is just a reparametrization and no generality is gained. If $q < d$, then we enter the world of curved exponential families.

Definition 3 Let

$$\mathcal{GE}(\nu, \boldsymbol{S}) = \{P_{\boldsymbol{\lambda}} \,:\, \boldsymbol{\lambda} \in \Lambda^e(\nu_{\boldsymbol{S}})\}$$

be a General Exponential Family, $\Xi \subset \mathbb{R}^q$ be an open set and $h : \Xi \longrightarrow \Lambda^e(\nu_{\boldsymbol{S}})$ be a function such that:

1. h is a smooth, one-to-one and of rank q everywhere in Ξ and
2. h is a homeomorphism onto its image $h(\Xi) \subset \Lambda^e(\nu_{\boldsymbol{S}})$.

Then the subfamily

$$\mathcal{GE}_0(\nu, \boldsymbol{S}) = \left\{ P_{h(\boldsymbol{\xi})} \in \mathcal{GE}(\nu, \boldsymbol{S}) \,:\, \boldsymbol{\xi} \in \Xi \right\}$$

is called a *curved exponential family (CEF) of* $\mathcal{GE}(\nu, \boldsymbol{S})$.

Define $\Lambda_0^e(\nu_{\boldsymbol{S}}) = h(\Xi)$. Under these conditions, according to standard geometrical terminology, h is an *embedding* and $\Lambda_0^e(\nu_{\boldsymbol{S}})$ is said to be embedded in $\Lambda^e(\nu_{\boldsymbol{S}})$. We may also call the family $\mathcal{GE}_0(\nu, \boldsymbol{S})$ an embedded subfamily of $\mathcal{GE}(\nu, \boldsymbol{S})$. This definition does not depend on the parametrization used.

A Curved Exponential family can be itself a GEF but of lower dimension.

Theorem 4 *Let $\mathcal{GE}(\nu, \boldsymbol{S})$ be a General Exponential Family of dimension d. An embedded subfamily $\mathcal{GE}_0(\nu, \boldsymbol{S}) = \left\{ P_{h(\boldsymbol{\xi})} \in \mathcal{GE}(\nu, \boldsymbol{S}) \,:\, \boldsymbol{\xi} \in \Xi \subset \mathrm{I\!R}^q \right\}$ is a q-dimensional General Exponential Family if and only if there exist a proper affine subspace of $\mathrm{I\!R}^d$ (of dimension q) that contains $h(\Xi)$. An equivalent condition is that h can be written in the form*

$$h(\boldsymbol{\xi}) = \boldsymbol{\lambda}_0 + \boldsymbol{B}\boldsymbol{\xi}$$

for some $\boldsymbol{\lambda}_0 \in \mathrm{I\!R}^d$ and $\boldsymbol{B}$ is $d \times q$ matrix of rank q.

In this case, the embedding h is an affine mapping and also defines an affine subspace. See equation (B.5). So, in the sense described above, an embedded subfamily of a GEF is a GEF itself if it corresponds to an affine subspace of the bigger GEF. See section 4.2 of Kass and Vos (1997) for more details.

2.1.5 The *log-likelihood*

As described in Appendix B, any affine space can be identified with its associated vector space using a particular mapping. It turns out to be that, for the type of affine spaces described above, the identification mapping is equivalent to the well known log-likelihood mapping used in statistical inference. In this way, for the case of exponential families, the log-likelihood acquires the geometrical interpretation of an isomorphism between vector spaces. The following Corollary goes into details.

Corollary 2 *Let $\mathcal{GE}(\nu, \boldsymbol{S})$ be a GEF. Given the arbitrary choice of an element $\bar{\eta}_0 \in \overline{\mathcal{EE}}(\nu, \boldsymbol{S})$ as the origin, there exists an identification between $\overline{\mathcal{EE}}(\nu, \boldsymbol{S})$ and $\overline{\mathcal{V}}_\nu(\boldsymbol{S})$ via the log-likelihood mapping of $\overline{\mathcal{EE}}(\nu, \boldsymbol{S})$, which is defined by $\ell_{\bar{\eta}_0} : \overline{\mathcal{EE}}(\nu, \boldsymbol{S}) \to \overline{\mathcal{V}}_\nu(\boldsymbol{S})$ and*

$$\ell_{\bar{\eta}_0}(\bar{\eta}) = \left[\log\left(\frac{d\eta}{d\eta_0}\right)\right]^* = \left\{\log\left(\frac{d\eta}{d\eta_0}\right) + c \cdot \mathbf{1} \,:\, c \in \mathrm{I\!R}\right\}.$$

Moreover, $\ell_{\bar{\eta}_0}$ is an isomorphism.

Proof: This is just the identification result between affine spaces and its associated vector spaces in Appendix B. ■

Note that this notion of log-likelihood in exponential families is consistent with its statistical meaning. The log-likelihood of any parametric model is defined up to the addition of a constant. It does not make any difference if we use any representative of $\ell_{\bar{\eta}_0}(\bar{\eta})$ for making inferences about the unknown parameters. In this case, the unknown parameters correspond to the natural parameters. For details see Cox and Hinkley (2000) and McCullagh (1999).

The difference with the statistical literature is that the definition of the log-likelihood mapping is given between an extended family of measures (not necessarily the extended exponential family $\overline{\mathcal{EE}}(\nu, \boldsymbol{S})$) and the vector space $\overline{\mathcal{V}}_\nu$, therefore it is not necessarily an isomorphism. This later notion of log-likelihood is the one used throughout this thesis.

2.2 Geometry of Mixture Families

2.2.1 General Mixture Families

Given a σ-finite measure ν on $(\mathrm{I\!R}, \mathcal{B}(\mathrm{I\!R}))$ consider the following spaces

$$\mathcal{D}^m_\nu := \left\{g \in L^1(\nu) \,:\, \int_{\mathrm{I\!R}} g \, d\nu = 1\right\}$$

and

$$\mathcal{V}_\nu^0 := \left\{ s \in L^1(\nu) \; : \; \int_{\mathbb{R}} s \, d\nu = 0 \right\} .$$

For any $g \in \mathcal{D}_\nu^m$ and $s \in \mathcal{V}_\nu^0$, the translation operation will be defined just as the sum of those functions, that is $g \oplus^m s := g + s$.

Theorem 5 $(\mathcal{D}_\nu^m, \mathcal{V}_\nu^0, \oplus^m)$ *is an affine space.*

Proof: First note that, since $L^1(\nu)$ is a vector space then $g \oplus^m s = g + s \in L^1(\nu)$ and

$$\int_{\mathbb{R}} (g \oplus^m s) \, d\nu = \int_{\mathbb{R}} g \, d\nu + \int_{\mathbb{R}} s \, d\nu = 1 + 0 = 1,$$

then $g \oplus^m s \in \mathcal{D}_\nu^m$. Then $\mathcal{V}_\nu^0$ is clearly a real vector space with addition operator the usual one for functions. In this case, the addition for vectors is the same as the translation operator of the affine space but we still keep different symbols to mantain coherence in the expressions. Let $\vec{0}$ be the zero function in $L^1(\nu)$. For any $g \in \mathcal{D}_\nu^m$ and $s_2, s_2 \in \mathcal{V}_\nu^0$, showing that $g \oplus^m \vec{0} = g$ and $(g \oplus^m s_1) \oplus^m s_2 = g \oplus^m (s_1 + s_2)$ is trivial. Given any two functions $g_1, g_2 \in \mathcal{V}_\nu^0$ the only vector s for which $g_2 = g_1 \oplus^m s$ is clearly $s = g_2 - g_1 \in L^1(\nu)$ (the difference of two real functions) and obviously

$$\int_{\mathbb{R}} (g_2 - g_1) \, d\nu = \int_{\mathbb{R}} g_2 \, d\nu - \int_{\mathbb{R}} g_1 \, d\nu = 1 - 1 = 0 \, ,$$

thus $s = g_2 - g_1 \in \mathcal{V}_\nu^0$. ■

We now describe a natural counterpart of General Exponential Families. We will define a General Mixture Family just as the interior of a convex subset of a finite dimensional affine subspace of $(\mathcal{D}_\nu^m, \mathcal{V}_\nu^0, \oplus^m)$.

Take any $f \in \mathcal{D}_\nu^m$ and a set of functions $s_1, \ldots, s_d \in \mathcal{V}_\nu^0$, then clearly

$$\left\{ f \oplus^m \sum_{i=1}^{d} \lambda_i s_i \; : \; (\lambda_1, \ldots, \lambda_d)^t \in \mathbb{R}^d \right\} \qquad (2.5)$$

is an affine subspace of $(\mathcal{D}_\nu^m, \mathcal{V}_\nu^0, \oplus^m)$. Note that this includes the usual concept of convex mixtures in which $f, f_1, \ldots, f_d \in \mathcal{D}_\nu^m$ yield the mixture

$$\left(1 - \sum_{i=1}^d p_i\right) f + \sum_{i=1}^d p_i f_i = f + \sum_{i=1}^d p_i \left[f_i - f\right]$$

where $p_i > 0$ and $\sum_{i=1}^d p_i < 1$ but here both the weights and the f_i are not necessarily positive everywhere. Now consider the following theorem, which is the mixture counterpart of Theorem 1 and Corollary 1.

Theorem 6 *For any $f \in \mathcal{D}_\nu^m$ and $s_1, \ldots, s_d \in \mathcal{V}_\nu^0$ let $D^m(\nu_{\boldsymbol{S}})$ be the set of values $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^t \in \mathbb{R}^d$ for which*

$$g(x; \boldsymbol{\lambda}) = f(x) + \sum_{i=1}^d \lambda_i s_i(x)$$

is positive a.e.[ν]. Then

1. *$D^m(\nu_{\boldsymbol{S}})$ is a convex set in $\mathbb{R}^d$,*
2. *$\log(g(x; \boldsymbol{\lambda}))$ is concave as a function of $\boldsymbol{\lambda}$ a.e.[ν] and*
3. *if the functions $s_1, \ldots, s_d$ are linearly independent a.e.[ν] then $\log(g(x; \boldsymbol{\lambda}))$ is strictly concave as a function of $\boldsymbol{\lambda}$.*

Proof: If $D^m(\nu_{\boldsymbol{S}})$ is nonempty, just take two different values $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ in $D^m(\nu_{\boldsymbol{S}})$. Let $\boldsymbol{\lambda}(u) := u\boldsymbol{\lambda}_1 + (1-u)\boldsymbol{\lambda}_2$. Then, for $u \in [0, 1]$,

$$\begin{aligned} g(x; \boldsymbol{\lambda}(u)) &= (1 - u + u) f(x) + \sum_{k=1}^d \left[u\lambda_1^i + (1-u)\lambda_2^i\right] s_i(x) \\ &= u g(x; \boldsymbol{\lambda}_1) + (1-u) g(x; \boldsymbol{\lambda}_2) \end{aligned}$$

because all affine spaces are convex. Then

$$u g(x; \boldsymbol{\lambda}_1) + (1-u) g(x; \boldsymbol{\lambda}_2) > 0 \qquad \text{a.e.}[\nu]$$

if $g(x; \boldsymbol{\lambda}_1) > 0$ and $g(x; \boldsymbol{\lambda}_2) > 0$ a.e.[ν], which implies that $\boldsymbol{\lambda}(u) \in D^m(\nu_{\boldsymbol{S}})$ for all $u \in [0, 1]$.

Concavity follows from the fact that for each x we can see $g(x;\boldsymbol{\lambda})$ as a linear function of the vector $(1,\boldsymbol{\lambda})^t \in \mathbb{R}^{d+1}$ and linear functions are concave. Because log is a concave increasing function we have that $\log(g(x;\boldsymbol{\lambda}))$ is concave. Now, the Hessian of $\log(g(x;\boldsymbol{\lambda}))$ is given by

$$[H(\boldsymbol{\lambda};x)]_{ij} = -\frac{[s_i(x)s_j(x)]}{[g(x;\boldsymbol{\lambda})]^2}$$

and therefore

$$\boldsymbol{l}^t H(\boldsymbol{\lambda};x)\boldsymbol{l} = -\frac{\left(\sum_{i=1}^{d} l_i s_i(x)\right)^2}{[g(x;\boldsymbol{\lambda})]^2}$$

and so, if the functions $s_1,\dots,s_d$ are linearly independent a.e.$[\nu]$ then $H(\boldsymbol{\lambda};x)$ is negative definite and $\log(g(x;\boldsymbol{\lambda}))$ is strictly concave. ■

Parallel with the definition of General Exponential Families, define $\Lambda^m(\nu_{\boldsymbol{S}}) :=$ interior$(D^m(\nu_{\boldsymbol{S}}))$ and $\nu_f := \int f d\nu$. This leads to the following

Definition 4 The convex subset of positive a.e.$[\nu]$ functions

$$\mathcal{GM}(\nu_f,\boldsymbol{S}) := \left\{ g(x;\boldsymbol{\lambda}) = f(x) + \sum_{i=1}^{d} \lambda_i s_i(x) \ : \ \boldsymbol{\lambda} \in \Lambda^m(\nu_{\boldsymbol{S}}) \right\}$$

for some $f \in \mathcal{D}_\nu^m$ and $\boldsymbol{S} = (s_1,\dots,s_d)^t$, is called the *General Mixture Family* (GMF) generated by ν_f and $\boldsymbol{S}$ provided that

1. the functions $s_1,\dots,s_d \in \mathcal{V}_\nu^0$ are linearly independent a.e.$[\nu]$ and
2. $\Lambda^m(\nu_{\boldsymbol{S}}) \neq \emptyset$.

As in the exponential case, $\boldsymbol{\lambda}$ will be called the *natural parameter* and $\Lambda^m(\nu_{\boldsymbol{S}})$ the *natural parameter space* of the Mixture Family. The boundary of $\Lambda^m(\nu_{\boldsymbol{S}})$ will be called the *Hard Mixture Boundary.* The number d is the *dimension* of the family.

Note. We will assume that $f > 0$ a.e.$[\nu]$. In this case sometimes will be more convenient to work with the isomorphic vector space

$$\mathcal{V}_{\nu_f}^0 = \left\{ s \in L^1(\nu_f) \ : \ \int_{\mathbb{R}} s\, d\nu_f = 0 \right\} \tag{2.6}$$

(instead of $\mathcal{V}_\nu^0$). The isomorphism is clearly given by

$$s \mapsto \frac{s}{f}\,, \qquad s \in \mathcal{V}_\nu^0.$$

In this case it is possible to construct a General Mixture Family by choosing a set $s_1, \ldots, s_d \in \mathcal{V}_{\nu_f}^0$ of linearly independent functions. Then the set

$$\left\{ g(x; \boldsymbol{\lambda}) = f(x) \left[1 + \sum_{i=1}^{d} \lambda_i s_i(x) \right] \,:\, \boldsymbol{\lambda} \in \Lambda^m(\nu_{\boldsymbol{S}}) \right\}$$

is clearly the General Mixture Family generated by ν_f and $fs_1, \ldots, fs_d$.

2.2.2 Mixture Boundaries

Consider the following simple example of a General Mixture Family. Let $f(x), f_1(x)$ be two densities with respect to the measure ν with the same support. Consider the simple mixture

$$g(x; \lambda) = (1 - \lambda) f(x) + \lambda f_1(x) = f(x) + \lambda [f_1(x) - f(x)]$$

which is clearly a GMF. Here $s(x) = s_1(x) = f_1(x) - f(x)$. The natural parameter space $\Lambda^m(\nu_s)$ contains the interval $[0, 1]$ but can be bigger. To see this, take ν as the Lebesgue measure in $\mathbb{R}$ and two members of the negative exponential family defined in example 1

$$f(x) = \theta_0 e^{-\theta_0 x} \qquad \text{and} \qquad f_1(x) = \theta_1 e^{-\theta_1 x} \tag{2.7}$$

for some $\theta_0, \theta_1 > 0$. If $\theta_0 > \theta_1$ then

$$D^m(\nu_s) = \left[0, \frac{\theta_0}{\theta_0 - \theta_1} \right]$$

and if $\theta_0 < \theta_1$ then

$$D^m(\nu_s) = \left[\frac{\theta_0}{\theta_0 - \theta_1}, 1 \right].$$

Definition 5 Let $\mathcal{GM}(\nu_f, \boldsymbol{S})$ be a General Mixture Family with natural parameter space $\Lambda^m(\nu_{\boldsymbol{S}})$. The boundary of any closed set contained in

$$\Lambda^m(\nu_{\boldsymbol{S}})$$

will be called a *Soft Boundary* of the family.

Clearly, if $\theta_0 > \theta_1$ in the above example, the set $\partial(D^m(\nu_s)) = \{0, \theta_0/(\theta_0 - \theta_1)\}$ is the Hard Boundary and $\partial([0,1]) = \{0,1\}$ is an example of a soft boundary. As we will see later, soft boundaries have their own meaning and interpretation depending on the context.

As another example, consider a fixed $\theta_0 > 0$, ν de Lebesgue measure in $\mathbb{R}^+$ and the following general mixture family

$$g(x;\lambda_1,\lambda_2) := f(x;\theta_0) + \lambda_1 \frac{\partial f(x;\theta_0)}{\partial\theta} + \lambda_2 \frac{\partial^2 f(x;\theta_0)}{\partial\theta^2}$$

where $f(x;\theta)$ is again $\theta e^{-\theta x}$. Clearly the derivatives

$$s_i(x) = \frac{\partial^i f(x;\theta_0)}{\partial\theta^i}, \quad i = 1,2$$

belong to $\mathcal{V}_\nu^0$ as the family f is regular. We can write g as

$$f(x;\theta_0)\left[\frac{\theta + \lambda_1 - (\lambda_1 + 2\lambda_2)x + \lambda_2\theta x^2}{\theta}\right]$$

so $D^m(\nu_{\boldsymbol{S}})$ is clearly given by the set of all $(\lambda_1,\lambda_2)^t \in \mathbb{R}^2$ such that the quadratic inside the bracket is positive for all $x > 0$. First, the set $D^m(\nu_{\boldsymbol{S}})$ is given by

$$D^m(\nu_{\boldsymbol{S}}) = \left\{(\lambda_1,\lambda_2)^t \in \mathbb{R}^2 \,:\, \lambda_1 \in [-\theta_0, \theta_0], l(\lambda_1) \le \lambda_2 \le u(\lambda_1)\right\}$$

where

$$l(\lambda_1) = \begin{cases} 0 & \lambda_1 < 0 \\ \dfrac{\theta_0(\theta_0 - \sqrt{\theta_0^2 - \lambda_1^2})}{2} & \lambda_1 \ge 0 \end{cases}$$

and

$$u(\lambda_1) = \frac{\theta_0(\theta_0 + \sqrt{\theta_0^2 - \lambda_1^2})}{2}$$

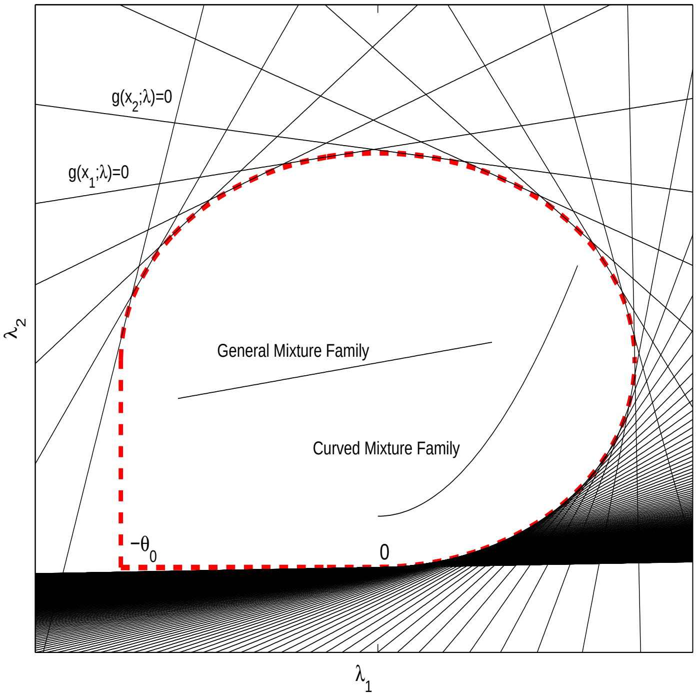

Figure 2.2: Polygon approximation of $\partial(D^m(\nu_{\boldsymbol{S}}))$

Also, the natural mixture families defined when we set $\lambda_2 = 0$ and $\lambda_1 = 0$ have natural parameter space given by $[-\theta_0, 0]$ and $[0, \theta_0^2]$ respectively, so the hard boundary in each case can be easily determined.

Hard Boundaries are not always easy to calculate as in the previous examples. However, given that $D^m(\nu_{\boldsymbol{S}})$ is always a convex set, hard boundaries can always be approximated in the following sense. Following Marriott (2002), for each fixed x, the set of allowable $\boldsymbol{\lambda}$ values forms a half hyperplane with boundary

$$g(x; \boldsymbol{\lambda}) = 0.$$

Thus the set $D^m(\nu_{\boldsymbol{S}})$ is the intersection of a family of half hyperplanes over all x. For a sufficiently dense grid over the support of $g(x; \boldsymbol{\lambda})$ we can approximate (from above) $D^m(\nu_{\boldsymbol{S}})$ using the corresponding polygon. See figure 2.2 which presents the hard boundary for $g(x;\lambda_1,\lambda_2)$ plotted in red and its polygonal approximation.

2.2.3 Curved Mixture Families

As in the exponential case, we can define Curved Mixture Families as embedded subfamilies of a General Mixture Family.

Definition 6 Let $\mathcal{GM}(\nu_f,\boldsymbol{S})$ be a General Mixture Family of dimension d, $\Xi \subset \mathbb{R}^q$ $(q<d)$ be an open set and $h:\Xi \to \Lambda^m(\nu_{\boldsymbol{S}})$ an embedding. Then the subfamily

$$\mathcal{GM}_0(\nu_f,\boldsymbol{S}) = \{g(\boldsymbol{x};h(\boldsymbol{\xi})) \in \mathcal{GM}(\nu_f,\boldsymbol{S}) \,:\, \boldsymbol{\xi}\in\Xi\}$$

is called a *Curved Mixture Family* (CMF) of dimension q.

Theorem 7 *Let $\mathcal{GM}(\nu_f,\boldsymbol{S})$ be a GMF of dimension d. An embedded subfamily is a GMF of dimension q if and only if it is contained in a q-dimensional proper affine subspace of $(\mathcal{D}^m_\nu,\mathcal{V}^0_\nu,\oplus^m)$. Equivalently, if and only if $h(\Xi)$ is contained in a q-dimensional proper affine subspace of $\mathbb{R}^d$.*

In this way, Curved mixture families are of the form

$$g(\boldsymbol{x};\boldsymbol{\xi}) = f_0(\boldsymbol{x}) + \sum_{i=1}^{d} \lambda_i(\boldsymbol{\xi})s_i(\boldsymbol{x})$$

for some functions $\lambda_i : \Xi \subset \mathbb{R}^q \to \mathbb{R}$ that satisfy the embedding assumption. For example, any regular curve inside $D^m(\nu_{\boldsymbol{S}})$ essentially defines a one dimensional curved mixture family. Figure 2.2 presents two examples of CMF's one of which is also a subset of GMF of dimension 1. Indeed, in general, if the image of the embedding h is a subset of a straight line, then this defines a GMF of dimension 1. For example, this is when we can write

$$\boldsymbol{\lambda}(\xi) = \boldsymbol{\lambda}_0 + r(\xi)\boldsymbol{\lambda}_1$$

for some $\boldsymbol{\lambda}_0,\boldsymbol{\lambda}_1 \in \mathbb{R}^d$ and a smooth full rank function $r:\Xi\subset\mathbb{R}\to\mathbb{R}$. This is the mixture counterpart of Theorems 2.3.4 and 4.2.2 of Kass and Vos

(1997). Finally, note that Hard Boundaries for Curved Mixture Families can be calculated as well.

2.2.4 Visualization

As we will consider our ambient space to be the affine space $(\mathcal{D}_\nu^m, \mathcal{V}_\nu^0, \oplus^m)$, it becomes helpful to have an idea concerning the *shape* of our models within this ambient space. Consider the following family affine projections proposed by Marriott (2005). First define, for any density g,

$$E_g[h(X)] := \int_{\mathbb{R}} h(x)g(x)\nu(dx)\,,$$

for any real valued integrable function h.

Theorem 8 *Define, for any integers n_1, n_2, n_3 for which the corresponding integrals converge, the map*

$$\begin{array}{rcl} (\mathcal{D}_\nu^m, \mathcal{V}_\nu^0, \oplus^m) & \rightarrow & \mathbb{R}^3 \\ g(x) & \mapsto & (E_g[X^{n_1}], E_g[X^{n_2}], E_g[X^{n_3}])^t. \end{array}$$

This map has the property that finite dimensional affine subspaces defined in $(\mathcal{D}_\nu^m, \mathcal{V}_\nu^0, \oplus^m)$ (General Mixture Families) map to finite dimensional affine subspaces of $\mathbb{R}^3$ (Affine Planes).

These "moment" projections respect the geometric structure. If a line is straight in $(\mathcal{D}_\nu^m, \mathcal{V}_\nu^0, \oplus^m)$, it will automatically be a straight line in the projection. Furthermore, a point which lies in a convex (or affine) hull in $(\mathcal{D}_\nu^m, \mathcal{V}_\nu^0, \oplus^m)$ will lie in the convex (affine) hull of the image in $\mathbb{R}^3$.

To see how this works, consider the negative exponential family $\mathcal{F}$ defined in example 1. This is a natural exponential family of dimension 1. Also consider the general mixture family $\mathcal{G}$ generated by two points in $\mathcal{F}$ as explained above in (2.7). As expected from Theorem 8, $\mathcal{G}$ appears as a straight line. In Figure 2.3, the hard boundary of $\mathcal{G}$ is represented by the two red points and a soft boundary by the two blue points. Note in this case one point of the boundary

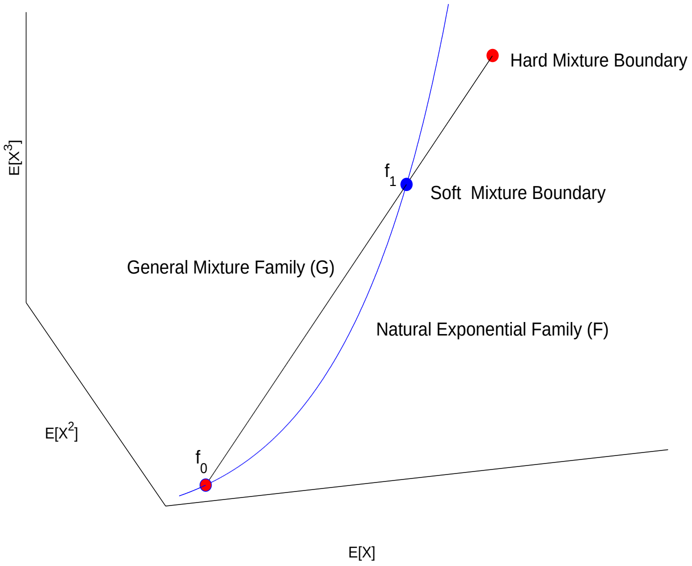

Figure 2.3: Boundaries in General Mixture Families

is hard and soft at the same time. Clearly, this soft boundary is delimiting the convex hull of the two points in $\mathcal{F}$.

Note also that the natural exponential family $\mathcal{F}$ is not a straight since it is not a general mixture family. In general, the intersection between GEF's (or NEF's) and GMF's is not empty, the multinomial distribution being a prominent example inside that intersection, see Amari (1990).

As another example of visualization consider the General Mixture Family

$$g(x;\lambda_1,\lambda_2) = f(x;\theta_0) + \lambda_1 \frac{\partial f(x;\theta_0)}{\partial\theta} + \lambda_2 \frac{\partial^2 f(x;\theta_0)}{\partial\theta^2}$$

described above. The visualization of this family is shown in Figure 2.4. The General Mixture Family corresponds to points in the plane generated by the vectors $\partial f(x;\theta_0)/\partial\theta$ and $\partial^2 f(x;\theta_0)/\partial\theta^2$ and the Hard Boundary for this family is plotted in red. The set inside the Hard Boundary is just the

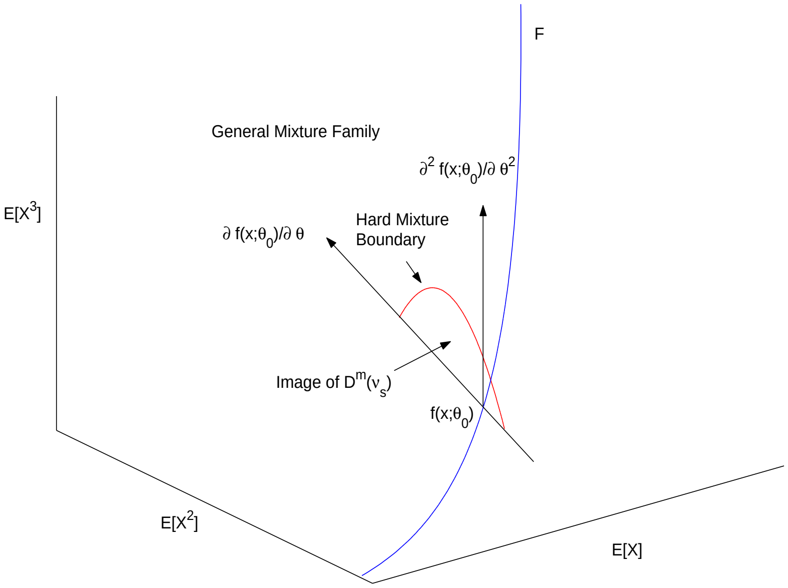

Figure 2.4: Boundaries in General Mixture Families

image of the set $D^m(\nu_{\boldsymbol{S}})$ under the mapping

$$(\lambda_1, \lambda_2) \mapsto g(x; \lambda_1, \lambda_2) \in \mathcal{D}^m_\nu.$$

All the three dimensional graphs presented in this thesis represent projections of the type described in Theorem 8 with respect to some moments.

2.3 Affine Bundles

In this section we consider a simple description of a geometrical object of interest in this thesis, an *Affine Bundle*. Local Mixture Models (introduced in the next chapter) are constructed as a particular subset of an Affine Bundle. In general, the basic idea behind the construction of an Affine Bundle is to attach a *fiber* (in our case an affine space: a General Mixture Family or Exponential Family) to each member of a given base object (in our case a given parametric family). Some related ideas appear in Amari (1987), Barndorff-Nielsen and Jupp (1989) and Barndorff-Nielsen et al. (1991). For general Fibre Bundles see Husemoller (1966).

Given a regular parametric family of densities

$$\mathcal{F} = \{f(x;\theta) \,:\, \theta \in \Theta \subset \mathrm{I\!R}\}\,.$$

with respect to the measure ν defined on $(\mathrm{I\!R}, \mathcal{B}(\mathrm{I\!R}))$, consider a probability density $g \notin \mathcal{F}$ and assume g has the same support (say $\mathcal{S}$) as the members of $\mathcal{F}$. For any $\theta \in \Theta$ and $x \in \mathcal{S}$ we can write the simple formula

$$g(x) = f(x;\theta)\,[1 + s^m(x;\theta)]$$

where

$$s^m(x;\theta) = \frac{g(x) - f(x;\theta)}{f(x;\theta)}$$

is the *relative deviation* between $g(x)$ and $f(x;\theta)$. Clearly,

$$E_\theta\,[s^m(x;\theta)] := \int s^m(x;\theta)\,\nu_\theta(dx) = 0\,, \quad \forall\,\theta \in \Theta\,,$$

where ν_θ the probability measure defined by

$$\nu_\theta(A) := \int_A f(x;\theta)\,\nu(dx)\,, \quad A \in \mathcal{B}(\mathrm{I\!R}) \cap \mathcal{S}.$$

Now, for each fixed $\theta \in \Theta$, consider the space of densities

$$\mathcal{P}^m_\theta := \left\{f(x;\theta)\,[1 + s(x)] \,:\, s(x) \in \mathcal{V}^0_{\nu_\theta} \cap \mathcal{N}^m_\theta\right\}\,,$$

where $\mathcal{V}^0_{\nu_\theta}$ is the vector space (defined in (2.6)) of all possible relative deviations from $f(x;\theta)$ and

$$\mathcal{N}^m_\theta := \left\{s(x) \in L^1(\nu_\theta) \,:\, 1 + s(x) > 0\,, \forall\, x \in \mathcal{S}\right\}.$$

Another formulation of the same idea is to consider that for $\theta \in \Theta$ and $x \in \mathcal{S}$ we can also write

$$\begin{aligned} g(x) &= f(x;\theta)\,\frac{g(x)}{f(x;\theta)} \\ &= f(x;\theta)\exp\{s^e(x;\theta) - \log(1)\} \\ &= f(x;\theta)\exp\{s^e(x;\theta) - \Psi_\theta(s^e(x;\theta))\}\,, \end{aligned}$$

where

$$s^e(x;\theta) = \log\left[\frac{g(x)}{f(x;\theta)}\right]$$

is the *log deviation* between $g(x)$ and $f(x;\theta)$, and

$$\Psi_{\nu_\theta}(s(x)) := \log E_\theta[\exp(s(x))]$$

is the cumulant generating functional of ν_θ evaluated at $s(x)$. Then, for each fixed $\theta \in \Theta$, it also makes sense to consider the following space of densities

$$\mathcal{P}^e_\theta := \left\{ f(x;\theta)\exp\left\{s(x) - \Psi_\theta(s(x))\right\} \,:\, s(x) \in \mathcal{V}^0_{\nu_\theta} \cap \mathcal{N}^e_\theta \right\}$$

where

$$\mathcal{N}^e_\theta := \left\{ s(x) \in L^1(\nu_\theta) \,:\, E_\theta[\exp(s(x))] < \infty \right\}.$$

In practical Statistics, sometimes attention is focused in a particular set of deviations. For example, if μ_θ is the mean under $f(x;\theta)$, the vector

$$s(x;\theta) = (x-\mu_\theta)^2 - E_\theta[(x-\mu_\theta)^2] \in \mathcal{V}^0_{\nu_\theta}$$

represents changes of the variance. Analogously, if m_θ is the median of $f(x;\theta)$, then the vector

$$s(x;\theta) = \mathbb{I}(x \geq m_\theta) - \frac{1}{2} \in \mathcal{V}^0_{\nu_\theta}\,,$$

represents changes in the median. Here $\mathbb{I}(x \in A)$ is the indicator function defined as

$$\mathbb{I}(x \in A) := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Other examples of deviations in $\mathcal{V}^0_{\nu_\theta}$ (assuming some extra conditions as in Appendix A) are the *higher order scores* defined as

$$e_i(x;\theta) := \frac{\dfrac{d^i f(x;\theta)}{d\theta^i}}{f(x;\theta)}\,, \quad i = 1, 2, \ldots$$

The most general type of directions we can consider are completely moving from one density to another as given by

$$s^m(x;\theta) = \frac{g(x) - f(x;\theta)}{f(x;\theta)}$$

and by

$$s^e(x;\theta) = \log\left[\frac{g(x)}{f(x;\theta)}\right] - E_\theta\left[\log\left[\frac{g(x)}{f(x;\theta)}\right]\right].$$

For each $\theta \in \Theta$ consider now a measurable function $\boldsymbol{S}_\theta : \mathbb{R}^d \to \mathbb{R}$ given by $\boldsymbol{S}_\theta(x) = (s_1(x;\theta), \ldots, s_d(x;\theta))^t$ for a set of linearly independent vectors $s_1(x;\theta), \ldots, s_d(x;\theta) \in \mathcal{V}^0_{\nu_\theta}$. Define the following sets

$$\begin{aligned} \Lambda^m(\nu_{\boldsymbol{S}_\theta}) &:= \text{interior}\left\{\boldsymbol{\lambda} \in \mathbb{R}^d \,:\, \langle \boldsymbol{\lambda}, \boldsymbol{S}_\theta(x)\rangle \in \mathcal{N}^m_\theta\right\} \\ \Lambda^e(\nu_{\boldsymbol{S}_\theta}) &:= \text{interior}\left\{\boldsymbol{\lambda} \in \mathbb{R}^d \,:\, \langle \boldsymbol{\lambda}, \boldsymbol{S}_\theta(x)\rangle \in \mathcal{N}^e_\theta\right\} \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product on $\mathbb{R}^d$. These will induce subsets

$$\mathcal{GE}(\nu_\theta, \boldsymbol{S}_\theta) =$$

$$\{f(x;\theta)\exp\left[\langle \boldsymbol{\lambda}(\theta), \boldsymbol{S}_\theta(x)\rangle - \Psi_\theta\left(\langle \boldsymbol{\lambda}(\theta), \boldsymbol{S}_\theta(x)\rangle\right)\right] \,:\, \boldsymbol{\lambda}(\theta) \in \Lambda^e(\nu_{\boldsymbol{S}_\theta})\} \subset \mathcal{P}^e_\theta$$

which is the *General Exponential Family* generated by ν_θ and $\boldsymbol{S}_\theta$ and

$$\mathcal{GM}(\nu_\theta, \boldsymbol{S}_\theta) := \{f(x;\theta)\left[1 + \langle \boldsymbol{\lambda}(\theta), \boldsymbol{S}_\theta(x)\rangle\right] \,:\, \boldsymbol{\lambda}(\theta) \in \Lambda^m(\nu_{\boldsymbol{S}_\theta})\} \subset \mathcal{P}^m_\theta$$

which is the *General Mixture Family* generated by ν_θ and $\boldsymbol{S}_\theta$. Then the families

$$\mathcal{BM}(\mathcal{F}, \mathcal{L}(\boldsymbol{S})) = \bigcup_{\theta\in\Theta} \mathcal{GM}(\nu_\theta, \boldsymbol{S}_\theta)$$

$$\mathcal{BE}(\mathcal{F}, \mathcal{L}(\boldsymbol{S})) = \bigcup_{\theta\in\Theta} \mathcal{GE}(\nu_\theta, \boldsymbol{S}_\theta)$$

represent the set of all possible densities deviating from $\mathcal{F}$ in the directions given by the family of vector spaces $\mathcal{L}(\boldsymbol{S}) = \{\mathcal{L}(\boldsymbol{S}_\theta) \,:\, \theta \in \Theta\}$ where

$$\mathcal{L}(\boldsymbol{S}_\theta) := \left\{\langle \boldsymbol{\lambda}, \boldsymbol{S}_\theta\rangle \,:\, \boldsymbol{\lambda} \in \mathbb{R}^d\right\}.$$

In Figure 2.5, the boundaries of the sets $\Lambda^e(\nu_{\boldsymbol{S}_\theta}), \Lambda^m(\nu_{\boldsymbol{S}_\theta}) \subset \mathbb{R}^d$ are respectively, the Hard exponential and mixture boundaries at θ.

To specify a point in $\mathcal{BE}(\mathcal{F}, \mathcal{L}(\boldsymbol{S}))$ or $\mathcal{BM}(\mathcal{F}, \mathcal{L}(\boldsymbol{S}))$ we can first specify the value of θ from the family $\mathcal{F}$ and then the particular deviation vector $s(x;\theta)$ from it. This latter vector can be specified with the unique linear combination

$$s(x;\theta) = \langle \boldsymbol{\lambda}(\theta), \boldsymbol{S}_\theta\rangle.$$

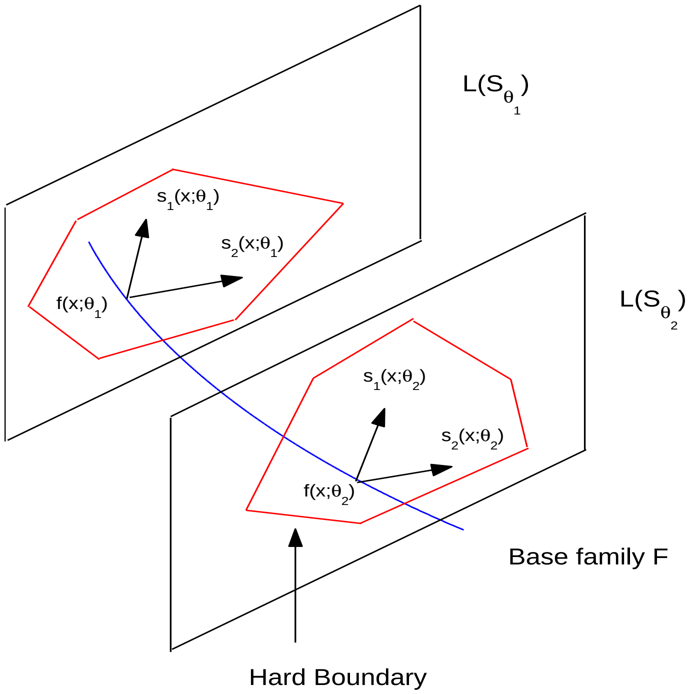

Figure 2.5: Construction of the Family $\mathcal{BE}(\mathcal{F}, \mathcal{L}_{\mathcal{F}})$

Note that $\boldsymbol{\lambda}$ is a function of θ, this is because the set $\{s_1(x;\theta), \ldots, s_d(x;\theta)\}$ is a basis with respect to $\mathcal{V}^0_{\nu_\theta}$, which depends on θ. Therefore, under some further conditions, we can identify any point $g \in \mathcal{BE}(\mathcal{F}, \mathcal{L}(\boldsymbol{S}))$ or $g \in \mathcal{BM}(\mathcal{F}, \mathcal{L}(\boldsymbol{S}))$ with a vector $(\theta, \boldsymbol{\lambda}(\theta)) \in \mathbb{R}^{d+1}$.

For example, in the case where

$$s_i(x;\theta) = \frac{g_i(x;\theta) - f(x;\theta)}{f(x;\theta)}\,,$$

for some parametric families

$$\mathcal{G}_i = \{g_i(x;\theta)\,;\, \theta \in \Theta\}$$

with $i = 1, 2, \ldots, d$, the family $\mathcal{BM}(\mathcal{F}, \mathcal{L}(\boldsymbol{S}))$ has elements of the form

$$\begin{aligned} & f(x;\theta) + \sum_{i=1}^{k} \lambda_i(\theta)\left[g_i(x;\theta) - f(x;\theta)\right] \\ = & \left[1 - \sum_{i=1}^{k} \lambda_i(\theta)\right] f(x;\theta) + \sum_{i=1}^{k} \lambda_i(\theta) g_i(x;\theta)\,, \end{aligned}$$

which, for a fixed θ, is an affine combination of $f(x;\theta)$ $g_1(x;\theta), \ldots, g_d(x;\theta)$ and, when $\lambda_i(\theta) > 0$ for all i and their sum is less than one, we clearly have a linear mixture.

Analogously, in the case where

$$s_i(x;\theta) = \log\left[\frac{g_i(x;\theta)}{f(x;\theta)}\right]\,,$$

the family $\mathcal{BE}(\mathcal{F}, \mathcal{L}(\boldsymbol{S}))$ has elements of the form

$$\left\{E_\theta\left[\prod_{i=1}^{d}\left[\frac{g_i(x;\theta)}{f(x;\theta)}\right]^{\lambda_i(\theta)}\right]\right\}^{-1} f(x;\theta)^{1-\sum_{i=1}^{d}\lambda_i(\theta)} \prod_{i=1}^{d} g_i(x;\theta)^{\lambda_i(\theta)}\,, \tag{2.8}$$

which, for a fixed θ, $\lambda_i(\theta) > 0$ for all i with their sum is less than one, is a geometric mixture.

In the next chapter we introduce a new class of statistical models which is of the form $\mathcal{BM}(\mathcal{F}, \mathcal{L}(\boldsymbol{S}))$. The motivation is the following. If $Q(\theta)$ is a continuous cumulative distribution function defined on Θ which assign most of its mass to a neighbourhood of a fixed $\vartheta \in \Theta$ then

$$\int_\Theta f(x;\theta) dQ(\theta) \;\approx\; f(x;\vartheta)\left[1 + \sum_{i=1}^{d} \lambda_i(\vartheta)\, e_i(x;\vartheta)\right]$$

where

$$\lambda_i(\vartheta) \approx \frac{E_Q[(\theta - \vartheta)^i]}{i!} \tag{2.9}$$

and $e_i(x;\vartheta)$ are the higher order scores defined above.

2.4 Euclidean Structure

So far, we have not considered any further geometrical structure apart from the affine structure. We can make our affine spaces Euclidean (or more properly Riemannian) by defining an inner product on the associated vector spaces. In our case, it is enough to define an inner product on $\mathcal{V}^0_{\nu_f}$, defined in (2.6) as

$$\mathcal{V}^0_{\nu_f} = \left\{ s \in L^1(\nu_f) \ : \ \int_{\mathbb{R}} s(x)\, f(x)\, \nu(dx) = 0 \right\}.$$

Definition 7 The *Fisher Information* inner product $\langle \cdot, \cdot \rangle_f : \mathcal{V}^0_{\nu_f} \times \mathcal{V}^0_{\nu_f} \to \mathbb{R}$ is defined as

$$\langle u, v \rangle_f := \int u(x) v(x)\, f(x)\, \nu(dx).$$

This is clearly bilinear symmetric and positive definite. This permits the introduction of important concepts like orthogonality. It turns out that the orthogonality induced by this inner product is the same as the usual Fisher orthogonality used in parametric statistical inference , see for example Cox and Reid (1987) and Barndorff-Nielsen and Cox (1994). As we will see in the next chapter it is convenient statistically to construct the fibers in our bundles to be Fisher orthogonal to the tangent space of the base family $\mathcal{F}$. In particular, Fisher orthogonality implies Local Mixture models are well identified also the well known asymptotic inferential separation between interest and nuisance parameters.

2.4.1 Frequentist interpretation of the space $\mathcal{V}^0_{\nu_\vartheta}$

Clearly, the space $\mathcal{V}^0_{\nu_\vartheta}$ is the vector space of random variables that have null expectation with respect to the density $f(x;\vartheta)$. That, if $s(x;\vartheta) \in \mathcal{V}^0_{\nu_\vartheta}$ and X is a random variable with density $f(x;\vartheta)$ then the random variable $S(X;\vartheta)$ will have zero expectation. This view of the space $\mathcal{V}^0_{\nu_\vartheta}$ has been used many times in the Frequentist Statistics literature concerning the issue of testing for the presence of mixing. Sometimes it appears under a different name such as tests for over-dispersion, tests of homogeneity (heterogeneity) or frailty tests.

The following is mainly based on the following references: Zelterman (1988), Zelterman and Chen (1988), Dean (1992), Chesher (1984) and Cox (1983). The general idea is that if a random sample $X_1, X_2, \ldots, X_n$ is generated by the density $f(x;\vartheta)$ for some known value $\vartheta \in \Theta$ then the average

$$\frac{1}{n}\sum_{i=1}^{n} S(X_i;\vartheta)$$

should have zero expectation. So, given a particular data set $x_1, \ldots, x_n$, large values of the statistic

$$T(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} S(x_i;\vartheta) \tag{2.10}$$

represent empirical evidence against the original hypothesis that the data was generated by $f(x;\vartheta)$. If the distribution of $T(\boldsymbol{x})$ can be calculated under such hypothesis then classical statistical inference can be performed. As this is not usually the case (or it is difficult to do so), usually, asymptotic inference is used. Assuming that hypothesis as true, by the Central Limit Theorem,

$$\frac{\sqrt{n}\,T(\boldsymbol{x})}{\sqrt{v(\vartheta)}} \xrightarrow{d} N(0,1)$$

as n tends to infinity, where

$$v(\vartheta) := \int [S(x;\vartheta)]^2 f(x;\vartheta)\,\nu(dx).$$

For the validity of this approach we need to assume a particular value of ϑ which does not always makes too much sense. Most of the time, the important assumption to be empirically tested is that the data we are observing come from $\mathcal{F}$, that is, come from $f(x;\vartheta)$ for some unknown $\vartheta \in \Theta$. In that case, the unknown ϑ in (2.10) is replaced by a "good" estimate of it.

To the best of our knowledge, all the literature concerning testing for the presence of mixing concentrates on the so-called *dispersion score statistic* defined as

$$DS(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} \frac{f^{(2)}(x_i;\hat{\vartheta})}{f(x_i;\hat{\vartheta})}$$

where $\hat{\vartheta}$ is an asymptotically efficient estimator of ϑ under the model $\mathcal{F}$. We will use the results when $\hat{\vartheta}$ is the maximum likelihood estimate of the unknown ϑ.

It can be shown that

$$\frac{\sqrt{n}\, DS(\boldsymbol{x})}{\sqrt{v(\hat{\vartheta})}} \xrightarrow{d} N(0,1)$$

when $n \to \infty$, where now

$$v(\hat{\vartheta}) = v_{22}(\hat{\vartheta}) - \frac{[v_{12}(\hat{\vartheta})]^2}{v_{11}(\hat{\vartheta})}$$

and where

$$\begin{aligned} v_{11}(\vartheta) &= \int \left(\frac{f^{(1)}(x;\vartheta)}{f(x;\vartheta)} \right)^2 f(x;\vartheta)\, \nu(dx) \\ v_{22}(\vartheta) &= \int \left(\frac{f^{(2)}(x;\vartheta)}{f(x;\vartheta)} \right)^2 f(x;\vartheta)\, \nu(dx) \\ v_{12}(\vartheta) &= \int \frac{f^{(1)}(x;\vartheta)}{f(x;\vartheta)} \cdot \frac{f^{(2)}(x;\vartheta)}{f(x;\vartheta)}\, f(x;\vartheta)\, \nu(dx) \end{aligned}$$

In statistical terms, $v(\vartheta)$ is the residual variance after taking into account the variability induced by $\hat{\vartheta}$. In geometrical terms $v(\vartheta)$ has also a very clear meaning. It is easy to show that

$$v(\vartheta) = \left\| \frac{f^{(2)}(x;\vartheta)}{f(x;\vartheta)} - \frac{v_{12}(\vartheta)}{v_{11}(\vartheta)} \frac{f^{(1)}(x;\vartheta)}{f(x;\vartheta)} \right\|_{\vartheta}^2$$

where $\| \cdot \|_{\vartheta}$ is the norm induced by $\langle \cdot, \cdot \rangle_{f_\vartheta}$. So, $v(\vartheta)$ is just the squared norm of the residual vector after projecting $f^{(2)}(x;\vartheta)/f(x;\vartheta)$ in $f^{(1)}(x;\vartheta)/f(x;\vartheta)$ orthogonally.

This is quite clear from the statistical point of view because it is well known that the asymptotic properties of the maximum likelihood estimator in the regular case are basically derived from the asymptotic behavior of the so-called score statistic

$$\frac{f^{(1)}(x;\hat{\vartheta})}{f(x;\hat{\vartheta})}$$

So, a simple asymptotic statistical test of the assumption that, an observed data set $x_1, x_2, \ldots, x_n$ comes from the density $f(x;\vartheta)$, for some unknown ϑ, is to reject that assumption if the observed absolute value of the standardized dispersion score statistic

$$\frac{\sqrt{n}\, DS(\boldsymbol{x})}{\sqrt{v(\hat{\vartheta})}}$$

exceeds some critical value of the standard normal distribution. Obviously this test is expected to work well only for large sample sizes. It can be corrected in may ways to improve its performance but what we really want to emphasize here is the use of the vectors in $\mathcal{V}^0_{\nu_\vartheta}$ as random variables to perform this kind of statistical tests.

In this part of the thesis we propose the use of a novel approach of looking at the same problem based on the underlying geometrical structure rather than on frequentist properties. The basic idea is very simple. Instead of treating the set of vectors

$$\left\{ \frac{f^{(2)}(x;\vartheta)}{f(x;\vartheta)}, \frac{f^{(3)}(x;\vartheta)}{f(x;\vartheta)}, \ldots, \frac{f^{(d)}(x;\vartheta)}{f(x;\vartheta)} \right\} \in \mathcal{V}^0_{\nu_\vartheta}$$

as random variables we are going to exploit the affine structure present in the space $(\mathcal{D}^m_\nu, \mathcal{V}^0_\nu, \oplus^m)$ and use families of densities of the form of general mixture families

$$\mathcal{G}_{\mathcal{F}} = \left\{ f(x;\vartheta) \left[1 + \sum_{i=2}^{k} \lambda_i(\vartheta) \frac{f^{(i)}(x;\vartheta)}{f(x;\vartheta)} \right] \,:\, \vartheta \in \Theta \,, \{\lambda_i\}_{i=2}^{k} \right\}.$$

This type of parametric families will be taken to be low dimensional, flexible and interpretable. The parameters ϑ and $\lambda_i(\vartheta)$ will be estimated from the data in a simple way using standard geometrical and classical statistical arguments.

If the elements in the above set of vectors are linearly independent as functions of x then they form a base for a subspace of $\mathcal{V}^0_{\nu_\vartheta}$. Then it will be clear that the $\lambda_i(\vartheta)$ are simply the affine coordinates with respect to that base and to the choice of the origin $f(x;\vartheta)$. So, for the elements of $\mathcal{G}_{\mathcal{F}}$ to be probability densities we need to restrict the values that the λ_i can take.

Chapter 3

Local Mixture Models

In this chapter the class of *Local Mixture Models* is introduced. We first develop detailed asymptotic expansions to show that mixtures of regular parametric families have an *affine-type* behavior when the mixing distribution is small. Motivated on this affine behavior, local mixture models are then defined as a bundle of general mixture families and some of its properties are described. *True Local Mixture Models* are also introduced as a smaller set of local mixture models that represent genuine mixtures more faithfully. The properties of NEF-QVF's are exploited to construct local mixture models that are identifiable, flexible and interpretable statistically. *Regular Proper Dispersion Models* are chosen to be the workhorse class of small mixing distributions.

3.1 Introduction

Mixture models can be found in a wide variety of statistical applications. Important general references are Titterington, Smith, and Makov (1985), Lindsay (1995) and McLachlan and Peel (2001) and the particular references therein. The geometry of this type of models has been studied in essentially two distinct ways: from the point of view of Convex Geometry, see for example the excellent account by Lindsay (1995) and from the point of view of Differential Geometry, see Amari (1990) or Kass and Vos (1997). But the Convex Geometry approach has been by far more studied than the Differential Geometric one. The seminal work on purely differential-geometric ideas applied to inference in mixture models is Marriott (2002). We basically follow Marriott's work.

We begin with some mathematical definitions. Let $\mathcal{F}$ be a family of cumulative distribution functions (cdf's) on $\mathbb{R}$ and let $\mathcal{A}$ be a family of probability measures defined on a Borel σ-algebra of subsets of $\mathcal{F}$. Then, if $\lambda \in \mathcal{A}$,

$$\int_{\mathcal{F}} g(F)\, d\lambda(F)$$

is defined in the usual manner for measurable mappings $g : \mathcal{F} \to \mathbb{R}$. If $g = g_x(F) = F(x)$ for some x, the integral above becomes

$$G(x;\lambda) := \int_{\mathcal{F}} F(x)\, d\lambda(F). \tag{3.1}$$

Definition 8 The resultant distribution function G will be called a *mixture* or more specifically a λ-*mixture* of $\mathcal{F}$, provided the *mixing measure* λ does not assign measure one to a particular member of $\mathcal{F}$. For given $\mathcal{F}$ and $\mathcal{A}$, the family

$$\mathcal{G}(\mathcal{F}) = \{G(x;\lambda) \,:\, \lambda \in \mathcal{A}\}$$

will be called the class of $\mathcal{A}$-mixtures over $\mathcal{F}$.

Thus, the term mixture, as employed here, means a genuine weighted average of cdf's. In particular, there may exist a parametrization of the family $\mathcal{F}$ with image an open region $\Theta \subset \mathbb{R}$, that is

$$\mathcal{F} = \{F(x;\theta) \,:\, \theta \in \Theta\}\,.$$

Let $\mathcal{Q} = \{Q(\theta)\}$ denote a class of d-dimensional cdf's with support Θ and assume $F(x;\theta)$ is measurable on $\mathbb{R}^{d+1}$. Then $\mathcal{A}$ may be taken to be the corresponding class of Lebesgue-Stieljes measures $\{\lambda_Q \,:\, Q \in \mathcal{Q}\}$ on $\mathbb{R}^d$ induced by $\mathcal{Q}$ and (3.1) becomes

$$G(x;Q) := \int_{\Theta} F(x;\theta)\, dQ(\theta). \tag{3.2}$$

If $\mathcal{F} = \{F(x;\theta) \,:\, \theta \in \Theta\}$ is a family of discrete distributions whose discontinuity points are independent of θ, then the resultant mixture will be discrete, inheriting the common points of discontinuity. On the other hand if $F(x;\theta)$ is absolutely continuous for every θ then $f(x;\theta) := dF(x;\theta)/dx$ is measurable on $\mathbb{R}^{d+1}$ and, by Fubini's Theorem, the resultant mixture $G(x;Q)$ is absolutely continuous with density

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

Definition 9 Given a family $\mathcal{F}$ of absolutely continuous cdf's and the corresponding family of densities (which we will also call $\mathcal{F}$). The density $g(x;Q)$ will be called a *Q-mixture density* of $\mathcal{F}$, provided the *mixing distribution* Q does not assign measure one to a particular point in Θ. For a given $\mathcal{F}$, the family

$$\mathcal{G}(\mathcal{F}) := \{g(x;Q) \,:\, Q \in \mathcal{Q}\}$$

will be called the class of $\mathcal{Q}$-mixtures over $\mathcal{F}$.

Appealing to the change of variable theorem, note that this definition does not depend on the way we parametrize $\mathcal{F}$.

We will be interested in mixtures when the mixing distribution Q is *small* in the following sense. Roughly, we will call a mixing distribution *local* when its associated density is close to a delta function, that is, when it is highly peaked around some value of its support and assign almost all of its mass to a small neighborhood around the same value.

From the point of view of Statistics we are interested in such kind of mixtures because they represent an important type of *deviations* from the family $\mathcal{F}$. Usually, practitioners of Statistics use a simple (low dimensional) parametric family of distributions $\mathcal{F}$ to model the random behavior of a particular phenomenon they are interested in. This is mainly because of: matter of simplicity, some of the parameters have a "real world meaning" or the parameters represent a systematic part of the phenomenon which is of interest. Often, some incompatibility is found between what has been observed and what $\mathcal{F}$ can predict, so a more flexible family than $\mathcal{F}$ is needed. For example, when dealing with univariate data and if $\mathcal{F}$ is one-dimensional, it is often found that the data have more variability than the variability predicted under the model. Applied Statisticians call this situation *over-dispersion.* This can happen, for example, because $\mathcal{F}$ assumes some kind of homogeneity in the population which is not really present and that is why one observes more variability than expected under $\mathcal{F}$. An attractive generalization of a family $\mathcal{F}$ that can handle such situation is precisely mixing over some of the parameters in $\mathcal{F}$. Under some conditions, the resulting mixtures will always have larger variances than those of the elements of $\mathcal{F}$. Apart from some particular instances, mixtures are considerably more complicated to manage (in statistical terms) compared to $\mathcal{F}$. So mixtures of $\mathcal{F}$ have more flexibility but are in general less tractable than $\mathcal{F}$.

The small mixing assumption is intended to be responsible for retaining some of the tractability in $\mathcal{F}$ lost by mixing. The geometry of mixtures can be very complicated whenever a reasonable large class of mixing distributions is allowed. When the mixing distribution is continuous or discrete but with an unknown number of components, the Geometry is essentially infinite dimensional and typically the mixtures have singular or boundary points. In contrast, the Geometry of an Exponential Family is that of a finite-dimensional smooth affine manifold. It is of course no coincidence that in the second case, inferential theories are well understood and in the first case difficult. It might be expected that any restriction of the general mixture family structure which maintains a simple geometry will simplify inferential problems.

The most fruitful form of simplification in both Statistics and Geometry is that of *local analysis*, for example when using an asymptotic expansion based on some form of Taylor's Theorem in Statistics or studying the tangent space of a manifold in Geometry. However, there are at least two ways of localising a mixture family. The most obvious is to look at a local neighborhood of $g(x;Q)$; that is, to look only at densities which are close to $f(x;Q)$ and then make appropriate approximations. This approach has problems because of the infinite-dimensional nature mixtures families, since an open subset of an infinite dimensional space is still infinite dimensional. A more productive approach is to assume the mixing distribution has only local support in the parameter space. Thus the localizing is done at the mixing distribution level.

Based mainly on the work of Marriott (2002), we propose new families of distributions that behave like mixtures of a given simple parametric family $\mathcal{F}$

when the mixing is small. These families gain flexibility but at the same time are statistically tractable and have similar properties to that of exponential family models.

We will focus on $\mathcal{Q}$-mixtures over $\mathcal{F}$ where $\mathcal{F}$ will be a family of the type called Real Natural Exponential Families with Quadratic Variance Functions (NEF-QVF) (see Morris (1982) and Letac (1992)) and $\mathcal{Q}$ will be a two dimensional class of mixing distributions inside the so-called *Proper Dispersion Models* (see Appendix C). The usual way to parametrize this latter class (which is particularly suitable for us here) is the standard form:

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

where the parameter ϑ is an indicator of the *position* of the mixing distribution and the parameter ϵ controls the *dispersion* of the mixing distribution around ϑ. Clearly, a suitable choice of the parameter ϑ will depend on the way the family $\mathcal{F}$ is parametrized. If the family $\mathcal{F}$ is parametrized in a different way, for example using a diffeomorphism $\mu = h(\theta)$, then the family $\mathcal{Q}$ will be parametrized as

$$\mathcal{Q} = \{Q(\mu; m, \epsilon) \,:\, m \in h(\Theta),\, \epsilon > 0\} \,,$$

where $m = h(\vartheta)$, which is again a proper dispersion model in standard form but with different support.

We are going to deal only with families of densities $\mathcal{F}$ that satisfy the regularity conditions on Appendix A . Those conditions basically ensure that $\mathcal{F}$ can be treated as a smooth manifold. In particular, they imply the existence of a σ-finite measure ν that dominates all elements in $\mathcal{F}$. Also we are going to deal only with families $\mathcal{Q}$ of continuous cdf's. An approach to the discrete case can be found in Anaya-Izquierdo and Marriott (2006) and for general discrete mixture models see McLachlan and Peel (2001).

3.2 Taylor-type expansions

Let us now study the behavior of the mixture

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

when Q is small. Assume Q has all the necessary moments and denote by $f^{(i)}(x;\vartheta)$ the i-th partial derivative of $f(x;\theta)$ with respect to θ evaluated at ϑ. It is well know that, for any $k \in \mathbb{N}$, the Taylor polynomial

$$f(x;\vartheta) + \sum_{i=1}^{d} \frac{f^{(i)}(x;\vartheta)}{i!}\,(\theta-\vartheta)^i \tag{3.3}$$

is a good approximation to $f(x;\theta)$, provided that θ is close to ϑ. Denote by $\bar{\Theta}$ the neighborhood of ϑ for which this approximation is good in a specific sense, for example, using certain bounds for the remainder. Integrating this expression with respect to $dQ(\theta)$ we get

$$f(x;\vartheta) + \sum_{i=1}^{d} \frac{f^{(i)}(x;\vartheta)}{i!}\,E_Q[(\theta-\vartheta)^i]\,. \tag{3.4}$$

This last expression will be a good approximation to $g(x;Q)$ if Q is small in the sense that it assigns negligible probability to $\Theta - \bar{\Theta}$.

Suppose now we reparametrize $\mathcal{F}$ using a diffeomorphism $h : \Theta \to \Phi$ and proceed as before. Define $\tilde{f}(x;\phi) := f(x;h^{-1}(\phi))$ and $\varphi := h(\vartheta)$. Denote by $\bar{\Phi}$ the set of values of ϕ for which the Taylor polynomial

$$\tilde{f}(x;\varphi) + \sum_{i=1}^{d} \frac{\tilde{f}^{(i)}(x;\varphi)}{i!}\,(\phi-\varphi)^i$$

is a good approximation to $\tilde{f}(x;\phi)$. Now the approximation of

$$\tilde{f}(x;\varphi) + \sum_{i=1}^{d} \frac{\tilde{f}^{(i)}(x;\varphi)}{i!}\,E_{\tilde{Q}}[(\phi-\varphi)^i] \tag{3.5}$$

to $g(x;\tilde{Q}) = g(x;Q)$ will be good if the transformed mixing distribution $\tilde{Q}$ assigns negligible probability to the set $\Phi - \bar{\Phi}$. Note that, by continuity of h, $\bar{\Phi}$ is clearly contained in $h(\bar{\Theta})$.

Let us look to a particular situation. Assume $\Theta = \mathbb{R}$ and consider the following family of mixing distributions

$$\mathcal{Q}_{ls} = \left\{ Q(\theta;\vartheta,\epsilon) = Q_0\left(\frac{\theta-\vartheta}{\sqrt{\epsilon}}\right) \,:\, \epsilon > 0 \right\}\,,$$

where ϑ is some real number and Q_0 is a (nondegenerate) continuous cdf that has all the necessary moments. By this construction $\theta = \vartheta + \sqrt{\epsilon} z$, where z is random variable with cdf Q_0. There is no loss of generality if we assume z has variance equal to one. Clearly, θ converges in distribution to the constant ϑ as $\epsilon \downarrow 0$. Because, convergence in distribution is preserved by diffeomorphisms, we also have that $\phi = h(\theta)$ converges in distribution to $\varphi = h(\vartheta)$ as $\epsilon \downarrow 0$.

We have in this case

$$E_Q[(\theta - \vartheta)^i] = \epsilon^{i/2} E_{Q_0}[z^i]\,,$$

where $z = (\theta - \vartheta)/\sqrt{\epsilon}$. If we take ϑ to be the expectation of θ and denote it by ϑ^* then clearly z has zero mean and

$$\begin{aligned} Var_Q[\theta] &= \epsilon \\ E[(\theta - \vartheta^*)^i] &= \epsilon^{i/2} E_{Q_0}[z^i]\,, \quad i \geq 3\,, \end{aligned}$$

So, for small enough ϵ we should expect a good approximation to the mixture $g(x; Q)$ by

$$f(x; \vartheta^*) + \sum_{i=2}^{d} \frac{f^{(i)}(x; \vartheta^*)}{i!}\, E_Q[(\theta - \vartheta^*)^i]$$

because the distribution of θ is "concentrating" around its mean ϑ^* for small values of ϵ. Note the vanishing of the first derivative term in the Taylor expansion, because we are assuming ϑ^* is the mean of θ. It is important to remind the reader about the fact that here, and actually all along this thesis, we are assuming the existence of all the necessary moments.

Now consider a reparametrization of the family $\mathcal{F}$ by using the diffeomorphism $h : \Theta \to \Phi$. As before, define $\phi = h(\theta)$ and $\varphi^* = h(\vartheta^*)$. Using Taylor's Theorem, we have the approximations

$$\begin{aligned} E_{\tilde{Q}}[\phi] &\approx \varphi^* + \frac{h''(\vartheta^*)}{2}\epsilon \\ Var_{\tilde{Q}}[\phi] &\approx [h'(\vartheta^*)]^2 \epsilon \end{aligned}$$

valid again for small ϵ. Note these simple approximations are telling us that, the transformed random variable $\phi = h(\theta)$ is behaving like $\phi^* + \sqrt{\epsilon} w$ for some (zero mean) random variable w when ϵ is small and this now implies a good approximation to the mixture $g(x;\tilde{Q}) = g(x;Q)$ by

$$\tilde{f}(x;\varphi^*) + \sum_{i=2}^{d} \frac{\tilde{f}^{(i)}(x;\varphi^*)}{i!}\, E_{\tilde{Q}}[(\phi-\varphi^*)^i]\,.$$

Also note this argument is valid for any diffeomorphism h. For example, if Q_0 is a standard normal distribution, then θ is also Gaussian with mean ϑ and variance ϵ. Now assume $\phi = h(\theta) = \exp(\theta)$. The distribution of ϕ is called the *log-normal distribution*. It is well known that

$$\frac{\phi-\varphi}{\varphi\sqrt{\epsilon}} \overset{d}{\rightarrow} W, \qquad \text{as } \epsilon \downarrow 0\,,$$

where W has standard normal distribution. Here $\overset{d}{\rightarrow}$ means convergence in distribution. So, for small enough ϵ we can approximate the distribution of ϕ by a normal with mean φ and variance $\varphi^2\,\epsilon$. This means that if ϵ is sufficiently small we can ensure the distribution $\tilde{Q}$ for ϕ assigns negligible probability to $\Phi - \bar{\Phi}$. Now, more generally, using the so called *Delta-Method* (Serfling (1980)) we have that convergence in distribution to normality is preserved by diffeomorphisms. Then

$$\frac{h(\theta)-h(\vartheta^*)}{\sqrt{[h'(\vartheta^*)]^2\epsilon}} \overset{d}{\rightarrow} W, \qquad \text{as } \epsilon \downarrow 0\,.$$

So, as long as we have convergence to normality in some parametrization, we will have convergence to normality in any diffeomorphic reparametrization.

Summarizing, when the family $\mathcal{F}$ has parametrization θ with image $\Theta = \mathbb{R}$ then the family of mixing distributions

$$\mathcal{Q}_{ls} = \left\{ Q(\theta;\vartheta,\epsilon) = Q_0\left(\frac{\theta-\vartheta^*}{\sqrt{\epsilon}}\right) \,:\, \epsilon > 0 \right\}\,,$$

represents a geometrically convenient way of modeling local $\mathcal{Q}$-mixtures of $\mathcal{F}$ in the sense that if the parameter ϵ is small, the distributions in $\mathcal{Q}_{ls}$ tend to be small no matter how $\mathcal{F}$ is being reparametrized. Note the important fact that all the previous discussion does depend on Q_0 only through some of its first moments. The subscript "*ls*" stands for *location-scale* as the parameters ϑ^* and $\sqrt{\epsilon}$ act on Q_0 changing its location and scale.

Approximations like (3.4) and (3.5) basically rely on the possibility of integrating term by term an expansion of the integrand. This would require some further assumptions, like the integrand being exponentially decaying.

To acknowledge this, consider now the following family of mixing distributions

$$\mathcal{Q}_{ld} = \left\{ dQ(\theta;\vartheta,\epsilon) = c(\epsilon)\exp\left(-\frac{d_0(\theta-\vartheta)}{2\epsilon}\right) d\theta \,:\, \epsilon \in (0,\epsilon_0) \right\} ,$$

where $\epsilon_0 > 0$, ϑ is some real number and $d_0(z)$ is a non-negative smooth function such that

$$\begin{aligned} d_0(0) &= 0 \\ d_0(z) &> 0\,, \quad \text{for } z \neq 0. \end{aligned} \tag{3.6}$$

Clearly, we can write the associated mixtures by

$$g(x;Q) = \int_\Theta f(x;\theta) dQ(\theta;\vartheta,\epsilon) = c(\epsilon)\int_{\mathbb{R}} f(x;\vartheta+z)\exp\left(-\frac{d_0(z)}{2\epsilon}\right) dz\,, \tag{3.7}$$

where $\theta = \vartheta + z$. Formally using Laplace's Method (see for example Wong (2001)), we are going to obtain an asymptotic expansion of the form

$$g(x;Q) \sim f(x;\vartheta) + \sum_{i=1}^{k} A_i(\epsilon) f^{(i)}(x;\vartheta) + R(x,\vartheta,\epsilon) \tag{3.8}$$

as $\epsilon \downarrow 0$, for some set of functions $\{A_i(\epsilon)\,,\, i = 1,2,\ldots\}$. Here, the $\sim$ symbol means that the right hand side of (3.8) behaves like $g(x;Q)$ when ϵ goes to zero and x and ϑ are kept fixed. This is a formal expansion and essentially does not need any further assumption, apart from the regularity assumption of $\mathcal{F}$ and conditions in (3.6). The restriction here of having ϵ defined in some bounded interval is to ensure the existence of appropriate moments of the mixing distribution.

Moreover, we will also show that the functions $A_i(\epsilon)$ have now an asymptotic relation with the moments of the mixing distribution Q. Explicitly we will show that

$$\frac{E_Q[(\theta-\vartheta)^i]}{i!} \sim A_i(\epsilon) + R_i(\epsilon) \qquad i = 1,2,\ldots$$

as $\epsilon \downarrow 0$. Here we note that R_i is a different remainder from that in expression (3.8). If we reparametrize $\mathcal{F}$ using the diffeomorphism $h(\theta) = \phi$ then clearly we can write

$$\int_{\Phi} \tilde{f}(x;\phi) d\tilde{Q}(\phi;\varphi,\epsilon) = \int_{\Phi} \tilde{f}(x;\phi) \left| \frac{dh^{-1}(\phi)}{d\phi} \right| c(\epsilon) \exp\left(-\frac{\tilde{d}(\phi;\varphi)}{2\epsilon} \right) d\phi \, ,$$

where $\varphi = h(\vartheta)$ and

$$\tilde{d}(\phi;\varphi) := d_0(h^{-1}(\phi) - h^{-1}(\varphi)) \, ,$$

then, by making the further change of variable $\phi = z + \varphi$ and defining

$$\begin{aligned} \tilde{V}^{-1/2}(\phi) &:= \left| \frac{dh^{-1}(\phi)}{d\phi} \right| \\ \tilde{d}_0(z) &:= \tilde{d}(\varphi + z;\varphi) \, , \end{aligned}$$

we turn the mixture into

$$g(x;\tilde{Q}) = \int_{\Phi - \varphi} \tilde{f}(x;\varphi + z) \tilde{V}^{-1/2}(\varphi + z) \exp\left(-\frac{\tilde{d}_0(z)}{2\epsilon} \right) d\phi$$

which is of the same form as (3.7) but now with an additional function multiplying the density. So, again we can formally apply Laplace's Method to obtain an expansion of the form

$$g(x;\tilde{Q}) \sim \tilde{f}(x;\varphi) + \sum_{i=1}^{d} \tilde{A}_i(\varphi,\epsilon) \tilde{f}^{(i)}(x;\varphi) + \tilde{R}(x,\varphi,\epsilon) \qquad (3.9)$$

as $\epsilon \downarrow 0$, for some set of functions $\tilde{A}_i(\varphi,\epsilon)$ which now can depend on φ. The relation with the moments of $\tilde{Q}$ is preserved as we will show that

$$\frac{E_{\tilde{Q}}[(\phi - \varphi)^i]}{i!} \sim \tilde{A}_i(\varphi,\epsilon) + \tilde{R}_i(\varphi,\epsilon) \qquad i = 1, 2, \ldots$$

as $\epsilon \downarrow 0$.

As in the Taylor expansions above, we will show that here we can also have control over the orders (for small ϵ) of the functions $\tilde{A}_i$ and on the remainder terms. For example, we will show that $\tilde{A}_1(\varphi,\epsilon)$ and $\tilde{A}_2(\varphi,\epsilon)$ are both $O(\epsilon)$ for each fixed φ, $\tilde{A}_3(\varphi,\epsilon)$ and $\tilde{A}_4(\varphi,\epsilon)$ are both $O(\epsilon^2)$ and so on. This clearly differs from the Taylor expansions above. Specifically, we are going to work with $d = 4$ and then the remainder terms will be shown to be of order ϵ^3. Finally, under mild conditions, mixing distributions in the family $\mathcal{Q}_{ld}$ are asymptotically normal. Explicitly,

$$\frac{\theta-\vartheta}{\sqrt{\epsilon}} \xrightarrow{d} N(0,1) \qquad \text{as } \epsilon \downarrow 0$$

and clearly, as before, this inherits the asymptotic normality under transformations, obtaining

$$\frac{\phi-\varphi}{\sqrt{\epsilon\,\tilde{V}(\varphi)}} \xrightarrow{d} N(0,1) \qquad \text{as } \epsilon \downarrow 0\,,$$

and justifying the definition of $\tilde{V}(\phi)$.

More generally, asymptotic expansions like (3.9) will not only be valid for families of mixing distributions like $\mathcal{Q}_{ld}$, but for the more general class of *Regular Proper Dispersion Models* of which $\mathcal{Q}_{ld}$ is just a particular subfamily known as the *location dispersion family.* The standard form of the densities in this general class is

$$dQ(\theta;\vartheta,\epsilon) = a(\epsilon)V^{-1/2}(\theta)\,\exp\left(-\frac{d(\theta;\vartheta)}{2\epsilon}\right)\,,$$

where $d(\theta;\vartheta)$ is a regular unit deviance function and $V(\theta)$ is the associated unit variance function (see Appendix C). The parameters ϑ and ϵ are called the *position* and *dispersion* parameters respectively. From the observations in the example above, Laplace's method can be formally applied over any diffeomorphic parametrization. The asymptotic normality result also applies to general proper dispersion models. Finally, note the obvious locality of the densities in the regular proper dispersion class. As ϵ decreases, the density becomes unimodal with mode in the interior of its support and also highly peaked around that mode. As before, this does not depend on the parametrization.

Summarizing, the family of regular proper dispersion mixing distributions, which we will denote by $\mathcal{Q}_{PDM}$, represents a geometrically convenient and much more general way of modeling local $\mathcal{Q}$-mixtures of $\mathcal{F}$, in the sense that if the dispersion parameter ϵ is small, the distributions in $\mathcal{Q}_{PDM}$ tend to be small no matter how $\mathcal{F}$ is being reparametrized. The use of these families as mixing distributions is an important contribution (innovation) of this thesis.

3.3 Laplace Expansions

3.3.1 Motivating Examples

As a motivating example, consider the situation of a mixture of a negative exponential family of densities

$$\mathcal{F} = \left\{ f(x;\mu) = \frac{1}{\mu} \exp\left(-\frac{1}{\mu} \right) \,:\, \mu > 0 \right\} ,$$

with an Inverse Gaussian family of mixing distributions

$$\mathcal{Q} = \left\{ dQ_1(\mu; m, \epsilon) = \frac{1}{\sqrt{2\pi\epsilon}}\, \mu^{-3/2} \exp\left(-\frac{1}{2\epsilon} \frac{(\mu - m)^2}{\mu m^2} \right) d\mu \,:\, m > 0, \epsilon > 0 \right\} .$$

Note that, this is clearly a proper dispersion model and in this particular case μ is the unknown mean of the distribution of the random variable X and m is the mean of the distribution of μ. An explicit expression of the mixture density is available and given by

$$\begin{aligned} g(x; Q_1(\mu; m, \epsilon)) &= \int_0^\infty f(x;\mu)\, dQ_1(\mu; m, \epsilon) \\ &= \frac{\left(m\epsilon + \sqrt{1 + 2\epsilon x}\right) \exp\left(-\frac{\sqrt{1+2\epsilon x} - 1}{m\epsilon} \right)}{m\,(1 + 2\epsilon x)^{3/2}} \qquad (3.10) \end{aligned}$$

It is possible to show that for all $x > 0$ and all $m > 0$,

$$\lim_{\epsilon \to 0} g(x; Q_1(\mu; m, \epsilon)) = f(x; m).$$

This is not surprising because, as in any dispersion model, the mixing distribution Q_1 becomes degenerate at its mean m when ϵ goes to zero. Moreover, the mixing density becomes more and more peaked around m as ϵ gets smaller. We should expect this limit to be the leading term of an asymptotic expansion of $g(x; Q_1(\mu; m, \epsilon))$ as $\epsilon \to 0$ while the higher order terms will give us more information about the behavior of g when ϵ is small, that is, when the mixture is local.

As another example of a proper dispersion mixing distribution, consider now the Gamma family of mixing distributions

$$\mathcal{Q} = \left\{ dQ_2(\mu; m, \epsilon) = \frac{(\epsilon\, e)^{-1/\epsilon}\, \mu^{-1}}{\Gamma(1/\epsilon)} \exp\left(-\frac{1}{\epsilon} \left[\frac{\mu}{m} - \log \frac{\mu}{m} - 1 \right] \right) d\mu \;:\; m, \epsilon > 0 \right\}$$

with mean $m > 0$. Also, we have here an explicit representation of the mixture density,

$$\begin{aligned} g(x; Q_2(\mu; m, \epsilon)) &= \int_0^\infty f(x; \mu)\, dQ_2(\mu; m, \epsilon) d\mu \\ &= \frac{2(\epsilon\, m)^{-\frac{1+\epsilon}{2\epsilon}} x^{\frac{1-\epsilon}{2\epsilon}}}{\Gamma(1/\epsilon)} K_{\frac{\epsilon - 1}{\epsilon}} \left(\sqrt{\frac{4x}{\epsilon m}} \right), \qquad (3.11) \end{aligned}$$

where $K_\nu(z)$ is the modified Bessel function of the second kind with index ν. It is also possible to show that for all $x > 0$ and $m > 0$

$$\lim_{\epsilon \to 0} g(x; Q_2(\mu; m, \epsilon)) = f(x; m).$$

The parametrization for the mixing distributions in terms of m and ϵ plays an important role here. Both mixtures have the form of a product of a function of ϵ and the generic Laplace-type integral

$$I_x(\epsilon) := \int_a^b H(x; \mu) \exp\left(-\frac{h(\mu)}{\epsilon} \right) d\mu\,,$$

where $h(\mu)$ is a smooth function with absolute minimum $m \in (a, b)$. In our cases $H(x; \mu)$ will always be a product of the base density $f(x; \mu)$ with other function of μ.

Assume $h(m) = 0$. As ϵ gets smaller the region where the integrand of $I_x(\epsilon)$ is significantly different from zero becomes a smaller and smaller neighborhood of $\mu = m$. Thus, in the determination of the asymptotic behavior of I_ϵ

as $\epsilon \to 0$, we need only be concerned with the behavior of H and h in an arbitrarily small neighborhood of $\mu = m$. Because of the properties of proper dispersion models, this does not depend on the way $\mathcal{F}$ is parametrized.

In the following, $f^{(r)}(x; m)$ will denote the rth-partial derivative of $f(x; \mu)$ with respect to μ evaluated at $\mu = m$. For each $x > 0$, a Laplace asymptotic expansion (see Wong (2001)) for the Inverse Gaussian mixing case yields

$$\begin{aligned} g(x; Q_1(\mu; m, \epsilon)) \quad \sim \quad & f(x; m) + \frac{\epsilon m^3}{2} f^{(2)}(x; m) + \frac{3\epsilon^2 m^5}{6} f^{(3)}(x; m) \\ & + \frac{3\epsilon^2 m^6}{24} f^{(4)}(x; m) + O_{x,m}(\epsilon^3) \end{aligned} \tag{3.12}$$

as $\epsilon \to 0$. The function $O_{x,m}(\epsilon^3)$ means an $O(\epsilon^3)$ constant for each value of x and m. In this particular case, the simple form of the factor $(2\pi\epsilon)^{-1/2}$ facilitates the expression.

The interpretation of this kind of expansions is the following. For each fixed value of x and m, the mixture $g(x; Q_1(\mu; m, \epsilon))$ (viewed now as a function of ϵ only) behaves like the right hand side of (3.12) as ϵ goes to zero. So, the symbol $\sim$ will be used to denote that, in order to avoid confusion with the equality or approximation symbol. Here $O_{x,m}(\epsilon^3)$ is a function that can include higher order derivatives of $f(x; \mu)$ (evaluated at m), but for fixed x and m it is of order ϵ^3.

Now, a Laplace asymptotic expansion for the gamma mixing case yields

$$\begin{aligned} g(x; Q_2(\mu; m, \epsilon)) \quad \sim \quad & f(x; m) + m^2 \left[\frac{\epsilon}{2}\right] f^{(2)}(x; m) + m^3 \left[\frac{2\epsilon^2}{6}\right] f^{(3)}(x; m) \\ & + m^4 \left[\frac{3\epsilon^2}{24}\right] f^{(4)}(x; m) + O_{x,m}(\epsilon^3) \end{aligned} \tag{3.13}$$

as $\epsilon \to 0$. Here, it was also necessary to expand the normalization factor $\Gamma(1/\epsilon)/(\epsilon\, e)^{-1/\epsilon}$ and divide both expansions to get the expression.

Note that truncation in this expansion may be used to define a curved mixture family in the following sense. In the Gamma case, consider the following density

$$g(x;m,\epsilon) := f(x;m) + \frac{\epsilon\, m^2}{2} f^{(2)}(x;m) + \frac{\epsilon^2 m^3}{3} f^{(3)}(x;m) + \frac{\epsilon^2 m^4}{8} f^{(4)}(x;m)\,.$$

For each fixed m, this is a curved mixture family with parameter ϵ which behaves like $g(x;Q_2)$ for small ϵ in the sense described above. It is embedded in the Natural Mixture Family

$$g(x;\boldsymbol{\lambda},m) = f(x;m) + \lambda_1 f^{(2)}(x;m) + \lambda_2 f^{(3)}(x;m) + \lambda_3 f^{(4)}(x;m)\,,$$

for each fixed $m > 0$. We might also consider the curved mixture family

$$g(x;m,\epsilon) := f(x;m) + \frac{\epsilon\, m^2}{2} f^{(2)}(x;m), \tag{3.14}$$

which is, in fact, for each fixed $m > 0$, a one-dimensional natural mixture family (with natural parameter $\epsilon\, m^2/2$). It behaves like $g(x;Q_2)$ for small ϵ but in a different way compared to the former curved mixture family. The natural parameter space $D_m(\nu)$ (for any $m > 0$) of this latter family has a very interesting and simple interpretation. It is easy to check in this case that $D_m(\nu) = [0, m^2/2]$, and this means that $0 \leq \epsilon \leq 1$. So, the parameter space of this natural mixture family is in one-to-one correspondence to those gamma mixing distributions with mean m (fixed) and $\epsilon \in [0,1]$. Nicely, those values of ϵ correspond to the case when the distribution has a unique mode inside the interval $(0,\infty)$. As discussed before, as $\epsilon \longrightarrow 0$ the mixing distribution tends to degenerate towards its mean m. For $\epsilon > 1$ the distribution has a mode at 0 and tends to be more dispersed if ϵ increases. This makes very clear the interpretation of the natural parameter space. It corresponds to those values of the parameter for which the mixing distribution is *local* in the sense defined above.

Now, for the Inverse Gaussian example, we have

$$\begin{aligned}
E_{Q_1}[(\mu - m)] &= 0 \\
\frac{E_{Q_1}[(\mu - m)^2]}{2!} &= \frac{\epsilon\, m^3}{2} \\
\frac{E_{Q_1}[(\mu - m)^3]}{3!} &= \frac{3\epsilon^2 m^5}{6} \\
\frac{E_{Q_1}[(\mu - m)^4]}{4!} &= \frac{3\epsilon^2 m^6 + 15\epsilon^3 m^7}{24}\,.
\end{aligned}$$

Using Laplace's method, we obtain the following expansions,

$$\begin{aligned}
\frac{E_{Q_1}[(\mu-m)^2]}{2!} &\sim \frac{\epsilon m^3}{2} + O_m(\epsilon^3)\\
\frac{E_{Q_1}[(\mu-m)^3]}{3!} &\sim \frac{3\epsilon^2 m^5}{6} + O_m(\epsilon^3)\\
\frac{E_{Q_1}[(\mu-m)^4]}{4!} &\sim \frac{3\epsilon^2 m^6}{24} + O_m(\epsilon^3)
\end{aligned}$$

as $\epsilon \rightarrow 0$. For the Gamma example, we have

$$\begin{aligned}
E_{Q_2}[(\mu-m)] &= 0\\
\frac{E_{Q_2}[(\mu-m)^2]}{2!} &= m^2\left[\frac{\epsilon}{2}\right]\\
\frac{E_{Q_2}[(\mu-m)^3]}{3!} &= m^3\left[\frac{2\epsilon^2}{6}\right]\\
\frac{E_{Q_2}[(\mu-m)^4]}{4!} &= m^4\left[\frac{3\epsilon^2+6\epsilon^3}{24}\right].
\end{aligned}$$

and using Laplace's method we obtain the following expansions:

$$\begin{aligned}
\frac{E_{Q_2}[(\mu-m)^2]}{2!} &\sim m^2\left[\frac{\epsilon}{2} + O(\epsilon^3)\right]\\
\frac{E_{Q_2}[(\mu-m)^3]}{3!} &\sim m^3\left[\frac{2\epsilon^2}{6} + O(\epsilon^3)\right]\\
\frac{E_{Q_2}[(\mu-m)^4]}{4!} &\sim m^4\left[\frac{3\epsilon^2}{24} + O(\epsilon^3)\right].
\end{aligned}$$

So, the coefficients of the derivative terms in expansions (3.12) and (3.13), also behave like the exact moments of the corresponding mixing distribution, when ϵ is small. In these particular examples, we obtained that the leading terms of the expansions of the second and third central moments, have the same expression of the exact moments. This will not happen in general as we will see now. Consider, for example, the change of variable $\theta = 1/\mu$ in the integral (3.10). This yields the following Laplace expansion,

$$\begin{aligned}\int_0^\infty f(x; 1/\theta)\ dQ_1(1/\theta; 1/\vartheta, \epsilon) \quad &\sim \quad \tilde{f}(x;\vartheta) + \epsilon \tilde{f}^{(1)}(x;\vartheta) \\ &+ \left[\frac{\epsilon\vartheta}{2} + \frac{3\epsilon^2}{2}\right] \tilde{f}^{(2)}(x;\vartheta) + \epsilon^2\vartheta \tilde{f}^{(3)}(x;\vartheta) \\ &+ \frac{\epsilon^2\vartheta^2}{8}\tilde{f}^{(4)}(x;\vartheta) + O_{x,\vartheta}(\epsilon^3),\end{aligned}$$

where $\vartheta = 1/m$ and $\tilde{f}(x;\theta) := f(x;1/\theta)$. Note that the integral on the left hand side is the same as (3.12), by the change of variable theorem. We are just finding an asymptotic expansion now in terms of the derivatives of the reparametrized density $\tilde{f}(x;\theta)$.

Here we emphasize that $\bar{Q}_1(\theta;\vartheta,\epsilon) := Q_1(1/\theta; 1/\vartheta, \epsilon)$ is a new proper dispersion model with position ϑ and the same dispersion parameter ϵ, see Appendix C. Note that it was necessary to reparametrize via $m \mapsto \vartheta$ in order to express the model in standard form. This transformed model is obviously called the reciprocal Inverse Gaussian model. In general, the dispersion model structure on the mixing density is preserved under arbitrary monotone differentiable transformations (diffeomorphisms). This means that the *locality* of the mixing density is also preserved. Now we have,

$$\begin{aligned} E_{\bar{Q}_1}[(\theta - \vartheta)] &= \epsilon \\ \frac{E_{\bar{Q}_1}[(\theta - \vartheta)^2]}{2!} &= \frac{\epsilon\vartheta + 3\epsilon^2}{2} \\ \frac{E_{\bar{Q}_1}[(\theta - \vartheta)^3]}{3!} &= \frac{6\epsilon^2\vartheta + 15\epsilon^3}{6} \\ \frac{E_{\bar{Q}_1}[(\theta - \vartheta)^4]}{4!} &= \frac{3\epsilon^2\vartheta^2 + 45\epsilon^3\vartheta + 105\epsilon^4}{24} \end{aligned}$$

and using Laplace's method we obtain:

$$\begin{aligned}
E_{\bar{Q}_1}[(\theta-\vartheta)] &\sim \epsilon + O_\vartheta(\epsilon^3) \\
\frac{E_{\bar{Q}_1}[(\theta-\vartheta)^2]}{2!} &\sim \frac{\epsilon\vartheta+3\epsilon^2}{2} + O_\vartheta(\epsilon^3) \\
\frac{E_{\bar{Q}_1}[(\theta-\vartheta)^3]}{3!} &\sim \frac{6\epsilon^2\vartheta}{6} + O_\vartheta(\epsilon^3) \\
\frac{E_{\bar{Q}_1}[(\theta-\vartheta)^4]}{4!} &\sim \frac{3\epsilon^2\vartheta^2}{24} + O_\vartheta(\epsilon^3)
\end{aligned}$$

as $\epsilon \to 0$. In the Gamma case we have the expansion

$$\begin{aligned}
\int_0^\infty \tilde{f}(x;\theta)\; d\bar{Q}_2(\theta;\vartheta,\epsilon) \;\sim\; & \tilde{f}(x;\vartheta) + \vartheta\left[\epsilon+\epsilon^2\right]\tilde{f}^{(1)}(x;\vartheta) \\
& +\vartheta^2\left[\frac{\epsilon}{2}+\frac{5\epsilon^2}{2}\right]\tilde{f}^{(2)}(x;\vartheta) + \vartheta^3\left[\frac{7\epsilon^2}{6}\right]\tilde{f}^{(3)}(x;\vartheta) \\
& +\vartheta^4\left[\frac{\epsilon^2}{8}\right]\tilde{f}^{(4)}(x;\vartheta) + O_{x,\vartheta}(\epsilon^3) \qquad (3.15)
\end{aligned}$$

and the exact moments and its expansions are the following

$$\begin{aligned}
E_{\bar{Q}_2}[(\theta-\vartheta)] &= \vartheta\left[\frac{\epsilon}{1-\epsilon}\right] \\
\frac{E_{\bar{Q}_2}[(\theta-\vartheta)^2]}{2!} &= \vartheta^2\left[\frac{\epsilon(1+2\epsilon)}{2(1-2\epsilon)(1-\epsilon)}\right] \\
\frac{E_{\bar{Q}_2}[(\theta-\vartheta)^3]}{3!} &= \vartheta^3\left[\frac{\epsilon^2(7+6\epsilon)}{6(1-3\epsilon)(1-2\epsilon)(1-\epsilon)}\right] \\
\frac{E_{\bar{Q}_2}[(\theta-\vartheta)^4]}{4!} &= \vartheta^4\left[\frac{\epsilon^2(3+46\epsilon+24\epsilon^2)}{24(1-4\epsilon)(1-3\epsilon)(1-2\epsilon)(1-\epsilon)}\right] \qquad (3.16)
\end{aligned}$$

$$\begin{aligned}
E_{\bar{Q}_2}[(\theta-\vartheta)] &\sim \vartheta\left[\epsilon+\epsilon^2+O(\epsilon^3)\right] \\
\frac{E_{\bar{Q}_2}[(\theta-\vartheta)^2]}{2!} &\sim \vartheta^2\left[\frac{\epsilon+5\epsilon^2}{2}+O(\epsilon^3)\right] \\
\frac{E_{\bar{Q}_2}[(\theta-\vartheta)^3]}{3!} &\sim \vartheta^3\left[\frac{7\epsilon^2}{6}+O(\epsilon^3)\right] \\
\frac{E_{\bar{Q}_2}[(\theta-\vartheta)^4]}{4!} &\sim \vartheta^4\left[\frac{3\epsilon^2}{24}+O_\vartheta(\epsilon^3)\right].
\end{aligned}$$

To check that these expansions describe the behavior of the exact moments (3.16) for small ϵ, just expand them as a series in ϵ to get

$$\begin{aligned}
\vartheta\left[\frac{\epsilon}{1-\epsilon}\right] &\sim \vartheta\left[\epsilon+\epsilon^2+O(\epsilon^3)\right] \\
\vartheta^2\left[\frac{\epsilon(1+2\epsilon)}{2(1-2\epsilon)(1-\epsilon)}\right] &\sim \vartheta^2\left[\frac{\epsilon+5\epsilon^2}{2}+O(\epsilon^3)\right] \\
\vartheta^3\left[\frac{\epsilon^2(7+6\epsilon)}{6(1-3\epsilon)(1-2\epsilon)(1-\epsilon)}\right] &\sim \vartheta^3\left[\frac{7\epsilon^2}{6}+O(\epsilon^3)\right] \\
\vartheta^4\left[\frac{\epsilon^2(3+46\epsilon+24\epsilon^2)}{24(1-4\epsilon)(1-3\epsilon)(1-2\epsilon)(1-\epsilon)}\right] &\sim \vartheta^4\left[\frac{3\epsilon^2}{24}+O_\vartheta(\epsilon^3)\right],
\end{aligned}$$

which is exactly what we have got using Laplace's method. As above, we can define curved mixture families by truncating this expansions. Consider, for example, the curved family

$$g(x;\vartheta,\epsilon) := f(x;1/\vartheta)+\vartheta\left[\epsilon+\epsilon^2\right]f^{(1)}(x;1/\vartheta)+\vartheta^2\left[\frac{\epsilon}{2}+\frac{5\epsilon^2}{2}\right]f^{(2)}(x;1/\vartheta)$$

defined for each $\vartheta > 0$. The parameter space for this family is $[0, 0.4891]$ for any $\vartheta > 0$. Again, this has an interpretation. The variance of the mixing distribution exists for $\epsilon \in (0, 1/2)$. But we have to be careful about this parameter space, because it includes points for which some other moments of the mixing distribution do not exist. If $\epsilon = 1/k$ $(k \geq 3)$ then the mixing distribution does not have moments of order $l \geq k$ (that is $E[\theta^l]$ does not exist). Thus, we need to further restrict the parameter space for the curved mixture family to behave like the mixture $g(x;\bar{Q}_2)$ for small values of ϵ. In this case it is enough to restrict the values of ϵ to be in the interval $(0, 1/4)$.

It is important to note here that the vanishing of the first derivative terms in the expansions in the variable μ is due to the fact that m is the mean of the proper dispersion model and therefore

$$E[\mu - m] = 0.$$

In the expansions with the variable θ, the new parameter ϑ is no longer the mean of the distribution and therefore, the moments obtained are no longer central moments. It is statistically more convenient, to obtain an expansion, now in terms of the mean of the new variable θ. To see this, consider the following.

For the reciprocal Gamma mixing example we have that

$$E_{\bar{Q}_2}[\theta] \sim \vartheta \left[1 + \delta(\epsilon)\right] ,$$

where $\delta(\epsilon) := \epsilon + \epsilon^2 + O(\epsilon^3)$. Then, the mean of $\bar{Q}_2$ behaves like ϑ for small ϵ. Define the following function

$$\begin{aligned} u(x;\vartheta,\epsilon) &:= \tilde{f}(x;\vartheta\left[1+\delta(\epsilon)\right]) \\ &+ \vartheta^2 \left[\frac{\epsilon + 4\epsilon^2}{2}\right] \tilde{f}^{(2)}(x;\vartheta\left[1+\delta(\epsilon)\right]) \\ &+ \vartheta^3 \left[\frac{4\epsilon^2}{6}\right] \tilde{f}^{(3)}(x;\vartheta\left[1+\delta(\epsilon)\right]) \\ &+ \vartheta^4 \left[\frac{3\epsilon^2}{24}\right] \tilde{f}^{(4)}(x;\vartheta\left[1+\delta(\epsilon)\right]). \end{aligned}$$

By expanding $\tilde{f}$ and its derivatives in a Taylor series (as a function of $\delta(\epsilon)$), it is easy to obtain that, for small $\delta(\epsilon)$,

$$\begin{aligned} u(x;\vartheta,\epsilon) \sim\; & \tilde{f}(x;\vartheta) + \vartheta\left[\epsilon+\epsilon^2\right]\tilde{f}^{(1)}(x;\vartheta) + \vartheta^2\left[\frac{\epsilon}{2}+\frac{5\epsilon^2}{2}\right]\tilde{f}^{(2)}(x;\vartheta) \\ & +\vartheta^3\left[\frac{7\epsilon^2}{6}\right]\tilde{f}^{(3)}(x;\vartheta) + \vartheta^4\left[\frac{\epsilon^2}{8}\right]\tilde{f}^{(4)}(x;\vartheta) + O_{x,\vartheta}(\epsilon^3) \end{aligned}$$

for fixed $x, \vartheta > 0$. This is exactly the same expansion obtained in (3.15). So, we can write

$$\int_0^\infty \tilde{f}(x;\theta)\ d\bar{Q}_2(\theta;\vartheta,\epsilon) \sim u(x;\vartheta,\epsilon) + O_{x,\vartheta}(\epsilon^3).$$

To interpret the new coefficients, consider the expressions for the centered moments

$$\begin{aligned}
\frac{E_{\bar{Q}_2}[(\theta - E_{\bar{Q}_2}[\theta])^2]}{2!} &= \vartheta^2 \left[\frac{\epsilon}{2(1-2\epsilon)(1-\epsilon)^2}\right] \\
\frac{E_{\bar{Q}_2}[(\theta - E_{\bar{Q}_2}[\theta])^3]}{3!} &= \vartheta^3 \left[\frac{4\epsilon^2}{6(1-3\epsilon)(1-2\epsilon)(1-\epsilon)^3}\right] \\
\frac{E_{\bar{Q}_2}[(\theta - E_{\bar{Q}_2}[\theta])^4]}{4!} &= \vartheta^4 \left[\frac{3\epsilon^2(1+5\epsilon)}{24(1-4\epsilon)(1-3\epsilon)(1-2\epsilon)(1-\epsilon)^4}\right],
\end{aligned}$$

which can be expanded as functions of ϵ to get

$$\begin{aligned}
\vartheta^2 \left[\frac{\epsilon}{2(1-2\epsilon)(1-\epsilon)^2}\right] &\sim \vartheta^2 \left[\frac{\epsilon + 4\epsilon^2}{2} + O(\epsilon^3)\right] \\
\vartheta^3 \left[\frac{4\epsilon^2}{6(1-3\epsilon)(1-2\epsilon)(1-\epsilon)^3}\right] &\sim \vartheta^3 \left[\frac{4\epsilon^2}{6} + O(\epsilon^3)\right] \\
\vartheta^4 \left[\frac{3\epsilon^2(1+5\epsilon)}{24(1-4\epsilon)(1-3\epsilon)(1-2\epsilon)(1-\epsilon)^4}\right] &\sim \vartheta^4 \left[\frac{3\epsilon^2}{24} + O(\epsilon^3)\right].
\end{aligned}$$

These are, up to the order indicated, the coefficients of $u(x;\vartheta,\epsilon)$. So what we have obtained, is a new asymptotic expansion but now with the derivatives of the density $\tilde{f}(x;\theta)$ evaluated at $\vartheta\,[1+\delta(\epsilon)]$, which behaves like the mean for small ϵ.

To consider visually some of the previous facts, we have plotted in figure 3.1 the projections of both mixtures $g(x;Q_1)$ (blue) and $g(x;Q_2)$ (green) on the mean, variance and skewness scale. This scale has been chosen because it is easy to interpret statistically. Moreover, this new scale corresponds to the following affine projections which also respect the geometry in $(\mathcal{D}_\nu^m, \mathcal{V}_\nu^0, \oplus^m)$

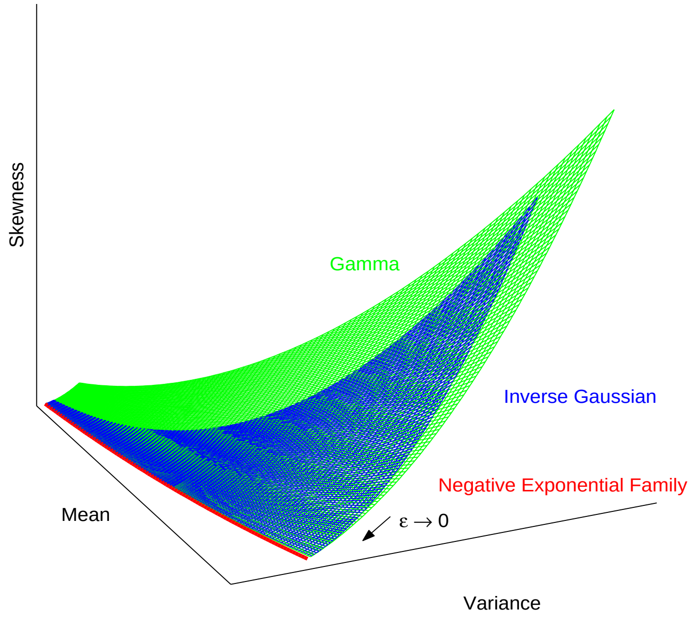

Figure 3.1: Gamma and Inverse Gaussian mixtures of $\mathcal{F}$

for every fixed value of m. The mean, variance and skewness are defined respectively as:

$$
\begin{aligned}
E[X] &:= \int_0^\infty xg(x;Q)dx \\
E[(X-m)^2] &:= \int_0^\infty (x-m)^2 g(x;Q)dx \\
E[(X-m)^3] &:= \int_0^\infty (x-m)^3 g(x;Q)dx.
\end{aligned}
$$

Note that, in this examples we have $E[X] = m$ because the mixing models are central (that is, $E_Q[\mu] = m$) when parametrized with μ. Both mixtures appear as two-dimensional surfaces and the negative exponential family (unmixed one) appear as the one dimensional curve plotted in red. From the previous facts both surfaces degenerate into the red curve as $\epsilon \to 0$ for any value of their mean m. It is worth pointing out that this surface can be treated in the usual geometric way as a *parameter free* entity. But the previous observations show that parametrization has a very important meaning as, for example, in one parametrization we have identification with a mixing distribution with essentially all moments, but in other parametrizations we don't.

As a final example consider a familiar situation in Bayesian analysis. For a given regular parametric family $\mathcal{F}$, let $\mathcal{P}$ be another parametric family with support Θ that describes the experimenter's prior beliefs about θ. This family is called the family of prior densities. Given a random sample $x_1, \ldots, x_n$ from $\mathcal{F}$, the predictive density for a new observation X_{n+1} from $\mathcal{F}$ is given by the following integral

$$g(x_{n+1}|q) = \int_\Theta f(x_{n+1}|\theta) q(\theta|\pi, x_1, \ldots, x_n) d\theta\,,$$

where q is the so-called posterior distribution

$$q(\theta|\pi, x_1, \ldots, x_n) = \frac{L(\theta; x_1, \ldots, x_n)\, \pi(\theta)}{\displaystyle\int_\Theta L(\theta; x_1, \ldots, x_n)\, \pi(\theta) d\theta},$$

where $\pi \in \mathcal{P}$ and L is the *likelihood function*

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta).$$

Under some conditions, see Bernardo and Smith (1994), the posterior density is *asymptotically Gaussian.* The Gaussian distribution is the typical example of a proper dispersion model. More importantly, as the sample size n becomes large, the posterior density will become very concentrated around some value that is very close to the posterior maximum. In this example, $1/n$ has essentially the same role as ϵ in the previous examples. So, in this case, we expect to have an expansion in powers of $1/n$ and coefficients which are functions of the posterior moments.

Before finish this section, it is worth mentioning that we obtained the Laplace asymptotic expansions using the method proposed by Fabijonas (2002) and the corresponding MAPLE code kindly provided by Professor Fabijonas.

3.3.2 Formal Expansions

The motivation of this section is to find general asymptotic expansions for the *local mixtures* discussed by examples in the previous section. Those mixtures have the form of a proper dispersion mixture of a NEF-QVF, that is

$$g(x;Q(\theta,\vartheta,\epsilon)) = \frac{\displaystyle\int_a^b f(x;\theta)V^{-1/2}(\theta)\exp\left(-\frac{d(\theta,\vartheta)}{2\epsilon}\right)d\theta}{a(\epsilon)}, \tag{3.17}$$

where $V(\theta)$ and $d(\theta,\vartheta)$ are the unit variance and deviance function of the proper dispersion model and $\Theta = (a,b)$. Clearly $a(\epsilon)$ is the normalization function

$$a(\epsilon) = \int_a^b V^{-1/2}(\theta)\exp\left(-\frac{d(\theta,\vartheta)}{2\epsilon}\right)d\theta.$$

To find an asymptotic expansion for $g(x;Q(\theta,\vartheta,\epsilon))$ (for fixed x), it is tempting to expand $f(x;\theta)$ in a Taylor series around $\theta = \vartheta$ and then, perform termwise integration. We found in the examples in the previous section that this is actually the case, at least to find an asymptotic expression valid up to order ϵ^3. In general, such procedure is only justified when the integrand is exponentially decaying as can be shown by means of the so-called Watson's Lemma.

Theorem 9 *(Watson's Lemma) Let $H(t)$ be a function of the positive real variable t, such that*

$$H(t) \sim \sum_{k=0}^{\infty} a_k\, t^{(k-1)/2} \quad \text{as } t \to 0 \tag{3.18}$$

Then

$$\int_0^{\infty} H(t)\exp(-t/\epsilon)\,dt \sim \sum_{k=0}^{\infty}\Gamma\left(\frac{k+1}{2}\right)a_k\,\epsilon^{(k+1)/2} \tag{3.19}$$

provided there exist $\epsilon_0 > 0$ *such that this integral converges absolutely for all* $\epsilon < \epsilon_0$.

Note here that Watson's Lemma is stated for the Laplace transform of the function H. This is not a very restrictive assumption, as one can transform a given integral to *look like* the Laplace transform of some function. This is actually the path to follow in the proof of Laplace's method.

Theorem 10 *(Laplace's Method) For the integral*

$$I(\epsilon) := \int_a^b H(z) \exp\left(-\frac{h(z)}{\epsilon}\right) dz\,,$$

assume that

1. $h(z)$ *and* $H(z)$ *have Taylor series expansions at every point of* (a, b),
2. $h(z)$ *has a simple minimum on* $z_0 \in (a, b)$ *and* $H(z_0) \neq 0$,
3. *without loss of generality we can assume that* $z_0 = 0$ *and also that* $h(0) = 0 = h'(0)$ *and*
4. $I(\epsilon)$ *converges absolutely for* $\epsilon < \epsilon_0$ *for some* $\epsilon_0 > 0$. *Then*

$$I(\epsilon) \sim \sum_{k=0}^{\infty} \Gamma\left(k + \frac{1}{2}\right) 2\, c_{2k}\, \epsilon^{k+1/2}, \tag{3.20}$$

where the coefficients c_k *are those found in the series expansion of* $\bar{H}(t) = H(z)/h'(z)$ *about* $z = 0$ *where*

$$z \sim \sum_{k=1}^{\infty} b_k t^{k/2} \quad \textit{as } z \longrightarrow 0$$

is obtained by reverting the change of variable $t = h(z)$.

Laplace's method, as stated here, transforms the integral $I(\epsilon)$ to

$$\begin{aligned} I(\epsilon) &= \int_0^b H(z) \exp\left(-\frac{h(z)}{\epsilon}\right) dz + \int_0^{-a} H(-z) \exp\left(-\frac{h(-z)}{\epsilon}\right) dz \\ &= \int_0^{h(b)} \bar{H}(t) \exp\left(-\frac{t}{\epsilon}\right) dt + \int_0^{h(a)} \tilde{H}(t) \exp\left(-\frac{t}{\epsilon}\right) dt, \qquad (3.21) \end{aligned}$$

where $\tilde{H}(t) = H(-z)/h'(-z)$. This last equation is symbolically exact, however in most situations both $\bar{H}$ and $\tilde{H}$ are not known analytically. Rather, we only know its series expansions about $t = 0$. Upon substituting this expansion into (3.21) the upper limits can be replaced by ∞ by arguing that the integrands contribution to each integral is exponentially small on the intervals $(h(a), \infty)$ and $(h(b), \infty)$. Finally, both integrals are now in the correct form for Watson's Lemma to be applied. This shows more clearly why the odd coefficients c_{2k+1} vanish.

In the case of our mixtures (3.17), the transformation $\theta = \vartheta + z$ has to be applied first, both in the numerator and in the denominator, in order to use Laplace's method.

First, let us explore the Laplace asymptotic expansions of the moments of a proper dispersion model. The following Theorem is stated without proof as it is very tedious but analytically straightforward.

Definition 10 A proper dispersion model $\mathcal{Q} = \{dQ(\theta; \vartheta, \epsilon) \,:\, \vartheta \in \Theta, \epsilon > 0\}$ is said to be *central* if

$$E_Q[\theta] = \int_\Theta \theta dQ(\theta; \vartheta, \epsilon) = \vartheta \,, \qquad \forall\, \vartheta \in \Theta \,.$$

Theorem 11 *The first four moments (centered at ϑ) of a proper dispersion model*

$$q(\theta; \vartheta, \epsilon) = a(\epsilon) V^{-1/2}(\theta) \exp\left(-\frac{d(\theta, \vartheta)}{2\epsilon}\right) \,,$$

have the following asymptotic expansions:

$$\begin{aligned}
E_Q[\theta-\vartheta] &\sim B_1(\vartheta)\,\epsilon + B_2(\vartheta)\,\epsilon^2 + O_\vartheta(\epsilon^3)\,, \\
E_Q[(\theta-\vartheta)^2] &\sim 2\,C_1(\vartheta)\,\epsilon + 2\,C_2(\vartheta)\,\epsilon^2 + O_\vartheta(\epsilon^3) \\
E_Q[(\theta-\vartheta)^3] &\sim 6D_1(\vartheta)\epsilon^2 + O_\vartheta(\epsilon^3) \\
E_Q[(\theta-\vartheta)^4] &\sim 24E_1(\vartheta)\,\epsilon^2 + O_\vartheta(\epsilon^3)\,, \qquad \forall\,\vartheta\in\Theta
\end{aligned} \tag{3.22}$$

as $\epsilon \to 0$*, where*

$$\begin{aligned}
B_1(\vartheta) &= -\left[\frac{V^2 d_3 + 2V'}{4}\right] \\
B_2(\vartheta) &= -\frac{3(V')^3}{4V} + V'V'' - \frac{V}{16}\left\{4V''' + 5d_3(V')^2\right\} + \frac{V^2}{8}\left\{d_4V' + 2d_3V''\right\} \\
&\quad -\frac{V^3}{16}\left\{d_5 + 2V'd_3^2\right\} + \frac{V^4}{6}d_3d_4 - \frac{5V^5d_3^3}{64} \\
C_1(\vartheta) &= V/2 \\
C_2(\vartheta) &= \frac{3(V')^2}{8} - \frac{VV''}{4} + \frac{V^2V'd_3}{4} - \frac{V^3d_4}{8} + \frac{5V^4d_3^2}{32} \\
D_1(\vartheta) &= -\left[\frac{VV'}{4} + \frac{5V^3d_3}{24}\right] \\
E_1(\vartheta) &= \frac{[V(\vartheta)]^2}{8}.
\end{aligned}$$

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

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

It is clearly possible to obtain explicit asymptotic expansions up to an order higher than three. We only present our expansions up to order three for clarity (expressions become massively long and obscure) and because it is enough to describe the properties we are interested in. The important thing to highlight is the presence of an asymptotic *pairing* in the asymptotic orders of these moments in general. Namely

$$E_Q[(\theta - \vartheta)^i] \sim O_\vartheta(\epsilon^{u(i)})\,, \tag{3.23}$$

where

$$u(i) = \left\lfloor \frac{i+1}{2} \right\rfloor$$

and $\lfloor x \rfloor$ means rounding towards infinity. This behavior is preserved for the centered moments and the normalized moments. Following a suggestion from Paul Vos, it is straightforward to verify that the cumulants $C_Q^i(\theta)$ have the more "usual" behaviour

$$C_Q^i(\theta) = O(\epsilon^{i-1})\,, \qquad i \geq 2$$

for $i \geq 2$. But here we insist in the use of moments.

Corollary 3 *The second, third and fourth centered (at the mean) moments of a proper dispersion model have the following asymptotic expansions as* $\epsilon \longrightarrow 0$

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

Proof: Just note that

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

■

Note that, in particular we have that, as $\epsilon \to 0$,

$$E_Q[(\theta - E_Q[\theta])^4] \sim 3(E_Q[(\theta - E_Q[\theta])^2])^2 + O_\vartheta(\epsilon^3). \tag{3.24}$$

This will be particularly useful later on. Also note that,

$$\begin{aligned}
E_Q[\theta] &\sim \vartheta + O_\vartheta(\epsilon) \\
E_Q[(\theta - E_Q[\theta])^2] &\sim V(\vartheta)\epsilon + O_\vartheta(\epsilon^2)\,.
\end{aligned}$$

This is consistent with the asymptotic normality result for proper dispersion models, which states that

$$\frac{[\theta - \vartheta]}{\sqrt{\epsilon}} \xrightarrow{d} N(0, V(\vartheta))\,, \qquad \epsilon \to 0\,, \tag{3.25}$$

see Jorgensen (1997) page 30.

If the proper dispersion model is an exponential dispersion model, then automatically is central and therefore its variance is $\epsilon\, V(\vartheta)$. But this is a very restrictive class. By Daniel's Theorem (see Jorgensen (1997) page 188) the models that are both regular proper and exponential dispersion models are only the Gamma, Gaussian and Inverse Gaussian.

The following expansions for the normalized moments of a proper dispersion model will be used later on. In particular we give an expansion for the squared coefficient of variation.

Corollary 4 *The normalized moments of a proper dispersion model have the following asymptotic expansions as $\epsilon \longrightarrow 0$,*

$$\begin{aligned} \frac{E_Q[(\theta - E_Q[\theta])^2]}{(E_Q[\theta])^2} &\sim \frac{V(\vartheta)}{\vartheta^2}\epsilon + \frac{L_2(\vartheta)}{\vartheta^3}\epsilon^2 + O_\vartheta(\epsilon^3) \\ \frac{E_Q[(\theta - E_Q[\theta])^3]}{(E_Q[\theta])^3} &\sim \frac{L_3(\vartheta)}{\vartheta^3}\epsilon^2 + O_\vartheta(\epsilon^3) \\ \frac{E_Q[(\theta - E_Q[\theta])^4]}{(E_Q[\theta])^4} &\sim \frac{L_4(\vartheta)}{\vartheta^4}\epsilon^2 + O_\vartheta(\epsilon^3), \end{aligned}$$

where

$$\begin{aligned} L_2(\vartheta) &= \frac{4VV' + 2d_3V^3 + \vartheta V^4 d_3^2 - \vartheta V^3 d_4 + \vartheta V^2 V' d_3 + 2\vartheta(V')^2 - 2\vartheta VV''}{4} \\ L_3(\vartheta) &= -\frac{V^3(\vartheta)d_3}{2} \\ L_4(\vartheta) &= 3V^2(\vartheta). \end{aligned}$$

Proof: Just note that

$$\begin{aligned} \frac{E_Q[(\theta - E_Q[\theta])^2]}{(E_Q[\theta])^2} &= \frac{E_Q[(\theta - \vartheta)^2] - (E_Q[\theta - \vartheta])^2}{(\vartheta + E_Q[\theta - \vartheta])^2} \\ &\sim \frac{V(\vartheta)}{\vartheta^2}\epsilon + L(\vartheta)\epsilon^2 + O_\vartheta(\epsilon^3). \end{aligned}$$

The expansion is obtained by dividing the corresponding expansions derived from Theorem 11 and Corollary 3. A similar argument yields the other two expansions. ■

As a consequence of the previous discussions and results, we now present a Theorem (which is one of the main contributions of part I of this thesis) that formally specifies the form of the expansions of mixtures when the mixing model is local.

Theorem 12 *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 proper dispersion model which is defined on Θ . Then the $\mathcal{Q}$-mixture of $\mathcal{F}$ has the following asymptotic expansion:*

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

as $\epsilon \to 0$, for fixed $\vartheta \in \Theta$ and x and for some 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

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

for some 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: As said before, for clarity in our exposition, we only show 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 Theorem 11. 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 Theorem 11 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

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

as $\epsilon \to 0$. By making use of the formulae given in Corollary 3 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). ■

From now on, we will refer to the first form of the expansion given in Theorem 12 as the *ϑ-centered* expansion and the other expansion as the *mean-centered* expansion. Note that actually this latter expansion is not centered at the exact mean but at the function $M_1(\vartheta, \epsilon)$ which behaves like the exact mean when ϵ is small. We will call the function $M_1(\vartheta, \epsilon)$ the *pseudo-mean* of the proper dispersion model. Also, we call the functions $M_i(\vartheta, \epsilon)$ the central *pseudo-moments* of the proper dispersion model, as they behave like the exact moments for small ϵ. Of course, when the proper dispersion model is central, both expansions coincide and the pseudo-mean converts into the the exact mean. Also we will be more interested, from the statistical point of view, in the mean-centered expansion.

We are not very concerned about the behavior of the remainder terms because it is not our aim to use these expansions to local mixtures in any analytical sense. We are only interested in the behavior of local mixtures when ϵ is small. Up to an specific asymptotic order, the asymptotic behavior of this kind of mixtures depends on

1. the behavior of $f(x;\theta)$ near the mean of the mixing distribution through its higher order derivatives and
2. the mixing distribution only through the set of pseudo-moments.

The second point makes sense as when the mixing distribution is unimodal and sufficiently concentrated, it can be very much determined by its first few moments. See Johnson and Rogers (1951) and Janson (1988).

Although Theorem 12 is valid for any regular family and any parametrization of it, we will be interested in the case where $\mathcal{F}$ is a NEF-QVF and is parametrized by its mean. This is because these families have an orthogonal polynomials property that will be very useful from statistical point of view later on.

Given the ϑ-centered expansion on the θ parametrization it is possible to obtain the corresponding expansion of the new parameter $\mu = h(\theta)$ under the diffeomorphism h by simply multiplying the vector of derivatives of f by the transpose of the invertible matrix (for the case $r = 2$)

$$\boldsymbol{H}(\vartheta) = \begin{pmatrix} h_0^{(1)} & 0 & 0 & 0 \\ h_0^{(2)} & \left[h_0^{(1)}\right]^2 & 0 & 0 \\ h_0^{(3)} & 3h_0^{(1)}h_0^{(2)} & \left[h_0^{(1)}\right]^3 & 0 \\ h_0^{(4)} & 4h_0^{(1)}h_0^{(3)} + 3\left[h_0^{(2)}\right]^2 & 6h_0^{(2)}\left[h_0^{(1)}\right]^2 & \left[h_0^{(1)}\right]^4 \end{pmatrix}$$

where

$$h_0^{(k)} := \frac{d^k}{d\theta^k} h(\theta)\bigg|_{\theta=\vartheta}.$$

That is, if we write

$$g(x, Q(\theta, \vartheta, \epsilon)) = f(x;\vartheta) + \boldsymbol{A}^t(\vartheta, \epsilon)\boldsymbol{f}(x;\vartheta) + R_a(x;\vartheta,\epsilon),$$

where

$$\begin{aligned} \boldsymbol{A}(\vartheta,\epsilon) &= (A_1(\vartheta,\epsilon), A_2(\vartheta,\epsilon), A_3(\vartheta,\epsilon), A_4(\vartheta,\epsilon))^t \\ \boldsymbol{f}(x;\vartheta) &= \left(f^{(1)}(x;\vartheta), f^{(2)}(x;\vartheta), f^{(3)}(x;\vartheta), f^{(4)}(x;\vartheta)\right)^t, \end{aligned}$$

then, if $m = h(\vartheta)$, we can write

$$g(x; Q(\mu; m, \epsilon)) = \tilde{f}(x; m) + \tilde{\boldsymbol{A}}^t(m, \epsilon)\tilde{\boldsymbol{f}}(x; m) + \tilde{R}_a(x; \vartheta, \epsilon)$$

where

$$\begin{aligned}
\tilde{f}(x; m) &= f(x; h^{-1}(m)) \\
\tilde{\boldsymbol{A}}(m, \epsilon) &= \boldsymbol{H}^t(h^{-1}(m))\boldsymbol{A}(h^{-1}(m)) \\
\tilde{\boldsymbol{f}}(x; \vartheta) &= \left(\tilde{f}^{(1)}(x; m), \tilde{f}^{(2)}(x; m), \tilde{f}^{(3)}(x; m), \tilde{f}^{(4)}(x; m)\right)^t .
\end{aligned}$$

To sketch this result consider that, for θ near ϑ,

$$\begin{aligned}
[h(\theta) - h(\vartheta)]^2 &= \left[h'(\vartheta)(\theta - \vartheta) + \frac{h''(\vartheta)}{2}(\theta - \vartheta)^2 + \frac{h'''(\vartheta)}{3!}(\theta - \vartheta)^3 + \cdots\right]^2 \\
&= [h']^2 (\theta - \vartheta)^2 + h'h''(\theta - \vartheta)^3 + \left[\frac{(h'')^2}{4} + \frac{h'h'''}{3}\right](\theta - \vartheta)^4 + \cdots
\end{aligned}$$

and therefore,

$$\begin{aligned}
\frac{E[(h(\theta) - h(\vartheta))^2]}{2} &= \left[h_0^{(1)}\right]^2 \frac{E[(\theta - \vartheta)^2]}{2} + 3\, h_0^{(1)} h_0^{(2)} \frac{E[(\theta - \vartheta)^3]}{3!} \\
&\quad + \left[3\left[h_0^{(2)}\right]^2 + 4\, h_0^{(1)} h_0^{(3)}\right] \frac{E[(\theta - \vartheta)^4]}{4!} + O_\vartheta(\epsilon^3).
\end{aligned}$$

Doing something similar, we can obtain the other rows of the matrix $\boldsymbol{H}$.

Note also that Theorem 12 is just a one-dimensional version of Theorem 9 in Marriott (2002) with the additional (but statistically important) assumption that the mixing distribution is a proper dispersion model. This assumption states more clearly the meaning of locality in the mixing distribution.

The theorem of Marriott essentially assumes (in the one dimensional case) that the mixing distribution is a location dispersion model (with a constant variance function) independently of its support. In this sense, Theorem 12 is slightly more general, as we allow for nonconstant variance functions for the mixing model and the support is properly defined. In theory, we can always get a mixing proper dispersion model with constant variance function by reparametrizing using the so-called *variance-stabilizing transformation* (see Jorgensen (1997)). However, our results and Marriott's are asymptotically (in ϵ) equivalent because any proper dispersion model converges in distribution to a Gaussian which is a location dispersion model and therefore has constant variance function (recall (3.25)). One of the prices to pay is that the accuracy of the approximation is not the same as in Marriott's. For example, if we keep terms until the second derivative of f, then Marriott's Theorem states we neglect terms of order ϵ^3 in the approximation and Theorem 12 states we neglect terms of order ϵ^2.

For the case where $\Theta = \mathbb{R}^+$, some important and useful simplifications arise if we further assume the proper dispersion model is in fact a scale dispersion model. We know in such a case that the unit variance function of the mixing model is of the form $V(\theta) = \theta^2$. This appears to be quite restrictive but, as we will see later, the family of scale dispersion models is very flexible and contains many families assumed in practice. This implies also flexibility for the resultant mixed distributions.

Corollary 5 *Let $\mathcal{F} = \left\{f(x;\theta) \,:\, \theta \in \Theta = \mathbb{R}^+\right\}$ be a regular family and also let $\mathcal{Q}_{sd} = \{dQ(\theta;\vartheta,\epsilon) \,:\, \vartheta \in \Theta, \epsilon > 0\}$ be a scale dispersion model defined on Θ. Then the $\mathcal{Q}_{sd}$-mixture of $\mathcal{F}$ has the following expansion:*

$$\begin{aligned} g(x;Q(\theta,\vartheta,\epsilon)) &= \frac{\displaystyle\int_\Theta f(x;\theta)\,\theta^{-1}\exp\left(-\frac{d_0(\theta/\vartheta)}{2\epsilon}\right)d\theta}{\displaystyle\int_\Theta \theta^{-1}\exp\left(-\frac{d_0(\theta/\vartheta)}{2\epsilon}\right)d\theta} \\ &\sim f(x;\vartheta) + \sum_{i=1}^{2r}\vartheta^i A_i^*(\epsilon) f^{(i)}(x;\vartheta) + O_{x,\vartheta}(\epsilon^{r+1}) \end{aligned}$$

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

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

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 some functions M_i^* such that

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

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

Proof: Proceeding as in the proof of Theorem 12 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 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 12, and therefore will be omitted. We only give the formulae for the M_i^* functions:

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

■

Corollary 6 *The normalized moments of a scale dispersion model have the following asymptotic expansions as* $\epsilon \to 0$,

$$\begin{aligned} \frac{E_Q[(\theta - E_Q[\theta])^2]}{(E_Q[\theta])^2} &\sim \epsilon + \left[\frac{12 + 4d_0^{(3)} - d_0^{(4)} + [d_0^{(3)}]^2}{4} \right] \epsilon^2 + O(\epsilon^3) \\ \frac{E_Q[(\theta - E_Q[\theta])^3]}{(E_Q[\theta])^3} &\sim -\frac{d_0^{(3)}}{2} \epsilon^2 + O(\epsilon^3) \\ \frac{E_Q[(\theta - E_Q[\theta])^4]}{(E_Q[\theta])^4} &\sim 3\,\epsilon^2 + O(\epsilon^3). \end{aligned}$$

Clearly, the normalized moments of a scale dispersion model does not depend on ϑ. We also state, without proof, the equivalent result of Corollary 5 for location dispersion models.

Corollary 7 *Let* $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta = \mathrm{I\!R}\}$ *be a regular family and also let* $\mathcal{Q}_{ld} = \{dQ(\theta;\vartheta,\epsilon) : \vartheta \in \Theta, \epsilon > 0\}$ *be a location dispersion model defined* *on Θ . Then the $\mathcal{Q}_{ld}$-mixture of $\mathcal{F}$ has the following expansion,*

$$\begin{aligned} g(x; Q(\theta, \vartheta, \epsilon)) &= \frac{\displaystyle\int_{\Theta} f(x;\theta)\,\exp\left(-\frac{d_0(\theta-\vartheta)}{2\epsilon}\right) d\theta}{\displaystyle\int_{\Theta} \exp\left(-\frac{d_0(\theta-\vartheta)}{2\epsilon}\right) d\theta} \\ &\sim f(x;\vartheta) + \sum_{i=1}^{2r} A_i^*(\epsilon) f^{(i)}(x;\vartheta) + O_{x,m}(\epsilon^{r+1}) \end{aligned}$$

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

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

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 some functions M_i^* such that

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

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

Proof: We only state the formulae for the functions A_i^* and M_i^*:

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

where now $d_0(u)$ is 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

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

■

3.4 Local Mixture Models

Motivated by the definition of a General Mixture Family and the expansions developed in Section 3.3.2 let us define Local Mixture Models formally.

Let $\mathcal{F}=\{f(x;\theta)\,:\,\theta\in\Theta\subset\mathbb{R}\}$ be a regular family of densities with respect to the measure ν. We can embed $\mathcal{F}$ in the affine space $(\mathcal{D}_\nu^m,\mathcal{V}_\nu^0,\oplus^m)$ using the embedding $E : \Theta \to \mathcal{D}_\nu^m$ defined by

$$\theta \mapsto f(x;\theta).$$

The regularity conditions on $\mathcal{F}$ imply the differentiability under the integral sign, and therefore

$$\int_\Theta \left.\frac{df(x;\theta)}{d\theta}\right|_{\theta=\theta_0} d\nu(x) = \frac{d}{d\theta}\int_\Theta f(x;\theta)\,d\nu(x)\left.\vphantom{\int}\right|_{\theta=\theta_0} = \frac{d}{d\theta}(1) = 0.$$

This means that the functions

$$\frac{df(x;\theta_0)}{d\theta} := \left.\frac{df(x;\theta)}{d\theta}\right|_{\theta=\theta_0} \in \mathcal{V}_\nu^0$$

for any $\theta_0 \in \Theta$. Clearly, the same happens with higher order derivatives, that is

$$f^{(k)}(x;\theta_0) := \frac{d^k f(x;\theta_0)}{d\theta^k} \in \mathcal{V}_\nu^0\,, \quad k \in \mathbb{N},$$

for any $\theta_0 \in \Theta$. Now, for $d \geq 1$, consider the following subset of $\mathcal{D}_\nu^m$

$$\mathcal{G}'_{\theta_0} = \left\{ g(x;\theta_0,\boldsymbol{\lambda}) = f(x;\theta_0) + \sum_{k=1}^{d} \lambda_k f^{(k)}(x;\theta_0) \;:\; \boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_d)^t \in \mathbb{R}^d \right\},$$

which is clearly an affine subspace of $(\mathcal{D}_\nu^m, \mathcal{V}_\nu^0, \oplus^m)$. Recall expression (2.5).

If the set of vectors

$$\left\{ f^{(1)}(x;\theta_0),\ldots,f^{(d)}(x;\theta_0) \right\} \subset \mathcal{V}_\nu^0$$

are linearly independent (as functions of x), then $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_d)^t$ is nothing but an affine parametrization (see Appendix B) relative to the origin $f(x;\theta_0)$ and also relative to the base conformed by that set.

Also, we can identify $\mathcal{G}'_{\theta_0}$ with the Jet space of order r of $\mathcal{F}$ at the point $f(x;\theta_0)$ (see Murray and Rice (1993) and Marriott (2002)).

Now, we can restrict the values of $\boldsymbol{\lambda}$ to get the general mixture family

$$\mathcal{G}_{\theta_0} = \left\{ g(x;\theta_0,\boldsymbol{\lambda}(\theta_0)) = f(x;\theta_0) + \sum_{k=1}^{d} \lambda_k(\theta_0) f^{(k)}(x;\theta_0) \;:\; \boldsymbol{\lambda}(\theta_0) \in \Lambda_{\theta_0}(\nu) \right\},$$

where $\Lambda_\theta(\nu) := \Lambda^m(\nu_{\boldsymbol{S}_\theta})$. Obviously, we can do the same for every $\theta_0 \in \Theta$ and then glue them all together to create a Local Mixture Model.

Definition 11 Let $\mathcal{F} = \{f(x;\theta) \,:\, \theta \in \Theta \subset \mathbb{R}\}$ be a regular parametric family. The *Local Mixture Model* of order d of the family $\mathcal{F}$ is the parametric family

$$\mathcal{G}_{\mathcal{F}} = \left\{ g(x;\theta,\boldsymbol{\lambda}(\theta)) = f(x;\theta) + \sum_{k=1}^{d} \lambda_k(\theta) f^{(k)}(x;\theta) : \theta \in \Theta, \boldsymbol{\lambda}(\theta) \in \Lambda_\theta(\nu) \right\}, \tag{3.26}$$

when

$$\left\{ f^{(1)}(x;\theta), \ldots, f^{(d)}(x;\theta) \right\}$$

is a linearly independent set of functions for every $\theta \in \Theta$ and $\Lambda_\theta(\nu)$ is non-empty for all $\theta \in \Theta$.

The *natural parametrization* in a Local Mixture Model is defined as the vector $(\theta, \boldsymbol{\lambda}(\theta))^t$ and the *Hard Boundary* of the local mixture model is defined to be the set

$$\bigcup_{\theta \in \Theta} \partial(\Lambda_\theta(\nu))\,.$$

That is, the union of all the Hard Mixture Boundaries.

It is not the aim of local mixture models to approximate genuine local mixtures in any analytical sense. The philosophy is that we can capture some information of dispersion mixing structure (if present) in the data by modeling using local mixture models of even order.

Note that the natural parametrization in this case corresponds to the given parametrization θ for $\mathcal{F}$ together with the affine parametrization of $\mathcal{G}_\theta$ when we think of it as a subset of $\mathcal{G}'_\theta$. This latter parametrization is expressed as a function of θ because, for each θ, we have a particular affine space and we are expressing its points with respect to the particular origin $f(x;\theta)$. Affine coordinates do not make any sense if we do not specify the origin with respect to which they are constructed. Therefore, the natural parameter space for a Local Mixture Family is not in general a Cartesian product unless the vector parameter function $\boldsymbol{\lambda}$ is constant.

This is the usual way to parametrize this kind of structures in Differential Geometry. For example, the Tangent Bundle of a finite dimensional smooth manifold is parametrized in this way. In fact, the Tangent Bundle is itself a smooth manifold. As explained in section 2.3, we can think of a Local Mixture Model as a *Fiber Bundle* as we are attaching to each point of $\mathcal{F}$ a convex subset of an affine space which is in fact a General Mixture Family. We can also see a Local Mixture model as a manifold with a boundary. These added geometrical structures entail some theoretical difficulties which really do not help too much, at least for the statistical purposes we have. They are clearly an interesting subject to explore in the future. The only thing we are going to take some care about is the fact that the original model $\mathcal{F}$ can lie in the boundary of the parameter space of its associated local mixture family.

We can visually consider the bundle construction of a Local Mixture Model in Figure 3.2 as the union of general mixture families.

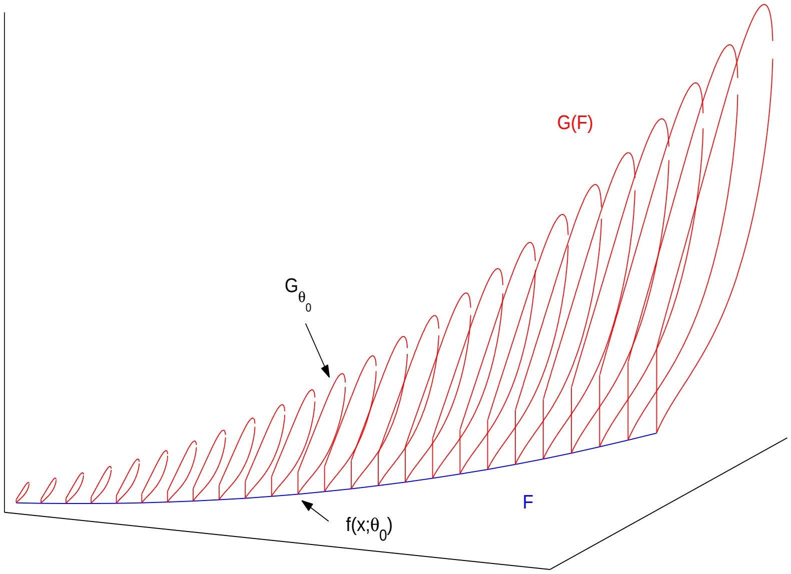

Figure 3.2: Construction of a Local Mixture

Also, at the heart of this construction of a local mixture model is the fact that we are really trying to mimic the behavior of genuine mixtures when the mixing distribution is local. In Theorem 12 it is shown that, if a mixture is local in that sense, the role played by the coefficients of the derivative terms in the expansion is exactly that of functions of ϑ. We will further discuss this issue when we define true local mixtures later on.

One important thing to state is that the definition of a Local Mixture Model does not really depend on the way $\mathcal{F}$ is parametrized. This is just a consequence of the underlying geometrical structure. Let $h : \Theta \to \Phi$ be a diffeomorphism. If we reparametrize the original family $\mathcal{F}$ using h then we end up with the same Local Mixture Model.

Theorem 13 *Let $\mathcal{G}_{\mathcal{F}}$ be a Local Mixture Model of the regular parametric family*

$$\mathcal{F} = \{f(x;\theta) \,:\, \theta \in \Theta\} \,.$$

Then $\mathcal{G}_{\mathcal{F}}$ is invariant to reparametrizations on $\mathcal{F}$. That is, if we reparametrize $\mathcal{F}$ using a diffeomorphism $h : \Theta \to \Phi$ then the Local Mixture Model obtained from this new specification is $\mathcal{G}_{\mathcal{F}}$. That is,

$$\begin{aligned}\mathcal{G}_{\mathcal{F}} &= \{g(x;\theta,\boldsymbol{\lambda}(\theta)) \,:\, \theta \in \Theta, \boldsymbol{\lambda}(\theta) \in \Lambda_\theta(\nu)\} \\ &= \{g(x;\phi,\boldsymbol{\eta}(\phi)) \,:\, \phi \in \Phi, \boldsymbol{\eta}(\phi) \in D_\phi(\nu)\} \,.\end{aligned}$$

Moreover, the change of parameter formula is given by

$$\left(\begin{array}{c}\theta \\ \boldsymbol{\lambda}(\theta)\end{array}\right) \mapsto \left(\begin{array}{c}\phi \\ \boldsymbol{A}^t(h^{-1}(\phi))\boldsymbol{\lambda}(h^{-1}(\phi))\end{array}\right) ,$$

where $\boldsymbol{A}$ is a change-of-basis matrix.

Proof: Define, for any $\theta_0 \in \Theta$ and $\phi_0 \in \Phi$ such that $\phi_0 = h(\theta_0)$,

$$\begin{aligned}\tilde{f}(x;\phi_0) &:= f(x;h^{-1}(\phi_0)) \\ \boldsymbol{f}(x;\theta_0) &:= \left(f^{(1)}(x;\theta_0),\ldots,f^{(d)}(x;\theta_0)\right)^t \\ \tilde{\boldsymbol{f}}(x;\phi_0) &:= \left(\tilde{f}^{(1)}(x;\phi_0),\ldots,\tilde{f}^{(d)}(x;\phi_0)\right)^t \,.\end{aligned}$$

By a simple application of the chain rule we can write,

$$\boldsymbol{f}(x;\theta_0) = \boldsymbol{A}(\theta_0)\,\tilde{\boldsymbol{f}}(x;h(\theta_0))\,,$$

where $\boldsymbol{A}(\theta_0)$ is the invertible lower triangular matrix

$$\begin{pmatrix} h_0^{(1)} & 0 & 0 & 0 & \cdots & 0 \\ h_0^{(2)} & \left[h_0^{(1)}\right]^2 & 0 & 0 & \cdots & 0 \\ h_0^{(3)} & 3h_0^{(1)}h_0^{(2)} & \left[h_0^{(1)}\right]^3 & 0 & \cdots & 0 \\ h_0^{(4)} & 4h_0^{(1)}h_0^{(3)} + 3\left[h_0^{(2)}\right]^2 & 6h_0^{(2)}\left[h_0^{(1)}\right]^2 & \left[h_0^{(1)}\right]^4 & & \vdots \\ \vdots & & & & \ddots & 0 \\ h_0^{(d)} & \cdots & \cdots & & & \left[h_0^{(1)}\right]^d \end{pmatrix},$$

where

$$h_0^{(k)} := \left.\frac{d^k}{d\theta^k}h(\theta)\right|_{\theta=\theta_0}.$$

Note that this equality implies that

$$\left\{f^{(1)}(x;\theta_0),\ldots,f^{(d)}(x;\theta_0)\right\}^t$$

are linearly independent as elements in $\mathcal{V}_\nu^0$, if and only, if

$$\left\{\tilde{f}^{(1)}(x;\phi_0),\ldots,\tilde{f}^{(d)}(x;\phi_0)\right\}$$

are linearly independent. Then,

$$\begin{aligned} g(x;\theta_0,\boldsymbol{\lambda}(\theta_0)) &= f(x;\theta_0) + \boldsymbol{\lambda}^t(\theta_0)\boldsymbol{f}(x;\theta_0) \\ &= f(x;\theta_0) + \boldsymbol{\lambda}^t(\theta_0)\boldsymbol{A}(\theta_0)\tilde{\boldsymbol{f}}(x;h(\theta_0)) \\ &= \tilde{f}(x;\phi_0) + \boldsymbol{\lambda}^t(h^{-1}(\phi_0))\boldsymbol{A}(h^{-1}(\phi_0))\tilde{\boldsymbol{f}}(x;\phi_0). \end{aligned}$$

This clearly gives the change of parameter formula

$$\begin{pmatrix} \theta \\ \boldsymbol{\lambda}(\theta) \end{pmatrix} \mapsto \begin{pmatrix} \phi \\ \boldsymbol{A}^t(h^{-1}(\phi))\boldsymbol{\lambda}(h^{-1}(\phi)) \end{pmatrix}.$$

■

The invariance of a local mixture family can be visualized in Figure 3.3. When working in the θ parametrization, any point in the family can be expressed as an affine combination in terms of the origin $f(x;\theta_0)$ and the base

$$\left\{f^{(1)}(x;\theta_0),\ldots,f^{(d)}(x;\theta_0)\right\}^t.$$

And, when working in the ϕ parametrization, can be expressed as an affine combination in terms of the same origin $f(x;\phi_0)$ and the base

$$\left\{\tilde{f}^{(1)}(x;\phi_0),\ldots,\tilde{f}^{(d)}(x;\phi_0)\right\}.$$

From this geometrical point of view, the invariance result is quite clear, it only corresponds to the invariance of representation of points in an Affine Space.

Another important aspect is that, for any fixed $\theta_0 \in \Theta$, Local Mixture Models are closed under general mixing as the following theorem shows. This appears also in Marriott (2002).

Theorem 14 *The parametric family of densities $\mathcal{G}_{\theta_0}$ is closed under mixing.*

Proof: Let Q be a distribution in $\mathbb{R}^d$ with support Λ_{θ_0}. Then,

$$\begin{aligned}\int_{\Lambda_{\theta_0}} g(x;\theta_0,\boldsymbol{\lambda})\,dQ(\boldsymbol{\lambda}) &= \int_{\Lambda_{\theta_0}} \left[f(x;\theta_0)+\sum_{k=1}^{d}\lambda_i f^{(k)}(x;\theta_0)\right] dQ(\boldsymbol{\lambda}) \\ &= f(x;\theta_0)+\sum_{k=1}^{d}\eta_i f^{(k)}(x;\theta_0),\end{aligned}$$

where

$$\eta_i = \int \lambda\, dQ_i(\lambda), \qquad i \in \{1,2,\ldots,d\},$$

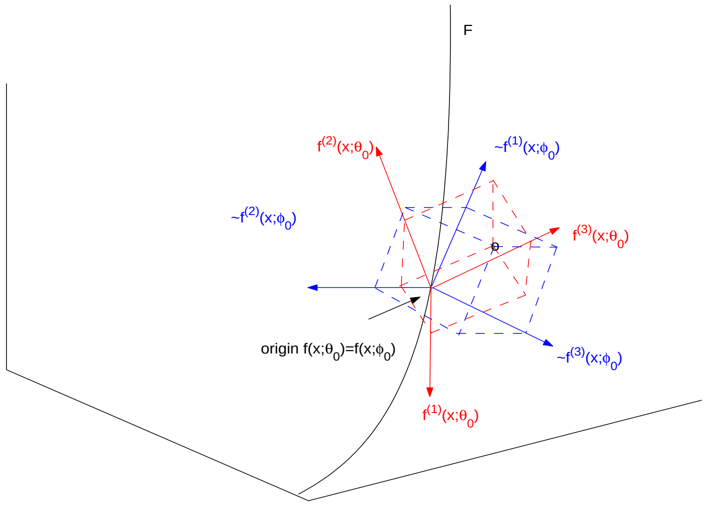

Figure 3.3: Invariance of Local Mixture Families

with Q_i the marginal distribution of the i entrance of $\boldsymbol{\lambda}$. Clearly $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)^t \in \Lambda_{\theta_0}(\nu)$ because of the convexity of $\Lambda_{\theta_0}(\nu)$. ■

This result is mainly a consequence of the fact that General Mixture Families are convex.

3.4.1 Moments of a Local Mixture Model

The fact that we are working with regular parametric families makes the computation of the moments of a Local Mixture Model very straightforward.

Actually, if $E[r(X);\theta]$ exist for all θ and is smooth as a function of θ, then

$$\begin{aligned}
E[r(X);\theta,\boldsymbol{\lambda}(\theta)] &= \int r(x)\, g(x;\theta,\boldsymbol{\lambda}(\theta))\,\nu(dx) \\
&= \int r(x)\left\{f(x;\theta)+\sum_{k=1}^{d}\lambda_k(\theta)f^{(k)}(x;\theta)\right\}\nu(dx) \\
&= \int r(x)f(x;\theta)\,\nu(dx)+\sum_{k=1}^{d}\lambda_k(\theta)\int r(x)f^{(k)}(x;\theta)\nu(dx) \\
&= \int r(x)f(x;\theta)\,\nu(dx)+\sum_{k=1}^{d}\lambda_k(\theta)\frac{d^k}{d\theta^k}\int r(x)f(x;\theta)\nu(dx) \\
&= E[r(X);\theta]+\sum_{k=1}^{d}\lambda_k(\theta)\frac{d^k}{d\theta^k}E[r(X);\theta].
\end{aligned}$$

3.5 True Local Mixture Models

So far, Local Mixture Families have been viewed simply as multidimensional parametric families constructed using a base parametric family $\mathcal{F}$. But the motivation of their definition is to mimic the behavior of genuine continuous (local) mixtures of $\mathcal{F}$. It is then natural to ask the question of whether a local mixture model can be a true mixture. As we will see in the next chapter, when studying mixtures of the negative exponential distribution, the answer to this question can be negative and in such case makes sense to modify in some way local mixture models to be much more like genuine mixtures.

In this section we study how we can restrict the parameter values of a local mixture to mimic the behavior of a mixture in a reasonable way. Let us start with a definition.

Definition 12 Let $\mathcal{F}=\{f(x;\theta)\,:\,\theta\in\Theta\}$ be a regular parametric family of densities. Then the Local Mixture Model

$$\mathcal{G}_{\mathcal{F}}=\{g(x;\theta,\boldsymbol{\lambda}(\theta))\,:\,\theta\in\Theta,\,\boldsymbol{\lambda}(\theta)\in\Lambda_\theta(\nu)\}$$

is a *True Local Mixture Model* if there exists $Q(\theta)$, a proper distribution function defined on Θ, such that

$$\int_{\Theta} f(x;\theta)dQ(\theta) \sim g(x;\vartheta,\boldsymbol{\lambda}^{*}(\vartheta,\epsilon)) + O_{x,\vartheta}(\epsilon^{r+1})$$

as $\epsilon \to 0$, with $\vartheta \in \Theta$ and

$$\boldsymbol{\lambda}^{*}(\vartheta,\epsilon) = (A_1(\vartheta,\epsilon),\ldots,A_{2r}(\vartheta,\epsilon))^t \in \Lambda_{\vartheta}(\nu),$$

where the functions A_i are the same as those defined in Theorem 12. Equivalently, the Local Mixture Model is a *True Local Mixture Model* if there exists $Q(\theta)$ such that

$$\int_{\Theta} f(x;\theta)dQ(\theta) \sim g(x;M_1(\vartheta,\epsilon),\boldsymbol{\lambda}^{**}(\vartheta,\epsilon)) + O_{x,\vartheta}(\epsilon^{r+1})$$

as $\epsilon \to 0$, with $M_1(\vartheta,\epsilon) \in \Theta$ and

$$\boldsymbol{\lambda}^{**}(\vartheta,\epsilon) = (0, M_2(\vartheta,\epsilon),\ldots,M_{2r}(\vartheta,\epsilon))^t \in \Lambda_{\vartheta}(\nu),$$

where the functions M_i are the same as those defined in Theorem 12.

The definition says that a local mixture model is a true local mixture model if it is capable of be having locally (that is, for small ϵ), like a Q-mixture of $\mathcal{F}$ in the sense that its parameter space admits values that can be realized as the derivative coefficients in either one of the two asymptotic expansions in Theorem 12. This is always true at least for sufficiently small values of ϵ.

Theorem 15 *There exists $\epsilon_0 > 0$ such that $\boldsymbol{\lambda}^{*}(\vartheta,\delta) \in \Lambda_{\vartheta}(\nu)$ for all $\delta \in [0,\epsilon_0]$.*

Proof: *Note that $(0,\ldots,0)^t \in \partial(\Lambda_{\vartheta}(\nu))$. The result is true because $\Lambda_{\vartheta}(\nu)$ is an open set in $\mathbb{R}^{2r}$ and $\boldsymbol{\lambda}^{*}(\vartheta,\epsilon)$ is an embedding and therefore is continuous as a function of ϵ for each fixed ϑ.* ■

As an example, just recall the Gamma mixing distribution example above where $\epsilon_0 = 1$.

Also, the definition of a true local mixture model emphasizes the fact that the parameter $\boldsymbol{\lambda}(\theta)$ must be really treated as a function of θ (supporting our notation), as the pseudo-moments are functions of ϑ (for any ϵ, small or not) and ϑ varies over Θ.

Equivalence in the definition follows from the fact that $\vartheta, A_1, M_2, \ldots, M_{2r}$ and $\vartheta, A_1, A_2, \ldots, A_{2r}$ are in one-to-one correspondence (asymptotically) for each fixed value of ϵ. As said before, we will focus from now on True Local Mixture Models in terms of the central pseudo-moments M_i. From the statistical point of view, those are clearly easier to interpret.

This means that the true local mixtures we will be interested in are of the form

$$\mathcal{G}_{\mathcal{F}} = \left\{ g(x;\theta;\boldsymbol{\lambda}(\theta)) = f(x;\theta) + \sum_{i=2}^{d} \lambda_i(\theta)\, f^{(i)}(x;\theta) \;:\; \theta \in \Theta\, \boldsymbol{\lambda}(\theta) \in \Lambda_\theta(\nu) \right\} \tag{3.27}$$

for d even. One has to keep in mind that, in this model, θ is now playing the role of the pseudo-mean $M_1(\vartheta, \epsilon)$, when ϑ varies over Θ and a fixed value of ϵ. In the same way, $\lambda_2(\theta), \ldots, \lambda_{2r}(\theta)$ are playing the role of the central pseudo-moments $M_2(\vartheta, \epsilon), \ldots, M_{2r}(\vartheta, \epsilon)$, respectively.

Note also that the parameter space has been reduced in one dimension. This means that knowing the values of the pseudo-moments $M_1, \ldots, M_{2r}$ does not imply the knowledge of $\vartheta, A_1, A_2, \ldots, A_{2r}$. The values of ϑ and A_1 will be undetermined as they can take any value as long as $\vartheta + A_1 = M_1$ (up to order ϵ^3). From the statistical point of view, this is not a serious restriction as we are more interested in the mean of a distribution instead of its components (additive components, in this case). This is also related to an identifiability issue to be discussed later. Moreover, the change of parameter formula in Theorem 13 is still valid if we can recover the values of the functions $\vartheta, A_1, \ldots, A_{2r}$.

It seems reasonable to treat the central pseudo-moments defined before like true central moments of some distribution, as they must behave like that at least for sufficiently small ϵ (that is, locally). Given that, observe that there can be values in $\Lambda_\vartheta(\nu)$ (for some ϑ) that cannot be realizable as the central moments. For example, negative values of $M_2(\vartheta, \epsilon)$ cannot be allowed as it should behave like a variance. Moreover, Rohatgi and Székely (1989) prove that for any distribution with finite first four moments it is true that

$$M_3^2 \leq M_4 M_2 - M_2^3, \tag{3.28}$$

where M_i is the ith central moment. Moreover, if the distribution is unimodal, Klaassen, Mokveld, and van Es (2000), then

$$M_3^2 \leq M_4 M_2 - \frac{3}{2} M_2^3. \tag{3.29}$$

Then it is reasonable to not allow values on the parameter space which do not satisfy such inequality. Note that the previous inequality can be easily expressed in terms of the normalized moments of the distribution, just dividing both sides by M_1^6. That is

$$\left[\frac{M_3}{M_1^3}\right]^2 \leq \frac{M_4}{M_1^4}\frac{M_2}{M_1^2} - \frac{3}{2}\left[\frac{M_2}{M_1^2}\right]^3. \tag{3.30}$$

According to Shaked (1980), if $\mathcal{F}$ is an exponential family and the mixing density $Q(\theta;\vartheta,\epsilon)$ is a central dispersion model, then

$$E_f[X] = \int x f(x;\vartheta)\nu(dx) = \vartheta$$

implies,

1. The function
$$R(x) = \frac{g(x;Q(\theta;\vartheta,\epsilon))}{f(x;\vartheta)} - 1$$
is convex and has the sign sequence $(+,-,+)$ as x transverses the real axis and

2. For any convex function $C(x)$ it is true that
$$E_g[C(X)] = \int c(x)g(x;Q(\theta;\vartheta,\epsilon))\nu(dx) \geq E_f[C(X)].$$

In particular, we have the well known result that, if mixed and unmixed models have the same mean, then the variance should increase by mixing, that is

$$V_g[X] \geq V_f[X].$$

Then, it is reasonable to restrict the parameter values of a local mixture model to satisfy conditions like the ones above.

3.6 Local Mixture Models of Natural Exponential Families

Natural Exponential Families are a very important class in statistics because they contain some of the most common and widely used models in the practice of statistics.

There exist some important simplifications when the underlying regular family $\mathcal{F}$ is a natural exponential family. Let $\mathcal{F}$ be a steep regular natural exponential family in its natural parametrization, that is

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

with respect to the σ-finite measure ν on $\mathbb{R}$. Then the derivatives of the densities have a simple form,

$$\begin{aligned} \frac{df(x;\theta)}{d\theta} &= f(x;\theta)\,[x - k_\nu'] \qquad (3.31)\\ \frac{d^2 f(x;\theta)}{d\theta^2} &= f(x;\theta)\left[(x-k_\nu')^2 - k_\nu''\right]\\ \frac{d^3 f(x;\theta)}{d\theta^3} &= f(x;\theta)\left[(x-k_\nu')^3 - 3(x-k_\nu')k_\nu'' - k_\nu'''\right]\\ \frac{d^4 f(x;\theta)}{d\theta^4} &= f(x;\theta)\left[(x-k_\nu')^4 - 6(x-k_\nu')^2 k_\nu'' - 4(x-k_\nu')k_\nu''' + 3(k_\nu'')^2 - k_\nu''''\right], \end{aligned}$$

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 we use the mean parametrization $\mu = k_\nu'(\theta)$ and denote the variance function by $V_f(\mu)$ then, these expressions turn into

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

If the variance function $V_f(\mu)$ is quadratic, then these expressions have a very special property.

Theorem 16 *Let $\mathcal{F}$ be a regular natural exponential family. Let μ be the mean parametrization and assume the variance function $V_f(\mu)$ is a polynomial of degree at most 2. Then, for each μ, the system of polynomials*

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

for $k=0,1,\ldots$ is orthogonal with respect to $f(x;\mu)$ (in the sense of definition 7). Moreover, $P_k(x;\mu)$ has exact degree k in both x and μ with leading term x^k.

Proof: See Morris (1982, 1983). ■

Clearly, those polynomials are linearly independent as functions of x. Because of this property, we get for free a local mixture model of any order. The orthogonality property will play an important statistical role later.

Corollary 8 *Let $\mathcal{F}$ be a regular natural exponential family and let μ be its mean parametrization. Assume the variance function $V_f(\mu)$ is a polynomial of degree at most 2. Then the Local Mixture Model $\mathcal{G}_{\mathcal{F}}$ of order r can be written as*

$$g(x;\mu,\boldsymbol{\eta}(\mu)) = f(x;\mu)\left[1+\sum_{k=1}^{d}\eta_k(\mu)\frac{P_k(x;\mu)}{V_f^k(\mu)}\right], \tag{3.32}$$

where $\{P_k(x;\mu)\}$ is the orthogonal system of polynomials described above.

It will be useful to consider the following simple reparametrization suggested in the formal expansions developed in Section 3.3.2

$$g(x;\mu,\boldsymbol{\eta}(\mu)) = f(x;\mu)\left[1+\sum_{k=1}^{d}\frac{\eta_k(\mu)}{k!}\frac{P_k(x;\mu)}{V_f^k(\mu)}\right], \tag{3.33}$$

where we keep the same notation in terms of $\boldsymbol{\eta}$ to avoid overloading the notation. For the case $d=4$, we have the following expressions for the mean, variance and third central moment functions,

$$\begin{aligned}
E_g[\mu, \boldsymbol{\eta}(\mu)] &= \mu + \eta_1(\mu) && (3.34)\\
V_g[\mu, \boldsymbol{\eta}(\mu)] &= V_f(\mu) + \eta_2(\mu) - \eta_1^2(\mu)\\
&= +\eta_1(\mu)V_f'(\mu) + \frac{\eta_2(\mu)}{2}V_f''(\mu) && (3.35)\\
S_g[\mu, \boldsymbol{\eta}(\mu)] &= E[(X - \mu - \eta_1(\mu))^3]\\
&= V_f(\mu)V_f'(\mu) + 2\eta_1^3(\mu) + \eta_3(\mu) - 3\eta_1(\mu)\eta_2(\mu)\\
&\quad +\eta_1(\mu)\left[[V_f'(\mu)]^2 + V_f(\mu)V_f''(\mu)\right]\\
&\quad +\eta_2(\mu)\left[3V_f'(\mu) + \frac{3}{2}V_f'(\mu)V_f''(\mu)\right]\\
&\quad +\eta_3(\mu)\left[\frac{1}{2}[V_f''(\mu)]^2 + \frac{3}{2}V_f''(\mu)\right]\\
&\quad -3\eta_1^2(\mu)V_f'(\mu) - \frac{3}{2}\eta_1(\mu)\eta_2(\mu)V_f''(\mu), && (3.36)
\end{aligned}$$

which does not depend on $\eta_u(\mu)$ for $u \geq 4$.

3.7 Identification

Identification, in a statistical sense, is related to the possibility of the model parameters being uniquely determined from the distribution of the observed random variables. When the parameter θ is known, the identification of a Local Mixture Model of order r is quite clear because of the underlying affine structure. The analogue is quite clear also for exponential families. For an extensive treatment of the identification issue, see Mimoso and de Bragança (1994).

Definition 13 Let $\mathcal{G} = \left\{g(x;\psi) \,:\, \psi \in \Psi \subset \mathrm{I\!R}^k\right\}$ be a parametric family of densities with respect to some σ-finite measure ν. The points $\psi_1, \psi_2 \in \Psi$ are said to be *observationally equivalent* (we write $\psi_1 \smile \psi_2$) if

$$f(x;\psi_1) = f(x;\psi_2) \qquad \text{a.e.}[\nu]$$

The relation $\smile$ is an equivalence relation and thus induce a partition on Ψ into equivalence classes $[\psi_0] = \{\psi \in \Psi \,:\, \psi \smile \psi_0\}$.

Definition 14 The point $\psi_0 \in \Psi$ is said to be *globally identifiable* if $[\psi_0] = \{\psi_0\}$. The model $\mathcal{G}$ is called identifiable if $[\psi] = \{\psi\}$ for all $\psi \in \Psi$. The point ψ_0 is said to be *locally identifiable* if, for some open neighborhood $U(\psi_0)$ of ψ_0, $[\psi_0] \cup U(\psi_0) = \{\psi_0\}$.

Now, we can prove the following.

Theorem 17 *Let $\mathcal{G}_{\mathcal{F}}$ be a Local Mixture Model of order d of the family $\mathcal{F}$. If the parameter θ is known then the family $\mathcal{G}_\theta$ is identifiable.*

Proof: Assume θ is known to be $\vartheta \in \Theta$. Suppose there exist a pair of values $\boldsymbol{\lambda}_1 \neq \boldsymbol{\lambda}_2$ such that

$$g(x;\vartheta,\boldsymbol{\lambda}_1) = g(x;\vartheta,\boldsymbol{\lambda}_2)$$

for all x. Then

$$\sum_{k=1}^{d} (\lambda_1^k - \lambda_2^k) f^{(k)}(x;\vartheta) = 0$$

for all x and this contradicts the linear independence of the derivative vector. ■

When θ is not known the problem is more difficult. To see why, consider the following. As noted by Marriott (2002), given the regular parametric family $\mathcal{F}$ let us expand $f(x;\theta)$ in a Taylor series around ϑ for any fixed x, that is,

$$f(x;\theta) = f(x;\vartheta) + \sum_{k=1}^{d} \frac{(\theta-\vartheta)^k}{k!} f^{(k)}(x;\vartheta) + r(x;\xi)$$

where

$$r(x;\xi) = \frac{(\theta-\vartheta)^{d+1}}{(d+1)!} f^{(d+1)}(x;\xi)$$

for some $\xi \in (\vartheta, \theta)$. In this way, the truncated Taylor's Series defines a curved mixture family for each fixed ϑ. It is embedded in the general mixture family

$$\mathcal{G}_\vartheta = \left\{ g(x;\vartheta,\boldsymbol{\lambda}) = f(x;\vartheta) + \sum_{k=1}^{d} \lambda_k f^{(k)}(x;\vartheta)\,, \boldsymbol{\lambda} \in \Lambda_\vartheta(\nu) \right\}$$

and the corresponding embedding $h : \Theta \to \Lambda_\vartheta(\nu)$ is clearly given by

$$h(\theta) = \left(\frac{(\theta-\vartheta)^1}{1!}, \ldots, \frac{(\theta-\vartheta)^d}{d!} \right)^t .$$

If we glue together all these curved mixture families, we get the family $\mathcal{G}_{app}$ defined as

$$\left\{ g_{app}(x;\vartheta,\theta) = f(x;\vartheta) + \sum_{k=1}^{d} \frac{(\theta-\vartheta)^k}{k!} f^{(k)}(x;\vartheta) \;:\; \vartheta \in \Theta\,,\, \theta \in h^{-1}(\Lambda_\vartheta(\nu)) \right\}$$

which is a subset of a local mixture model.

Clearly, this parametric family contains good approximations to $\mathcal{F}$ at each point ϑ in the sense of the Taylor's expansion above. That is,

$$f(x;\theta) \approx g_{app}(x;\vartheta,\theta)$$

for θ close to ϑ. To get an idea of how good those approximations can be, consider the case when $\mathcal{F}$ is a NEF-QVF (parametrized by its mean μ and with variance function $V_f(\mu)$). Then it is easy to prove the following result.

Theorem 18 *Assume $\mathcal{F}$ is NEF-QVF. For each fixed m, the variance function of the curved mixture family*

$$g(x;m,\mu) = f(x;m) + \sum_{k=1}^{d} \frac{(\mu-m)^k}{k!} f^{(k)}(x;m)$$

is exactly $V_f(\mu)$ and the mean is μ.

To check this is true, just use expression in (3.35). Recall that a NEF-QVF is characterized (within the class of exponential families) by its variance function together with its domain of means. In this case, the curved mixture family is not even an exponential family, but the mean is μ (use formula (3.34)) and the domain of the means is $h^{-1}(\Lambda_\vartheta(\nu))$, which is a proper subset of the domain of means of the complete family $\mathcal{F}$. This result is simple but important in its own right and can also give some insights into the theory of NEF-QVF's. Moreover, the result is also true for the so-called Natural real exponential families with cubic variance function, see for example Letac and Mora (1990).

The previous discussion means that local mixture models have local identification problems when we let the coefficient of the first derivative to be nonzero. Explicitly, we have

$$\begin{aligned} g(x;\theta,0,\ldots,0) &= f(x;\theta) \\ &\approx g_{app}(x;\vartheta,\theta) \\ &= g\left(x;\vartheta,\theta-\vartheta,\frac{(\theta-\vartheta)^2}{2!},\ldots,\frac{(\theta-\vartheta)^d}{d!}\right), \end{aligned}$$

for θ close to ϑ and this implies nearly non-identifiability of the local mixture family $\mathcal{G}_\mathcal{F}$. We note here that the point $(\theta,0,0,0,0)^t$ can lie in the boundary of $\Lambda_\theta(\nu)$, but this is not a problem as we can continuously extend the coordinates of a local mixture model to the boundaries of the parameter space.

On the other hand, it is possible to confound local mixing around some $\vartheta \in \Theta$ with a small displacement of that value in the original family $\mathcal{F}$. That is, it is possible to have

$$\left(\vartheta,\theta-\vartheta,\frac{(\theta-\vartheta)^2}{2},\ldots,\frac{(\theta-\vartheta)^d}{d!}\right)^t \approx (\vartheta,A_1(\vartheta,\epsilon),A_2(\vartheta,\epsilon),\ldots,A_d(\vartheta,\epsilon))^t$$

for small ϵ, and θ sufficiently close to ϑ. So, again it seems reasonable to not allow parameter values of a true local mixture model such as those the form

$$\left(\vartheta,y,\frac{y^2}{2},\ldots,\frac{y^d}{d!}\right)^t$$

for small y. This makes sense for a true local mixture model, since the order (for small ϵ) of the functions A_1, A_2 should be $O(\epsilon)$, for A_3, A_4 should be

$O(\epsilon^2)$ and so on. As shown by Marriott (2005) it is enough (locally) to consider the restriction $M_2(\vartheta, \epsilon) \geq 0$ for this purpose.

Finally, it is possible to ensure global identifiability in the case of local mixtures of NEF-QVF's.

Theorem 19 *Let $\mathcal{G}_{\mathcal{F}}$ be a Local Mixture Model of the form (3.27) of a NEF-QVF $\mathcal{F}$ parametrized by its mean μ. Then the family is identified in all its parameters $(\mu, \eta_2, \eta_3, \ldots, \eta_d)$.*

Proof: For each μ, the mean of $g(x; \mu, \boldsymbol{\eta})$ is exactly μ, hence it is sufficient to show identifiability for each fiber as it is shown by Theorem 17. ■

This supports the use of true local mixture models of the form (3.27) based on the mean-centered expansion of Theorem 12.

3.8 Local Scale and Local Location mixture models

Motivated by mean centered expansions of Corollaries 5 and 7, we now define the following statistical models and call them Local scale and Local location mixture models respectively. One of the main features in both cases is the fact that the pseudo-moments have a specific form as functions of the parameter.

Definition 15 Let $\mathcal{F} = \left\{ f(x; \theta) \,:\, \theta \in \Theta = \mathrm{I\!R}^+ \right\}$ be a regular parametric family. The *Local scale mixture model* of order d of the family $\mathcal{F}$ is defined as the parametric family

$$\mathcal{G}_{\mathcal{F}}^{scale} = \left\{ g(x; \theta, \boldsymbol{\gamma}) = f(x; \theta) + \sum_{k=2}^{d} \frac{\theta^k \, \gamma_k}{k!} \, f^{(k)}(x; \theta) \,:\, \theta \in \Theta \,,\, \gamma \in \Gamma_\theta \right\} ,$$

when

$$\left\{ f^{(2)}(x; \theta), \ldots, f^{(r)}(x; \theta) \right\}$$

is a linearly independent set of functions for every $\theta \in \Theta$ and $\Gamma_\theta \subset \mathrm{I\!R}^{d-1}$ is nonempty for all θ. The natural parametrization in a Local scale mixture model is defined as $(\theta, \boldsymbol{\gamma})$.

Definition 16 Let $\mathcal{F} = \{f(x;\theta) \,:\, \theta \in \Theta = \mathrm{I\!R}\}$ be a regular parametric family. The *Local location mixture model* of order r of the family $\mathcal{F}$ is defined as the parametric family

$$\mathcal{G}_{\mathcal{F}}^{loc} = \left\{ g(x;\theta,\boldsymbol{\gamma}) = f(x;\theta) + \sum_{k=2}^{d} \frac{\gamma_k}{k!} f^{(k)}(x;\theta) \,:\, \theta \in \Theta \,,\, \gamma \in \Gamma_\theta \right\} \,,$$

when

$$\left\{ f^{(2)}(x;\theta), \ldots, f^{(r)}(x;\theta) \right\}$$

is a linearly independent set of functions for every $\theta \in \Theta$ and $\Gamma_\theta \subset \mathrm{I\!R}^{d-1}$ is nonempty for all θ. The natural parametrization in a Local location mixture model is defined as $(\theta, \boldsymbol{\gamma})$.

First note that these models are simple reparametrizations of local mixture models of the form (3.27). Also note that their definitions are restricted to the case where the parameter space for $\mathcal{F}$ is $\mathrm{I\!R}^+$ (scale) and $\mathrm{I\!R}$ (location). Here, $\boldsymbol{\gamma}$ is considered constant as a function of θ, but for each $\theta \in \Theta$ the parameter space Γ_θ for $\boldsymbol{\gamma}$ depends on θ. This is just a consequence 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 for example, the case where $\mathcal{F}$ is a Poisson family with mean parameter θ. It is well known that a Poisson distribution with mean ϑ is close to a Normal distribution with mean and variance ϑ. If the mixing distribution is a scale dispersion model with position ϑ and dispersion parameter ϵ sufficiently small such that the normal approximation (3.25) holds, then the distribution of θ is close to a normal with mean ϑ and variance ϑ^2. Then, for the mixing distribution to be local, it is reasonable to require that $\epsilon\,\vartheta^2 < \vartheta$. Otherwise, the mixing distribution will have more variability around ϑ than X.

In the special case where $\mathcal{F}$ is a scale family with scale parameter $\theta \in \mathrm{I\!R}^+$ then it is straightforward to check that θ is still a scale parameter for the family $\mathcal{G}_{\mathcal{F}}^{scale}$ and therefore Γ_θ does not depend on θ. Similarly, when $\mathcal{F}$ is a location family with location parameter $\theta \in \mathbb{R}$ then θ is still a location parameter for the family $\mathcal{G}_{\mathcal{F}}^{loc}$ and therefore Γ_θ does not depend on θ. So, in these special cases the smallness of the mixing distribution depends only on the dispersion parameter ϵ. This is one of the most important aspects of the negative exponential distribution that we are going to exploit in the next chapter.

The interpretation of the local scale and local location models is clear. When we have a regular family $\mathcal{F}$ parametrized in such a way that $\Theta = \mathbb{R}^+$ ($\Theta = \mathbb{R}$) and the mixing distribution on that parametrization is a scale (location) dispersion model with small ϵ, then the above models mimic the behavior of the mean-centered expansions in Corollaries 5 and 7, 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.

Note also that, if a mixture of $\mathcal{F}$ has mixing density as a scale dispersion model, then any diffeomorphism which maps $\mathbb{R}^+$ to $\mathbb{R}$ (for example the natural logarithm), converts the mixing model into a scale mixture model and the above definitions are related in the same sense. Our definition of a Local location mixture models coincides with the first definition of a local mixture model given by Marriott (2002).

3.9 Estimation in Local Mixture Models of NEF-QVFs

In this section, we discuss simple likelihood inference for Local Mixture Models of NEF-QVFs. For the relevance of the likelihood function in statistical inference see, for example, Barndorff-Nielsen and Cox (1994) or Pace and Salvan (1997) and the references therein. We will mainly restrict ourselves to the case of Local location and Local scale mixture models of NEF-QVF. This is mainly because of the statistical advantages that they present and also because they are flexible for practical statistical purposes.

Given a random sample $\boldsymbol{x} = (x_1, \ldots, x_n)^t$ from a local mixture model (3.26), the log-likelihood of the natural parameters $(\theta, \boldsymbol{\lambda}(\theta))$ is given by:

$$\ell_{lm}(\theta, \boldsymbol{\lambda}(\theta); \boldsymbol{x}) = \sum_{i=1}^{n} \log \left(f(x_i; \theta) \left[1 + \sum_{k=1}^{d} \lambda_k(\theta) \frac{f^{(k)}(x_i; \theta)}{f(x; \theta)} \right] \right) . \qquad (3.37)$$

Here, we use the subscript "lm" which stands for local mixture. Note that, in general, this is a non-parametric likelihood in the sense that $\boldsymbol{\lambda}$ is an unknown function of θ that we want to estimate from the data. Note that, we can write

$$\ell_{lm}(\theta, \boldsymbol{\lambda}(\theta); \boldsymbol{x}) = \ell_f(\theta; \boldsymbol{x}) + \sum_{i=1}^{n} \log \left(1 + \sum_{k=1}^{d} \lambda_k(\theta) \frac{f^{(k)}(x_i; \theta)}{f(x; \theta)} \right) , \qquad (3.38)$$

where

$$\ell_f(\theta; \boldsymbol{x}) = \ell_{lm}(\theta, \mathbf{0}) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

is the log-likelihood under the original model $\mathcal{F}$. We begin with a general result which applies to any local mixture model. Analogous to Exponential Families we have an important property of the log-likelihood of a local mixture model, a property which is the mixture counterpart of Corollary 1.

Theorem 20 *For each fixed $\theta \in \Theta$, the log-likelihood of the parameter $\boldsymbol{\lambda}(\theta)$ from a sample of size n from a Local Mixture model (3.26) is concave on its convex domain $\Lambda_\theta(\nu)$.*

Proof: It is enough to prove that

$$\log \left(1 + \sum_{k=1}^{d} \lambda_k(\theta) \frac{f^{(k)}(x; \theta)}{f(x; \theta)} \right)$$

is concave as the sum of concave functions is always concave. This is true because the function in the argument of the logarithm is linear in $\boldsymbol{\lambda}(\theta)$ and therefore concave, and the logarithm of a concave function is also concave. ■

Thus, the log-likelihood function of the affine parameters of a local mixture model has properties like those of an exponential family. Therefore, the problem of finding the maximum likelihood estimator of the parameter $\boldsymbol{\lambda}(\theta)$ for each fixed $\theta \in \Theta$ is well defined and well known in nonlinear programming. That is, finding the maximum of a concave function over a convex set. See for example Bazaraa, Sherali, and Sheety (1993). We will denote such a estimator by $\widehat{\boldsymbol{\lambda}}(\theta)$.

When one is only interested in the parameter θ of the original family $\mathcal{F}$ and consider the other parameters as nuisance, for each $\theta \in \Theta$ we can construct the *profile log-likelihood* of θ as follows

$$\ell_p(\theta) = \ell_{lm}(\theta, \hat{\boldsymbol{\lambda}}(\theta)) \,, \tag{3.39}$$

which can be used as an inference tool for the parameter θ. Sometimes it is more convenient to use the normalized version known as the *profile log-likelihood ratio*

$$\ell_p^0(\theta) = \ell_p(\hat{\theta}) - \ell_p(\theta) = \ell_{lm}(\hat{\theta},\hat{\boldsymbol{\lambda}}) - \ell_{lm}(\theta, \hat{\boldsymbol{\lambda}}(\theta)) \,, \tag{3.40}$$

where $\hat{\theta}$ and $\hat{\boldsymbol{\lambda}}$ are the overall mle's of θ and $\boldsymbol{\lambda}$ and clearly $\ell_p^0(\hat{\theta}) = 0$. Now, Theorem 20 clearly applies to the log-likelihood of the parameters $\boldsymbol{\gamma}$ in a local scale (or location) mixture model.

Corollary 9 *For each fixed $\theta \in \Theta$, the log-likelihood of the parameter $\boldsymbol{\gamma}$ from a sample of size n from a Local scale (or location) mixture model is concave on its convex domain Γ_θ.*

Here, it is important to emphasize the simplification implied by the adoption of a local scale mixture model. If the mixture has a mixing density which is a scale dispersion model, then the normalized pseudo-moments does not depend on the parameter that indexes $\mathcal{F}$. The same applies to the location case with the pseudo-moments instead of the normalized ones. In statistical terms this means that we need to estimate a constant vector instead of a vector function. We now establish a useful result which also links statistical and geometrical properties in a clear way.

Theorem 21 *Let $\mathcal{F} = \{f(x;\mu) \, : \, \mu \in M\}$ be a NEF-QVF expressed in its mean parametrization μ. Let*

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

be a local scale mixture model of order d of $\mathcal{F}$ and

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

be a local location mixture model of order d of $\mathcal{F}$. Then the parameters $(\mu, \gamma_2, \ldots, \gamma_d)$ are Fisher orthogonal at $(\mu, 0, \ldots, 0)$ for all $\mu \in M$.

Proof: We will prove the result for the scale case. The location case is similar. Simply note that

$$S_1(x;m) := \left. \frac{\partial \log g(x;\mu,\boldsymbol{\gamma})}{\partial \mu} \right|_{\mu=m,\boldsymbol{\gamma}=\mathbf{0}} = \frac{f^{(1)}(x;m)}{f(x;m)}$$

$$S_k(x;m) := \left. \frac{\partial \log g(x;\mu,\boldsymbol{\gamma})}{\partial \gamma_k} \right|_{\mu=m,\boldsymbol{\gamma}=\mathbf{0}} = \frac{m^k}{k!} \frac{f^{(k)}(x;m)}{f(x;m)}$$

for $k = 2, \ldots, r$. Then the entry (i,j) of the Fisher's Matrix is

$$E_f[S_i(x;m) \, S_j(x;m)] = u_{i,j} \, I_m \left(\frac{f^{(i)}(x;m)}{f(x;m)}, \frac{f^{(j)}(x;m)}{f(x;m)} \right),$$

where $I_m(\cdot,\cdot)$ is the Fisher's information inner product from Definition 7. Using the formulae in Section 8 of Morris (1982), we have

$$I_m \left(\frac{f^{(i)}(x;m)}{f(x;m)}, \frac{f^{(j)}(x;m)}{f(x;m)} \right) = \delta_{i,j} \, \frac{a_i}{V_f^i(m)},$$

where $\delta_{i,j}$ is the Kronecker delta and

$$a_i = i! \prod_{s=0}^{i-1} (1 + sc) \tag{3.41}$$

and c is the coefficient of m^2 in the variance function $V_f(m)$. Then we have

$$E_f[S_i(x;m)\,S_j(x;m)] = \begin{cases} 0 & i \neq j \\ \dfrac{1}{V_f(m)} & i = j = 1 \\ \left[\dfrac{m^i}{i!}\right]^2 \dfrac{a_i}{V_f^i(m)} & i = j \geq 2\,. \end{cases}$$

This concludes the proof. ■

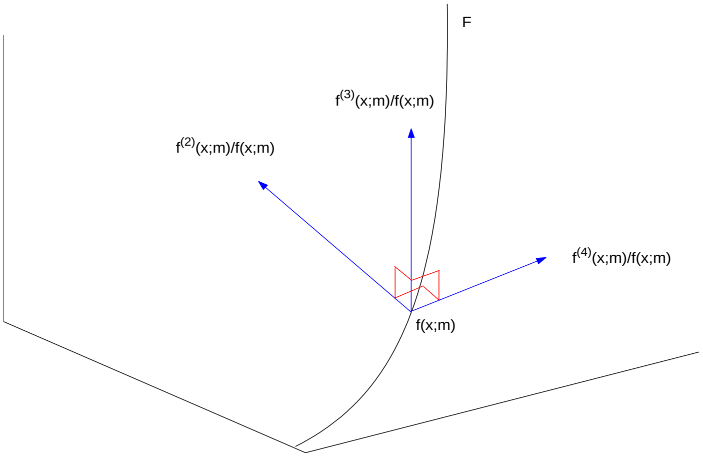

Figure 3.4: Fisher Orthogonality

The orthogonality property in Theorem 16 has two interpretations. The familiar one described graphically in Figure 3.4 and the statistical one. This latter interpretation implies, among other things, that the maximum likelihood estimators of $(\mu, \gamma_2, \ldots, \gamma_d)^t$ are asymptotically (in sample size) independent, when the true model is a member of $\mathcal{F}$. For more details about Fisher's orthogonality, see Barndorff-Nielsen and Cox (1994). We note here that the exclusion of the first derivative term in the local mixture definition was crucial, supporting again the use of this kind of local mixture model.

Note also that the generality of the previous theorem is somewhat restricted. Morris (1982) showed that there exist only six NEF-QVF's (modulo certain transformations). These are the normal distribution with variance 1, the Poisson, the Bernoulli, the geometric, the negative exponential and the hyperbolic secant distributions. The normal and the hyperbolic secant distributions have domain of the means equal to $\mathbb{R}$, so only for those families we can apply Theorem 21 with $M = \mathbb{R}$. The Poisson, geometric and negative exponential have domain of the means equal to $\mathbb{R}^+$, so only for those families we can apply Theorem 21 with $M = \mathbb{R}^+$.

Chapter 4

Local Mixture Models of the Negative Exponential Distribution

Mixtures of the negative exponential distribution have received considerable attention in the statistical literature. For example, Jewell (1982) discusses a characterization of mixtures of Weibull distributions (of which the negative exponential is a particular case) and also nonparametric maximum likelihood estimation of the mixing distribution. Keilson and Steutel (1974) discuss scale and power mixtures, and applied their results to construct some important inequalities and to show that the squared coefficient of variation of the mixing distribution is a measure of distance in the space of mixtures of exponential distributions.

Mixtures of the negative exponential are characterized as being completely monotone (see Jewell (1982) and Heckman, Robb, and Walker (1990)). This implies in particular that the density has to be monotone decreasing, so they do not exhibit multimodality. Alsoimplies that simple diagnostics, such as histograms of the data, are not useful to detect this kind of mixture structure.

From the point of view of reliability analysis, mixtures of general distributions with decreasing failure rate are always of the same type. Therefore, mixtures of exponentials have always decreasing failure rate. Moreover, Heckman, Robb, and Walker (1990) show that the failure rate of a model being a completely monotone function is a sufficient condition for the model to be represented as a mixture of exponentials.

As said before, we are only interested in the case where the mixing distributions are continuous.

4.1 Scale dispersion mixtures

Consider the family of negative exponential densities

$$\mathcal{F} = \left\{ \frac{1}{\mu} \exp\left(-\frac{x}{\mu} \right) \, : \, \mu > 0 \right\}$$

with respect to the Lebesgue on $\mathbb{R}^+$ and parametrized by its mean μ. This is clearly a NEF-QVF as its variance is μ^2. We are interested in the following type of mixtures.

Definition 17 We define the family of mixtures of $\mathcal{F}$ of the form

$$\begin{aligned} g(x; Q(\mu; m, \epsilon)) &= \int_0^\infty \frac{1}{\mu} \exp\left(-\frac{x}{\mu} \right) dQ(\mu; m, \epsilon) \\ &= \int_0^\infty \frac{1}{\mu} \exp\left(-\frac{x}{\mu} \right) a(\epsilon) \frac{1}{\mu} \exp\left(-\frac{d_0(\mu/m)}{2\epsilon} \right) d\mu \end{aligned}$$

as the *Family of Scale dispersion mixtures of* $\mathcal{F}$ when Q is a regular scale dispersion model, that is, when the function $d_0(u)$ is smooth and nonnegative with $d_0(u) = 0$ if and only if $u = 1$.

Clearly, the generality in this definition comes from the generality of the function $d_0(u)$. To see how the members of this family are, consider the following examples. The Generalized Inverse Gaussian distribution can be parametrized in the following way,

$$q_\beta(\mu; m, \epsilon) = \frac{\left(\dfrac{1+\beta}{1-\beta}\right)^{\frac{\beta}{2\epsilon}} e^{-1/\epsilon}}{2\mu K_{\beta/\epsilon}\left(\dfrac{\sqrt{1-\beta^2}}{\epsilon}\right)} \exp\left(-\frac{d_\beta(\mu; m)}{2\epsilon}\right), \qquad \mu > 0\,, \quad (4.1)$$

where

$$d_\beta(\mu; m) = 2\beta \log\left(\frac{m}{\mu}\right) + \frac{\mu}{m}(1+\beta) + \frac{m}{\mu}(1-\beta) - 2$$

with $m, \epsilon > 0$ and $\beta \in [-1, 1]$. $K_\nu(z)$ is the modified Bessel function of the third kind with index ν. To be clear (since there seems to be some confusion in the literature), we mean the Bessel function with integral representation

$$K_\nu(z) = \frac{1}{2}\int_0^\infty u^{\nu-1} \exp\left(-\frac{z}{2}(u + u^{-1})\right) du.$$

This family defines a scale dispersion model for each fixed value of β and with unit deviance d_β. The values $\beta = \pm 1$ correspond (by taking the appropriate limits) to the Gamma distribution

$$q_1(\mu; m, \epsilon) = \frac{\epsilon^{-1/\epsilon} e^{-1/\epsilon}}{\Gamma(1/\epsilon)} \frac{1}{\mu} \exp\left(-\frac{1}{\epsilon}\left(\frac{\mu}{m} - \log\frac{\mu}{m} - 1\right)\right)$$

and Reciprocal Gamma distribution

$$q_{-1}(\mu; m, \epsilon) = \frac{\epsilon^{-1/\epsilon} e^{-1/\epsilon}}{\Gamma(1/\epsilon)} \frac{1}{\mu} \exp\left(-\frac{1}{\epsilon}\left(\frac{m}{\mu} - \log\frac{m}{\mu} - 1\right)\right),$$

respectively. The value $\beta = 0$ corresponds to the Hyperbola distribution

$$q_0(\mu; m, \epsilon) = \frac{e^{-1/\epsilon}}{2\,K_0(1/\epsilon)} \frac{1}{\mu} \exp\left(-\frac{(\mu - m)^2}{2\epsilon\mu m}\right).$$

See Jorgensen (1997) page 194 for details. And finally, the mean is given by

$$E[\mu] = m\sqrt{\frac{1-\beta}{1+\beta}}\; \frac{K_{\beta/\epsilon+1}\left(\sqrt{\dfrac{1-\beta^2}{\epsilon^2}}\right)}{K_{\beta/\epsilon}\left(\sqrt{\dfrac{1-\beta^2}{\epsilon^2}}\right)}.$$

Now, the Q_β-mixture

$$g(x;Q_\beta) = \int_0^\infty \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) \, dQ_\beta(\mu;m,\epsilon)$$

has a closed form given by

$$\frac{\left(\frac{1+\beta}{1-\beta}\right)^{\frac{\beta}{2\epsilon}} K_{\frac{\beta}{\epsilon}-1}\left(\frac{1}{\epsilon}\sqrt{\frac{(1+\beta)[2x\epsilon+m(1-\beta)]}{m}}\right) \left[\frac{2x\epsilon+m(1-\beta)}{m(1+\beta)}\right]^{\frac{\beta-\epsilon}{2\epsilon}}}{mK_{\beta/\epsilon}\left(\frac{\sqrt{1-\beta^2}}{\epsilon}\right)}.$$

This family appears to be related to the Bessel function family described in Johnson, Kotz, and Balakrishnan (1994). Also, this family contain, for example, the Pareto distribution of the second kind when the mixing distribution is reciprocal Gamma. The mixture density in such case is given by

$$g(x;Q_{-1}) = \frac{1}{m}\left(1+\frac{x\epsilon}{m}\right)^{-(1+1/\epsilon)},$$

see Embrechts, Kluppelberg, and Mikosch (1997) and Johnson, Kotz, and Balakrishnan (1994) for details about this distribution. The corresponding mixture densities for the cases $\beta = 0$ and $\beta = 1$ are given by

$$g(x;Q_0) = \frac{K_1\left(\frac{1}{\epsilon}\sqrt{1+\frac{2\epsilon x}{m}}\right)}{m\sqrt{1+\frac{2\epsilon x}{m}}K_0\left(\frac{1}{\epsilon}\right)}$$

and

$$g(x;Q_1) = \frac{2(\epsilon\, m)^{-\frac{1+\epsilon}{2\epsilon}} x^{\frac{1-\epsilon}{2\epsilon}}}{\Gamma(1/\epsilon)} K_{\frac{\epsilon-1}{\epsilon}}\left(\sqrt{\frac{4x}{\epsilon m}}\right).$$

As another example of a scale dispersion model, consider the lognormal distribution

$$q(\mu;m,\epsilon) = \frac{1}{\sqrt{2\pi\epsilon}}\frac{1}{\mu}\exp\left(-\frac{(\log(\mu)-\log(m))^2}{2\epsilon}\right).$$

There is no closed form expression for the corresponding mixture density and it is precisely in these cases that our local mixtures appear to be most useful.

4.2 Models

To keep focused on the presentation, we will only be interested in the local scale mixture model of $\mathcal{F}$ of order $d = 4$ which, according to Definition 15, has density given by

$$g_0(x;\mu,\boldsymbol{\gamma}) = f(x;\mu) + \sum_{k=2}^{4} \frac{\gamma_k}{k!} f^{(k)}(x;\mu) = \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right) p_4(x;\mu,\boldsymbol{\gamma}) \qquad (4.2)$$

where $p_4(x;\mu,\boldsymbol{\gamma})$ is the quartic polynomial

$$\begin{aligned} p_4(x;\mu,\boldsymbol{\gamma}) &= \left(\frac{\gamma_4}{24}\right)\frac{x^4}{\mu^4} + \left(\frac{\gamma_3}{6} - \frac{2\gamma_4}{3}\right)\frac{x^3}{\mu^3} + \left(\frac{\gamma_2}{2} - \frac{3\gamma_3}{2} + 3\gamma_4\right)\frac{x^2}{\mu^2} \\ &+ \left(-2\gamma_2 + 3\gamma_3 - 4\gamma_4\right)\frac{x}{\mu} + \left(1 + \gamma_2 - \gamma_3 + \gamma_4\right). \end{aligned}$$

We will call this model Model 0 from now on. Note that μ is a scale parameter for this model. This is just inherited from the fact that a scale dispersion mixture of negative exponentials has the same property. Specifically,

$$\begin{aligned} g(x;Q(\mu;m,\epsilon)) &= \int f(x;\mu)a(\epsilon)\mu^{-1}\exp\left(-\frac{d_0(\mu/m)}{2\epsilon}\right)d\mu \\ &= \int f(x;mw)a(\epsilon)w^{-1}\exp\left(-\frac{d_0(w)}{2\epsilon}\right)dw \\ &= \int m^{-1}f(x/m;w)a(\epsilon)w^{-1}\exp\left(-\frac{d_0(w)}{2\epsilon}\right)dw \\ &= m^{-1}g(x/m;Q(w;1,\epsilon)) \end{aligned}$$

where $w = \mu/m$. Note here we are exploiting the fact that the mean of the negative exponential distribution is a scale parameter.

The submodels of (4.2) we will be interested in are:

Model	Restriction on the parameters
1	$\gamma_3 = \gamma_4 = 0$
2	$\gamma_4 = 3\,\gamma_2^2$
3	$\gamma_4 = 0$

4.2.1 Model 1

This corresponds to the local scale mixture model of order $d = 2$ and therefore its density is given by

$$g_1(x; \mu, \gamma_2) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) \left[1 + \gamma_2 \left[1 - \frac{2x}{\mu} + \frac{x^2}{2\mu^2}\right]\right]. \qquad (4.3)$$

Model 1 has the following interpretation from Corollary 5. It mimics the behavior of a scale dispersion mixture of $\mathcal{F}$ when ϵ is small and we retain terms of order ϵ. That is

$$\int_0^\infty f(x;\mu)dQ(\mu; m, \epsilon) \sim f(x; m(1+M_1^*(\epsilon)))+m^2 M_2^*(\epsilon) f^{(2)}(x; m)+R_b(x, m, \epsilon)$$

as $\epsilon \to 0$, with $R_b(x, m, \epsilon) = O_{x,m}(\epsilon^2)$.

Here μ plays the role of the pseudo-mean but the parameter γ_2 will play the role of the pseudo-squared coefficient of variation of the mixing distribution, that is

$$\frac{m^2 M_2^*(m, \epsilon)}{m^2(1 + M_1^*(\epsilon))^2} = \frac{M_2^*(m, \epsilon)}{(1 + M_1^*(\epsilon))^2}.$$

To see this, just note that

$$f(x; m(1 + M_1^*(\epsilon))) + m^2 M_2^*(\epsilon) f^{(2)}(x; m) + O_{x,m}(\epsilon^2) =$$

$$f(x; m(1 + M_1^*(\epsilon))) + m^2(1 + M_1^*(\epsilon))^2 \tfrac{M_2^*(m,\epsilon)}{(1+M_1^*(\epsilon))^2} f^{(2)}(x; m) + O_{x,m}(\epsilon^2).$$

Note also that this does not affect the order of the expansion, since we are dividing and multiplying by the function $(1+M_1^*(\epsilon))^2$ which is $O(1)$ as $\epsilon \to 0$.

Corollary 6 gives an expression of an asymptotic expansion of the exact squared coefficient of variation. The expression is very simple in this case,

$$\frac{E_Q[(\mu - E_Q[\mu])^2]}{(E_Q[\mu])^2} \sim \epsilon + O(\epsilon^2)$$

as $\epsilon \downarrow 0$. Clearly, if the mixing model is a scale dispersion then, this term does not depend on m.

The natural parameter space D_1 for model 1 is given by

$$\{(\mu, \gamma_2) \, : \, c_0 > 0, c_1 > 0, c_2 > 0\} \cup \{(\mu, \gamma_2) \, : \, c_0 > 0, c_1 + 2\sqrt{c_0 c_2} > 0, c_2 > 0\} \ ,$$

where c_0, c_1, c_2 the coefficients in the quadratic factor of (4.3). In this case, we have the simple expression

$$D_1 = \{(\mu, \gamma_2) \, : \, \mu > 0 \, , 0 < \gamma_2 < 1\} \ .$$

Note that this already includes the positivity of the pseudo-squared coefficient of variation.

The Hard boundary for model 1 has a nice and simple interpretation, it says that the model is a proper density when the pseudo-squared coefficient of variation γ_2 is on the interval $(0, 1)$. Keilson and Steutel (1974) show that the squared coefficient of variation of the mixing distribution in the negative exponential case is a distance, in the formal sense, from a mixture of negative exponentials to the unmixed $\mathcal{F}$.

The mean and variance for model 1 are given by

$$\begin{aligned} E[X; \mu, \gamma_2] &= \mu \\ V[X; \mu, \gamma_2] &= \mu^2[1 + 2\gamma_2] \, . \end{aligned}$$

In fact, it is easy to check that

$$E[(X - \mu)^k; \mu, \gamma_2] = \mu^k[c_{1,k} + \gamma_2 c_{2,k}], \quad k \geq 2 \, , \tag{4.4}$$

for some positive constants $c_{1,k}$ and $c_{2,k}$. Note that the behavior of the moments of this local scale mixture model is that of inflating the corresponding moments of the negative exponential distribution. This argument is clearly

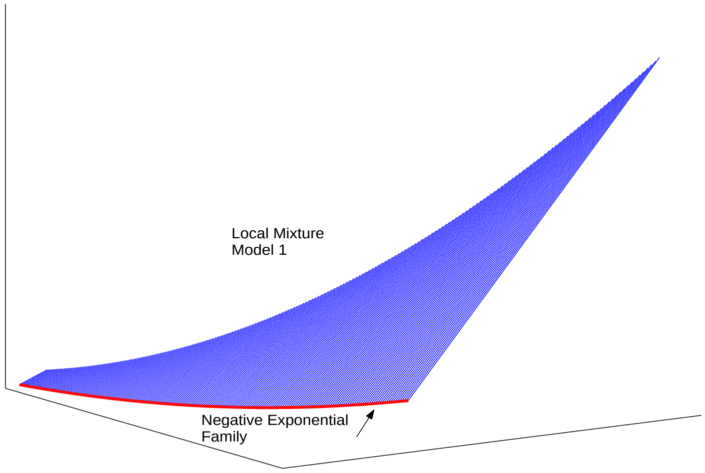

Figure 4.1: Visualization of Model 1

local as γ_2 is bounded above by 1. So, local mixture model 1 is capable of recognizing the increasing behavior not only in the variance but also in other central moments.

It is important to note that, locally, the central moments of any scale dispersion mixture have the form (4.4). To see this, just apply Corollary 5 to obtain the expression

$$\begin{aligned}\int x^k g(x;Q)dx \quad &\sim \quad \int x^k f(x;m(1+M_1^*(\epsilon)))dx \\ &\quad +m^2 M_2^*(\epsilon)\int x^k f^{(2)}(x;m(1+M_1^*(\epsilon)))dx \\ &\quad +\int x^k R_b(x;m,\epsilon)dx \\ &= \quad k!\,m^k[1+M_1^*(\epsilon)]^k + m^2\,M_2^*(\epsilon)\left\{k!k(k-1)m^{k-2}[1+M_1^*(\epsilon)]^{k-2}\right\} \\ &\quad +\int x^k R_b(x;m,\epsilon)dx \\ &= \quad m^k[1+M_1^*(\epsilon)]^k\,W^*(\epsilon,k)+\int x^k R_b(x;m,\epsilon)dx\end{aligned}$$

for $k \geq 2$ and for some function $W^*(\epsilon, k)$. For $k = 1$, it is clear that

$$\int xg(x;Q)dx \sim m(1 + M_1^*(\epsilon)) + \int xR_b(x;m,\epsilon)dx.$$

Recall that, in this case $R_b(x;m,\epsilon)$ is $O_{x,m}(\epsilon^2)$. Then, it is easy to show that,

$$\int (x - E_g[X])^k g(x;Q)dx \sim m^k[1 + M_1^*(\epsilon)]^k \, V^*(\epsilon, k) + O_m(\epsilon^2)\,,$$

for some function $V^*(\epsilon, k)$, and provided that all the previous integrals of the remainders exist. Recall also that the pseudo-mean is $m[1 + M_1^*(\epsilon)]$ not m.

Note that the variance of model 1 is bounded in the following way:

$$\mu^2 < V[X;\mu,\eta_2(\mu)] < 3\mu^2$$

To understand the constraints imposed by the hard boundary and the locality implicit in our models, assume that the mixing model is a Gamma distribution with mean m and dispersion ϵ, as in Section 3.1. The variance of such mixing distribution is $m^2\epsilon$ and therefore, the squared coefficient of variation is ϵ. Being inside the hard boundary in this case means that, the local mixture is only going to be able to model the behavior of this mixture when the squared coefficient of variation of the mixing distribution is less than one. Moreover, if $\epsilon > 1$, the family of Gamma distributions with mean m and dispersion ϵ has a unique mode at zero, but if $0 < \epsilon < 1$ then the mode is inside the interval $(0, \infty)$ and, as said before, the density shrinks to m as $\epsilon \longrightarrow 0$.

Now, assume the mixing model is a Lognormal distribution with position parameter $\log(m)$ and dispersion parameter ϵ. The mean and variance for this model are $m\exp(\epsilon/2)$ and $m^2\exp(\epsilon)(\exp(\epsilon)-1)$, so the squared coefficient of variation is

$$\exp(\epsilon) - 1$$

and therefore the local mixture is only going to be able to model the behavior of this mixture when $0 < \epsilon < \log(2)$.

More generally, recall that the variance function of the negative exponential distribution with mean μ is $V_f(\mu) = \mu^2$. Then, we can interpret the hard boundary for model 1 as follows. Model 1 is going to be able to model the behavior of the scale dispersion mixture when the variance function of the mixing distribution (in this case is approximately $\mu^2\gamma_2$) is smaller than the variance function of the negative exponential distribution, that is $\mu^2\gamma_2 \leq \mu^2$ for all $\mu > 0$. As any variance has to be nonnegative, this gives the other inequality. Obviously, we are assuming the mean of the negative exponential model to be the mean of the mixing distribution.

To understand the simplification implied by the scale dispersion mixing assumption, now suppose that the mixing model is not a scale dispersion model. For example, if it is an Inverse Gaussian distribution with mean m and dispersion ϵ, the variance of the corresponding mixing distribution is $m^3\,\epsilon$ and the boundary means that model 1 is only going to be able to model the behavior of this mixture for values of m and ϵ such that $\epsilon\, m < 1$. If the mixing model is a reciprocal Inverse Gaussian distribution with position m and dispersion ϵ, the mean and variance are given by $m + \epsilon$ and $\epsilon(m + 2\epsilon)$, respectively. Then model 1 is going to be able to model the behavior of this mixture only for values of the mean and ϵ, such that

$$0 < \frac{\epsilon(m + 2\epsilon)}{(m + \epsilon)^2} < 1\,,$$

and both last inequalities depend on m and ϵ at the same time.

Now, the following question arises: How close local mixture models can be from genuine mixtures?. Recall our definition of a True local mixture model. Essentially, we require for a local mixture model to be a true local mixture that the affine parameters of the local mixture model can be realized as the pseudo-moments of some proper dispersion mixing distribution as stated in the expansions in Theorem 12 and its Corollaries 5 and 7. But even in such case, the resulting true local mixture model mimics the behavior of the mixture only locally. To measure the closeness of a true local mixture to a genuine mixture (for our case of negative exponential mixtures), consider the following important Theorem from Feller (1970).

Theorem 22 *Suppose that X is a positive random variable such that $S(0) = 1$, where*

$$S(x) = Pr(X > x)$$

is the survival function. Then,

$$S(x) = \int_0^\infty \exp\left(-\frac{x}{\mu}\right) dQ(\mu)$$

for some proper probability distribution Q if and only if

$$(-1)^k \frac{\partial^k S(x)}{\partial x^k} \geq 0\,, \tag{4.5}$$

for all $x \geq 0$ and all $k \in \mathbb{N}$.

Proof: See Feller (1970) page 439. ■

So we can use this theorem to check if our local mixtures can be genuine mixtures.

Corollary 10 *Model 1 can never be a genuine mixture of exponentials.*

Proof: We need to check that the survival function of model 1 satisfies conditions (4.5) of the theorem. Note that, for $k = 1$, the condition is just the positivity condition that defines the hard boundary and therefore is always satisfied. It is easy to check that the other conditions are equivalent to

$$0 \leq \gamma_2 \leq \frac{2}{k+1} \qquad k = 2, 3, \ldots.$$

So, model 1 can never be a genuine mixture, as the only parameter values for that to happen are any $\mu > 0$ but $\gamma_2 = 0$, which corresponds to the original model which, by definition, is not a mixture. ■

This is not a surprising result since we construct local mixture models to mimic the behavior of a genuine mixture only for small values of the dispersion parameter of some proper dispersion mixing density. However, we can take advantage of this last theorem to impose new soft boundaries to further restrict true local mixture models.

Recall that the squared coefficient of variation is a measure of the distance between mixtures of exponentials and unmixed exponentials. Then we can formally say: the larger the k, the closer we are to non-mixing.

For example, the $k = 2$ condition is saying that a mixture of exponentials must have a non-increasing density. In Figure 4.2, we plot some densities of model 1 for a fixed value of μ. Any value of μ will give the same shape, as μ is a scale parameter. As can be seen from the plots, some of them have a bump, and therefore they are not non-increasing. This happens for values of γ_2 close to 1. Then, restricting the local mixture model using the soft boundary imposed by the $k = 2$ condition on Theorem 22, allows only for non-increasing densities, that is, now using the parameter space

$$D_1^* = \left\{ (\mu, \eta_2(\mu)) \,:\, \mu > 0,\, 0 < \gamma_2 \leq \frac{2}{3} \right\}.$$

We can further restrict the natural parameter space but the interpretation of the resulting boundary is not so easy.

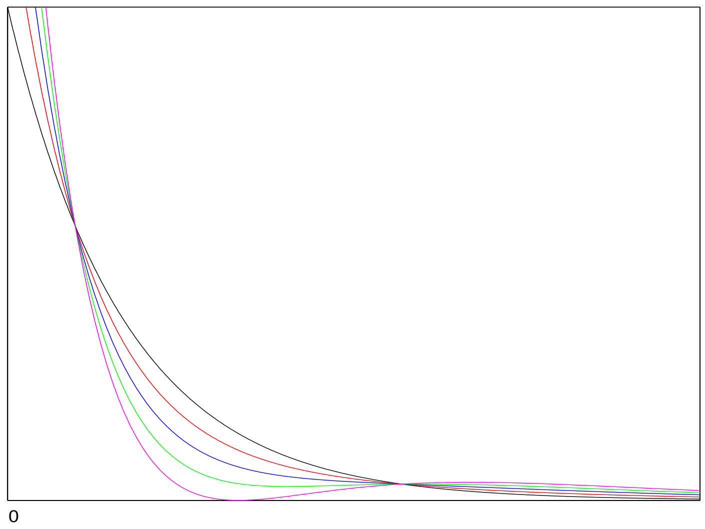

Figure 4.2: Some densities of model 1 (unmixed model in black)

Even though they are not genuine mixtures, local mixtures keep some important properties of a genuine mixture. As said before, if $g(x; Q)$ is a mixture of $f(x; \mu)$ such that $E_Q[\mu] = \bar{m}$, then the ratio

$$R(x) = \frac{g(x; Q)}{f(x; \bar{m})} - 1$$

is convex and has the sign sequence $(+,-,+)$ as x transverses the real axis. For model 1 we have,

$$R(x) = \frac{\mu^2\gamma_2}{2}\left(\frac{f^{(2)}(x;\mu)}{f(x;\mu)}\right) = \gamma_2\left(1 - \frac{2x}{\mu} + \frac{x^2}{2\mu^2}\right) ,$$

which is clearly convex. The sign changes also follows easily. In fact, we can visually see the sign changes in Figure 4.2. The unmixed negative exponential density is plotted for reference in black.

As another genuine mixture property of model 1, recall that we must have

$$E_g[X^k] \geq E_f[X^k] , \qquad k \geq 2 ,$$

when g is a genuine mixture with the same mean as f. This is true in this case as the function x^k for $k \geq 1$ is convex on the positive real line. For model 1 we have

$$E_{g_1}[X^k] = \mu^k\left[k! + \frac{k(k-1)}{2}\gamma_2\right] ,$$

so the previous set of inequalities translates to $\gamma_2 \geq 0$ which is always true. So, for the local mixture model 1, that set of moments is always bigger than the corresponding set of the unmixed model.

As mixtures of exponentials are important in the analysis of positive data, consider the following. It is well known that if a density $f(x)$ is a mixture of exponentials, then it must have decreasing hazard function for all $x > 0$, see for example Barlow and Proschan (1975). The hazard function is defined in the continuous case as

$$h(x) := \frac{f(x)}{S(x)}$$

which for model 1 turns out to be

$$h(x;\mu,\eta_2(\mu)) = \frac{2 + 2\gamma_2\mu^2 - 4\gamma_2\mu x + \gamma_2 x^2}{\mu(2\mu^2 + \gamma_2 x^2 - 2\gamma_2\mu x)}. \qquad (4.6)$$

It is easy to show that it is never monotone and actually always has a minimum at

$$x^* = \frac{\mu}{\gamma_2}\left(\gamma_2 + \sqrt{\gamma_2(2-\gamma_2)}\right) .$$

On the other hand, this hazard function has the property that

$$\lim_{x\to\infty} h(x;\mu,\eta_2(\mu)) = \frac{1}{\mu}$$

for all $\mu > 0$. This mimics the behavior of a genuine mixture. In general, the hazard function of a mixture of distributions tends to follow, in the long run, the lowest hazard function among the hazard functions of the distributions that are being mixed. See Shaked and Spizzichino (2001).

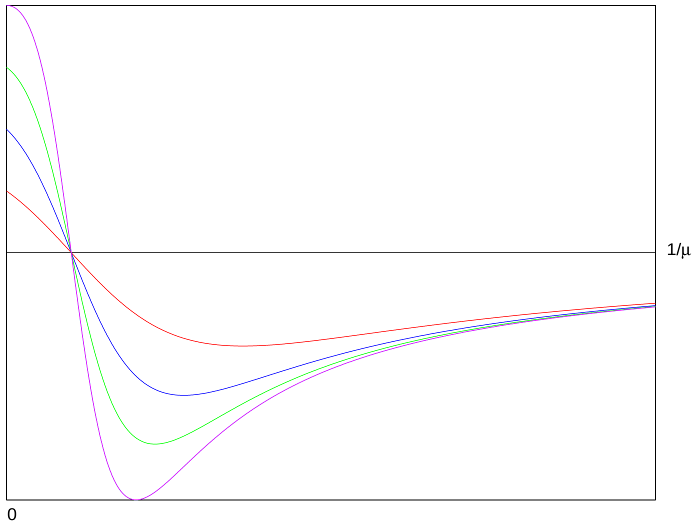

Figure 4.3: Some hazard functions of model 1

Model 1 has been used indirectly in many articles related to testing for the presence of mixing in a negative exponential model. Some of them are Mosler and Seidel (2001), Jaggia (1997), Chang and Suchindran (1997) and Kiefer (1984). They used Model 1 to construct the dispersion score test statistic. For example, Kiefer assumes a mixture model (in our notation) of the form

$$g(x;Q_1) = \int \tilde{f}(x;\varphi+u)dQ_1(u)\,,$$

where u has mean zero under Q_1 and $\tilde{f}(x;\phi) = f(x;e^{-\phi})$. That is, ϕ is the logarithm of the reciprocal of the mean. Clearly, φ is the unknown mean of the random variable ϕ. Jaggia assumes a mixture of the form

$$g(x;Q_2) = \int \tilde{\tilde{f}}(x;\vartheta\, v)dQ_2(v)\,,$$

where now v has mean one under Q_2 and $\tilde{\tilde{f}}(x;\theta) = f(x;1/\theta)$. That is, θ is the rate parameter. Using a Taylor expansion argument (as in the introduction), they obtain the following approximations to $g(x;Q_1)$ and $g(x;Q_2)$ respectively ,

$$\begin{aligned} \tilde{f}(x;\varphi) + \frac{Var_{Q_1}[\phi]}{2}\,\tilde{f}^{(2)}(x;\varphi) &= \tilde{f}(x;\varphi)\left[1 + \frac{Var_{Q_1}[\phi]}{2}\left\{1 - 3xe^{\varphi} + x^2e^{2\varphi}\right\}\right] \\ \tilde{\tilde{f}}(x;\vartheta) + \frac{Var_{Q_2}[\theta]}{2}\,\tilde{\tilde{f}}^{(2)}(x;\vartheta) &= \tilde{\tilde{f}}(x;\vartheta)\left[1 + \frac{Var_{Q_2}[\theta]}{2}\left\{\frac{\vartheta x^2 - 2x}{\vartheta}\right\}\right]. \end{aligned}$$

Using these approximations, the authors obtain the corresponding dispersion score statistics, which are given by

$$\begin{aligned} DS_1(\boldsymbol{x}) &= \frac{1}{n}\sum_{i=1}^{n}\left[1 - 3x_ie^{\hat{\varphi}} + x_i^2e^{2\hat{\varphi}}\right] = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\bar{x}^2} - 1 \\ DS_2(\boldsymbol{x}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{\vartheta}x_i^2 - 2x_i}{\hat{\vartheta}}\right] = (1/n)\sum_{i=1}^{n}(x_i - \bar{x})^2 - \bar{x}^2, \end{aligned}$$

where $\hat{\varphi}$ and $\hat{\vartheta}$ are the maximum likelihood estimates of φ and ϑ, respectively, under the assumption of no mixing, that is, under the assumption that each observation in the sample follows a negative exponential distribution with unknown mean $e^{-\varphi}$ and $1/\vartheta$, respectively.

Both statistics have a very nice and simple interpretation. $DS_1(\boldsymbol{x})$ is the relative difference between the sample variance and the variance under the model, and $DS_2(\boldsymbol{x})$ is the absolute difference between the sample variance and the variance under the model. Recall that the variance of a negative exponential distribution is the square of the mean. So, when the sample variance exceeds the variance under the model, we have empirical evidence of mixing. If fact, from (3.31) it is clear that if $\mathcal{F}$ is an exponential family expressed in its natural parametrization, then the dispersion score always has the form

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 - V_f(\bar{x}),$$

where $V_f(\mu)$ is the variance function of $\mathcal{F}$. Also recall that in the case of the negative exponential distribution, the rate parameter is essentially the natural parameter.

Lindsay (1989) shows that in the case when $\mathcal{F}$ is a NEF-QVF, it is more informative to use the mean parametrization in the following sense. Lindsay shows that, for any known μ_0 and $k = 1, 2, 3, \ldots$,

$$\widehat{m}_{0,k} = \frac{1}{n}\sum_{i=1}^{n}\frac{k!}{a_k}\,P_k(x_i;\mu_0)$$

is an unbiased estimator of

$$m_{0,k} := E_Q[(\mu-\mu_0)^k]\,,$$

for any mixing distribution Q. The constants a_k were defined in (3.41). For example, in our negative exponential case,

$$\widehat{m}_{0,2} = \frac{1}{n}\sum_{i=1}^{n}\left[2\mu_0^2 - 4\mu_0 x_i + x_i^2\right]$$

is an unbiased estimator of $E_Q[(\mu-\mu_0)^2]$. Since we want to estimate the variance, substituting μ_0 by $\bar{x}$ we obtain

$$\frac{DS_2(\boldsymbol{x})}{2}\,.$$

So, the dispersion score $DS_2(\boldsymbol{x})$ can be regarded as an estimator of the variance of the mixing distribution.

Here is convenient to consider the following reparametrization of Model 1. Define $\eta_2 = \mu^2\gamma_2$. Note that the model in this parametrization is a general mixture family for each fixed value of μ. Clearly, η_2 will play the role of the pseudo-variance of the mixing distribution. The mean and variance of model 1 now take the form

$$\begin{aligned} E[X;\mu,\eta_2] &= \mu \\ V[X;\mu,\eta_2] &= \mu^2 + 2\eta_2\,. \end{aligned}$$

Simple moment estimators of μ and η_2 can be obtained, namely

$$\begin{aligned}\hat{\mu}^{mom} &= \bar{x}\\ \widehat{\eta}_2^{mom} &= \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2-\bar{x}^2\right]=\frac{DS_2(\boldsymbol{x})}{2},\end{aligned}$$

where the superscript *mom* stands for method of moments. Clearly, we also have

$$\widehat{\gamma}_2^{mom} = \frac{1}{2}\left[\frac{1}{n}\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\bar{x}^2}-1\right]=\frac{DS_1(\boldsymbol{x})}{2}.$$

Now it is clear that $DS_1(\boldsymbol{x})$ can be considered as an estimator of the squared of the coefficient of variation of the mixing distribution. However, to be consistent with our definition of a local mixture model, we can restrict the values of this estimator to be inside the interval $[0, 2/(k+1)]$ although this is not part of the definition of the dispersion score. Moreover, Darling (1953) shows that

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\bar{x}^2}=n[DS_1(\boldsymbol{x})+1]$$

defines a (right sided) locally most powerful test against mixtures of exponentials and also derives its asymptotic distribution under the hypothesis of no mixing. See also O'Reilly and Stephens (1982) for tests of fit in the negative exponential case.

4.2.2 Model 2

Recall Model 0 has the following form

$$g_0(x;\mu,\boldsymbol{\gamma})=\frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right)p_4(x;\mu,\boldsymbol{\gamma}),$$

where $p_4(x;\mu,\boldsymbol{\gamma})$ is the quartic polynomial given by,

$$\begin{aligned}p_4(x;\mu,\boldsymbol{\gamma}) &= \left(\frac{\gamma_4}{24}\right)\frac{x^4}{\mu^4}+\left(\frac{\gamma_3}{6}-\frac{2\gamma_4}{3}\right)\frac{x^3}{\mu^3}+\left(\frac{\gamma_2}{2}-\frac{3\gamma_3}{2}+3\gamma_4\right)\frac{x^2}{\mu^2}\\ &+ \left(-2\gamma_2+3\gamma_3-4\gamma_4\right)\frac{x}{\mu}+\left(1+\gamma_2-\gamma_3+\gamma_4\right).\end{aligned}$$

Proceeding as in model 1, it is clear that here μ will continue to play the role of the pseudo-mean and the parameters γ_i will play the role of the pseudo normalized moments of the mixing distribution.

The first soft boundary we are going to impose to get a true local mixture model is the obvious one $\gamma_2 \geq 0$, since the squared coefficient of variation of any distribution must be positive. Note that in model 1 we got this boundary for free.

From Corollary 6, we know that when the mixing distribution is a scale dispersion model, then the normalized moments are independent of the parameter and have the following asymptotic expansions as $\epsilon \downarrow 0$,

$$\begin{aligned}
\frac{E_Q[(\mu - E_Q[\mu])^2]}{(E_Q[\mu])^2} &\sim \epsilon + \left[\frac{12 + 4d_0^{(3)} - d_0^{(4)} + [d_0^{(3)}]^2}{4}\right] \epsilon^2 + O(\epsilon^3) \\
\frac{E_Q[(\mu - E_Q[\mu])^3]}{(E_Q[\mu])^3} &\sim -\frac{d_0^{(3)}}{2}\, \epsilon^2 + O(\epsilon^3) \\
\frac{E_Q[(\mu - E_Q[\mu])^4]}{(E_Q[\mu])^4} &\sim 3\,\epsilon^2 + O(\epsilon^3).
\end{aligned}$$

To construct the new model, recall expression (3.24)

$$E_Q[(\mu - E_Q[\mu])^4] \sim 3(E_Q[(\mu - E_Q[\mu])^2])^2 + O_m(\epsilon^3).$$

Also, from the expansions of the normalized moments above we get

$$\frac{E_Q[(\mu - E_Q[\mu])^4]}{(E_Q[\mu])^4} \sim 3\left[\frac{E_Q[(\mu - E_Q[\mu])^2]}{(E_Q[\mu])^2}\right]^2 + O(\epsilon^3).$$

So, the normalized moments behave like that for small ϵ. We can therefore restrict the parameter values by

$$\gamma_4 = 3\,\gamma_2^2.$$

It is interesting to note that this restriction is forcing the usual coefficient of kurtosis of the mixing distribution to be zero, as in the normal distribution. Recall that the coefficient of skewness is defined, in our notation, as

$$\frac{\gamma_4}{\gamma_2^2} - 3\,.$$

Clearly, for each μ, the resulting model (after a simple reparametrization) is a curved mixture model of dimension 2 embedded in model 0. Note that this restriction automatically makes γ_4 positive, as it should be in a genuine mixture.

The density for model 2 is then

$$g_2(x;\mu,\gamma_2,\gamma_3) = \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right) p_4^{\dagger}(x;\mu,\gamma_2,\gamma_3)\,, \tag{4.7}$$

where $p_4^{\dagger}(x;\mu,\gamma_2,\gamma_3)$ is now the quartic polynomial given by

$$\begin{aligned} p_4^{\dagger}(x;\mu,\gamma_2,\gamma_3) \;&=\; \left(\frac{\gamma_2^2}{8}\right)\frac{x^4}{\mu^4} + \left(\frac{\gamma_3}{6} - 2\gamma_2^2\right)\frac{x^3}{\mu^3} + \left(\frac{\gamma_2}{2} - \frac{3\gamma_3}{2} + 9\gamma_2^2\right)\frac{x^2}{\mu^2} \\ &+\; \left(-2\gamma_2 + 3\gamma_3 - 12\gamma_2^2\right)\frac{x}{\mu} + \left(1 + \gamma_2 - \gamma_3 + 3\gamma_2^2\right). \end{aligned}$$

Using the results in Ulrich and Watson (1994), it is possible to characterize the parameter space D_2 for model 2. It is clear that, because μ is a scale parameter, then D_2 is of the form

$$D_2 = \left\{(\mu,\gamma_2,\gamma_3)\;:\;\mu > 0\,,(\gamma_2,\gamma_3)\in\Gamma_2^{\dagger}\right\}\,,$$

i.e. a Cartesian product. There is no closed form expression for the boundary of $\Gamma_2^{\dagger}$ in terms of the parameters (γ_2,γ_3), but it is very easy to calculate using the results in Ulrich and Watson (1994). The Hard boundary for the parameters (γ_2,γ_3) is plotted in red in Figure 4.4.

Also in Figure 4.4, we plotted in blue the $k = 2$ soft boundary implied by Theorem 22. Recall this boundary implies monotone decreasing densities.

The parametrization in terms of (γ_2,γ_3) has some problems, as it takes into account values of the parameter space which corresponds to small variance but relatively high skewness of the mixing distribution. One such case is indicated with a cross in Figure 4.4. Clearly, those parameters values are not compatible with the small mixing assumption.

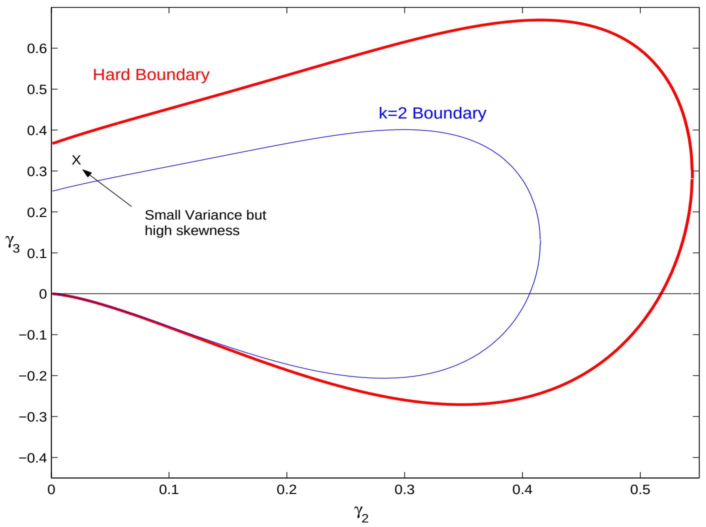

Figure 4.4: Boundaries of Model 2

Note that a mixing distribution can be small in two quite different ways. It is convenient to denote the distinction with two separate classes. The first will be called *Laplace type mixing* where the mixing distribution has small variance and unimodal. The second type will be called *Contamination type mixing* where, although the variance is small, there can be a small proportion of the realized values a long way from the mean. Such a class can have more than one mode and in general show high skewness. It is clear that, for small enough ϵ, proper dispersion models are of the Laplace type. This is because they are asymptotically normal. We distinguished the Contamination type because empirically, we want to avoid this latter kind of mixing in our models by imposing soft boundaries, such as the following.

Consider the following reparametrization of model 2,

$$\begin{aligned} \gamma_2 &= \alpha \\ \gamma_3 &= (3-\beta)\,\alpha^2\,, \end{aligned} \qquad (4.8)$$

which is clearly a diffeomorphism because $\alpha > 0$. Note that, for each fixed value of μ and β, the induced subfamily is a Curved Mixture Family embedded in Model 0 via the mapping

$$\alpha \mapsto (\alpha, (3-\beta)\alpha^2, 3\alpha^2).$$

Note that this is exactly mimicking the behavior of the second, third and fourth normalized moments of a scale dispersion mixing model for small ϵ according to Corollary 6. That is, the third normalized moment is asymptotically proportional to the squared of the second normalized moment with constant of proportionality equal to minus half of the third derivative of the deviance evaluated at its minimum, and the fourth normalized moment is asymptotically 3 times the squared of the second normalized moment.

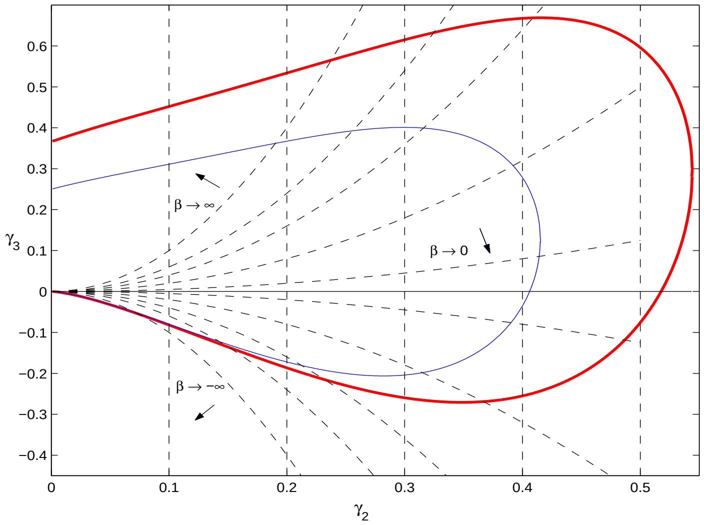

Figure 4.5: Reparametrization for Model 2

Those subfamilies are plotted in dashed black for a range of values of β in Figure 4.5. Then it is clear, from the figure, that restricting the values of the parameters β to a specific interval of the form $[K_1, K_2]$ will avoid contamination type mixing distributions. One important fact is that each of these curved mixture families is converging very slowly to the boundary $\gamma_2 = 0$, so not very extreme values of K_i are enough to capture the local behavior of a large range of scale dispersion mixing models.

For instance, the interval $[K_1, K_2] = [-1, 1]$ covers the behavior of the Generalized Inverse Gaussian distribution. To see this, consider the Generalized Inverse Gaussian distribution. Recall that, for this family we have

$$d_0(u) = 2\beta \log(1/u) + u(1+\beta) + \frac{1}{u}(1-\beta) - 2$$

which implies

$$d_0^{(3)} = 2\beta - 6.$$

Recall from Corollary 6 that

$$\frac{E_{Q_\beta}[(\mu - E_{Q_\beta}[\mu])^3]}{(E_{Q_\beta}[\mu])^3} \sim -\frac{d_0^{(3)}}{2}\left[\frac{E_{Q_\beta}[(\mu - E_{Q_\beta}[\mu])^2]}{(E_{Q_\beta}[\mu])^2}\right]^2 + O(\epsilon^3).$$

This justifies the particular form of parametrization (4.8) above. But parametrization (4.8) is just a matter of convenience. Actually, a single particular value of $d_0^{(3)}$ might correspond to different distributions. For example, if $\beta = 0$, the value of $d_0^{(3)} = -6$ corresponds to either the Lognormal (which is not included in the Generalized Inverse Gaussian family) or the Hyperbola distribution.

Finally, the moments for Model 2 are given by

$$\begin{aligned} E_g[X] &= \mu \\ E_g[(X-\mu)^2] &= \mu^2(1+2\gamma_2) \\ E_g[(X-\mu)^3] &= 2\mu^3(1+6\gamma_2+3\gamma_3). \end{aligned}$$

Note that the expressions for the mean and variance are exactly the same as in Model 1, and that the expression for the third central moment is the corresponding one for the negative exponential distribution, inflated by a factor which depends linearly on γ_2 and γ_3.

Finally, we can further impose soft boundaries like the one in expression (3.30), which states an inequality for the second, third and fourth normalized moments of any unimodal distribution. For Model 2 this inequality results in the cusp

$$\gamma_3^2 \leq \gamma_4\gamma_2 - \frac{3}{2}\gamma_2^3 = \frac{3}{2}\gamma_2^3\,. \tag{4.9}$$

This soft boundary also avoids contamination type mixing distributions.

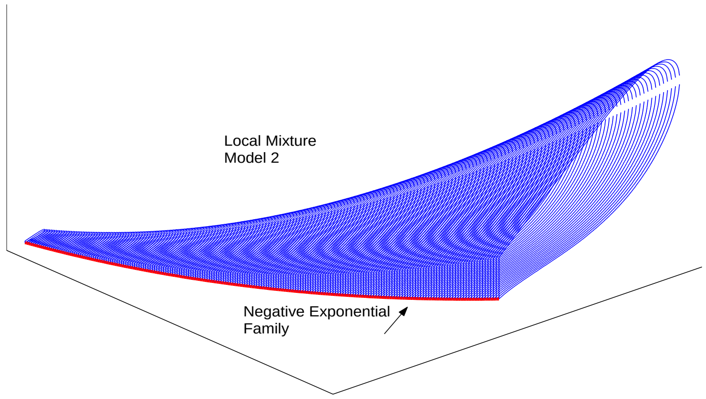

Figure 4.6: Visualization of Model 2

4.2.3 Model 3

Here we introduce a model which perhaps does not have a clear interpretation in terms of the asymptotic expansions derived from Theorem 12, but nevertheless can be useful in practice. Model 3 is generated when we impose the restriction $\gamma_4 = 0$. The corresponding density has the following form

$$g_3(x; \mu, \gamma_2, \gamma_3) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) p_3(x; \mu, \gamma_2, \gamma_3),$$

where $p_3(x; \mu, \gamma_2, \gamma_3)$ is the cubic polynomial

$$p_3(x; \mu, \gamma_2, \gamma_3) = \left(\frac{\gamma_3}{6}\right) \frac{x^3}{\mu^3} + \left(\frac{\gamma_2}{2} - \frac{3\gamma_3}{2}\right) \frac{x^2}{\mu^2} \left(-2\gamma_2 + 3\gamma_3\right) \frac{x}{\mu} + \left(1 + \gamma_2 - \gamma_3\right).$$

Using the results of Schmidt and Heß (1988) it easy to verify that the parameter space for this model is given by

$$D_3 = \{(\mu, \gamma_2, \gamma_3) \,:\, \mu > 0\,,\, \gamma_3 > 0\,,\, u_3(\mu, \gamma_2, \gamma_3) < 0\}\,,$$

where

$$\begin{aligned} u_3(\mu, \lambda_2, \lambda_3) &= 36\gamma_3^2\gamma_2^2 + 18\gamma_3\gamma_2^2 - 20\gamma_3\gamma_2^3 - 36\gamma_3^3\gamma_2 - 18\gamma_3^2\gamma_2 \\ &\quad +18\gamma_3^3 - 6\gamma_2^3 + 18\gamma_3^4 + 6\gamma_2^4 - 9\gamma_3^2. \end{aligned}$$

Note this model is restricting the third normalized moment of the mixing distribution to be positive ($\gamma_3 > 0$), but not restricting the values of the squared coefficient of variation. In this way, we can impose again the obvious soft boundary $\gamma_2 > 0$. Mimicking Model 2, we can also impose boundaries of the type

$$(3 - K_2)\gamma_2^2 \le \gamma_3 \le (3 - K_1)\gamma_2^2.$$

In Figure 4.7, the hard boundary is plotted in red and the $k = 2$ boundary is plotted in blue.

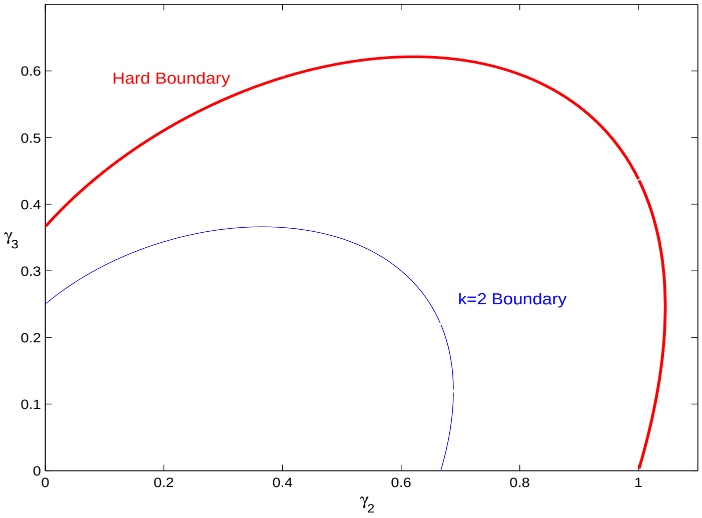

Figure 4.7: Boundaries of Model 3

4.2.4 Relationship with other models

In this section we review some relationships that we found between Local Mixture Models of the Negative Exponential and some other statistical models in the literature. These relationships can be exploited in the future to further development and understanding of local mixture models in general.

Weibull distribution

Suppose we reparametrize the generalized inverse Gaussian distribution (4.1) in a different way as follows

$$q_\gamma(\mu; m, \epsilon) = \frac{m^{-\gamma}\exp(-m^{2\gamma}/\epsilon)\mu^{\gamma-1}}{2K_\gamma(m^{2\gamma}/\epsilon)}\exp\left(-\frac{d_\gamma(\mu; m)}{2\epsilon}\right),$$

where

$$d_\gamma(\mu; m) = \frac{(\mu - m)^2}{\mu m^{1-2\gamma}}$$

and $\mu, \epsilon > 0$ and $\gamma \in \mathbb{R}$. Then this family defines again a proper dispersion model for three specific values of γ, namely $\pm 1/2$ and zero. The values $\pm 1/2$ give rise to the Reciprocal Inverse Gaussian and Inverse Gaussian distribution, respectively. The Hyperbola distribution is again obtained with $\gamma = 0$.

The relationship with the Weibull distribution is the following. If a random variable θ has a positive stable distribution with index α and parameter δ (see Feller (1970)) then its Laplace transform is given by

$$\int_0^\infty \exp(-s\,\theta)dQ_\alpha(\theta) = \exp\left(-\frac{\delta\, s^\alpha}{\alpha}\right).$$

This means that if μ has a reciprocal positive stable distribution with index α and parameter δ, then the mixture

$$g(x; Q_{\alpha,\delta}) = \int \frac{1}{\mu}\exp\left(-\frac{1}{\mu}\right)dQ_{\alpha,\delta}(1/\mu)$$

is a Weibull distribution with scale α/δ and shape α. To see this, just consider

$$\begin{aligned} P[X > x; Q_{\alpha,,\delta}] &= \int_0^\infty P[X > x; \mu]dQ_{\alpha,\delta}(1/\mu) \\ &= \int_0^\infty \exp\left(-\frac{x}{\mu}\right)dQ_{\alpha,\delta}(1/\mu) \\ &= \exp\left(-\frac{\delta\, x^\alpha}{\alpha}\right), \end{aligned}$$

which gives the desired result. Denote by $T_p(\mu, \epsilon)$ the reproductive exponential dispersion model with unit variance functions of the form

$$V_p(\mu) = \mu^p$$

for $p \in \mathbb{R}$ (see Appendix C). This is the so-called Tweedie Family. Note that this class of models includes the Gamma ($p = 2$) and the Inverse Gaussian ($p = 3$). Denote by $T_p(\infty, \epsilon)$ the limit of the distributions $T_p(\mu, \epsilon)$ when μ goes to infinity. From Jorgensen (1997) we have the following Theorem.

Theorem 23 *The distribution $T_p(\infty, \epsilon)$ is positive stable with index $\alpha \in (0, 1)$ if $p > 2$, where $\alpha = (p - 2)/(p - 1)$.*

The distributions $T_p(\infty, \epsilon)$ are positive stable with $\alpha = (p - 1)/(p - 2)$ and $\delta = \alpha\sqrt{2/\epsilon}$. This means that the family of Weibull distributions with scale $\sqrt{\epsilon/2}$ and shape α for $\alpha \in (0, 1)$ can be constructed as mean mixtures of the negative exponential distribution when the mean has a mixing distribution which is the reciprocal of a $T_p(\infty, \epsilon)$ dispersion model. For example, for $p = 3$ we have $\alpha = 1/2$. The corresponding $T_3(\infty, \epsilon)$ distribution is clearly the limit of the Inverse Gaussian distribution when the mean goes to infinity, which in this case is

$$q_{1/2}(\theta; \epsilon) = \frac{1}{\sqrt{2\pi\epsilon\theta^3}} \exp\left(-\frac{1}{2\epsilon\,\theta}\right).$$

The corresponding mixture is clearly a Weibull with scale $\sqrt{\epsilon/2}$ and shape $\alpha = 1/2$. Equivalently, by making the change of variable $\mu = 1/\theta$, this distribution transforms to

$$q_2(\mu, \epsilon) = \frac{1}{\sqrt{2\pi\epsilon\mu}} \exp\left(-\frac{\mu}{2\epsilon}\right),$$

which is the limit of a reciprocal Gamma distribution when the position parameter m goes to zero.

Smooth goodness-of-fit tests

The idea of smooth goodness-of-fit testing was introduced by Neyman (1937). A good review can be found in Rayner and Best (1989) and the references therein.

Consider a parametric model $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ on which we would like to perform an omnibus goodness-of-fit test. One way of doing this is to construct an embedding in a finite dimensional space on which we can perform a standard score type test. The usual form of the embedding space is given by

$$g(x;\theta,\lambda_1,\ldots,\lambda_k) = \frac{f(x;\theta)\exp\left\{\sum_{i=1}^{d}\lambda_i h_i(x)\right\}}{M(\theta,\lambda_1,\ldots,\lambda_k)}\,, \tag{4.10}$$

where M is a normalizing constant and the functions $h_i(x)$ are suitably chosen; often they are a set of orthogonal polynomials with respect to the base model.

The form (4.10) is mimicking the behavior of an exponential family instead of a mixture family. This is clearly a Bundle of General Exponential Families as described in section 2.3 .On the other hand, in the papers of Rayner and Best (1986) and Koziol (1987) essentially local mixture models are used in a different context. The specific form is given by

$$g(x;\theta,\lambda_1,\ldots,\lambda_k) = f(x;\theta)\left\{1+\sum_{i=1}^{d}\lambda_i h_i(x)\right\}, \tag{4.11}$$

where the functions $h_i(x)$ must be chosen such that

$$\int h_i(x)f(x;\theta)dx = 0\,,$$

exactly as we do.

Phase type distributions

Consider the time until absorption on a continuous time Markov chain with a finite state space. The distribution of this time is called a phase-type distribution. This type of distributions were introduced in Neuts (1994), and have densities of the form

$$f_X(x|A,\alpha) = -\alpha^t \exp\{xA\}A\mathbf{1}, \tag{4.12}$$

where $\boldsymbol{\alpha}$ is a vector of initial probabilities, A is a $k \times k$ matrix giving the transition rates, and $\mathbf{1}$ is the k-dimensional vector of 1s. Note that we use matrix exponentiation in this formula. Also note that this is positive, since the sum of the rows of A are all negative.

Phase type distributions always have rational Laplace transforms. We define the *degree* of the distribution as the order of the polynomial in the denominator written when in irreducible form. A representation of a phase type distribution is given by the (A, α) form given in equation 4.12. Note that there can be many different representations for the same distribution. The *order* is then defined as the minimum of the orders (i.e. k in equation 4.12) of all possible representations. Note that the order is always greater than or equal to the degree.

The *triangular* order is defined as the minimum of k over all representations where A is upper triangular. If this exists then this is greater than or equal to the order. See, for example, O'Cinneide (1990) for more details. Note that if A is in upper triangular from this has a direct interpretation in lifetime data analysis that the item is degrading, i.e. moving monotonically through the different states, see Aalen (1995).

Consider the local mixture of an exponential of the form

$$f_X(x|\mu, \nu_1, \nu_2) = \frac{1}{\mu}e^{-x/\mu} + \nu_1 \frac{\partial}{\partial \mu}\left(\frac{1}{\mu}e^{-x/\mu}\right) + \nu_2 \frac{\partial^2}{\partial \mu^2}\left(\frac{1}{\mu}e^{-x/\mu}\right). \qquad (4.13)$$

This can be expressed as a linear combination of Gamma densities of the form

$$p(\mu, \nu_1, \mu_2)\Gamma(3, \mu) + q(\mu, \nu_1, \mu_2)\Gamma(2, \mu) + r(\mu, \nu_1, \mu_2)\Gamma(1, \mu)\,, \qquad (4.14)$$

where

$$\Gamma(i, \mu) = \frac{1}{\Gamma(i)\mu^i}x^{i-1}e^{-x\mu}.$$

The density (4.14) has been explored in the paper O'Cinneide (1993).

The hard boundary for this local mixture is given by

$$p, r \geq 0, q \geq -\sqrt{2pr}.$$

The following results shows when the local mixture is phase type.

Theorem 24 *The local mixture given by (4.13) is a phase type distribution if and only if it lies strictly inside the hard boundary.*

Proof: Direct calculation shows that the Laplace transform of the density give by (4.13) has only one real pole. Hence we can apply the characterisation theorem for phase type distributions, see O'Cinneide (1990). ■

We can use Theorem 7.5 from O'Cinneide (1993) to characterise the triangular order of the local mixture.

Theorem 25 *The triangular order of 4.13 is given by*

1. 3 *if* $q \geq 0$ *and this agrees with the order;*
2. $3 + \lceil \frac{\xi}{2-\xi} \rceil$*, where* $\xi = q^2/pr$*.*

This has the interpretation that, as we approach the hard boundary, the number of states needed in the representation grows without bound. Alternatively, if we are modelling with a three state model we have a soft boundary in the parameter space. Also note that case 1 is precisely the case where the density given by 4.14 is exactly a convex mixture of three Gammas. For estimation in phase type models see Asmussen, Nerman, and Olsson (1996), and in the censored case see Olsson (1996).

4.3 Statistical Inference

In this section, we show empirically that first order asymptotic inference over the parameters in a local mixture model, seems to apply when the true data generation process is a scale dispersion mixture with small dispersion parameter ϵ.

We will focus first on Model 1. We generated 10,000 independent replications of a random sample of size n from a scale dispersion mixture of the negative

Mixing Distribution		
Gamma	G1 $\gamma_2 = 0.1$ $n = 100$	G2 $\gamma_2 = 0.1$ $n = 1000$
	G3 $\gamma_2 = 0.5$ $n = 100$	G4 $\gamma_2 = 0.5$ $n = 1000$
Reciprocal Gama	RG1 $\gamma_2 = 0.1$ $n = 100$	RG2 $\gamma_2 = 0.1$ $n = 1000$
	RG3 $\gamma_2 = 0.5$ $n = 100$	RG4 $\gamma_2 = 0.5$ $n = 1000$

Table 4.1: Parameter values used in the simulations for Model 1

exponential distribution with mean 5 and according to the values in Table 4.1.

Here γ_2 denotes the squared coefficient of variation of the mixing distribution. The labels G1, RG1 and so on are just simply identifiers of that particular combination of the parameters. We generated the samples in the usual two-step way, that is, first generating a value μ from the mixing distribution and then generating a value of x from a negative exponential distribution with mean μ.

First and second rows of figures 4.8 to 4.15 show the histograms of the maximum likelihood estimators (mle's) $(\hat{\mu}^{mle}, \hat{\gamma}_2^{mle})$ of the parameters (μ, γ_2) of model 1 defined in (4.3) and the corresponding normal probability plots, respectively. Also plotted are the histograms and normal probability plots of the moment estimator of γ_2, that is

$$\hat{\gamma}_2^{mom} = \frac{DS_1(\boldsymbol{x})}{2},$$

where $DS_1(\boldsymbol{x})$ is the dispersion score defined above.

The mle's were calculated using the function `fmincon` of the Optimization Toolbox in MATLAB. Recall, the parameter space for model 1 is the product $(0,\infty)\times(0,1)$. Motivated by the usual asymptotic confidence interval for the mean of a negative exponential distribution, and because we are interested only on inferences in small mixing distributions, we restricted the search for the values of μ to the interval $\bar{\boldsymbol{x}}(1 \pm 2k/\sqrt{n})$ for an specific value of the inflation factor k. We found in our simulations that $k = \sqrt{6}$ always was adequate good enough to capture of all the relevant information in the corresponding likelihoods.

In all the simulations, the mle's were calculated using the $k = 2$ boundary which in this case restricts the parameter γ_2 to be on the interval $[0, 2/3]$. Therefore, we restricted our search for the mle's for model 1 within the compact set

$$[\bar{\boldsymbol{x}}(1 - 2k/\sqrt{n}), \bar{\boldsymbol{x}}(1 + 2k/\sqrt{n})] \times [0, 2/3]$$

For comparison purposes, we only show both estimators of γ_2 when they yield a value inside the interval $(0, 2/3)$. So, our statements about the distribution of both estimators of γ_2 will be conditional on being in the interval $(0, 2/3)$. It is important to mention that we should take that restriction into account when interpreting the corresponding normal probability plot. When there is a visible scree in the corresponding normal probability plot, it means that the estimator has been truncated at that value and the intersection with the vertical axis shows the proportion of values for which this happens.

Also, in the third row of each of the figures, we show the scatter plots of $(\hat{\mu}^{mle}, \hat{\gamma}_2^{mle})$, $(\hat{\mu}^{mle}, \hat{\mu}_2^{mom})$ where $\hat{\mu}^{mom} = \bar{\boldsymbol{x}}$, the sample mean, and finally the scatter plot of $(\hat{\gamma}_2^{mle}, \hat{\gamma}_2^{mom})$.

As can be seen from the plots, for the cases where $\gamma_2 = 0.1$, the distribution of the mle $\hat{\mu}^{mle}$ of the mixture mean μ is slightly skewed to the right compared to a normal distribution. This is mainly because of some few extreme points that appear in the right tail. This skewness is clearly increased when $\gamma_2 = 0.5$, specially when the mixing distribution is reciprocal Gamma. We will explain that behavior later. Also note that, in all cases this skewness decreases a little bit when increasing the sample from $n = 100$ to $n = 1000$. Also, all the histograms appear to be centered at the correct true value of the mean of the mixture which equals 5 in all cases.

A similar situation occurs for the the mle $\hat{\gamma}_2^{mle}$ of γ_2, although this estimator can reach either the positivity boundary or the $k = 2$ boundary. Now, the distribution of $\hat{\gamma}_2^{mle}$ is not always centered at the correct value. In simulations G3 and G4, the corresponding histogram is clearly centered at around 0.35, a wrong value, the correct value is supposed to be 0.5. The situation is even worse for the Reciprocal Gamma distribution. In simulations RG3 and RG4, the corresponding histogram of $\hat{\gamma}_2^{mle}$ is clearly centered at around 0.25, which is away from the correct value of 0.5. It is clear also that we cannot expect this to improve by increasing the sample size because only the dispersion of the histogram will change (actually decrease), not its location. In general, this bias is a result of the fact that local mixtures are designed to mimic genuine mixtures when the mixing distributions are small. This is saying that a Gamma distribution or a Reciprocal Gamma with $\gamma_2 = 0.5$ are not small mixing distributions in that sense. The local mixture model is trying to fit a mixing distribution with smaller squared coefficient of variation in each case.

It is interesting to note that in simulations RG3 and RG4, the value of 0.25 where the histograms of $\hat{\gamma_2}^{mle}$ are centered, corresponds to the value of the dispersion parameter ϵ for which the coefficient of variation is 0.5. In a reciprocal Gamma distribution with dispersion parameter ϵ, the squared coefficient of variation is

$$\gamma_2 = \frac{\epsilon}{1 - 2\epsilon}$$

which clearly behaves like ϵ when ϵ is small, as should happen according to Corollary 6 when we only retain terms of order ϵ. So, roughly, local mixture model 1 is trying to fit a value of γ_2 that corresponds to an approximation of that kind. But there is a limit for this behavior of the local mixture model. In simulations G3 and G4 where $\epsilon = \gamma_2 = 0.5$, because in a Gamma distribution with dispersion ϵ the squared coefficient of variation is ϵ, we see that the distributions of $\hat{\gamma}_2^{mle}$ concentrates at around 0.35. Now, it is clearer that local mixtures behave like true mixtures only when the mixing distribution is small. The smallness in the case of mean mixtures of the negative exponential distribution is dictated by the smallness of the squared coefficient of variation of the mixing distribution γ_2.

With respect to the moment estimate $\hat{\gamma}_2^{mom}$ of γ_2 , recall it is proportional to the dispersion score $DS_1(\boldsymbol{x})$ which has been used many times in the literature of testing for overdispersion in the negative exponential distribution. See, for example, page 234 of Mosler and Seidel (2001). From our simulations it is clear that the distribution of this moment estimator does look very much like a truncated normal distribution although the corresponding histogram is always nearly centered around the correct value of the squared coefficient of variation γ_2. In fact, from the result in page 70 of Lindsay (1995), it is not difficult to check, using the properties of convergence in probability, that $\hat{\gamma}_2^{mom} = DS_1(\boldsymbol{x})/2$ is a weakly consistent estimator of the squared coefficient of variation of the mixing distribution. This justifies the fact that the histograms of that estimator concentrate around the correct value.

In the last plot in each of the figures, we show the scatter plots of $\hat{\gamma}_2^{mle}$ vs $\hat{\gamma}_2^{mom}$, the red dashed line is the 45^o line plotted for reference. As expected from the discussion above, from those scatter plots we can see a good agreement between both estimators for small values of them, except in the case where $\gamma_2 = 0.5$. In those cases, $\hat{\gamma}_2^{mom}$ tends to be bigger than $\hat{\gamma}_2^{mle}$ and this is most clear for sample size $n = 1000$, where $\hat{\gamma}_2^{mom}$ is about to converge to the correct value. Thus, both estimators can complement each other in the following sense. If $\hat{\gamma}_2^{mom}$ crosses the hard boundary 2/3 then we have a simple diagnostic about the possible non-smallness of the mixing distribution, which may prompt us to reconsider our modeling assumptions.

Also clear from the plots, is the elliptically contoured joint distribution of $(\hat{\mu}^{mle}, \hat{\gamma_2}^{mle})$. Remarkably, we see that when the mixing is small, $(\hat{\mu}^{mle}, \hat{\gamma_2}^{mle})$ are nearly orthogonal (which under asymptotic normality will imply independence) extending, at least empirically, the orthogonality under non-mixing proved in Theorem 21. It is important to recall here that this contours can be centered at a wrong value.

Finally, we show in the figures the scatter plots of $\hat{\mu}^{mle}$ against $\bar{\boldsymbol{x}}$. The agreement between both estimates is quite good, specially for large sample sizes. This is not the case when the mixing distribution is reciprocal Gamma with $\gamma_2 = 0.5$. In simulations RG3 and RG3 we can clearly see that $\hat{\mu}^{mle}$ is slightly larger than the sample mean. It is also clear that if the variance of the mixture distribution exists then the sample mean is a consistent estimator of the true mean μ. So, the local mixture mle of μ tend to be biased when the mixing distribution is not small. This bias has a clear explanation. As shown in the proof of Theorem 12, the mean centered expansion of a mixture relies on the fact that the function $A_1(\vartheta, \epsilon) = O_\vartheta(\epsilon)$ is small. Then, clearly, if the scale dispersion mixing distribution does have a small value of the dispersion parameter ϵ (and therefore not a small squared coefficient of variation), the local mixture approximation is not going to be good and this is where the bias on the estimator of the mean comes from. When the mixing distribution is not so small, the local mixture model tries to fit a larger of value of the mean instead of increasing the variance. The bias is not present in simulations G3 and G4 because a Gamma mixing distribution has the property of being central and this implies $A_1(\vartheta, \epsilon) \equiv 0$.

In practice, applied statisticians are usually only interested in the mean of the distribution of the observed data and then treat the mixing distribution as nuisance. In this respect, we can use the profile log-likelihood ratio function (3.40) defined here as

$$\ell_p^0(\mu) = \ell_{lm}(\hat{\mu}, \hat{\gamma}_2) - \ell_{lm}(\mu, \hat{\gamma}_2(\mu))$$

to make inferences about the parameter of interest μ, treating the local mixture parameter γ_2 as nuisance. A couple of typical plots of such profile log-likelihood ratio functions are plotted in Figure 4.16. It is well known that under the usual assumptions of the classical first order asymptotic theory of parametric models, when a model is embedded in a larger one, twice the profile log-likelihood ratio evaluated at the true value of the parameter has an asymptotic χ^2 distribution with one degree of freedom provided the embedded model is one dimensional. Thus, in Figure 4.16 we show the Q-Q plots (under a $\chi^2_{(1)}$) of the profile log-likelihood ratios evaluated at the true mean of the mixing distribution from 10,000 replications of samples of size $n = 100$ and $n = 1000$, respectively, under a Gamma mixing distribution with mean 5 and squared coefficient of variation 0.1. It is clear from the plots that the $\chi^2_{(1)}$ provides a good fit to those profile ratios. This implies in particular that asymptotic confidence intervals for the true mean of the mixing distribution can be constructed using this empirical result and, consequently, they should have asymptotically the correct coverage. As expected, we also found empirically that this χ^2 approximation no longer holds if the mixing distribution has a larger squared coefficient of variation.

At this point, it is important to mention that, for all the previous simulations, we found very similar results for other scale dispersion mixtures, such as when the mixing distribution is lognormal or other members of the Generalized Inverse Gaussian like the Hyperbola distribution. These cover a wide range of mixing distributions.

To better understand the frequently mentioned smallness of the mixing distribution, we now consider the effect of extreme observations in the estimation of the parameters in a local mixture model. First of all, large observations are evidence of contamination mixing, as discussed above. If the mixing distribution is small and has finite variance, we cannot expect great deviations from a negative exponential in the sense that the largest order statistics cannot be way larger than those under the non mixing assumption. To give a concise example, consider a reciprocal Gamma distribution with position m and dispersion ϵ. This mixing distribution does not have finite variance for values of ϵ greater than 0.5. Moreover, the generated scale dispersion mixture is a Pareto distribution of the second kind whose density is given by

$$g(x;Q_{-1}) = \frac{1}{m}\left(1+\frac{x\epsilon}{m}\right)^{-(1+1/\epsilon)}.$$

Clearly, this is a heavy tailed distribution for positive values of ϵ. So, when we generate samples from this distribution, we expect to have large upper order statistics compared to the negative exponential. This is what is happening in Figures 4.8 to 4.15 when we observe some few distant points from the histograms or the scatter plots. Those points might correspond to a case where the mixing distribution yielded a large observed value compared to its position parameter and therefore, the negative exponential value has to be generated with that value of the mean, giving a high probability to observe a very large value compared to mean of the mixing distribution. In other words, sometimes the value of ϵ is not small enough for the normal approximation (3.25) of a dispersion model to hold. Also, this approximation only makes sense if the corresponding proper dispersion model has finite variance, which always occur for a sufficiently small value of ϵ.

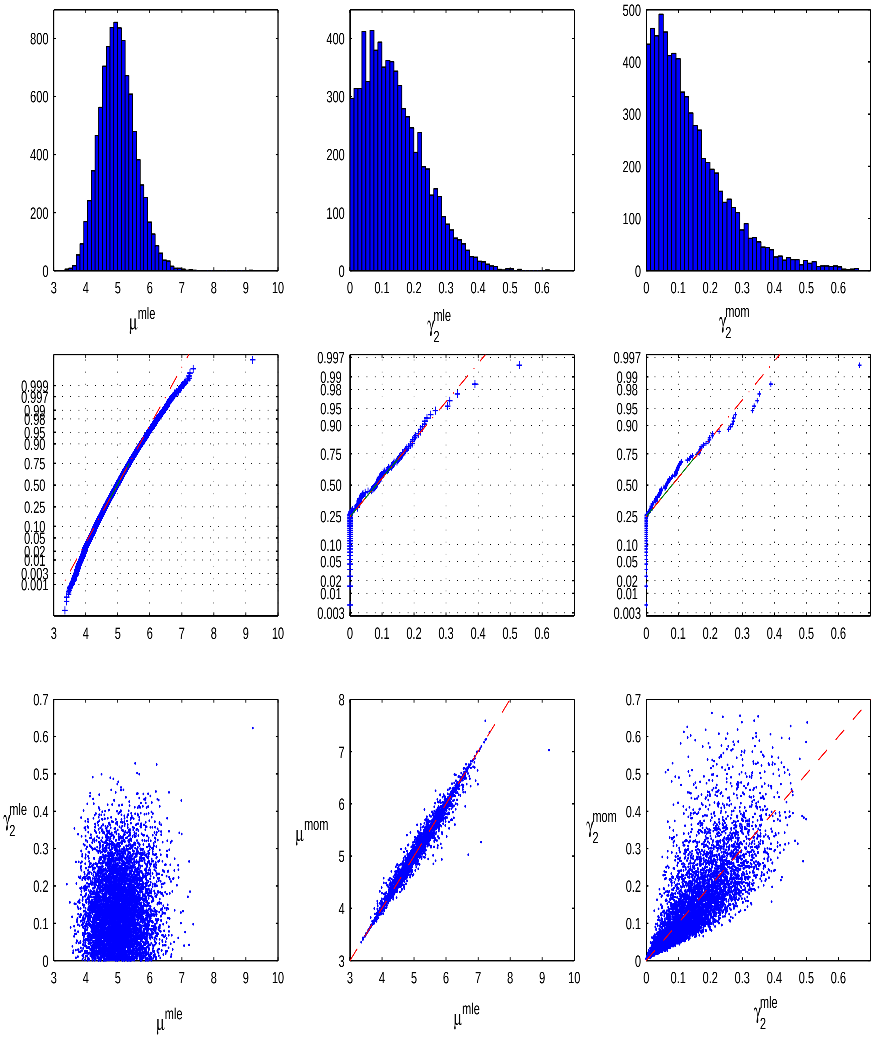

Figure 4.8: Gamma mixing with $\mu = 5$ and $\gamma_2 = 0.1$. Sample size 100

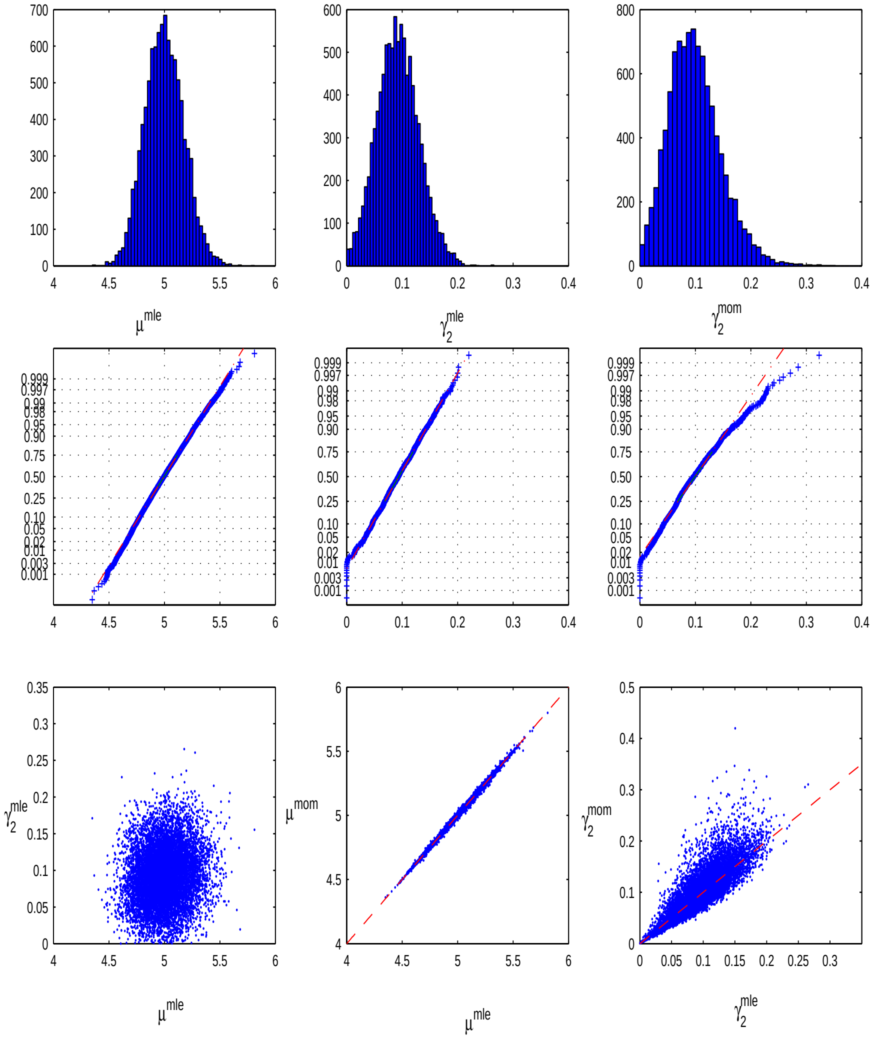

Figure 4.9: Gamma mixing with $\mu = 5$ and $\gamma_2 = 0.1$. Sample size 1000

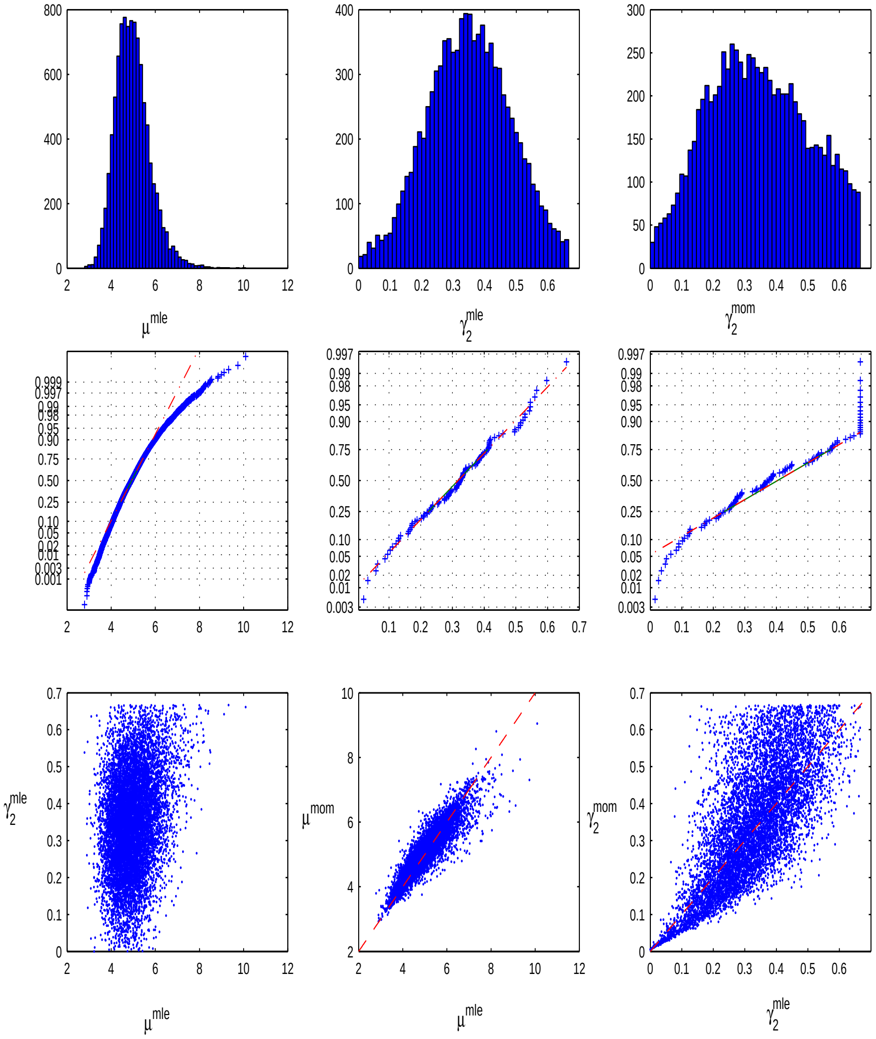

Figure 4.10: Gamma mixing with $\mu = 5$ and $\gamma_2 = 0.5$. Sample size 100

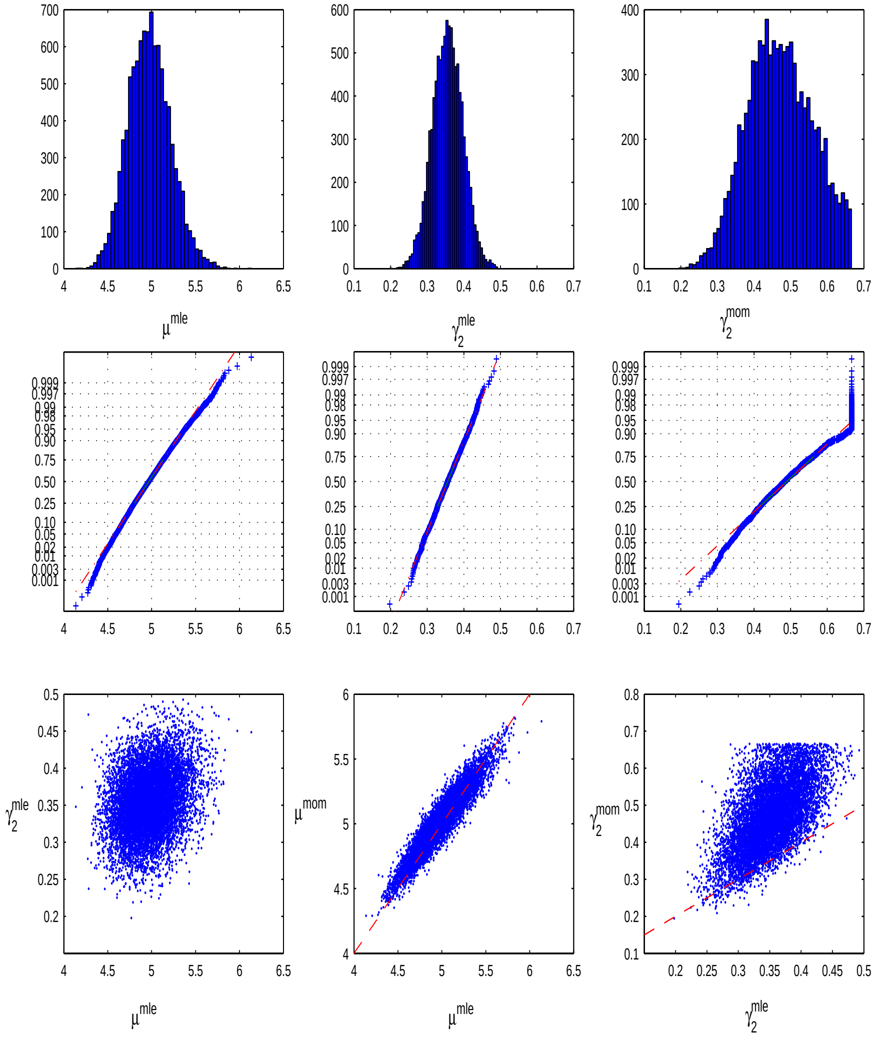

Figure 4.11: Gamma mixing with $\mu = 5$ and $\gamma_2 = 0.5$. Sample size 1000

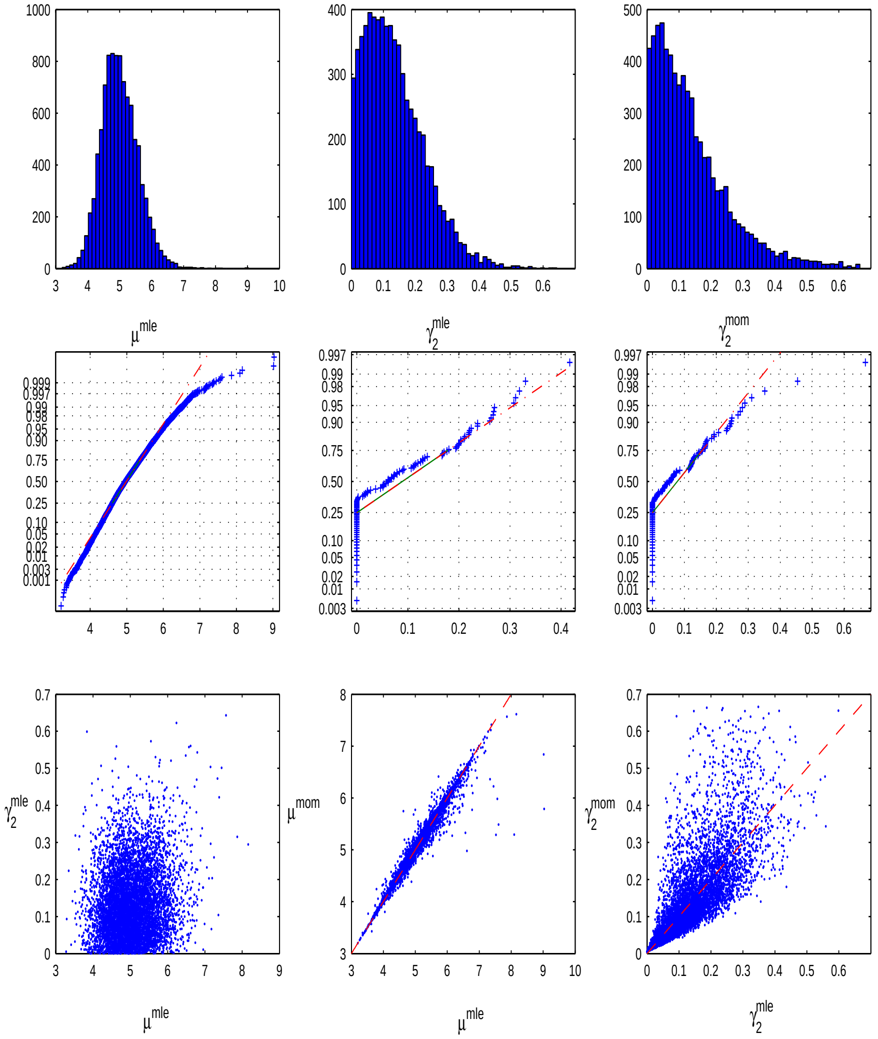

Figure 4.12: Reciprocal Gamma mixing with $\mu = 5$ and $\gamma_2 = 0.1$. Sample size 100

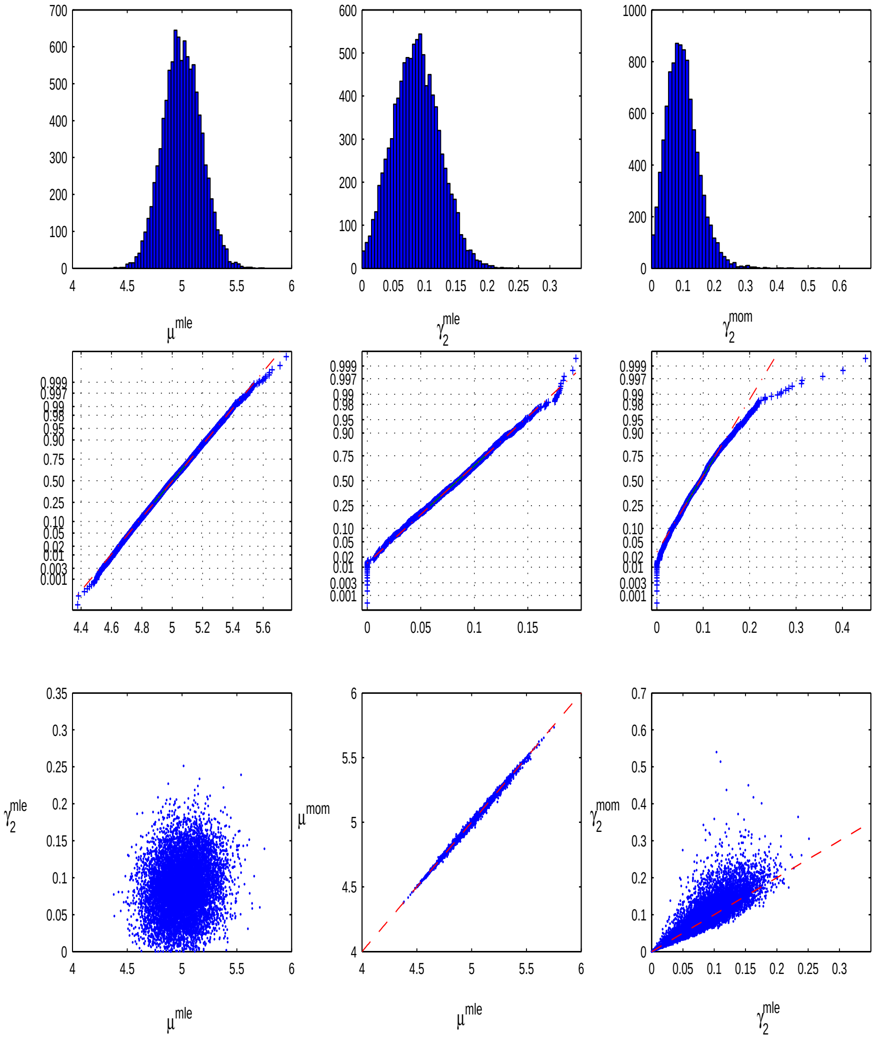

Figure 4.13: Reciprocal Gamma mixing with $\mu = 5$ and $\gamma_2 = 0.1$. Sample size 1000

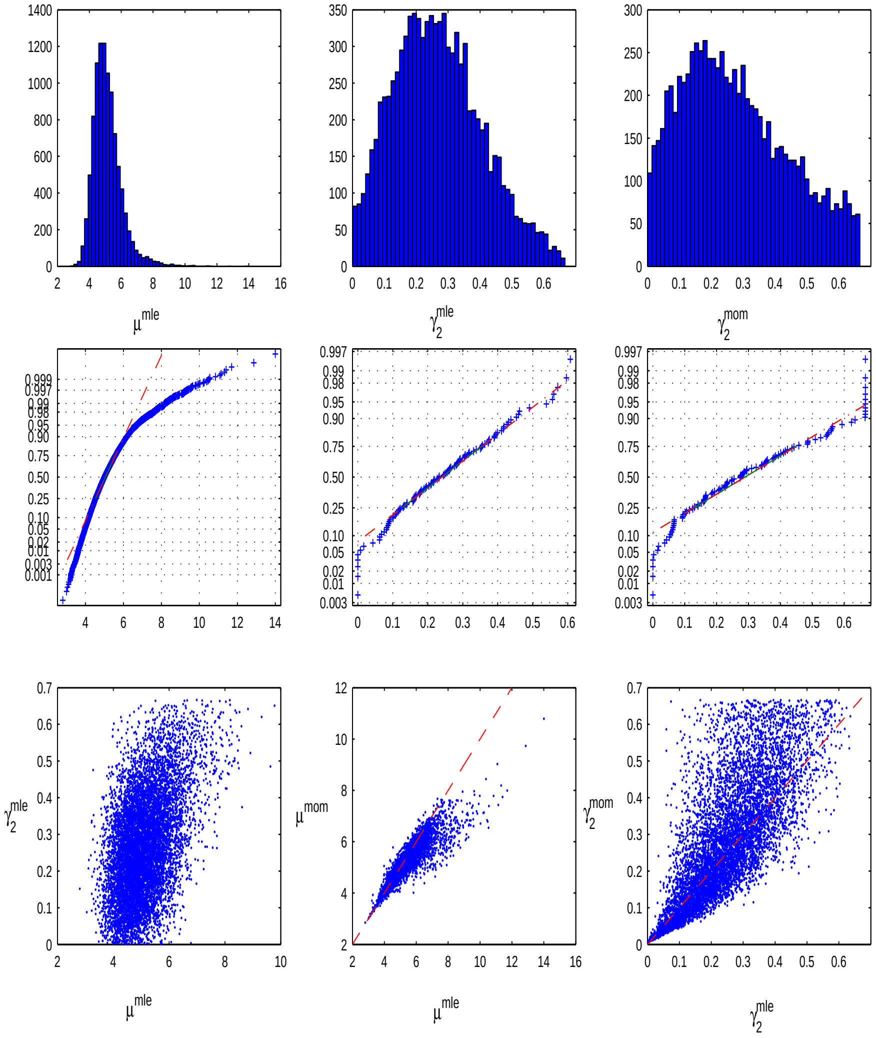

Figure 4.14: Reciprocal Gamma mixing with $\mu = 5$ and $\gamma_2 = 0.5$. Sample size 100

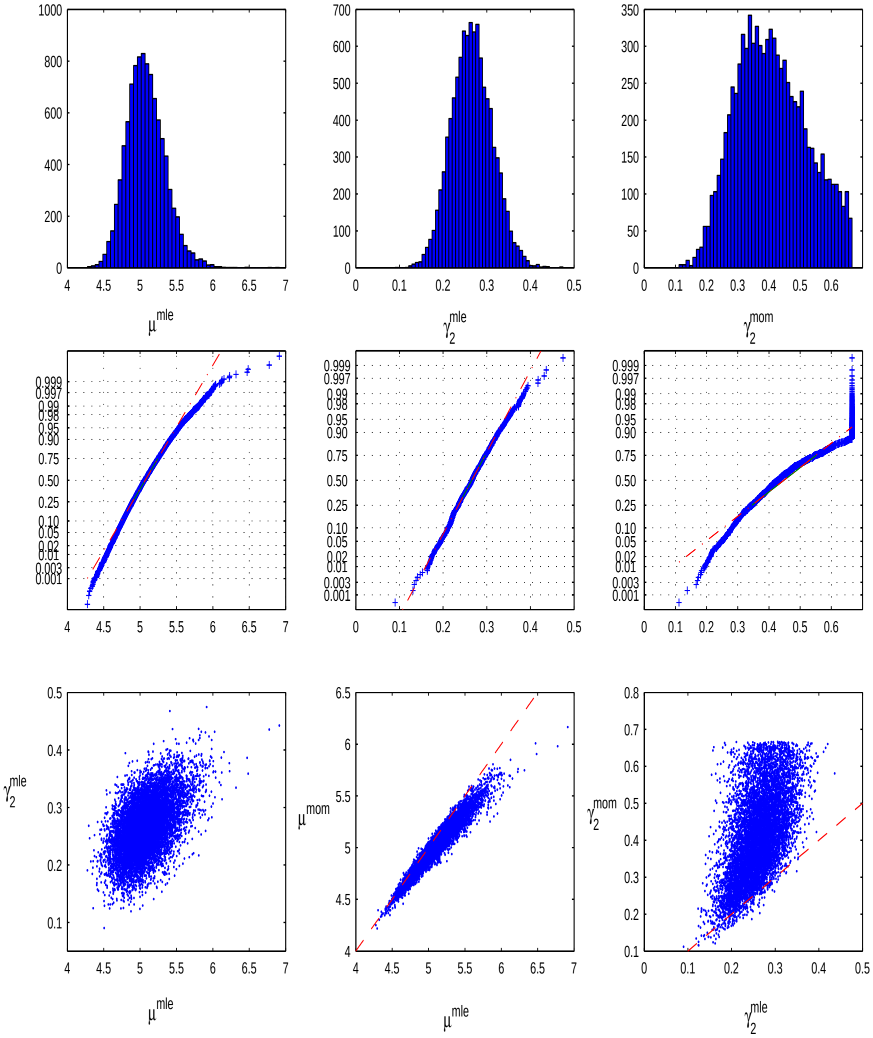

Figure 4.15: Reciprocal Gamma mixing with $\mu = 5$ and $\gamma_2 = 0.5$. Sample size 1000

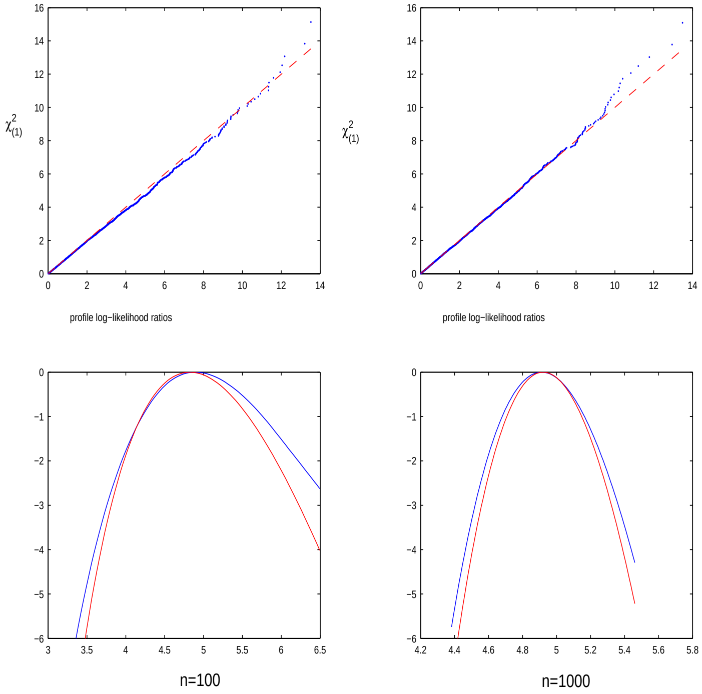

Figure 4.16: Profile log-likelihoods under Model 1

To be more graphical in this description, consider the log-likelihood contours plotted in Figure 4.17. We generated a sample of size 100 and 1000 from a reciprocal Gamma distribution with mean $\mu = 5$ and squared coefficient of variation $\gamma_2 = 0.5$. Plotted are the log-likelihoods ratios (up to a drop of 6) of this pair of sets of data under the true mixture model and using local mixture model 1. Clearly, the true likelihood is always unknown in practice, but here we it use as a gold standard for comparison. In the first data set, we observe a large maximum, which is clearly far away from the rest of the data. The local mixture likelihood is clearly not giving much inferential importance to the true point, which is in this case $(5, 0.5)$. The true likelihood is very flat in a strip which contains the true value. Something similar happens if we increase the sample size to 1000. In this case, we actually observe not only one but some other high values.

On the other hand, in Figure 4.18 we show that under a Gamma mixing distribution with coefficient of variation $\gamma_2 = 0.5$, the agreement between the local mixture and true likelihoods can be better. This is mainly because a Gamma distribution with squared coefficient of variation 0.5 is more "normal" than a reciprocal Gamma with the same mean and the same coefficient of variation. In general, we found a very good agreement between the likelihoods of the local mixture model 1 and the true likelihoods when the coefficient of variation was small (less that 0.1) and there were no discrepant observations in the data.

A very simple diagnostic of non local mixing is therefore, the presence of extreme observations that clearly affect the mle of the parameters in a local mixture model. Concrete diagnostics of this kind can be found in the work of Caroni and Kimber (2004).

Summarizing, we can get uniformly good approximations to the true likelihoods under a scale dispersion mixture, as long as the normal approximation to the corresponding mixing scale dispersion model is good. We know this always occurs for a sufficiently small value of ϵ. Now we are in a position to state an easier to understand definition of when a mixing distribution is small: a proper dispersion mixing distribution is small when the normal approximation to it is good.

A simple measure of the closeness to normality is the third normalized mo-

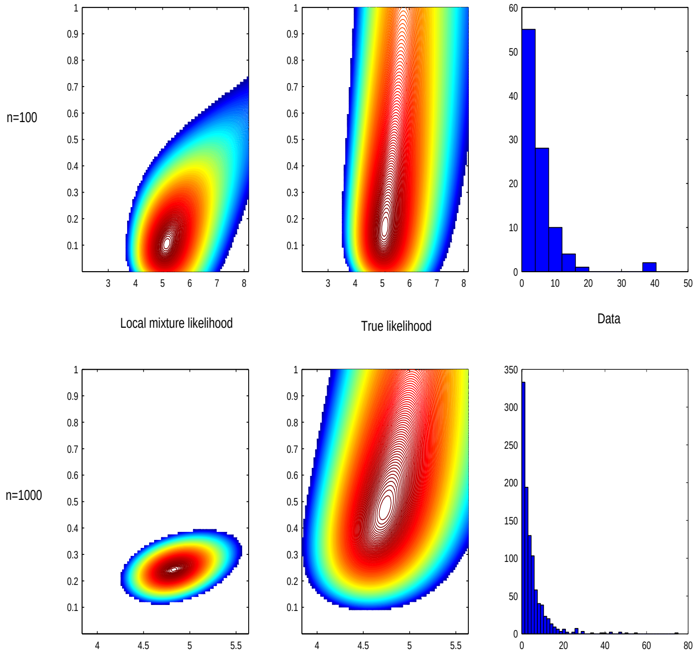

Figure 4.17: Local mixture and Reciprocal Gamma log-likelihoods

ment of the mixing distribution, which here we will denote by γ_3. This motivates the use of Model 2. Recall that this model forces the coefficient of kurtosis to be zero as in the normal distribution. Also, by imposing soft boundaries like

$$(3 - K_2)\gamma_2^2 \leq \gamma_3 \leq (3 - K_1)\gamma_2^2\,,$$

we can match the behavior of the third normalized moment, which is supposed to behave like a quadratic near the origin with the sign given by the third derivative of the deviance evaluated at its minimum. Clearly, γ_3 is a measure of closeness to normality in the following sense. First, it represents the skewness of the distribution so, if it is close to zero, we are close to a normal. If we write $\gamma_3 = (3 - \beta)\,\gamma_2^2$ as in (4.8), it is also clear that this skewness will converge faster to zero than the squared coefficient of variation. Together with the restriction $\gamma_4 = 3\gamma_2^2$, this means that the squared

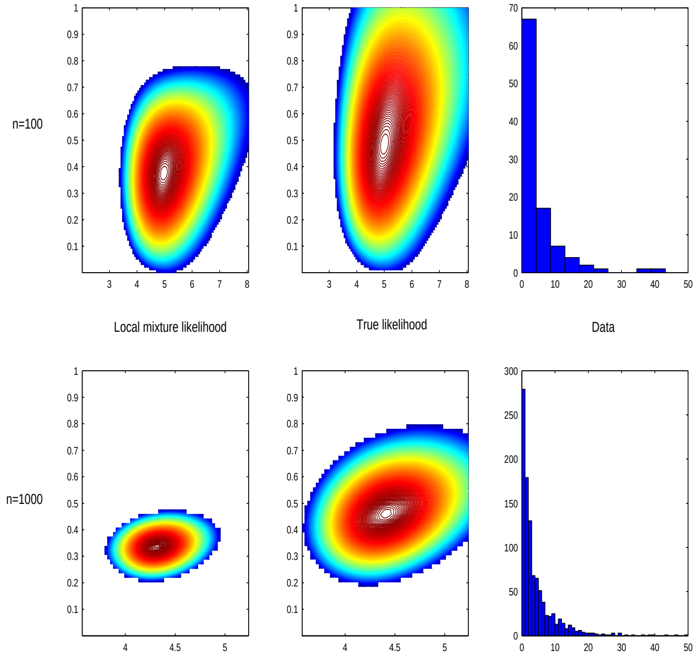

Figure 4.18: Local mixture and Gamma log-likelihoods

coefficient of variation is driving the convergence of the mixing distribution to the unmixed negative exponential. This argument has strong support not only on the asymptotic expansions derived from Theorem 12, but also on the fact that, the squared coefficient of variation of any mixing distribution of the negative exponential in its mean scale is a formal distance to the unmixed distribution, as proved by Keilson and Steutel (1974).

Now, in Figures 4.19 and 4.20 we present similar results to those shown for Model 1. We generated 10,000 independent replications of a random sample of size $n = 1000$ from scale dispersion mixtures with Gamma and Reciprocal Gamma mixing distributions with mean $\mu = 5$ and squared coefficient of variation $\gamma_2 = 0.1$. We used a small value of the squared coefficient of variation and a relatively large sample because, as shown for Model 1, local scale mixtures perform better in such cases.

In the first row of each figure are the histograms of $\hat{\mu}^{mle}$, $\hat{\gamma}_2^{mle}$ and $\hat{\gamma}_3^{mle}$ of the parameters of Model 2 defined in (4.7). We will not concentrate here on the normality of the estimators, as we are more interested in the effect of soft boundaries. The distributions of $\hat{\mu}^{mle}$ and $\hat{\gamma}_2^{mle}$ clearly appear to be centered at the correct values of $\mu = 5$ and $\gamma_2 = 0.1$. The same happens for $\hat{\gamma}_3^{mle}$, where the correct value under a Gamma mixing distribution is $\gamma_2 = 2\epsilon^2 = 2(0.1)^2 = 0.02$, while under the reciprocal Gamma is

$$\gamma_2 = \frac{4\epsilon^2}{(1-2\epsilon)(1-3\epsilon)} = \frac{4(0.0833)^2}{(1-2(0.0833))(1-3(0.0833))} = 0.044 \; .$$

In the second row of each figure, we show the scatter plots of $(\hat{\gamma}_2^{mle}, \hat{\gamma}_2^{mom})$ and $(\hat{\gamma}_3^{mle}, \hat{\gamma}_3^{mom})$. Similar to the results obtained under Model 1, here we also found a good agreement between both estimators only for small values of each of them. Also, in the same row, is the scatter plot of $(\hat{\gamma}_2^{mle}, \hat{\gamma}_3^{mle})$, however, to explain this plot we first need to mention how we calculated the mle's for model 2.

As in model 1, the mle's were calculated using the function `fmincon` of the Optimization Toolbox in MATLAB. Following the discussion concerning the reparametrization (4.8) of model 2, we judiciously chose as soft boundary for the parameter β the interval $[K_1, K_2] = [-20, 20]$. The generalized inverse Gaussian family is by far contained in such an interval. In practice, we will be required to fix an interval of that form if we really want to model local mixtures, so the interval we have chosen seems reasonable.

We plotted in red the hard boundary for (γ_2, γ_3) and in blue the $k = 2$ boundary. In dashed green is plotted the $-20 \leq \beta \leq 20$ boundary. Having this, we can see from our scatter plots that the joint distribution of the mle of (γ_2, γ_3) can reasonably be well approximated by a distribution with elliptic contours. It is also clear that there exists a positive correlation between both estimators. This is just a consequence of the fact that, under the mixing distribution, there exists a relationship of the same kind between the squared coefficient of variation and the third normalized moment, for small values of the dispersion parameter ϵ. The points that stick into the soft boundary $-20 \leq \beta \leq 20$ correspond to cases where the sample gave rise to high upper order statistics. In this sense, this boundary can be used as a simple diagnostic for detecting non-local mixing.

In the last row of each figure we present the scatter plots of $(\hat{\mu}^{mle}, \hat{\gamma}_2^{mle})$ and $(\hat{\mu}^{mle}, \hat{\gamma}_3^{mle})$. As in Model 1, here we found evidence of elliptically contoured joint distributions and approximate orthogonality between the estimators. Finally, in the last plot of each figure is the Q-Q plot of twice the profile log-likelihood ratio evaluated at the true mean for model 2. The $\chi^2_{(1)}$ fit to the empirical distribution is also quite good. Finally, we mention that we found very similar results to the previous simulations using Model 3.

Now, in Figures 4.21 to 4.23 we show the results of a simulation exercise of the performance of local mixture model 1 under no mixing. We generated 10,000 replications of samples of sizes $n = 100$ and $n = 1000$ from a negative exponential distribution with mean $\mu = 5$. First, from figure 4.21 it is clear that the sample mean is the mle of μ under model 1 for samples sizes over a thousand. This implies that it retains all the optimal properties which this estimator has, when we see the negative exponential as an exponential family.

From Figures 4.22 and 4.23, we can see that the distributions of $\hat{\gamma}_2^{mle}$ and $\hat{\gamma}_2^{mom}$ resemble a half normal distribution with zero mean specially for large sample sizes. The proportion of positive values for any of the estimators is tending to 0.5 when the sample increases. Also we can see much better agreement between both estimators for sample sizes like $n = 1000$ because the range of values they take is considerably smaller than the corresponding values under small mixing, for example with the Gamma distribution and with the same sample size. The orthogonality result in Theorem 21 is evident also from the scatter plots. Moreover, we can also observe a reasonable fit of the $\chi^2_{(1)}$ to the distribution of the profile log-likelihood ratio evaluated at the true mean. Extreme observations can also appear under no mixing, although this happen with very small probability.

Finally, to avoid repetition, we mention that we obtained similar results for model 3, conditional on the fact that $\gamma_3 \geq 0$ and imposing the same boundaries as in Model 2.

It is worth finishing this chapter with the following considerations. There

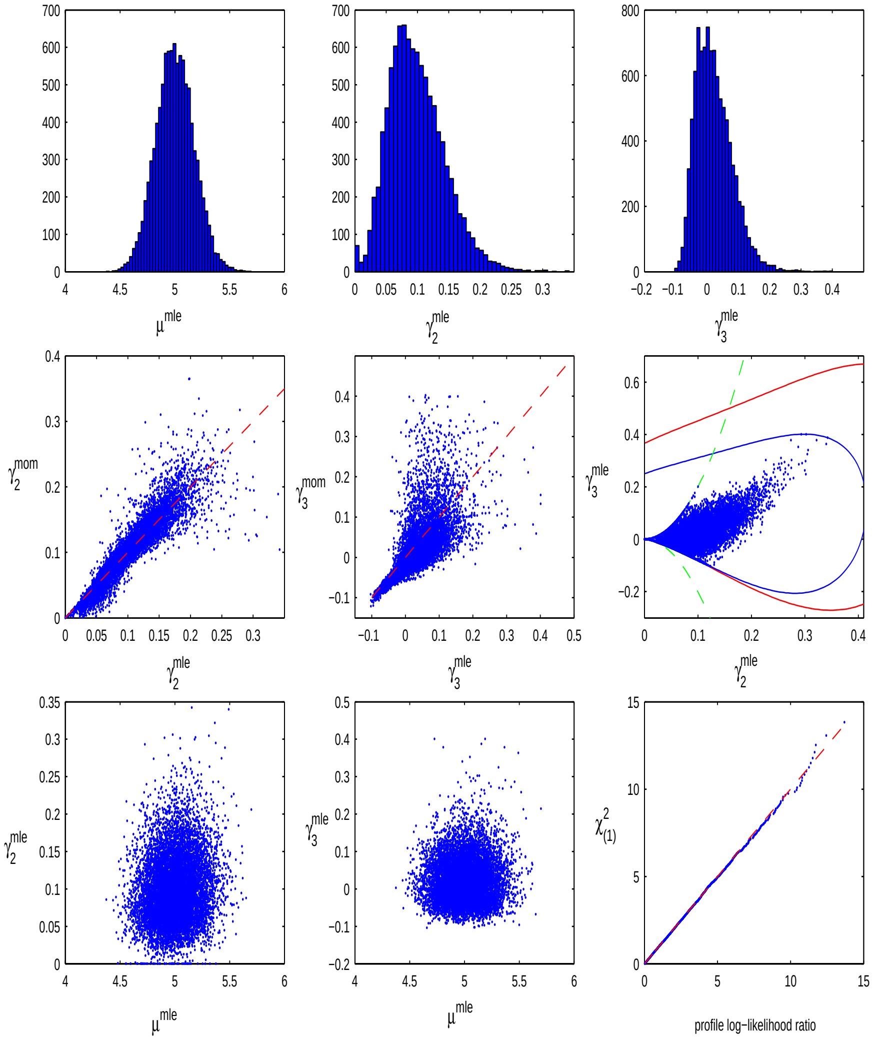

Figure 4.19: Gamma mixing with $\mu = 5$ and $\gamma_2 = 0.1$. Sample size 1000

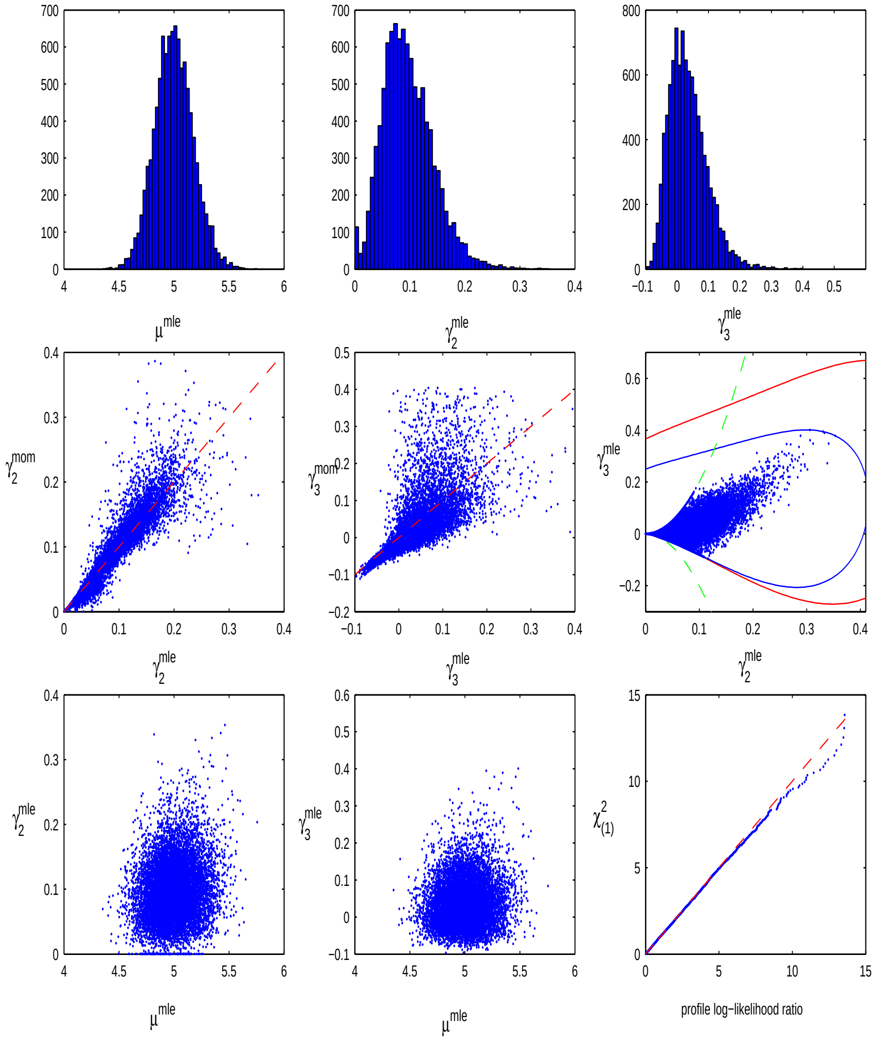

Figure 4.20: Reciprocal Gamma mixing with $\mu = 5$ and $\gamma_2 = 0.1$. Sample size 1000

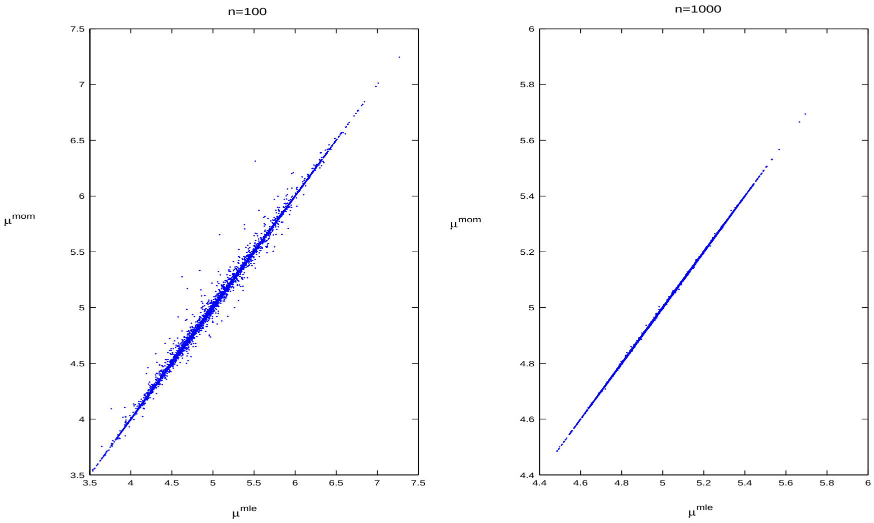

Figure 4.21: $\hat{\mu}^{mle}$ vs $\bar{\boldsymbol{x}}$ under no mixing

is trade-off between local scale mixing and estimation error in the following sense. It is well known that, for a negative exponential distribution, the maximum likelihood estimator of the mean parameter is the sample mean and that this estimator has the following asymptotic property:

$$\bar{\boldsymbol{x}} \xrightarrow{d} N\left(m, \frac{1}{nI(m)}\right),$$

when $n \to \infty$, where m is the true mean parameter and $I(m)$ is the expected Fisher's information evaluated at m. Then, we have

$$\bar{\boldsymbol{x}} \xrightarrow{d} N\left(m, \frac{m^2}{n}\right),$$

when $n \to \infty$. From the results in Appendix C, we have that if the distribution of μ is a scale dispersion model with position m and dispersion parameter ϵ, then it follows that

$$\mu \xrightarrow{d} N\left(m, \epsilon V(m)\right)$$

when $\epsilon \to 0$, but we also know that $V(m) = m^2$, so

$$\mu \xrightarrow{d} N\left(m, \epsilon\, m^2\right)$$

when $\epsilon \to 0$. Thus, there is a clear trade-off between the two types of asymptotics. If we are really observing mixtures of $\mathcal{F}$ with a very small value of ϵ but our sample size is clearly smaller than $1/\epsilon$, then we will make inferences with $\mathcal{F}$ without even notice that there was a small mixing (probably insignificant) in the sample. On the other hand, this justifies in some sense the fact that we have needed large sample sizes in order to recognize small mixing in our simulations.

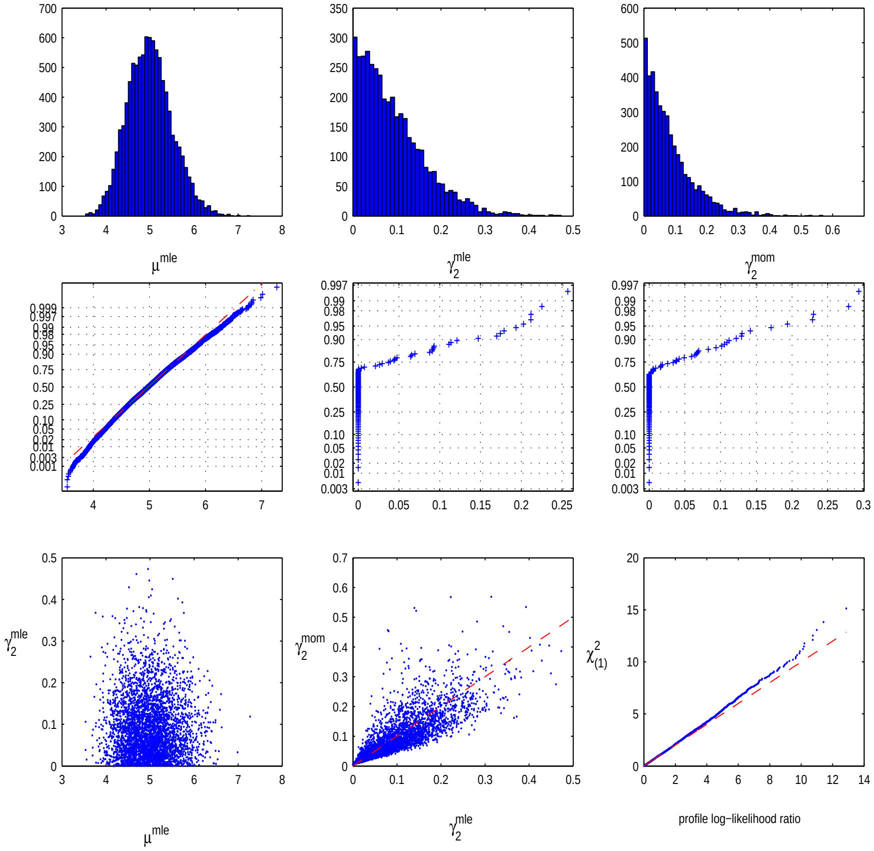

Figure 4.22: Model 1 under no mixing. Sample size $n = 100$

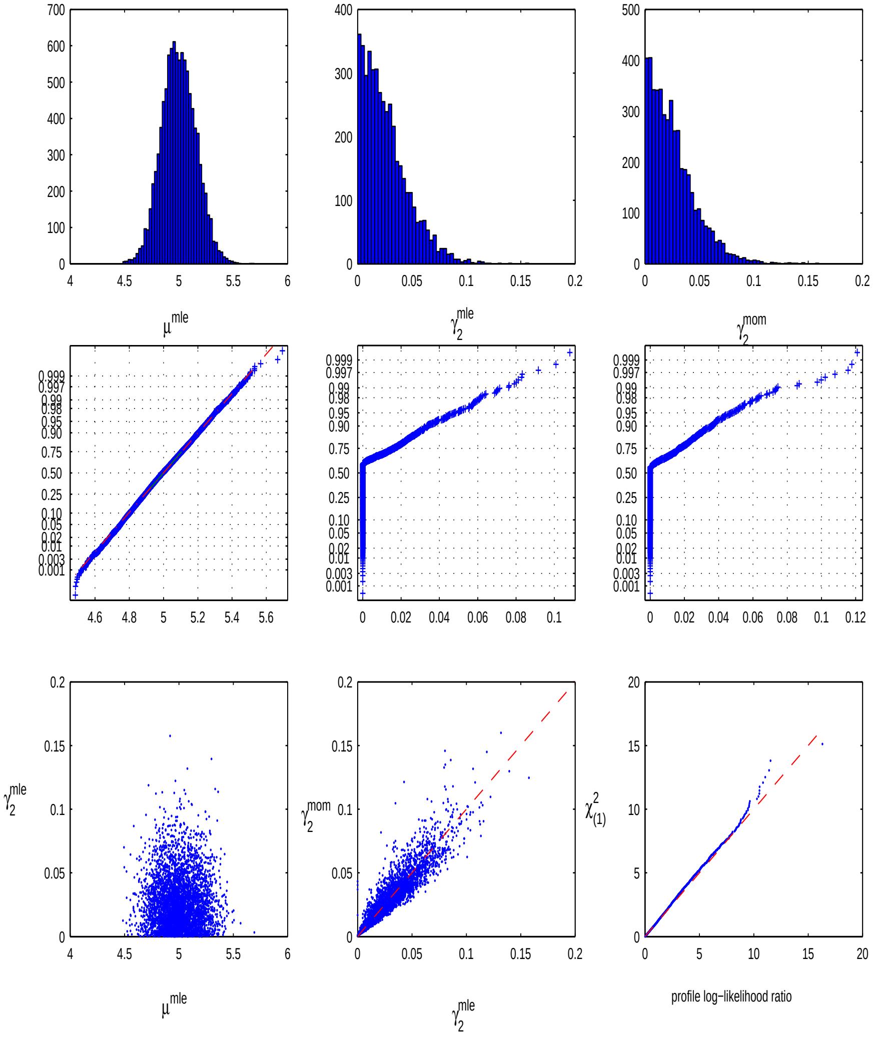

Figure 4.23: Model 1 under no mixing. Sample size $n = 1000$

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The conclusions of this part of the thesis are as follows:

1. The use of one of the simplest geometrical structures, Affine Spaces, provides a better understanding of the difficult statistical problem of estimation in mixture models.
2. A novel approach to the problem of detecting local mixing in one dimensional NEF-QVF's is proposed. Our approach is based on the underlying Affine Geometrical structure and is different from the usual approach of embedding in a space of random variables combined with the use of significance tests. This latter approach is more *data-driven* while ours is purely *model-driven*. This is the reason why we did not use any real data example.
3. Clear geometrical and statistical analogies between Exponential and Mixture models were given and used to propose new concepts like *General* and *Curved Mixture Families* that will certainly be useful for the future work in the area. We are aware that those analogies are a simple consequence of the Duality Theory mainly developed by Amari (see for

example Amari and Nagaoka (2000)) but the contribution here is the better statistical understanding we give to them from the point of view of mixture models.

4. Local Mixture models constitute a class of low dimensional, parameter interpretable and geometrically insightful models that appear to be an attractive alternative to non-parametric and semi-parametric methods commonly used in the estimation of mixing distributions.

5. The asymptotic expansions developed from Theorem 12 provide simple expressions that can be used not only in the mixture model setting but also in any other setting in which dispersion models are involved.

6. The use of Dispersion Models as mixing distributions combined with the asymptotic expansions developed from Theorem 12 constitutes a novel approach to the local analysis of mixture models, and generalizes the original proposal of Marriott (2002).

7. Also, the use of Dispersion Models as mixing distributions provides a better statistical understanding of the invariance geometrical properties inherent to continuous mixture models and an understanding of the smallness assumption inherent in local mixture models

8. Local scale and Local location mixture models provide a simple modeling framework when the mixing parameter is either scale or location. We used the former with the negative exponential but the latter invites, for example, to the analysis of mean mixtures of the normal distribution with known variance.

9. Local scale mixture models of the Negative Exponential Distribution were analyzed in some detail, both geometrically and statistically. We showed they provide a simple statistical tool to analyze general scale dispersion mixtures when the dispersion parameter ϵ is small. Our simulations showed that simple asymptotic statistical procedures perform reasonably well for such models when the sample size is over $n = 100$. Also shown by our simulations is the fact that local mixture models have a very clear statistical meaning only when the mixing distribution is small, with this smallness measured by the squared coefficient of variation of the mixing distribution. As negative exponential distributions are widely used in the analysis of reliability and lifetime data,

local scale mixtures of them provide a new tool for the analysis of such kind of data.

10. We showed that appropriate reparametrizations of a local mixture model can reveal nice underlying statistical configurations. For example, reparametrization (4.8) revealed that our Model 2 is in fact a bundle of curved mixture families.

11. Far beyond the mixture setting, local mixture models provide a geometrically insightful generalization of any regular parametric family of densities.

12. Finally, as aimed in the introduction, we think we have contributed by developing a mathematical tool which is accessible to both Statistical and Geometrical audiences.

5.2 Future Work

Geometry of Perturbations: Cook (1986) proposed a very interesting geometric approach to the assessment of local influence in parametric statistical models. The approach has been revisited recently by Critchley and Marriott (2004) and seems to be closely related to Local Mixture Models, thus providing an attractive area to work in the near future.

Local Mixture Models for small dependence: A very simple way to induce dependence in a pair of random variables is via mixtures combined with conditional independence. Suppose we have a pair of random variables (X, Y) which are conditionally independent given the value of a third random variable θ. If $f(x, y \mid \theta)$ is the conditional distribution then the marginal joint density of (X, Y) is given by

$$g(x, y; Q) = \int f(x, y \mid \theta)\, dQ(\theta)$$

and this clearly invites to a geometrical analysis using local mixtures.

Multivariate Local Mixture Models: This type of models were defined by Marriott (2002) but have not been used ever since. Combined with the multivariate dispersion models of Jorgensen and Lauritzen (2000), they appear to be an interesting topic to explore in the future.

Deeper Geometrical Analysis: As explained in section 2.3, Local Mixture models have the underlying geometrical structure of an Affine Bundle. Interesting geometrical notions like *Affine Connections* and *curvature* apply to this kind of structures. The Statistical understanding and implications of such notions is also a very interesting area to explore in future work.

Part II

Goodness of Fit

Chapter 6

Goodness of Fit for the Inverse Gaussian and Gamma distributions under censoring

6.1 Introduction

The modification of the Cramér-von Mises statistics to test the fit from a censored sample for a completetely specified distribution was first proposed by Pettitt and Stephens (1976). In Pettitt (1976) a generalization was made for the case where parameters are unknown and the results were applied to find the asymptotic distributions of Cramér-von Mises type statistics when testing for normality. Tests of fit based on EDF statistics for the inverse Gaussian were discussed by Pavur, Edgeman, and Scott (1992) and O'Reilly and Rueda (1992). Lockhart and Stephens (1983) and Pettitt and Stephens (1983) discuss the same problem, but for the gamma distribution. These latter papers refer to the uncensored case.

In this chapter we discuss the problem of testing the null hypothesis that a sample, which may be censored at one or both ends, comes from an inverse Gaussian distribution with unknown parameters. We also discuss the problem when the null hypothesis corresponds to a gamma distribution. The derived tests are based on the results of Pettitt (1976) and Durbin (1973).

In Section 6.2 the theory for the asymptotic distributions of the test statistics is summarized and in Section 6.3 formulae are given to evaluate the covariance functions of the underlying empirical processes. Some asymptotic percentiles are tabulated in Section 6.4. Section 6.5 describes the test procedures and a small Monte Carlo study is carried out in Section 6.6.

6.2 Asymptotic Theory

Consider a type I censored sample of size n, from a continuous density $g(x)$, with ordered observations

$$x_q < x_{(s+1)} < x_{(s+2)} < \ldots < x_{(n-r)} < x_p \tag{2.1}$$

where $s, r \geq 0$ and x_q and x_p are fixed known constants. Suppose we wish to test the null hypothesis

$$H_0: \; g(x) = f(x;\boldsymbol{\theta}) \qquad \text{for some } \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

where $\boldsymbol{\theta}$ is a vector of unknown parameters ($\boldsymbol{\Theta} \subset \mathbb{R}^2$). The case where $\boldsymbol{\theta}$ is completely known was studied by Pettitt and Stephens (1976) and will be referred to here as Case **0**. In this chapter we deal with two parametric families of densities: the inverse Gaussian family, with density of the form

$$f(x;\mu,\lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right) \qquad x > 0, \quad \mu,\lambda > 0, \tag{2.2}$$

and the gamma family with density

$$f(x;\alpha,\beta) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)}\, x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right) \qquad\qquad x > 0, \quad \alpha,\beta > 0. \tag{2.3}$$

At this point, we distinguish 6 different situations:

Inverse Gaussian $\boldsymbol{\theta}^T = (\mu, \lambda)$	**Gamma** $\boldsymbol{\theta}^T = (\alpha, \beta)$
Case **IG1**: μ known, λ unknown.	Case $\mathbf{\Gamma 1}$: α known, β unknown.
Case **IG2**: λ known, μ unknown.	Case $\mathbf{\Gamma 2}$: β known, α unknown.
Case **IG3**: μ and λ unknown.	Case $\mathbf{\Gamma 3}$: α and β unknown.

where the superscript T denotes the transpose of a vector. Denote by $F(x;\boldsymbol{\theta})$ the cumulative distribution function corresponding to $f(x;\boldsymbol{\theta})$. If $\boldsymbol{\theta}$ is known (Case **0**) and p, q satisfy

$$q = F(x_q\,;\boldsymbol{\theta}) \qquad \text{and} \qquad p = F(x_p\,;\boldsymbol{\theta}) \tag{2.4}$$

Pettitt and Stephens (1976) suggest the use of a modified form of Anderson-Darling's and Cramér-von Mises' statistic for testing the fit of $F(x;\boldsymbol{\theta})$ with the available censored sample (2.1). These modifications are:

$${}_{qp}W_n^2 = n\int_q^p \left(G_n(t) - t\right)^2 dt \quad \text{and} \quad {}_{qp}A_n^2 = n\int_q^p \frac{(G_n(t)-t)^2}{t(1-t)}dt,$$

where $G_n(t)$ is the EDF of the transformed ordered sample $\{t_{(s+1)}, \ldots, t_{(n-r)}\}$ where $t_{(i)} = F(x_{(i)};\boldsymbol{\theta})$. In their paper, Pettitt and Stephens provide percentage points for the asymptotic distribution of both statistics for various values of p ($q = 0$ and $q = 1-p$).

For the case where $\boldsymbol{\theta}$ is partially (or totally) unknown, let $\widehat{\boldsymbol{\theta}}_n$ be the maximum likelihood estimator (or another asymptotically efficient estimator) of $\boldsymbol{\theta}$ based on the sample (2.1). For testing H_0 in this case, we use the modification of Anderson-Darling's and Cramér-von Mises' statistic proposed by Pettitt (1976). The modifications are:

$${}_{qp}\widehat{W}_n^2 = n\int_q^p \left(\widehat{G}_n(t) - t\right)^2 dt \quad \text{and} \quad {}_{qp}\widehat{A}_n^2 = n\int_q^p \frac{(\widehat{G}_n(t)-t)^2}{t(1-t)}dt$$

where $\widehat{G}_n(t)$ is the EDF of the transformed ordered sample $\{\hat{t}_{(s+1)}, \ldots, \hat{t}_{(n-r)}\}$ where $\hat{t}_{(i)} = F(x_{(i)};\widehat{\boldsymbol{\theta}}_n)$. Note that, in order to properly define these statistics, one needs to know the values of p and q which satisfy (2.4). This situation is evidently not possible because $\boldsymbol{\theta}$ is unknown. In this sense, the latter expressions represents a pair of non-constructible statistics.

In this chapter we propose the use of the statistics ${}_{\hat{q}\hat{p}}\widehat{W}_n^2$ and ${}_{\hat{q}\hat{p}}\widehat{A}_n^2$, where $\hat{q} = F(x_q\,;\,\widehat{\boldsymbol{\theta}}_n)$ and $\hat{p} = F(x_p\,;\,\widehat{\boldsymbol{\theta}}_n)$ which are asymptotically equivalent to ${}_{qp}\widehat{W}_n^2$ and ${}_{qp}\widehat{A}_n^2$, since $\hat{q}$ and $\hat{p}$ converge in probability to q and p respectively. Following Durbin (1973), Pettitt (1976) shows that if the sequence of estimators $\{\widehat{\boldsymbol{\theta}}_n\}$ is asymptotically efficient then the process

$$\hat{\xi}_n(t) = \sqrt{n}(\widehat{G}_n(t) - t) \qquad t \in [q,p]$$

converges weakly to the Gaussian process defined over the interval $[q,p]$ with zero mean and covariance function

$$\rho(s,t) = \min(s,t) - st - g^T(t)\mathcal{I}^{-1}g(s), \tag{2.5}$$

where $\mathcal{I}$ is Fisher's expected information provided by a single observation and $g(t)$ is the vector of derivatives of $F(x;\boldsymbol{\theta})$ with respect to the unknown parameters evaluated at $t = F(x;\boldsymbol{\theta})$. If only one component of $\boldsymbol{\theta}$ is unknown $\mathcal{I}$ is scalar, and if both components are unknown $\mathcal{I}$ is a 2×2 matrix.

Once the covariance function $\rho(s,t)$ is obtained, the asymptotic distribution of ${}_{qp}\widehat{W}_n^2$ may be evaluated because it coincides with the distribution of

$$\sum_{i=1}^{\infty} \lambda_i Z_i^2,$$

where the $Z_i's$ are independent standard normal variables and $\lambda_1, \lambda_2, \ldots$ are the eigenvalues of the integral equation

$$\lambda\, \vartheta(t) = \int_q^p \rho(s,t)\vartheta(s)ds. \tag{2.6}$$

The asymptotic distribution of ${}_{qp}\widehat{A}_n^2$ can be obtained in a similar way by replacing the covariance function $\rho(s,t)$ by $\rho(s,t)/\{(s-s^2)(t-t^2)\}^{1/2}$.

6.3 Obtaining The Covariance Functions

We denote $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ to cover any of the two families of densities mentioned previously. The matrix $\mathcal{I}$ for cases **IG3** and **Γ3** is

$$\mathcal{I} = -\begin{pmatrix} E\left[\dfrac{\partial^2 \ln L(\boldsymbol{\theta}; x)}{\partial \theta_1^2}\right] & E\left[\dfrac{\partial^2 \ln L(\boldsymbol{\theta}; x)}{\partial \theta_1 \partial \theta_2}\right] \\ E\left[\dfrac{\partial^2 \ln L(\boldsymbol{\theta}; x)}{\partial \theta_2 \partial \theta_1}\right] & E\left[\dfrac{\partial^2 \ln L(\boldsymbol{\theta}; x)}{\partial \theta_2^2}\right] \end{pmatrix} \tag{3.1}$$

and for the other cases, namely **IG1**, **IG2**, **Γ1** and **Γ2**, the corresponding diagonal elements of the matrix are Fisher's expected information quantities. The likelihood function for a single observation x is given by

$$L(\boldsymbol{\theta}; x) = F(x_q; \boldsymbol{\theta})^{\gamma(x)} \times f(x; \boldsymbol{\theta})^{\delta(x)} \times [1 - F(x_p; \boldsymbol{\theta})]^{1-\delta(x)-\gamma(x)}, \tag{3.2}$$

where

$$(\delta(x), \gamma(x)) = \begin{cases} (0,0) & x \geq x_p \\ (0,1) & x \leq x_q \\ (1,0) & x_q < x < x_p \,. \end{cases}$$

One can express the entries of the matrix in (3.1) by the well known identity:

$$E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}; x)}{\partial \theta_i \partial \theta_j}\right] = (1-p) \cdot E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}; x)}{\partial \theta_i \partial \theta_j} \middle| x \geq x_p\right] + \tag{3.3}$$

$$q \cdot E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}; x)}{\partial \theta_i \partial \theta_j} \middle| x \leq x_q\right] + (p-q) \cdot E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}; x)}{\partial \theta_i \partial \theta_j} \middle| x_q < x < x_p\right] .$$

Next, we state a proposition that is useful in evaluating the covariance function for the inverse Gaussian case and in determining the asymptotic distribution of the test statistics used in this part of the thesis.

Proposition 1 *For cases* **IG1**, **IG2** *and* **IG3**, *the covariance function $\rho(s,t)$ defined in (2.5) depends only on $\eta = \lambda/\mu$.*

Proof: Let $\eta = \lambda/\mu$ and consider the case $x_q < x < x_p$. It is easy to show that

$$\begin{aligned}
-\frac{\partial^2 \ln f(x;\mu,\lambda)}{\partial\mu^2} &= \frac{1}{\mu^2}\left[\eta\left(\frac{3x}{\mu}-2\right)\right], \\
-\frac{\partial^2 \ln f(x;\mu,\lambda)}{\partial\lambda^2} &= \frac{1}{\mu^2}\left[\frac{1}{2\eta^2}\right] \qquad \text{and} \\
-\frac{\partial^2 \ln f(x;\mu,\lambda)}{\partial\mu\partial\lambda} &= -\frac{1}{\mu^2}\left[\frac{x}{\mu}-1\right].
\end{aligned}$$

Note that these derivatives depend on x as a linear function in x/μ. We shall prove next that the conditional expectation of x/μ given that $x_q < x < x_p$ depend only on η. We have that

$$E[x \mid x_q < x < x_p] = \frac{\displaystyle\int_{x_q}^{\infty} xf(x;\mu,\lambda)\,dx - \int_{x_p}^{\infty} xf(x;\mu,\lambda)\,dx}{F(x_p;\mu,\lambda) - F(x_q;\mu,\lambda)},$$

making the change of variable $y = \mu/x$ in the integrals one obtains

$$E[x \mid x_q < x < x_p] = \frac{\mu[F(\mu/x_q;1,\eta) - F(\mu/x_p;1,\eta)]}{F(x_p;\mu,\lambda) - F(x_q;\mu,\lambda)}.$$

We use the following property

$$F(x;\mu,\lambda) = F(x/\mu;1,\eta), \tag{3.4}$$

which implies that if $t = F(x;\mu,\lambda)$ then $x/\mu = F^{-1}(t;1,\eta)$. In this way,

$$E[x/\mu \mid x_q < x < x_p\,] = \frac{F(z_q^{-1};1,\eta) - F(z_p^{-1};1,\eta)}{p-q}$$

where $z_q = F^{-1}(q;1,\eta)$ and $z_p = F^{-1}(p\,;1,\eta)$. Note that if values are given for p, q and η one could evaluate $E[x/\mu \mid x_q < x < x_p\,]$. On the other hand, it is easy to show that

$$\frac{\partial F(x;\mu,\lambda)}{\partial\mu} = \frac{1}{\mu}\,h_1(\eta,t) \qquad \text{and} \qquad \frac{\partial F(x;\mu,\lambda)}{\partial\lambda} = \frac{1}{\mu}\,h_2(\eta,t)$$

at $t = F(x; \mu, \lambda)$, where

$$\begin{aligned} h_1(\eta, t) &= -\sqrt{\eta z_t}\phi(R_t) - 2\eta e^{2\eta}\Phi(L_t) + \sqrt{\eta z_t}e^{2\eta}\phi(L_t) \qquad \text{and} \\ h_2(\eta, t) &= \tfrac{R_t}{2\eta}\phi(R_t) + 2e^{2\eta}\Phi(L_t) + \tfrac{L_t}{2\eta}e^{2\eta}\phi(L_t) \end{aligned} \tag{3.5}$$

with $z_t = x/\mu = F^{-1}(t; 1, \eta)$, $R_t = \sqrt{\eta}(z_t^{1/2} - z_t^{-1/2})$ and $L_t = -\sqrt{\eta}(z_t^{1/2} + z_t^{-1/2})$. From these expressions one has that

$$\begin{aligned} \frac{\partial^2 F(x; \mu, \lambda)}{\partial \mu^2} &= -\frac{1}{\mu^2}(\eta\, h_1'(\eta, t) + h_1(\eta, t)), \\ \frac{\partial^2 F(x; \mu, \lambda)}{\partial \lambda^2} &= \frac{1}{\mu^2}\, h_2'(\eta, t) \qquad\qquad \text{and} \\ \frac{\partial^2 F(x; \mu, \lambda)}{\partial \lambda \partial \mu} &= \frac{1}{\mu^2}\, h_1'(\eta, t), \end{aligned}$$

where the dash denotes the derivative with respect to η. So, the conditional expectations for $x \geq x_p$ and $x \leq x_q$ in equation (3.3) take the form of a product of μ^{-2} and a function of η. It then follows that μ^{-2} can be factored out of any element in the matrix $\mathcal{I}$, and μ^{-1} from the vector $g(t)$ for all $t \in [q, p]$ leaving the quadratic form

$$g^T(t)\mathcal{I}^{-1}g(s)$$

to be a function of η only. ■

In order to simplify the following expressions, we denote by ΔF the difference $F(z_q^{-1}; 1, \eta) - F(z_p^{-1}; 1, \eta)$. The covariance function (2.5) in the **IG3** case can then be written as:

$$\rho_{\text{IG3}}(s, t) = \min(s, t) - st - \frac{A_{s\,t}}{D_{s\,t}}, \tag{3.6}$$

where

$$\begin{aligned}
A_{s\,t} = {} & h_1(\eta,s)h_1(\eta,t)\left[\frac{p-q}{2\eta^2}+\frac{(h_2(\eta,q))^2}{q}+\frac{(h_2(\eta,p))^2}{1-p}-\{h_2'(\eta,q)-h_2'(\eta,p)\}\right] \\
& + h_2(\eta,s)h_2(\eta,t)\Bigg[3\eta\Delta F-2\eta(p-q)+\frac{(h_1(\eta,q))^2}{q}+\eta\{h_1'(\eta,q)-h_1'(\eta,p)\} \\
& + \frac{(h_1(\eta,p))^2}{1-p}+h_1(\eta,q)-h_1(\eta,p)\Bigg]-\Bigg[p-q-\Delta F+\frac{h_1(\eta,q)h_2(\eta,q)}{q} \\
& + \frac{h_1(\eta,p)h_2(\eta,p)}{1-p}-\{h_1'(\eta,q)-h_1'(\eta,p)\}\Bigg]\Big[h_1(\eta,t)h_2(\eta,s)+h_1(\eta,s)h_2(\eta,t)\Big]
\end{aligned}$$

and

$$\begin{aligned}
D_{s\,t} = {} & \Bigg[3\eta\Delta F-2\eta(p-q)+\frac{(h_1(\eta,q))^2}{q}+\frac{(h_1(\eta,p))^2}{1-p}+h_1(\eta,q)-h_1(\eta,p) \\
& + \eta\{h_1'(\eta,q)-h_1'(\eta,p)\}\Bigg]\Bigg[\frac{p-q}{2\eta^2}+\frac{(h_2(\eta,q))^2}{q}+\frac{(h_2(\eta,p))^2}{1-p}-h_2'(\eta,q) \\
& + h_2'(\eta,p)\Bigg]-\Bigg[p-q-\Delta F+\frac{h_1(\eta,q)h_2(\eta,q)}{q}+\frac{h_1(\eta,p)h_2(\eta,p)}{1-p} \\
& - \{h_1'(\eta,q)-h_1'(\eta,p)\}\Bigg]^2.
\end{aligned}$$

The covariance functions for cases **IG1** and **IG2** can be expressed as:

$$\rho_{\text{IG1}}(s,t)=\min(s,t)-st-\frac{h_2(\eta,s)\,h_2(\eta,t)}{\mathcal{I}_2(\eta)} \tag{3.7}$$

$$\rho_{\text{IG2}}(s,t)=\min(s,t)-st-\frac{h_1(\eta,s)\,h_1(\eta,t)}{\mathcal{I}_1(\eta)} \tag{3.8}$$

where

$$\begin{aligned}\mathcal{I}_1(\eta) &= 3\eta\Delta F - 2\eta(p-q) + \frac{(h_1(\eta,q))^2}{q} + \eta\{h_1'(\eta,q) - h_1'(\eta,p)\} \\ &+ \frac{(h_1(\eta,p))^2}{1-p} + h_1(\eta,q) - h_1(\eta,p) \\ \\ \mathcal{I}_2(\eta) &= \frac{p-q}{2\eta^2} + \frac{(h_2(\eta,q))^2}{q} + \frac{(h_2(\eta,p))^2}{1-p} - \{h_2'(\eta,q) - h_2'(\eta,p)\}.\end{aligned}$$

Now, for the gamma case it is possible to show, as in the inverse Gaussian case, that the covariance function depends only on one quantity, the shape parameter α, using the fact that β is a scale parameter. In order to express the covariance function we define

$$\begin{aligned}E &= E[x/\beta \mid x_q < x < x_p] \\ k_1(w) &= \frac{1}{\Gamma(\alpha)}\int_0^w u^{\alpha-1}e^{-u}\log u\, du \\ k_2(w) &= \frac{1}{\Gamma(\alpha)}\int_0^w u^{\alpha-1}e^{-u}(\log u)^2\, du \\ z_t &= F^{-1}(t;\alpha,1) \\ f_t &= f(z_t;\alpha,1) \qquad t \in [q,p],\end{aligned}$$

note that E does not depend on β since

$$E = \alpha \left[\frac{F(z_p;\alpha+1,1) - F(z_q;\alpha+1,1)}{F(z_p;\alpha,1) - F(z_q;\alpha,1)}\right].$$

For case $\mathbf{\Gamma 3}$ one has that

$$g(t) = \begin{pmatrix} k_1(z_t) - t\,\psi(\alpha) \\ \\ -\,\dfrac{z_t\,f_t}{\beta} \end{pmatrix}, \qquad t \;\in\; [q,p]$$

where $\psi(\cdot)$ is the digamma function. Therefore, the covariance function (2.5) for this case, can be written as:

$$\rho_{\Gamma 3}(s,t) = \min(s,t) - st - \frac{A^\star_{s\,t}}{D^\star_{s\,t}}, \tag{3.9}$$

where

$$\begin{aligned}
A^\star_{s\,t} = z_t\,z_s\,f_t\,f_s \left[k_2(z_p) - k_2(z_q) + k_1^2(z_q) + \frac{k_1^2(z_p) - 2\psi(\alpha)k_1(z_p) + p\,\psi^2(\alpha)}{1-p} \right] \\
+ \left[k_1(z_t)k_1(z_s) - \psi(\alpha)\left\{ s\,k_1(z_t) + t\,k_1(z_s) \right\} + s\,t\,\psi^2(\alpha) \right] \Bigg[2E - \alpha(p-q) \\
+ \; z_q \left\{ \frac{z_q f_q^2}{q} + f_q(z_q - \alpha - 1) \right\} + z_p \left\{ \frac{z_p f_p^2}{1-p} - f_p(z_p - \alpha - 1) \right\} \Bigg] \\
- \left[z_s f_s \left\{ t\,\psi(\alpha) - k_1(z_t) \right\} + z_t f_t \left\{ s\,\psi(\alpha) - k_1(z_s) \right\} \right] \Bigg[p - q \\
+ \; z_q\,f_q \left\{ \ln z_q - \frac{k_1(z_q)}{q} \right\} + z_p\,f_p \left\{ -\ln z_p + \frac{\psi(\alpha) - k_1(z_p)}{1-p} \right\} \Bigg]
\end{aligned}$$

and

$$D^{\star}_{s\,t} = \left[k_2(z_p) - k_2(z_q) + k_1^2(z_q) + \frac{k_1^2(z_p) - 2\psi(\alpha)k_1(z_p) + p\,\psi^2(\alpha)}{1-p} \right] \Bigg[2E$$

$$-\alpha(p-q) + z_q \left\{ \frac{z_q f_q^2}{q} + f_q(z_q - \alpha - 1) \right\} + z_p \left\{ \frac{z_p f_p^2}{1-p} - f_p(z_p - \alpha - 1) \right\} \Bigg]$$

$$- \left[p - q + z_q\, f_q \left\{ \ln z_q - \frac{k_1(z_q)}{q} \right\} + z_p\, f_p \left\{ -\ln z_p + \frac{\psi(\alpha) - k_1(z_p)}{1-p} \right\} \right]^2 .$$

The covariance functions for cases **Γ1** and **Γ2** can be expressed as:

$$\rho_{\Gamma 1}(s,t) = \min(s,t) - st - \frac{z_s z_t f_s f_t}{\mathcal{I}_2(\alpha)} \tag{3.10}$$

$$\rho_{\Gamma 2}(s,t) = \min(s,t) - st - \frac{k_1(z_t)k_1(z_s) - \psi(\alpha)(sk_1(z_t) + tk_1(z_s)) + st\psi^2(\alpha)}{\mathcal{I}_1(\alpha)} \tag{3.11}$$

where

$$\begin{aligned} \mathcal{I}_1(\alpha) &= k_2(z_p) - k_2(z_q) + k_1^2(z_q) + \frac{k_1^2(z_p) - 2\psi(\alpha)k_1(z_p) + p\,\psi^2(\alpha)}{1-p} \\ \mathcal{I}_2(\alpha) &= 2E - \alpha(p-q) + z_q \left\{ \frac{z_q f_q^2}{q} + f_q(z_q - \alpha - 1) \right\} \\ &+ z_p \left\{ \frac{z_p f_p^2}{1-p} - f_p(z_p - \alpha - 1) \right\} . \end{aligned}$$

Finally, to evaluate $k_1(\cdot)$ and $k_2(\cdot)$ we use the following expansions truncated to 200 terms:

$$k_1(w) = \sum_{j=0}^{\infty} \frac{(-1)^j}{j!} \left[\frac{w^{\alpha+j} \left\{ (\alpha+j) \ln w - 1 \right\}}{(\alpha+j)^2} \right]$$

$$k_2(w) = \sum_{j=0}^{\infty} \frac{(-1)^j}{j!} \left[\frac{w^{\alpha+j} \left\{ (\alpha+j)^2 (\ln w)^2 - 2(\alpha+j) \ln w + 2 \right\}}{(\alpha+j)^3} \right].$$

In fact, it is not necessary to approximate such integrals. Computer algebra systems like MATHEMATICA easily compute such integrals using formal series without the need of truncation.

6.4 Percentiles of the Asymptotic Distributions

Percentiles of the asymptotic distribution of ${}_{qp}\widehat{W}_n^2$ and ${}_{qp}\widehat{A}_n^2$ for both the gamma and inverse Gaussian can be found accurately by solving the eigenvalue problem (2.6) numerically, first approximating the integral with a sum using (say) one hundred points in the interval $[q, p]$ and then evaluating the covariance functions (expression 3.6 to 3.11) in the 100×100 grid. Imhof's (1961) method can then be used to obtain the required percentiles.

Only for illustration, upper percentiles of ${}_{0,p}\widehat{A}_n^2$ were calculated for $p = 0.75$ in the inverse Gaussian cases for some values of η and upper percentiles of ${}_{qp}\widehat{A}_n^2$ were also evaluated for $(q, p) = (0, 0.75)$ and $(q, p) = (.25, 1)$ for the gamma cases for some values of α. Results are presented in the following tables.

Table I: Upper percentiles of ${}_{0,0.75}\widehat{A}_n^2$.

	Case **IG3**					Case **IG2**				
η	.25	.10	.05	.025	0.01	.25	.10	.05	.025	.01
2^{-10}	.373	.549	.691	.838	1.040	.668	1.063	1.387	1.726	2.191
2^{-8}	.372	.548	.689	.836	1.037	.666	1.060	1.383	1.722	2.185
2^{-6}	.369	.543	.682	.827	1.025	.661	1.050	1.369	1.703	2.160
2^{-4}	.360	.527	.660	.798	0.986	.641	1.014	1.320	1.639	2.077
2^{-2}	.338	.488	.606	.728	0.894	.591	0.921	1.191	1.473	1.860
2^{0}	.310	.438	.538	.639	0.777	.516	0.781	0.995	1.218	1.524
2^{2}	.293	.409	.498	.588	0.709	.456	0.672	0.842	1.018	1.259
2^{4}	.287	.399	.484	.570	0.686	.429	0.623	0.775	0.931	1.142
2^{6}	.286	.396	.481	.566	0.680	.419	0.607	0.752	0.902	1.104
2^{8}	.285	.396	.480	.565	0.678	.416	0.601	0.745	0.893	1.093
∞	.285	.397	.480	.563	0.675	.414	0.599	0.742	0.888	1.086

It is well known (Johnson, Kotz, and Balakrishnan (1994)) that if a random variable X follows the inverse Gaussian density (2.2) then the random variable $\sqrt{\eta}(X/\mu - 1)$ converges in distribution to a standard normal provided that $\eta = \lambda/\mu$ tends to infinity no matter how μ and λ behave. On the other hand, if η tends to zero in such a way that μ tends to infinity while λ remains fixed it is also well known that $Y = X^{-1}$ converges to a gamma distribution with shape $\alpha = 1/2$ and scale $\beta = 1/\lambda$. Table I exhibits percentiles of the asymptotic distribution of ${}_{0,0.75}\widehat{A}_n^2$ in cases **IG2** and **IG3** for some values of η. For case **IG3**, the last row indicating $\eta = \infty$ refers to the asymptotic percentiles of ${}_{0,0.75}\widehat{A}_n^2$ in the normal case (say **N3** case) with both mean and variance unknown (2) while for case **IG2** the same row refers to the normal case with unknown mean and known variance (say case **N2**). Note that, as long as η tends to infinity, the percentiles approach the corresponding ones in the normal case.

Table II: Upper percentiles of ${}_{0,0.75}\widehat{A}_n^2$.

	Case **IG1**				
η	.25	.10	.05	.025	0.01
0	0.522	0.806	1.038	1.281	1.613
2^{-10}	0.522	0.807	1.039	1.282	1.614
2^{-8}	0.523	0.808	1.041	1.284	1.617
2^{-6}	0.525	0.812	1.047	1.293	1.630
2^{-4}	0.532	0.829	1.073	1.328	1.676
2^{-2}	0.559	0.887	1.159	1.443	1.832
2^{0}	0.621	1.021	1.352	1.699	2.172
2^{2}	0.696	1.180	1.582	2.000	2.570
2^{4}	0.747	1.286	1.732	2.197	2.831
2^{6}	0.771	1.336	1.804	2.291	2.956
2^{8}	0.782	1.358	1.835	2.332	3.011
∞	0.791	1.378	1.863	2.368	3.057

In Table II the asymptotic percentiles of ${}_{0,0.75}\widehat{A}_n^2$, when $\eta \to 0$, identified in row $\eta = 0$, converge to the percentiles that appear in Table V in the $\mathbf{\Gamma 1}$ case with $\alpha = 0.5$. Whereas if $\eta \to \infty$, convergence is to the percentiles (row $\eta = \infty$) in the normal case with known mean and unknown variance (say case **N1**). The asymptotic percentiles for the three different normal cases mentioned above where obtained using the formulas in (2) and the method described at the beginning of this section.

Table III: Upper percentiles of ${}_{0.25,1}\widehat{A}_n^2$ and ${}_{0,0.75}\widehat{A}_n^2$ respectively.

	Case $\mathbf{\Gamma 3}$					Case $\mathbf{\Gamma 3}$				
α	.25	.10	.05	.025	0.01	.25	.10	.05	.025	.01
2^{-4}	.381	.565	.713	.867	1.077	.348	.505	.629	.758	.933
2^{-2}	.319	.455	.560	.669	0.817	.322	.460	.568	.680	.833
2^{-1}	.302	.425	.519	.615	0.745	.305	.430	.527	.627	.762
2^{0}	.293	.410	.498	.588	0.709	.295	.412	.502	.594	.717
2^{1}	.289	.402	.489	.576	0.693	.290	.403	.490	.578	.696
2^{2}	.286	.397	.482	.567	0.681	.287	.399	.484	.571	.686
∞	.285	.397	.480	.563	0.675	.285	.397	.480	.563	.675

Table IV: Upper percentiles of ${}_{0,0.75}\widehat{A}_n^2$.

	Case $\mathbf{\Gamma 2}$				
α	0.25	0.10	0.05	0.025	0.01
0	0.489	0.735	0.933	1.140	1.422
2^{-6}	0.469	0.704	0.892	1.090	1.360
2^{-4}	0.469	0.703	0.890	1.090	1.350
2^{-3}	0.467	0.699	0.884	1.080	1.340
2^{-2}	0.462	0.688	0.867	1.050	1.310
2^{-1}	0.450	0.665	0.833	1.010	1.240
2^{0}	0.435	0.636	0.794	0.955	1.170
2^{1}	0.424	0.616	0.765	0.918	1.130
2^{2}	0.418	0.605	0.751	0.900	1.100
2^{3}	0.415	0.601	0.745	0.892	1.090
∞	0.414	0.599	0.742	0.888	1.086

For the gamma, it is also well known that as α becomes larger then the gamma (standardized) distribution tends to the standard normal. Table III shows again convergence of the asymptotic percentiles of ${}_{0.25,1}\widehat{A}_n^2$ and ${}_{0,0.75}\widehat{A}_n^2$ in the **Γ3** case to the corresponding ones (row $\alpha = \infty$) in the **N3** case. In Table IV the same kind of convergence appears in the **Γ2** case but now to the **N2** case.

Also in Table IV, as $\alpha \to 0$ the percentiles tends to the ones in the exponential case studied by Sirvanci and Levent (1982).

Table V: Upper percentiles of ${}_{0.25,1}\widehat{A}_n^2$ and ${}_{0,0.75}\widehat{A}_n^2$ respectively.

	Case **Γ1**					Case **Γ1**				
α	.25	0.1	0.05	0.025	0.01	.25	0.1	0.05	0.025	0.01
2^{-4}	.771	1.270	1.680	2.120	2.710	.666	1.060	1.380	1.720	2.180
2^{-2}	.600	0.958	1.250	1.560	1.980	.597	0.938	1.220	1.510	1.910
2^{-1}	**.522**	**0.806**	**1.038**	**1.281**	**1.613**	.538	0.829	1.070	1.310	1.650
2^{0}	.469	0.704	0.892	1.090	1.360	.489	0.735	0.933	1.140	1.420
2^{1}	.440	0.648	0.813	0.983	1.210	.456	0.674	0.846	1.030	1.270
2^{2}	.425	0.621	0.774	0.931	1.150	.438	0.639	0.798	0.961	1.180
2^{3}	.418	0.608	0.755	0.907	1.110	.427	0.621	0.772	0.927	1.140
∞	.414	0.599	0.742	0.888	1.086	.414	0.599	0.742	0.888	1.086

Finally, Table V shows the convergence of the asymptotic percentiles of ${}_{0.25,1}\widehat{A}_n^2$ and ${}_{0,0.75}\widehat{A}_n^2$ in the **Γ1** case to the ones in the **N2** case.

Obviously, the formulae in Section 3 could be used to reproduce any uncensored case for both the inverse Gaussian and the gamma. Another application is for the Levy distribution (O'Reilly and Rueda (1998)) because of its relationship with the gamma distribution. If X is a random variable that follows a Levy distribution with scale σ, then the reciprocal $Y = 1/X$ follows a gamma distribution with $\alpha = 1/2$ and $\beta = 1/\sigma$. Thus, if we have a type I censored sample with known mean (μ), it is equivalent to test the inverse Gaussian hypothesis with the original sample or to test the gamma distribution (with $\alpha = 1/2$) with the sample of the corresponding reciprocals only observing that the censoring limits, $y_q = 1/x_p$ and $y_p = 1/x_q$ imply that the pair (q, p) is changed to $(1-p, 1-q)$.

6.5 Test Procedure

Consider a type I censored sample of size n

$$x_q < x_{(s+1)} < x_{(s+2)} < \ldots < x_{(n-r)} < x_p \qquad (5.1)$$

from a continuous density $g(x)$, where $s, r \geq 0$ and x_q and x_p are fixed known constants. To test the null hypothesis

$$H_0: \ g(x) = f(x; \boldsymbol{\theta}) \qquad \text{for some } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

where f is either of the form (2.2) or (2.3), follow the next steps:

1. Calculate $\widehat{\boldsymbol{\theta}}_n$, $\hat{q} = F(x_q\,; \widehat{\boldsymbol{\theta}}_n)$ and $\hat{p} = F(x_p\,; \widehat{\boldsymbol{\theta}}_n)$. For estimation in the inverse Gaussian case see Chao (1985) and for the gamma see Harter and Moore (1965).

2. Evaluate $z_{(i)} = F(x_{(i)}; \widehat{\boldsymbol{\theta}}_n)$ for $i = s+1, \ldots, n-r$ and ${}_{\hat{q}\hat{p}}\widehat{W}_n^2$ or ${}_{\hat{q}\hat{p}}\widehat{A}_n^2$ using the formulae (1):

$$_{\hat{q}\hat{p}}\widehat{W}_n^2 = {}_{\hat{p}}\widehat{W}_n^2 - {}_{\hat{q}}\widehat{W}_n^2 \qquad (5.2)$$

$$_{\hat{q}\hat{p}}\widehat{A}_n^2 = {}_{\hat{p}}\widehat{A}_n^2 - {}_{\hat{q}}\widehat{A}_n^2 \qquad (5.3)$$

where

$$\begin{aligned} {}_{\hat{a}}\widehat{W}_n^2 &= \sum_{i=1}^{R}\left(z_{(i)} - \frac{2i-1}{2n}\right)^2 - \frac{R(4R^2-1)}{12n^2} + n\hat{a}\left(\frac{R^2}{n^2} - \hat{a}\frac{R}{n} + \frac{\hat{a}^2}{3}\right) \\ {}_{\hat{a}}\widehat{A}_n^2 &= \sum_{i=1}^{R}\left(\frac{2i-1}{n}\right)\left[\log(1-z_{(i)}) - \log(z_{(i)})\right] - 2\sum_{i=1}^{R}\log(1-z_{(i)}) \\ + & \quad n\left[\frac{2R}{n} - \left(\frac{R}{n}\right)^2 - 1\right]\log(1-\hat{a}) + \frac{R^2}{n}\log(\hat{a}) - n\hat{a} \qquad (5.4) \end{aligned}$$

with

$$R = \begin{cases} n-r & \text{if} \quad \hat{a} = \hat{p} \\ s & \text{if} \quad \hat{a} = \hat{q} \end{cases}$$

3. Finally, compute the p-value associated to the observed value of ${}_{\hat{q}\hat{p}}\widehat{W}_n^2$ or ${}_{\hat{q}\hat{p}}\widehat{A}_n^2$ using the asymptotic distribution of ${}_{qp}\widehat{W}_n^2$ or ${}_{qp}\widehat{A}_n^2$ with the method described at the beginning of Section 6.4 and entering the estimated value of $\boldsymbol{\theta}$, that is $\widehat{\boldsymbol{\theta}}_n$. To do this step a numerical routine in MATLAB made by the author is available upon request.

6.6 Monte Carlo Study

We carried out a small Monte Carlo study to assess the finite sample properties of ${}_{\hat{q}\hat{p}}\widehat{A}_n^2$ when used to test H_0 using the procedure given in the previous section. We simulated ten thousand, type I censored samples of size n fixing $x_q = 0$, $p = 0.75$ and $x_p = F^{-1}(p\,;\boldsymbol{\theta})$ and then, evaluated ${}_{0,\hat{p}}\widehat{A}_n^2$ for each one of the samples using expression (5.3). When computing the test statistic, p and $\boldsymbol{\theta}$ were estimated, as we would do in practice. The actual Monte-Carlo percentiles were based on the 10,000 runs and are presented in the following table without any type of correction.

Table VI: Monte-Carlo upper percentiles of ${}_{0,\hat{p}}\widehat{A}_n^2$.

	Case **IG3** ($\eta = 1$)					Case **Γ3** ($\alpha = 1$)				
n	.25	.10	.05	.025	0.01	.25	.10	.05	.025	.01
10	.319	.458	.632	.731	0.799	.303	.444	.499	.533	.723
20	.313	.434	.523	.607	0.768	.302	.427	.536	.635	.748
40	.313	.466	.580	.622	0.708	.302	.435	.526	.614	.772
60	.324	.461	.498	.588	0.709	.308	.446	.514	.609	.709
80	.334	.422	.523	.590	0.778	.312	.435	.530	.611	.699
100	.316	.441	.516	.629	0.770	.301	.419	.529	.588	.705
∞	.310	.438	.538	.639	0.777	.295	.412	.502	.594	.717

The corresponding percentiles of the asymptotic distribution of ${}_{0,p}\widehat{A}_n^2$, which is the limiting distribution for the Monte-Carlo distributions if we had known the value of p and $\boldsymbol{\theta}$, are exhibited in the last row of Table VI.

In a practical situation, the above assumed knowledge of p and $\boldsymbol{\theta}$ is not possible and, for each simulated sample, one would compute a p-value for ${}_{0,\hat{p}}\widehat{A}_n^2$ using $\hat{p}$ and $\widehat{\boldsymbol{\theta}}_n$ in the asymptotic distribution. As a small verification that this procedure actually provides an accurate method, out of 500 simulated (inverse Gaussian) samples with the same parameters fixed previously, the p-value was computed for each sample as described here and a test for uniformity in the $(0, 1)$ interval was made using Anderson-Darling's statistic A^2. The results are shown in Table VII.

Table VII: Test for uniformity of the obtained p-values.

Sample size (n)	Value of A^2	Significance
10	4.326	0.006
20	3.301	0.019
30	2.617	0.043
100	2.409	0.055

As can be observed, uniformity of these p-values is still rejected with $n = 100$ which seems puzzling in the light of the very nice approximation suggested in Table VI. The explanation of this apparent contradiction is that, in Table VI, the Monte-Carlo distributions are quite well approximated by the asymptotic distribution *in the right tail*; that is, for tail areas below 0.25.

To double-check this observation on the good approximation of the asymptotic distribution on the right tail, a conditional test was performed for the simulated p-values, conditioning on the fact that these were below 0.25. The results on Table VIII confirm the good approximation.

Table VIII: Conditional test for uniformity of the obtained p-values (≤ 0.25).

Sample size (n)	Value of A^2	Significance
10	0.654	0.594
20	0.719	0.539
30	0.922	0.398
100	0.224	0.980

6.7 Conclusions

A procedure has been developed for testing the inverse Gaussian and, separately, for testing the gamma distribution. This procedure works under type I single or double censoring. Connections were established with other known cases (normal with one or both parameters unknown, exponential and Levy distribution). A procedure is suggested to compute p-values instead of constructing and using tables. This procedure is outlined and the corresponding routine was developed in MATLAB.

Appendix A

Regularity Conditions

In this Appendix we define the regularity conditions for the families of probability densities used in the first part of the thesis. Let

$$\mathcal{F} = \{f(x;\theta) : \theta \in \Theta \subset \mathbb{R}\}$$

be a one dimensional parametric family of densities defined on $\mathbb{R}$. The following conditions on $\mathcal{F}$ are basically taken from Amari (1990) page 16.

Definition 18 We will say that the parametric family $\mathcal{F}$ is *regular* if it satisfies the following conditions

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

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

(A3) Every density $f(x;\theta)$ is smooth[1] as a function of θ a.e.$[\nu]$ and the partial derivatives $\{\partial/\partial\theta^i\}$ and integration with respect to the measure ν are always commutative.

[1] means of class $\mathcal{C}^\infty(\Theta)$

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

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

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

(A5) The set of functions

$$\left\{ \frac{\partial^k f(x;\theta)}{\partial \theta^k} \,:\, k = 1, 2, \dots \right\}$$

is linearly independent a.e. $[\nu]$

(A6) For each $\theta_0 \in \Theta$ there exist ν integrable functions $h_k(x, \theta_0)$ for $k = 1, 2, \dots$, such that

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

and

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

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

Appendix B

Affine Spaces

In this Appendix we present a brief description of the important aspects of Affine Spaces used in Part I of the thesis. For more details the reader can consult for example Berger (1994).

Affine Geometry is just the "usual" Geometry but without the notions of distance and angles. An affine space can be thought of as a set which becomes a vector space by simply selecting a point to be the origin (the zero vector). One of the simplest examples is the plane which it is not a vector space itself but if one selects arbitrarily a point to be the origin then any other point can be regarded as the vector whose tip is precisely that point and is based at the selected origin.

It is well known that the structure of those arrows forms a real vector space in the sense that two arrows represent the same vector if they are parallel translates of each other and they are added and multiplied by scalars according to the parallelogram rules applied to the corresponding arrows. Given a vector v and a point x on the plane we can consider the particular arrow based at x which corresponds to v. Denote its tip by $x \oplus v$. Then each vector v defines an operation on the plane which sends the point x to the point $x \oplus v$. This operation is usually called the translation through v and denoted by $\oplus v$ applied to the right so that the value of $\oplus v$ acting on x is given in the usual way as $x \oplus v$.

Notice it is true that $(p \oplus v) \oplus w = p \oplus (v + w)$, where + is the addition operator between two arrows, and that given any two points x and y there is a unique vector v for which $y = x \oplus v$, namely the vector corresponding to the arrow from x to y. That is to say, the choice of an origin sets up a one-to-one correspondence between points and vectors. Obviously the same structure appears in three and higher dimensional euclidean spaces.

Definition 19 An *affine space* is either the empty set or a triplet $(\mathcal{X}, \mathcal{V}, \oplus)$ consisting of a nonempty set $\mathcal{X}$ (of *points*), a real vector space $\mathcal{V}$ (of *translations*) and a action $\oplus : \mathcal{X} \times \mathcal{V} \to \mathcal{X}$ satisfying the following conditions: We define $x \oplus v := \oplus(x, v)$

(AF1) Let $\vec{0}$ be the zero vector in $\mathcal{V}$. For all $x \in \mathcal{X}$

$$x \oplus \vec{0} = x$$

(AF2) For all $\vec{u}, \vec{v} \in \mathcal{V}$ and all $x \in \mathcal{X}$

$$(x \oplus \vec{u}) \oplus \vec{v} = x \oplus (\vec{u} + \vec{v})$$

(AF3) For any two points $x, y \in \mathcal{X}$ there is a unique $\vec{u} \in \mathcal{V}$ such that $x \oplus \vec{u} = y$. We will denote this unique vector $\vec{u}$ by $\overrightarrow{xy}$.

The *dimension of the affine space* $(\mathcal{X}, \mathcal{V}, \oplus)$ is defined to be the dimension of the vector space $\mathcal{V}$.

To avoid confusion we will always use a symbols like $\oplus, \boxplus, \ldots$ for the translation operator and keep the usual symbol + for the addition of vectors (or real numbers). It is important to keep in mind that when $x, y \in \mathcal{X}$ then $x + y$ it is not defined.

In group theory, conditions (AF1) and (AF2) say that the abelian group $\mathcal{V}$ acts on $\mathcal{X}$ and condition (AF3) is equivalent to say that $\mathcal{V}$ actually acts transitively and faithfully on $\mathcal{X}$.

For every fixed $x_0 \in \mathcal{X}$ consider the mapping $T_{x_0} : \mathcal{V} \to \mathcal{X}$ given by

$$T_{x_o}(\vec{v}) = x_0 \oplus \vec{v}$$

which is bijective in virtue of (AF3). Consider also the mapping $\ell_{x_0} : \mathcal{X} \to \mathcal{V}$ given by

$$\ell_{x_0}(x) = \overrightarrow{x_0 x} \tag{B.1}$$

which is also bijective in virtue of (AF3). Note that

$$\ell_{x_o}(T_{x_0}(\vec{v})) = \ell_{x_0}(x_0 \oplus \vec{v}) = \overrightarrow{x_0(x_0 \oplus \vec{v})} = \vec{v}$$

and

$$T_{x_0}(\ell_{x_0}(x)) = T_{x_0}(\overrightarrow{x_0 x}) = x_0 \oplus \overrightarrow{x_0 x} = x.$$

It is therefore established a one-to-one correspondence in which we can identify $\mathcal{X}$ with $\mathcal{V}$ via the isomorphism ℓ_{x_0}. In this way, we can consider $\mathcal{X}$ as the vector space obtained *by taking x_0 as the origin in $\mathcal{X}$*. This is called the *vectorialization* of $\mathcal{X}$ at x_0. Thus, an affine space $(\mathcal{X}, \mathcal{V}, \oplus)$ is a way of defining a vector space structure on a set of points $\mathcal{X}$, without making the commitment to a fixed origin in $\mathcal{X}$.

Some important properties are the following. Given $x_1, x_2, x_3 \in \mathcal{X}$, since $x_3 = x_1 \oplus \overrightarrow{x_1 x_3}$, $x_2 = x_1 \oplus \overrightarrow{x_1 x_2}$ and $x_3 = x_2 \oplus \overrightarrow{x_2 x_3}$ then

$$x_3 = x_2 \oplus \overrightarrow{x_2 x_3} = (x_1 \oplus \overrightarrow{x_1 x_2}) \oplus \overrightarrow{x_2 x_3} = x_1 \oplus (\overrightarrow{x_1 x_2} + \overrightarrow{x_2 x_3})$$

and thus, by (AF3)

$$\overrightarrow{x_1 x_2} + \overrightarrow{x_2 x_3} = \overrightarrow{x_1 x_3}$$

which is known as *Chasles' identity.* For $x, y \in \mathcal{X}$ we can always write $y = x \oplus \overrightarrow{xy} = (y \oplus \overrightarrow{yx}) \oplus \overrightarrow{xy}$ and using (AF1) and (AF3) we get $\vec{0} = \overrightarrow{yx} + \overrightarrow{xy}$ implying that $\overrightarrow{yx} = -\overrightarrow{xy}$.

B.1 Affine combinations

A fundamental concept in linear algebra is that of a linear combination. The corresponding concept in Affine spaces is that of affine combination. However there is a problem, the sum of two points in an affine space it is not well defined, actually the result is different depending on the selected origin. That is, as described above, we can identify any point in the affine space with a vector via the choice of an origin, then the sum of two points can be defined as the sum of the corresponding vectors but this depends on the selected origin.

Thus, some extra condition is needed for an affine combination to make sense. It turns out that if the scalars involved in the combination add up to unity then the definition is intrinsic as the following simple lemma shows

Lemma 1 *Given an affine space $\mathcal{X}$, let $\{x_i\}_{i\in I}$ be a finite set of points in $\mathcal{X}$ and $\{\lambda_i\}_{i\in I}$ be a set of real numbers such that $\sum_{i\in I}\lambda_i = 1$ then the point*

$$x = x_o \oplus \sum_{i\in I} \lambda_i \, \overrightarrow{x_o x_i} \tag{B.2}$$

is independent of the choice of the origin $x_o \in \mathcal{X}$.

Definition 20 The unique point x in the previous Lemma is called the *affine combination* of $\{x_i\}_{i\in I}$ with coefficients $\{\lambda_i\}_{i\in I}$. It is usually denoted by

$$\sum_{i\in I} \lambda_i \, x_i.$$

Note that the affine combination x is the unique point such that

$$\overrightarrow{yx} = \sum_{i\in I} \lambda_i \, \overrightarrow{yx_i} \qquad \text{for every } y \in \mathcal{X}$$

that is, y playing the role of origin in (B.2) and, setting $y = x$, the unique point such that

$$\sum_{i\in I} \lambda_i \, \overrightarrow{xx_i} = 0.$$

B.2 Affine subspaces

In linear algebra, a vector subspace can be characterized as a nonempty subset of a vector space closed under linear combinations. In affine spaces the notion of subspace corresponds to the same but with affine combinations.

Definition 21 Given an affine space $(\mathcal{X}, \mathcal{V}, \oplus)$, a subset $\mathcal{Y}$ of $\mathcal{X}$ is an *affine subspace* if for every finite set of points $\{y_i\}_{i \in I} \subset \mathcal{Y}$ and for any set $\{\lambda_i\}_{i \in I}$ of real numbers such that $\sum_{i \in I} \lambda_i = 1$, the affine combination $\sum_{i \in I} \lambda_i y_i$ belongs to $\mathcal{Y}$.

As expected, affine subspaces can also be characterized in terms of subspaces of $\mathcal{V}$. Given any point $x \in \mathcal{X}$ and any subspace $\mathcal{U}$ of $\mathcal{V}$, define

$$x \oplus \mathcal{U} := \{x \oplus u \mid u \in \mathcal{U}\}.$$

Lemma 2 *Let $(\mathcal{X}, \mathcal{V}, \oplus)$ be an affine space.*

1. *A nonempty subset $\mathcal{Y}$ of $\mathcal{X}$ is an affine subspace if and only if for every point $y \in \mathcal{Y}$, the set*
$$\mathcal{U}_y := \{\overrightarrow{yz} \mid z \in \mathcal{Y}\}$$
is a subspace of $\mathcal{V}$. Consequently $\mathcal{Y} = y \oplus \mathcal{U}_y$. Furthermore,
$$\mathcal{U} := \{\overrightarrow{yz} \mid y, z \in \mathcal{Y}\}$$
is a subspace of $\mathcal{V}$ and $\mathcal{U}_y = \mathcal{U}$ for all $y \in \mathcal{X}$. Thus $\mathcal{Y} = y \oplus \mathcal{U}$.
2. *For any subspace $\mathcal{U}$ of $\mathcal{V}$ and for any $y \in \mathcal{X}$ the set $\mathcal{Y} = y \oplus \mathcal{U}$ is an affine subspace of $\mathcal{X}$.*

The subspace $\mathcal{U}$ is called the *direction space* of the affine subspace. The dimension of the affine subspace $\mathcal{Y}$ is defined to be the dimension of its direction space. Note that an affine subspace it is actually an affine space in its own right.

B.3 Affine Maps

Corresponding to linear maps there exist the notion of affine maps.

Definition 22 Given any two affine spaces $(\mathcal{X}, \mathcal{V}, \oplus)$ and $(\mathcal{Y}, \mathcal{W}, \boxplus)$, a function $h : \mathcal{X} \rightarrow \mathcal{Y}$ is an affine map if and only if for every family $\{x_i\}_{i \in I}$ of points in X and for every family $\{\theta_i\}_{i \in I}$ of scalars such that $\sum_{i \in I} \theta_i = 1$, we have

$$h\left(\sum_{i \in I} \theta_i \, x_i\right) = \sum_{i \in I} \theta_i \, f(x_i).$$

In other words, h preserves affine combinations.

Lemma 3 *Given a point $x \in \mathcal{X}$, a point $y \in \mathcal{Y}$ and a linear map $\vec{h} : \mathcal{V} \rightarrow \mathcal{W}$ the map $h : \mathcal{X} \rightarrow \mathcal{Y}$ defined as*

$$h(x \oplus \vec{v}) := y + \vec{h}(\vec{v}) \,, \quad \vec{v} \in \mathcal{V}$$

is an affine map. Conversely, given an affine map $h : \mathcal{X} \rightarrow \mathcal{V}$ there exist a unique linear map $\vec{h} : \mathcal{V} \rightarrow \mathcal{W}$ such that

$$h(x + \vec{v}) = h(x) + \vec{h}(\vec{v}) \,, \quad x \in \mathcal{X} \,, \vec{v} \in \mathcal{V}.$$

This unique linear map is called the linear map associated with h.

Let $\mathcal{X}_0$ be an affine subspace of $\mathcal{X}$. Since an affine map preserves affine combinations, and since the affine subspace $\mathcal{X}_0$ is closed under affine combinations, the image $h(\mathcal{X}_0)$ under the affine map h is an affine subspace of $\mathcal{Y}$.

B.4 Affine Independence and coordinates

Corresponding to the notion of linear independence in vector spaces we have the notion of affine independence.

Definition 23 Given an affine space $(\mathcal{X}, \mathcal{V}, \oplus)$ a family $\{x_i\}_{i \in I} \subset \mathcal{X}$ is *affinely independent* if the family $\left\{\overrightarrow{x_i x_j}\right\}_{j \in (I - \{i\})}$ is linearly independent for some $i \in I$.

This definition makes sense because it is easily shown that the independence of the family $\left\{\overrightarrow{x_i x_j}\right\}_{j \in (I - \{i\})}$ does not depend on the choice of x_i. Recall that given a k dimensional vector space $\mathcal{V}$ and an ordered basis $V_0 = \{\vec{v}_1 \ldots, \vec{v}_k\}$ for $\mathcal{V}$, the standard representation of $\mathcal{V}$ with respect to V_0 is the function $\psi_{V_0} : \mathcal{V} \to \mathbb{R}^k$ defined by $\psi_{V_0}(\vec{v}) = (a_1, \ldots, a_k)^t$ where

$$\vec{v} = \sum_{i=1}^{k} a_i\, \vec{v}_i. \tag{B.3}$$

In fact, ψ_{V_0} is an isomorphism. Now for a point $x_0 \in \mathcal{X}$ consider now the following composition

$$\mathcal{C}_{x_0,V_0} := \psi_{V_0} \circ \ell_{x_0} : \mathcal{X} \to \mathbb{R}^k$$

where ℓ_{x_0} is the bijection defined in (B.1). The map $\mathcal{C}_{x_0,V_0}$ is therefore a bijection and can be expressed as

$$\mathcal{C}_{x_0,V_0}(x) = (\theta_1, \ldots, \theta_k)^t \text{ when } x = x_0 \oplus \sum_{i=1}^{k} \theta_i \vec{v}_i.$$

Definition 24 The function $\mathcal{C}_{x_0,V_0}$ is called an *affine parametrization* of the affine space $(\mathcal{X}, \mathcal{V}, \oplus)$ with respect to the point x_0 and the basis V_0.

We can choose another point x_1 and another basis $V_1 = \{\vec{w}_1, \ldots, \vec{w}_k\}$ and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^t$ then

$$\begin{aligned}
\mathcal{C}_{x_0,V_0} \circ \mathcal{C}^{-1}_{x_1,V_1}(\boldsymbol{\theta}) &= \mathcal{C}_{x_0,V_0}(\ell^{-1}_{x_1} \circ \psi^{-1}_{V_1}(\boldsymbol{\theta})) \\
&= \psi_{V_0} \circ \ell_{x_0}\left(x_1 \oplus \psi^{-1}_{V_1}(\boldsymbol{\theta})\right) \\
&= \psi_{V_0}\left(\overrightarrow{x_0\left[x_1 \oplus \psi^{-1}_{V_1}(\boldsymbol{\theta})\right]}\right) \\
&= \psi_{V_0}\left(\overrightarrow{x_0x_1} + \psi^{-1}_{V_1}(\boldsymbol{\theta})\right) \\
&= \psi_{V_0}\left(\overrightarrow{x_0x_1}\right) + \psi_{V_0} \circ \psi^{-1}_{V_1}(\boldsymbol{\theta}) \\
&= \boldsymbol{\theta}_0 + \boldsymbol{M}\boldsymbol{\theta} \qquad (B.4)
\end{aligned}$$

where $\boldsymbol{M}$ is the change of basis matrix from V_1 to V_0 and $\boldsymbol{\theta}_0^t = (\theta_1^0, \ldots, \theta_k^0)$ where

$$\overrightarrow{x_0x_1} = \sum_{i=1}^{k} \theta_i^0 \vec{v}_i.$$

Let $\mathcal{Y}$ be a r dimensional affine subspace of the k dimensional affine space $(\mathcal{X}, \mathcal{V}, \oplus)$. Also let y_0 be a point in $\mathcal{Y}$ and $W = \{\vec{w}_1, \ldots, \vec{w}_r\}$ a basis of the direction space of $\mathcal{Y}$. For $\boldsymbol{\eta} \in \mathrm{I\!R}^r$ we have

$$\begin{aligned}
\mathcal{C}_{x_0,V_0} \circ \mathcal{C}_{y_0,W}^{-1}(\boldsymbol{\eta}) &= \mathcal{C}_{x_0,V_0}(\ell_{y_0}^{-1} \circ \psi_W^{-1}(\boldsymbol{\eta})) \\
&= \psi_{V_0}(\overrightarrow{x_0 y_0}) + \psi_{V_0} \circ \psi_W^{-1}(\boldsymbol{\eta}) \\
&= \boldsymbol{\eta}_0 + \boldsymbol{B}\boldsymbol{\eta} \qquad \text{(B.5)}
\end{aligned}$$

where $\boldsymbol{\eta}_0 \in \mathrm{I\!R}^k$ and $\boldsymbol{B}$ is a $k \times r$ matrix of rank r.

Lemma 4 *Given an affine space $(\mathcal{X}, \mathcal{V}, \oplus)$ let $\{x_0, x_1, \ldots, x_m\}$ be a set of $m+1$ points in $\mathcal{X}$. Let $x \in \mathcal{X}$ be such that $x = \sum_{i=0}^m \theta_i x_i$ where $\sum_{i=0}^m \theta_i = 1$. Then the set $\{\theta_0, \theta_1, \ldots, \theta_m\}$ is unique if and only if $\left\{\overrightarrow{x_0 x_1}, \ldots, \overrightarrow{x_0 x_m}\right\}$ is linearly independent*

Definition 25 An *Affine Frame* in an affine space $(\mathcal{X}, \mathcal{V}, \oplus)$ of dimension k is a set of $k+1$ points $\{x_0, x_1, \ldots, x_k\}$ such that $\left\{\overrightarrow{x_0 x_1}, \ldots, \overrightarrow{x_0 x_k}\right\}$ is a base for $\mathcal{V}$.

Corollary 11 *Let $(\mathcal{X}, \mathcal{V}, \oplus)$ be an affine space of dimension k. Given an affine frame $F_0 = \{x_0, x_1, \ldots, x_k\}$ the mappings $\mathcal{C}_{F_0} : \mathcal{X} \to \mathrm{I\!R}^k$ and $\mathcal{B}_{F_0} : \mathcal{X} \to \mathrm{I\!R}^k$ defined by*

$$\mathcal{C}_{F_0}(x) = (\theta_1, \ldots, \theta_k)^t \text{ when } x = x_0 \oplus \sum_{i=1}^k \theta_i \overrightarrow{x_0 x_i}$$

$$\mathcal{B}_{F_0}(x) = (\theta_0, \theta_1, \ldots, \theta_k)^t \text{ when } x = \sum_{i=0}^k \theta_i x_i \text{ and } \sum_{i=0}^k \theta_i = 1$$

are bijections.

Thus, given an affine space $(\mathcal{X}, \mathcal{V}, \oplus)$ for any $m \geq 1$ it is equivalent to consider a set of $m+1$ points $\{x_0, x_1, \ldots, x_m\}$ in $\mathcal{X}$ or a pair $(x_0, \{\vec{v}_1, \ldots, \vec{v}_m\})$ where the $\vec{v}_i$ are vectors in $\mathcal{V}$.

Consider an affine space $(\mathcal{X}, \mathcal{V}, \oplus)$ of dimension k. Let x_0 be a point in $\mathcal{X}$ and $\{\vec{v_1}, \ldots, \vec{v_k}\}$ be a base for $\mathcal{V}$. Then clearly this a base for the vectorialization of X at x_0. Therefore any point x in $\mathcal{X}$ can be written as

$$x = x_0 + \sum_{i=1}^{k} \theta_i \vec{v}_i$$

for some real numbers $\theta_1, \ldots, \theta_k$.

Appendix C

Dispersion Models

In this Appendix we present a brief of the most important aspect of Dispersion Models that we use in the first part of the thesis. For details, the reader can consult Jorgensen (1997).

The main idea behind dispersion models is that the notions of location and scale may be generalized to *position* and *dispersion* respectively. Similarly, the residual sum of squares from analysis of variance may be generalized to the notion *deviance*. As an important motivation, consider first the density (with respect to the Lebesgue measure on $\mathbb{R}$) function of the normal distribution $N(\mu, \sigma^2)$

$$f(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \qquad x \in \mathbb{R}.$$

From the point of view of dispersion models, the crucial properties of this density function are:

1. The exponent $(x-\mu)^2/(2\sigma^2)$ is a negative constant times the squared distance between x and μ.
2. The factor $(2\pi\sigma^2)^{-1/2}$ does not depend on μ.

The crucial step is to generalize the notion squared distance to the notion of unit deviance. In the following we consider a family of probability distributions for a real-valued random variable Θ. We confine ourselves here to the case where Θ is a continuous random variable. Denote by $\mathcal{S}$ the interior of the smallest interval containing the union of the support of all the members in the family. We let ϑ be a parameter with domain $\mathcal{S}$.

Definition 26 A function $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is called a *unit deviance* if it satisfies

$$\begin{aligned} d(\theta;\theta) &= 0 \qquad \forall\,\theta \in \mathcal{S} \\ d(\theta;\vartheta) &> 0 \qquad \forall\,\theta \neq \vartheta. \end{aligned}$$

A unit deviance d is called *regular* if $d(\theta, \vartheta)$ is twice continuously differentiable on $\mathcal{S} \times \mathcal{S}$ and staisfies

$$\frac{\partial^2 d(\vartheta,\vartheta)}{\partial\vartheta^2} > 0 \qquad \forall\,\vartheta \in \mathcal{S}.$$

The *unit variance function* $V : \mathcal{S} \to \mathbb{R}^+$ of a regular unit deviance is defined by

$$V(\vartheta) = \frac{2}{\dfrac{\partial^2 d(\vartheta,\vartheta)}{\partial\vartheta^2}}$$

The normal distribution unit deviance $d(\theta, \vartheta) = (\theta - \vartheta)^2$ is obviously regular with unit variance function $V(\vartheta) = 1$. Note that the definition does not require the unit deviance to be symmetric in its arguments, although it can be proved it is locally symmetric for θ near ϑ.

Let ϑ_0 be a fixed value in $\mathcal{S}$. Note that if d is a regular unit deviance, then $d(\theta, \vartheta_0)$ has a unique minimum at ϑ_0 implying that for all $\vartheta \in \mathcal{S}$

$$\frac{\partial d(\vartheta,\vartheta)}{\partial\theta} = 0$$

differentiating with respect to ϑ one has

$$\frac{\partial(\vartheta,\vartheta)}{\partial\theta^2} + \frac{\partial(\vartheta,\vartheta)}{\partial\vartheta\partial\theta} = 0$$

and thus

$$V(\vartheta) = \frac{2}{\dfrac{\partial^2 d(\vartheta, \vartheta)}{\partial \vartheta^2}} = \frac{2}{-\dfrac{\partial(\vartheta, \vartheta)}{\partial \vartheta \partial \theta}}.$$

Furthermore we have the following expansion of d near its minimum

$$d(\vartheta_0 + \delta x, \vartheta_0 + \delta m) = \frac{\delta^2}{V(\vartheta_0)}(x - m)^2 + o(\delta^2) \tag{C.1}$$

This expansion shows that a regular unit deviance behaves approximately as the normal unit deviance near its minimum, with curvature given by the reciprocal of the unit variance function.

Definition 27 A *reproductive dispersion model* $DM(\mu, \sigma^2)$ with *position parameter* ϑ and *dispersion parameter* ϵ is a family of probability distributions whose density functions with respect to a suitable measure may be written in the form

$$f(\theta; \vartheta, \epsilon) = a(\theta; \epsilon) \exp\left\{-\frac{1}{2\epsilon} d(\theta; \vartheta)\right\}, \qquad \theta \in \mathcal{S} \tag{C.2}$$

where $a \geq 0$ is a suitable function, d is a regular unit deviance on $\mathcal{S} \times \mathcal{S}$, $\vartheta \in \mathcal{S}$ and $\epsilon > 0$. A dispersion model density of the form (C.2) is said to be expressed in *standard form.*

Note that the support of any $DM(\vartheta, \epsilon)$ depends on ϵ only. Note also that we require (C.2) to hold on $\mathcal{S}$ wich implies that $a(\theta, \epsilon)$ is zero outside the support of $DM(\vartheta, \epsilon)$.

Definition 28 We call (C.2) a *proper dispersion model* $PD(\vartheta, \epsilon)$ if the unit deviance d is regular and (C.2) takes the form

$$f(\theta; \vartheta, \epsilon) = a(\epsilon) V^{-1/2}(\theta) \exp\left\{-\frac{1}{2\epsilon} d(\theta; \vartheta)\right\} \tag{C.3}$$

for $\theta, \vartheta \in \mathcal{S}$, a suitable function a and V the unit variance function.

We call (C.2) a *reproductive exponential dispersion model* $ED(\vartheta, \epsilon)$ if the unit deviance takes the form

$$d(\theta, \vartheta) = \theta h_1(\vartheta) + h_2(\vartheta) + h_3(\theta) \tag{C.4}$$

for some suitable functions h_1 h_2 and h_3.

Note that a natural exponential family is a reproductive exponential dispersion model with h_1 the identity and $\epsilon = 1/2$. Conversely note that a reproductive exponential dispersion model with known ϵ gives a natural exponential family. For example, the normal distribution is a proper dispersion model and also a reproductive exponential dispersion model.

Let $d : \mathbb{R} \to \mathbb{R}$ be a nonnegative twice continuously differentiable function satisfying $d(0) = 0$, $d(\theta) > 0$ for $\theta \neq 0$ and $d''(0) > 0$. Then $d(\theta - \vartheta)$ is a regular unit deviance.

Definition 29 If the integral

$$\frac{1}{a(\epsilon)} = \int_{\mathbb{R}} \exp\left\{-\frac{1}{2\epsilon} d(\theta - \vartheta)\right\} d\theta$$

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

$$f(\theta; \vartheta, \epsilon) = a(\epsilon) \exp\left\{-\frac{1}{2\epsilon} d(\theta - \vartheta)\right\} \tag{C.5}$$

is defined to be a *location-dispersion model* with location parameter ϑ.

Note the corresponding variance function is constant and therefore location-dispersion models are proper dispersion models. In fact, the unit deviance can be rescaled so that the unit variance function is $V(\theta) = 1$, the unit variance function of the normal distribution, showing that many different dispersion models may share the same unit variance function. In analogy to general location-scale models, scale dispersion models can also be defined.

Let $d : \mathbb{R} \to \mathbb{R}$ be a nonnegative twice continuously differentiable function satisfying $d(1) = 0$, $d(\theta) > 0$ for $\theta \neq 1$ and $d''(1) > 0$. Then $d(\theta/\vartheta)$ is a regular unit deviance.

Definition 30

$$f(\theta; \vartheta, \epsilon) = a(\epsilon)\theta^{-1} \exp\left\{-\frac{1}{2\epsilon} d(\theta/\vartheta)\right\} \tag{C.6}$$

is defined to be a *scale-dispersion model* with scale parameter ϑ.

Note this model has unit variance function proportional to θ^2 showing again that there are many proper dispersion models that have the same unit variance function.

We now consider transformations of dispersion models. Consider a regular unit deviance on $\mathcal{S} \times \mathcal{S}$ and a diffeomorphism $h : \mathcal{S} \to h(\mathcal{S})$. A new unit deviance may then be defined on $h(\mathcal{S}) \times h(\mathcal{S})$ by

$$d_h(\phi; \varphi) := d(h^{-1}(\phi), h^{-1}(\varphi))$$

Note that d_h is also regular and the corresponding unit variance function is

$$V_h(\varphi) = V(h^{-1}(\varphi))[h'(h^{-1}(\varphi))]^2.$$

If Θ follows a dispersion model with unit deviance d consider now the transformation $\Phi = h(\Theta)$ with h being the diffeomorphism defined above. Then Φ follows a dispersion model with unit deviance d_h as seen from the following expression of the density of Φ

$$f(\phi; \varphi, \epsilon) = A(\phi, \epsilon) \exp\left\{-\frac{1}{2\epsilon} d_h(\phi; \varphi)\right\} \tag{C.7}$$

where

$$A(\phi, \epsilon) = \frac{a(h^{-1}(\phi); \epsilon)}{|h'(h^{-1}(\phi))|}$$

Consider the particular transformation

$$h(\theta) = \int_{\theta_0}^{\theta} V^{-1/2}(\vartheta) d\vartheta \tag{C.8}$$

for fixed θ_0. This transformation yields a unit deviance d_h with constant variance function $V_h \equiv 1$. It is called the *variance stabilizing transformation.*

The general idea behind the definitions is that, with d being a measure of squared distance, the second factor of (C.2) tends to give a mode point of the density near ϑ. It actually gives a mode when a does not depend on θ, for example in a proper dispersion model with V constant. The smaller the value of ϵ, the higher and more narrow the peak of this mode will be. This makes the ϑ and ϵ somewhat analogous to location and scale respectively.

Another important aspect of (C.2) is that $a(\theta;\epsilon)$ does not depend on ϑ. This generalizes the second property of the normal distribution above, although in contrast to the normal case, $a(\theta;\epsilon)$ may depend on θ. Another important observation is that, in order to show that a given two parameter family is a dispersion model, we must find a parametrization (ϑ,ϵ) that brings the density into the form (C.2). The parameter ϵ is determined up to a constant, because multiplying d and ϵ by the same constant leaves the dispersion model unchanged.

The word *unit* in the terms unit deviance and unit variance function is used here both in the statistical sense of "observational units" and in the sense of corresponding to standardized forms of the functions $\epsilon^{-1}d(\theta;\vartheta)$ and $\epsilon V(\vartheta)$ respectively, with $\epsilon = 1$. The terminology *variance function* refers to the role of $\epsilon V(\vartheta)$ as the asymptotic variance of Θ. In fact, for reproductive exponential dispersion models the exact variance of Θ is $\epsilon V(\vartheta)$ for any value of ϵ. The unit variance function hence summarizes how the variance behaves as a function of ϑ. It is essentially the curvature of the deviance at its minimum. Furthermore, the shape of the unit variance function itself provides a useful summary of the degree and type of non-normality of the corresponding dispersion model. The terminology *reproductive model* refers to the property that the distribution of the sample mean $\bar{\theta}$ for a random sample of size n from the model belongs to the model itself. For an arbitrary reproductive dispersion model the asymptotic variance of the sample mean $\bar{\theta}$ behave like

$$\text{asymptotic variance}(\bar{\theta}) \sim \frac{\epsilon}{n}V(\vartheta) \qquad \text{as } \epsilon \to 0$$

We now introduce the saddlepoint approximation for dispersion model, a useful approximation of the density function.

Definition 31 The *saddlepoint* approximation for a dispersion model with regular unit deviance d is defined by

$$f(\theta;\vartheta,\epsilon) \sim [2\pi\epsilon\, V(\theta)]^{-1/2} \exp\left\{-\frac{1}{2\epsilon}d(\theta;\vartheta)\right\}\ ,\ \text{as } \epsilon \to 0 \tag{C.9}$$

The saddlepoint approximation is valid for an extensive range of models and is often very accurate. It is even exact in a very few special cases such as the normal. The saddlepoint approximation may be interpreted as being half way the original density and a normal approximation. In fact if we replace $V(\theta)$ by $V(\vartheta)$ and introduce a quadratic approximation like (C.1) we essentially get a normal approximation.

Note that the saddlepoint approximation only gives an approximation to the function $a(\theta;\epsilon)$. In fact this approximation is equivalent to

$$\epsilon^{1/2}\, a(\theta;\epsilon) \to [2\pi V(\theta)]^{-1/2} \text{ as } \epsilon \to 0.$$

Note also that the saddlepoint approximation on the right hand side of (C.9), while positive, is not in general a density function on $\mathcal{S}$. However it may be rescaled to become a density function, motivating the following definition.

Definition 32 Let d be a given regular unit deviance defined on $\mathcal{S}\times\mathcal{S}$ the corresponding *renormalized saddlepoint approximation* is the density function defined by

$$f_0(\theta;\vartheta,\epsilon) = a_0(\vartheta,\epsilon)V^{-1/2}(\theta)\exp\left\{-\frac{1}{2\epsilon}d(\theta;\vartheta)\right\} \tag{C.10}$$

where $a_0(\vartheta,\epsilon)$ is the normalizing function such that

$$\frac{1}{a_0(\vartheta,\epsilon)} = \int_{\mathcal{S}} V^{-1/2}(\theta)\exp\left\{-\frac{1}{2\epsilon}d(\theta;\vartheta)\right\} d\theta$$

Fortunately it is possible to show using a Laplace approximation that

$$a_0(\vartheta,\epsilon) \sim [2\pi\epsilon]^{-1/2} \text{ as } \epsilon \to 0.$$

Now we can state two important features of the class of proper dispersion models has:

1. the renormalized saddlepoint approximation is exact and

2. making the variance stabilizing transformation (C.8) we get a location dispersion model

$$f(\phi;\varphi,\epsilon) = a(\epsilon)\exp\left\{-\frac{1}{2\epsilon}d_h(\phi,\varphi)\right\}$$

   making φ the mode of the density.

If $d(\theta,\vartheta)$ is monotone as a function of θ on each side of ϑ, the density is unimodal, and become more and more peaked and concentrated around ϑ as ϵ decreases. In this sense, the interpretation of the parameters ϑ and ϵ as position and dispersion parameters is especially clear for proper dispersion models transformed via the stabilizing variance transformation.

Theorem 26 *Let Θ be a continuous random variable with density (C.10) of a renormalized saddlepoint approximation then it follows that*

$$\frac{\Theta-\vartheta}{\epsilon^{1/2}} \xrightarrow{d} N(0,V(\vartheta)) \text{ as } \epsilon\to 0 \tag{C.11}$$

References

Aalen, O. (1995). Phase type distributions in survival analysis. *Scandinavian Journal of Statistics 22*, 447–463.

Amari, S. and H. Nagaoka (2000). *Methods of information geometry.* American Mathematical Society.

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

Amari, S.-I. (1987). Differential geometrical theory of statistics. In S.-I. Amari, O. E. Barndorff-Nielson, R. E. Kass, S. L. Lauritzen, and C. R. . Rao (Eds.), *Differential Geometry in Statistical Inference*, pp. 19–94.

Amari, S.-I. (1990). *Lecture notes in statistics-28: Differential-geometrical methods in statistics.* Springer-Verlag Inc.

Anaya-Izquierdo, K. and P. Marriott (2006). Local mixture models of exponential families. *Submitted*.

Anaya-Izquierdo, K. A. (2001). Tests of fit for the inverse gaussian and gamma distributions under censoring. *Communications in Statistics: Theory and Methods 30*(4), 757–773.

Asmussen, S., O. Nerman, and M. Olsson (1996). Fitting phase-type distributions via the em algorithm. *Scandinavian Journal of Statistics 23*, 419–441.

Barlow, R. E. and F. Proschan (1975). *Statistical Theory of Reliability and Life Testing Probability Models.* Holt, Rinehart & Winston Inc.

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory.* John Wiley & Sons.

Barndorff-Nielsen, O., P. Blæsild, A. Carey, P. Jupp, M. Mora, and M. Murray (1991). Finite-dimensional algebraic representations of the infinite phylon group. In U. of Aarhus (Ed.), *Research report*, pp. 1–45.

Barndorff-Nielsen, O. E. and D. R. Cox (1994). *Inference and asymptotics.* Chapman & Hall Ltd.

Barndorff-Nielsen, O. E. and P. E. Jupp (1989). Approximating exponential models. *Annals of the Institute of Statistical Mathematics 41*, 247–267.

Bazaraa, M., H. Sherali, and C. Sheety (1993). *Nonlinear programming: Theory and algorithms.* Wiley, New York.

Berger, M. (1994). *Geometry I.* Springer-Verlag Inc.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian theory.* John Wiley & Sons.

Brown, L. D. (1986). *Fundamentals of statistical exponential families: with applications in statistical decision theory.* Institute of Mathematical Statistics.

Caroni, C. and A. C. Kimber (2004). Detection of frailty in weibull lifetime data using outlier tests. *J Statistical Computation and Simulation 74*, 15–23.

Cena, A. (2003). Geometric structures on the non-parametric statistical manifold. *PhD Thesis, University of Milan.*

Chang, H.-Y. and C. M. Suchindran (1997). Testing overdispersion in data with censoring using the mixture of exponential families. *Communications in Statistics: Theory and Methods 26*, 2945–2966.

Chao, S. T. (1985). Maximum likelihood estimates of parameters for the inverse Gaussian distribution based on censored samples (STMA V31 0135). *Journal of the Chinese Statistical Association 23*, 141–152.

Chesher, A. (1984). Testing for neglected heterogeneity. *Econometrica 52*, 865–872.

Cook, R. D. (1986). Assessment of local influence (C/R: p156-169). *Journal of the Royal Statistical Society, Series B, Methodological 48*, 133–155.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika 70*, 269–274.

Cox, D. R. and D. V. Hinkley (2000). *Theoretical statistics*. Chapman & Hall Ltd.

Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B: Methodological 49*, 1–18.

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

Darling, D. (1953). On a class of problems related to the random division of an interval. *Annals of Mathematical Statistics 24*, 239–253.

De Sanctis, A. (2002). The geometry of exponential families of probability distributions. *Statistica 62*(2), 317–321.

Dean, C. B. (1992). Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association 87*, 451–457.

Durbin, J. (1973). *Distribution Theory for Tests Based on the Sample Distribution Function*. SIAM [Society for Industrial and Applied Mathematics].

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics 3*(6), 1189–1242.

Eguchi, S. e. a. (2005). *Proceedings of the Second International Symposium on Information Geometry and its Applications*. University of Tokyo.

Embrechts, P., C. Kluppelberg, and T. Mikosch (1997). *Modelling extremal events: For insurance and finance*. Springer-Verlag Inc.

Fabijonas, B. R. (2002). Laplace's method on a computer algebra system with an application to the real valued modified bessel functions. *Journal of Computational and Applied Mathematics 146*, 323–342.

Feller, W. (1970). *An Introduction to Probability Theory and it Applications*. John Wiley and Sons: London.

Gibilisco, P. and G. Pistone (1998). Connections on non-parametric statistical manifolds by orlicz space geometry. *Infinite Dimensional Analysis, Quantum Probability and Related Topics 1*(2), 325–347.

Giri, N. (1996). *Group invariance in statistical inference*. World Scientific, Singapore.

Grasselli, M. R. (2005). Dual connections in nonparametric classical information geometry. *to appear in the Annals of the Institute for Statistical Mathematics*.

Harter, H. L. and A. H. Moore (1965). Maximum-likelihood estimation of the parameters of gamma and Weibull populations from complete and from censored samples (Corr: V9 p195; V15 p431). *Technometrics 7*, 639–643.

Heckman, J. J., R. Robb, and J. R. Walker (1990). Testing the mixture of exponentials hypothesis and estimating the mixing distribution by the method of moments. *Journal of the American Statistical Association 85*, 582–589.

Husemoller, D. (1966). *Fiber Bundles*. Springer.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika 48*, 419–426.

Jaggia, S. (1997). Alternative forms of the score test for heterogeneity in a censored exponential model. *The Review of Economics and Statistics 79*(2), 340–343.

Janson, S. (1988). Normal convergence by higher order semiinvariants with applications to sums of dependent random variables and random graphs. *Annals of Applied Probability 16*(1), 305–312.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Prc. Roy. Soc. London. Series A. 196*, 456–461.

Jewell, N. P. (1982). Mixtures of exponential distributions. *The Annals of Statistics 10*, 479–484.

Johnson, N., S. Kotz, and N. Balakrishnan (1994). *Continuous univariate distributions*. John Wiley & Sons.

Johnson, N. L. and C. A. Rogers (1951). The moment problem for unimodal distributions. *Ann. Math. Statist. 22*, 432–439.

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

Jorgensen, B. and S. Lauritzen (2000). Multivariate dispersion models. *Journal of Multivariate Analysis 74*, 267–281.

Kass, R. E. and P. W. Vos (1997). *Geometrical foundations of asymptotic inference*. John Wiley & Sons.

Keilson, J. and F. W. Steutel (1974). Mixtures of distributions, moment inequalities and measures of exponentiality and normality. *The Annals of Probability 2*, 112–130.

Kiefer, N. M. (1984). A simple test for heterogeneity in exponential models of duration. *Journal of Labor Economics 3*(4), 539–549.

Klaassen, C., P. Mokveld, and B. van Es (2000). Squared skewness minus kurtosis bounded by 186/125 for unimodal distributions. *Statistics & Probability Letters 50*, 131–135.

Koziol, J. A. (1987). An alternative formulation of Neyman's smooth goodness of fit tests under composite alternatives. *Metrika 34*, 17–24.

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.

Letac, G. and M. Mora (1990). Natural real exponential families with cubic variance functions. *The Annals of Statistics 18*, 1–37.

Lindsay, B. G. (1989). On the determinants of moment matrices. *The Annals of Statistics 17*(2), 711–721.

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

Lockhart, R. A. and M. A. Stephens (1983). Goodness of fit statistics with estimated shape parameters. In *Tecnical Report, Simon Fraser University.*

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.

Marriott, P. (2005). Inference with local mixtures.

Marriott, P. and P. Vos (2004). On the global geometry of parametric models and information recovery. *Bernoulli 10*, 639–649.

McCullagh, P. (1999). Quotient spaces and statistical models. *The Canadian Journal of Statistics 27*, 447–456.

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

Mimoso, C. D. and C. A. de Bragança (1994). On identifiability of parametric statistical models. *Journal of the Italian Statistical Society 3*, 125–151.

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.

Mosler, K. and W. Seidel (2001). Testing for homogeneity in an exponential mixture model. *Australian & New Zealand Journal of Statistics 43*(2), 231–247.

Murray, M. K. and J. W. Rice (1993). *Differential geometry and statistics.* Chapman & Hall Ltd.

Neuts, M. F. (1994). *Matrix-geometric Solutions in Stochastic Models: an Algorithmic Approach.* Dover Publications, Inc.

Neyman, J. (1937). Smooth test for goodness of fit. *Skand. Aktuartioskr. 20*, 149–199.

O'Cinneide, C. A. (1990). Characterization of phase-type distributions. *Communications in Statistics: Stochastic Models 6*, 1–57.

O'Cinneide, C. A. (1993). Triangular order of triangular phase-type distributions. *Communications in Statistics: Stochastic Models 9*, 507–529.

Olsson, M. (1996). Estimation of phase type distributions from censored data. *Scandinavian Journal of Statistics 23*, 443–446.

O'Reilly, F. and M. Stephens (1982). Characterizations and goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B 44*, 353–360.

O'Reilly, F. J. and R. Rueda (1992). Goodness of fit for the inverse Gaussian distribution. *The Canadian Journal of Statistics 20*, 387–397.

O'Reilly, F. J. and R. Rueda (1998). A note on the fit for the Lvy distribution. *Communications in Statistics, Part A – Theory and Methods [Split from: @J(CommStat)] 27*, 1811–1821.

Pace, L. and A. Salvan (1997). *Principles of statistical inference: from a neo-Fisherian perspective.* World Scientific Publishing Co. Pte. Ltd.

Pavur, R. J., R. L. Edgeman, and R. C. Scott (1992). Quadratic statistics for the goodness-of-fit test of the inverse Gaussian distribution. *IEEE Transactions on Reliability 41*, 118–123.

Pettitt, A. N. (1976). Cramer-von Mises statistics for testing normality with censored samples. *Biometrika 63*, 475–482.

Pettitt, A. N. and M. A. Stephens (1976). Modified Cramer-von Mises statistics for censored data. *Biometrika 63*, 291–298.

Pettitt, A. N. and M. A. Stephens (1983). Edf statistics for testing for the gamma distribution. In *Tecnical Report, Stanford University.*

Pistone, G. and M. P. Rogantin (1999). The exponential statistical manifold: Mean parameters, orthogonality and space transformations. *Bernoulli 5*, 721–760.

Pistone, G. and C. Sempi (1995). An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The Annals of Statistics 23*, 1543–1561.

Rao, C. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc. 37*, 81–89.

Rayner, J. C. W. and D. J. Best (1986). Neyman-type smooth tests for location-scale families. *Biometrika 73*, 437–446.

Rayner, J. C. W. and D. J. Best (1989). *Smooth Tests of Goodness of Fit.* Oxford University Press.

Rohatgi, V. and Székely (1989). Sharp inequalities between skewness and kurtosis. *Statistics and Probability Letters 8*, 297–299.

Schmidt, J. W. and W. Heß (1988). Positivity of cubic polynomials on intervals and positive spline interpolation. *BIT 28*, 340–352.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics.* John Wiley & Sons.

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

Shaked, M. and F. Spizzichino (2001). Mixtures and monotonicity of failure rate functions. In *Advances in reliability [Handbook of Statistics 18]*, pp. 185–198.

Sirvanci, M. and I. Levent (1982). Cramer-von Mises statistic for testing exponentiality with censored samples (Corr: V71 p220). *Biometrika 69*, 641–646.

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

Ulrich, G. and L. T. Watson (1994). Positivity conditions for quartic polynomials. *SIAM Journal on Scientific Computing 15*(3).

Wong, R. (2001). *Asymptotic approximations of integrals*. SIAM.

Zelterman, D. (1988). Likelihood ratio tests for central mixtures. *Statistics & Probability Letters 6*, 275–279.

Zelterman, D. and C.-F. Chen (1988). Homogeneity tests against central-mixture alternatives (Corr: V86 p837). *Journal of the American Statistical Association 83*, 179–182.