Applications of Geometry in Optimization and Statistical Estimation

by

Vahed Maroufy

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2015

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Geometric properties of statistical models and their influence on statistical inference and asymptotic theory reveal the profound relationship between geometry and statistics. This thesis studies applications of convex and differential geometry to statistical inference, optimization and modeling. We, particularly, investigate how geometric understanding assists statisticians in dealing with non-standard inferential problems by developing novel theory and designing efficient computational algorithms. The thesis is organized in six chapters as it follows.

Chapter 1 provides an abstract overview to a wide range of geometric tools, including affine, convex and differential geometry. It also provides the reader with a short literature review on the applications of geometry in statistical inference and exposes the geometric structure of commonly used statistical models. The contributions of this thesis are organized in the following four chapters, each of which is the focus of a submitted paper which is either accepted or under revision.

Chapter 2 introduces a new parametrization to general family of mixture models of the exponential family. Despite the flexibility and popularity of mixture models, their associated parameter spaces are often difficult to represent due to fundamental identification problems. Other related problems include the difficulty of estimating the number of components, possible unboundedness and non-concavity of the log-likelihood function, non-finite Fisher information, and boundary problems giving rise to non-standard analysis. For instance, the order of a finite mixture is not well defined and often can not be estimated from a finite sample when components are not well separated, or some are not observed in the sample. We introduce a novel family of models, called the discrete mixture of local mixture models, which reparametrizes the space of general mixtures of the exponential family, in a way that the parameters are identifiable, interpretable, and, due to a tractable geometric structure, the space allows fast computational algorithms. This family also gives a well-defined characterization to the number of components problem. The component densities are flexible enough for fitting mixture models with unidentifiable components, and our proposed algorithm only includes the components for which there is

enough information in the sample.
This paper is under revision in *Statistics and Computing Journal*.

Chapter 3 uses geometric concepts to characterize the parameter space of local mixture models (LMM), introduced in Marriott (2002) as a local approximation to continuous mixture models. Although LMMs are shown to satisfy nice inferential properties, their parameter space is restricted by two types of boundaries, called the hard boundary and the soft boundary. The hard boundary guarantees that an LMM is a density function, while the soft boundary ensures that it behaves locally in a similar way to a mixture model. The boundaries are shown to have particular geometric structures that can be characterized by geometry of polytopes, ruled surface and developable surfaces. As working examples the LMM of a normal model and the LMM of a Poisson distribution are considered. The boundaries described in this chapter have both discrete aspects, (i.e. the ability to be approximated by polytopes), and smooth aspects (i.e. regions where the boundaries are exactly or approximately smooth).
A version of this chapter has been published in *Geometric Science of Information*, Lecture Notes in Computer Science, 9389. p 577-585.

Chapter 4 uses the model space introduced in Chapter 2 for extending a prior model and defining a perturbation space in the Bayesian sensitivity analysis. This perturbation space is well-defined, tractable, and consistent with the elicited prior knowledge, the three properties that improve the methodology in Gustafson (1996). We study both local and global sensitivity in conjugate Bayesian models. In the local analysis the worst direction of sensitivity is obtained by maximizing the directional derivative of a functional between the perturbation space and the space of posterior expectations. For finding the maximum global sensitivity, however, two criteria are used; the divergence between posterior predictive distributions and the difference between posterior expectations. Both local and global analyses lead to optimization problems with a smooth boundary restriction.
Work from this chapter is in a paper under revision in *Statistics and Computing Journal*.

Chapter 5 studies Cox's proportional hazard model with an unobserved frailty for which no specific distribution is assumed. The likelihood function, which has a mixture structure with an unknown mixing distribution, is approximated by the model introduced in Chapter 2, which is always identifiable and estimable. The nuisance parameters in the approximat-

ing model, which represent the frailty distribution through its moments, lie in a convex space with a smooth boundary, characterized as a smooth manifold. Using differential geometric tools, a new algorithm is proposed for maximizing the likelihood function restricted by the smooth yet non-trivial boundary. The regression coefficients, the parameters of interest, are estimated in a two step optimization process, unlike the existed methodology in Klein (1992) which assumes a gamma assumption and uses Expectation-Maximization approach. Simulation studies and data examples are also included, illustrating that the new methodology is promising as it returns small estimation bias; however, it produces larger standard deviation compared to the EM method. The larger standard deviation can be the result of using no information about the shape of the frailty model, while the EM model assumes the gamma model in advance; however, there are still ways to improve this methodology. Also, the simulation section and data analysis in this chapter is rather incomplete and more work needs to be done.

Chapter 6 outlines a few topics as future directions and possible extensions to the methodologies developed in this thesis.

Acknowledgements

Although my name is alone on the front cover of this dissertation, I am by no means its sole contributor. Rather, there are a number of people behind this piece of work who deserve my greatest thanks and appreciation.

First and foremost, my sincere gratitude goes to my supervisor, Prof. Paul Marriott, for his inspiring guidance, kindness, and generosity. Without him this work would never have been possible. His support went far beyond technical insights into geometry and statistics and financial support; he lead me to a more rational and less stressful academic life. I learned from him how to work in a group with courage, patience and mutual respect, how to stay on my feet after disappointments, and how to learn from mistakes.

I would also like to extend my gratitude to Prof. Christopher Small, as my presence at the University of Waterloo was due to his kindness, generosity and trust in me. To my thesis committee, Prof. Shinto Eguchi at Institute of Statistical Mathematics, Tokyo, Prof. Pengfei Li, Prof. Martin Lysy, Prof. Tony Wirjanto and Prof. Dinghai Xu, at University of Waterloo, for their insightful comments on the earlier version of this thesis.

Thank you to all the faculty members and staff of the department of Statistics and Actuarial science for accepting me as a member of such a wonderful and friendly family. Special mention to Mary Lou Dufton, Lucy Simpson, Marg Feeney, Karen Richardson, and Anthea Dunne, who had answers to all my administrative inquiries and solutions to all the problems.

I would also like to thank my best friends Reza Ramezan, Gwyneth Ramezan, Kadir Mirza, Amirhosein Vakili and Sina Hajitaheri for being my true family in Canada- for being there with me at all the tough moments and for all the fun and memorable times we had during these years.

Finally, the biggest sacrifice I made to make this happen was to move away from my closest loved ones: my loving, strong, and protective parents Rahim Maroufy and Soda Ahmadzadeh, and my funny, generous and supportive brothers Ahad, Farhad, and Mohammad. I am forever indebted to them and I want to thank them for all the encouragement, love, and emotional support they provided me throughout my life.

To
my parents Rahim and Soda
who taught me the first lessons about life.
I could never have done this without your support and encouragement.

and

my three cherished brothers
Ahad, Farhad and Mohammad,
without you my life would be boring and incomplete.

Contents

List of Tables

List of Figures

Chapter 1

Convex and Differential Geometry in Statistics

1.1 Introduction

Geometric methods are frequently applied in statistics since various geometric concepts such as vector spaces, affine spaces, manifolds, polyhedrons and convex hulls commonly appear in statistical theory (Amari, 1985; Lindsay, 1995). The application of differential geometry to statistics was inaugurated by Rao (1945), Jeffreys (1946) and Efron (1975), then explicitly formulated by Amari (1985), Eguchi (1985), Barndorff-Nielsen (1987b) and Critchley et al. (1993). Essentially, a family of parametric models is characterized as a manifold on which a suitable geometry is imposed using tensors based on statistical objects such as the Fisher information. More applications of these approaches to statistical modeling and asymptotic theory can be found in Eguchi (1985, 1991), Barndorff-Nielsen (1988), Murray and Rice (1993), Critchley et al. (1994), Marriott and Salmon (2000) and Marriott (2002). In addition, convex geometry tools have been applied to the maximum likelihood estimation in mixture models (Lindsay, 1995), extended exponential family and log-linear models (Fienberg and Rinaldo, 2012; Eriksson et al., 2006), and graphical models (Wainwright and Jordan, 2006; Peng et al., 2012).

This introductory chapter intends to give an abstract overview of applications of geometry to statistics, and reviews the geometric tools necessary for developing the statistical theories and computational algorithms presented throughout this thesis. The chapter is organized as follows. Section 1.2 is an abstract review on the applications of differential and convex geometry in statistics. Section 1.3, introduces affine spaces and studies two important affine spaces in statistical theory. These spaces play an instrumental role in developing latter sections where we define a statistical manifold as an embedded manifold into an affine space, and define its geometry from the geometry of the affine spaces. Section 1.4 is a short overview on the theory of convex spaces, specifically cones and polyhedrons. Section 1.5 studies two specific surfaces: ruled surfaces and developable surfaces which are shown to be useful in characterizing the boundary of certain convex spaces in Chapter 3. Section 1.6 is devoted to differential geometry theory, where a statistical manifold is defined by embedding a parametric model into an affine space, and the required geometric concepts such as tangent spaces, metric tensors and affine connections, are explicitly defined. Finally, the chapter closes with Section 1.7, covering a number of commonly used statistical models and their essential geometric properties.

1.2 Geometry in Statistics

Differential and convex Geometry have been applied to statistical inference theory of commonly used statistical families, such as the exponential family, the mixture and local mixture family, log-linear and graphical models (Section 1.7). Efron (1975) introduces the statistical curvature of one-dimensional models, and studies its influence on statistical inference and efficiency. For example, exponential families have nice inferential properties due to zero statistical curvature inside the exponential affine space (Section 1.3), while non-exponential models have non-zero statistical curvature, hence their asymptotic theory is not as tractable. For higher-dimensional models Murray and Rice (1993, p.18) provide a similar criterion, which is called the second fundamental form, obtained by the orthogonal component of the derivative of score functions to the tangent space. Amari (1985, ch.4) studies the asymptotic theory of inference in a curved exponential model when seen as a submanifold embedded in the exponential family. The geometry of the embedded manifold

is obtained from the larger manifold and used to derive the joint asymptotic probability function of the maximum likelihood estimator (MLE) and associated ancillary statistics, as well as the conditional probability function of MLE given the ancillary, in the form of *Edgeworth* expansions. He also provides a geometric interpretation for consistency, first, second and third order efficiency of an estimator and shows that the MLE is consistent, first and second order efficient, and the bias corrected MLE is third order efficient (Amari, 1985, ch.5-8).

Further results of this type can be found in Eguchi (1983), using minimum contrast geometry, in Barndorff-Nielsen (1986a), Barndorff-Nielsen et al. (1986), Barndorff-Nielsen (1987a), Barndorff-Nielsen (1987b) and Barndorff-Nielsen (1988) using observed information geometry, and in Marriott (1989) and Critchley et al. (1993) using preferred-point geometry. In addition, Critchley and Marriott (2014) show how embedding statistical models in a simplex generalizes the above geometries to statistical models that are not manifolds because of boundary restrictions, and study the behavior of their likelihood functions close to the boundaries.

Studying the geometric aspects of the mixture family and local mixture family, such as convexity and flatness with respect to a suitable geometry, is also of great importance in exploiting the flexibility of these families in statistical inference and modeling (Amari, 1985; Lindsay, 1995; Marriott, 2002). In Lindsay (1995, ch.5), nonparametric maximum likelihood estimation (NPMLE) theorem estimates the mixing distribution by a unique nonparametric discrete distribution with a number of support points not more than the sample size. Essentially, a concave likelihood function is maximized over a convex feasible space of distributions which is defined by the convex hull of the unicomponent likelihood curve. Marriott (2002) and Anaya-Izquierdo and Marriott (2007a) study the local geometry of mixture models and show that a family of continuous mixture models with relatively small mixing variation can be approximated by a finite dimensional family to an arbitrary order. The approximating family, called the local mixture family, is a union of convex subspaces inside the mixture affine space (Section 1.3). These models are extended in Marriott (2006) and applied to information recovery and sensitivity analysis in Marriott and Vos (2004) and Critchley and Marriott (2004).

Understanding the geometry of convex polyhedrons is the key to maximum likelihood

estimation in log-linear models. Eriksson et al. (2006) shows that in hierarchical log-linear models, MLE exists, if and only if, the observed vector of margins lies in the interior of the marginal cone. Fienberg and Rinaldo (2012) and Rinaldo et al. (2009) develop similar results under the conditional Poisson sampling scheme. Using the theory of the extended exponential family, they derive the necessary and sufficient conditions for the existence of MLE, which depend on the sampling zeros in the observed table. They also provide an algorithm to obtain the extended MLE, which requires characterization of the boundary of the marginal cone using the geometry of faces, projection cones and normal cones. Their algorithm has two major steps; first, a unique face of a possibly low-dimensional polyhedron, on which the observed vector of margin lies, is obtained by repetitive linear programing; second, the loglikelihood is restricted to the face and maximized.

Parameter estimation in undirected graphical models is also a problem of optimizing an objective function on a marginal polytope. In some recent related works (Sontag and Jaakkola, 2007; Wainwright and Jordan, 2006; Peng et al., 2012) graphical models with cycles (i.e., not trees) are considered in which their marginal polytopes are not tractable. They replace the marginal polytope with an outer polytope or a semi-definite bound, then use a cutting plane algorithm for maximizing the objective function.

1.3 Affine Spaces

This section intends to give a brief overview of affine spaces and an introduction to two important affine spaces in statistics. The affine property leads to a nice inference and asymptotic theory of statistical models, for example in the exponential family (Murray and Rice, 1993).

Definition 1.1 *A geometrical space, $(X, V, +)$, consists of a set X, a vector space V and a translation operation $+$ is called an affine space if $\forall x \in X$ and $v_1, v_2 \in V$,*

$$x + v_1 \in X \quad \text{and} \quad (x + v_1) + v_2 = x + (v_1 + v_2)$$

and $\forall x_1, x_2 \in X \quad \exists v \in V$ such that $x_1 + v = x_2$.

In the following examples we review the two important affine spaces in statistical theory; the exponential affine space and mixture affine space. The earlier includes the exponential family as an affine subspace, while the later includes the mixture family with latent parameterization (Section 1.7.3) and the local mixture family. Also, both spaces contain the space of smooth densities as a subset and give a way of characterizing families of parametric models as embedded manifolds by embeddings $\theta \to \log f(x;\theta)$ and $\theta \to f(x;\theta)$, respectively (Murray and Rice, 1993, ch.4; Marriott, 2002).

Example 1.1 *The exponential affine space is the triplet $(\mathcal{M}, V_e, \oplus)$, where $\mathcal{M}$ is the space of all positive measures up to a positive finite scale, absolutely continuous with respect to a measure, i.e, all have the same support S, $V_e = \{g(x)|g \in C^\infty(S,\mathbb{R})\}$, and $\oplus$ is the transformation that for any $p \in \mathcal{M}$ and $g \in V_e$ returns $p \oplus g = p\, e^g$.*

Example 1.2 *The triplet $(X_m, V_m, +)$ is called mixture affine space in which*

$$X_m = \left\{ g(x) | \int g(x)\, d\tau = 1 \right\} \ , \ V_m = \left\{ g(x) | \int g(x)\, d\tau = 0 \right\},$$

for a measure τ on the support S, and $+$ is the usual addition of functions.

In Murray and Rice, 1993 the exponential affine space is used as the embedding manifold for a family of smooth density functions, and its simple geometry is exploited to define a geometry on the embedded family by projection. Marriott (2002) embeds a smooth family $f(x;\theta)$ in the mixture affine space and gives a local approximation to a family of continuous mixture models of $f(x;\theta)$ by a convex subspace of the linear embedding space at each point θ. This local approximation subspace is called the family of local mixture models, which has nice inferential properties (Section 1.7.4).

1.4 Convex Geometry

Convex sets, in their general and specific forms, such as convex hulls, polytopes and polyhedral cones, arise in statistical inference theory frequently (Section 1.2). This section

reviews the geometry of convex sets of different kinds and their related properties to statistical theory. Most of the definitions and theorems are taken from Berger (1987) and Matousek (2002). For basic geometric concepts such as hyperplane, half-space, open and close sets, distance, metric space refer to the aforementioned references and Bonnesen and Fenchel (1987). Throughout this section, by cones, polyhedrons and polytopes we shall mean the convex form of them.

Definition 1.2 *A subset C of an affine space A is called convex if $tx + (1-t)y \in C$ for any $x, y \in C$ and $t \in [0,1]$. A function $f : C \to \mathbb{R}$ is called strictly convex if for any $x, y \in C$ and $\gamma \in [0,1]$ we have $f(\gamma x + (1-\gamma)y) < \gamma f(x) + (1-\gamma)f(y)$. Also, f is concave if $-f$ is convex.*

A common example of a convex set is the convex hull of a set of points (or any subset $B \subset A$) which is defined as the smallest convex set containing all the points (subset B). Half-spaces and their intersections are also examples of convex sets. The dimension of a non-empty convex set C, $dim C$, is defined by the dimension of the affine subspace $\langle C \rangle$, the smallest linear subspace containing C, called the subspace spanned by C. Thus, $C \subset A$ is called full-dimensional if $dim\langle C \rangle = dim A$. Also, by imposing a suitable metric on A, the distance between C and any point $y \in A$ is defined by $\min\{d(x,y); x \in C\}$, for which following result holds.

Theorem 1.1 *Suppose C is a convex set in an Euclidean affine space E, and $x \in E$, then there exists at most one point $y \in C$ such that $d(x,y) = \min\{d(x,z)|z \in C\}$, where $d(\cdot,\cdot)$ is the Euclidean distance.*

An alternative way for characterizing convex closed sets is by supporting hyperplanes. Recall first that, for a vector $u \in A$ and a constant v the hyperplane $\mathcal{H} = \{x \in A | x \cdot u = v\}$ divides A into two half-spaces $\{x \in A | x \cdot u \geq v\}$ and $\{x \in A | x \cdot u \leq v\}$, and two subsets $B_1, B_2 \in A$ are said to be separated by the hyperplane $\mathcal{H}$, if they lie in different half-spaces created by $\mathcal{H}$.

Definition 1.3 *For a subset B of the affine space A, the supporting hyperplane of B at $x \in B$ is defined as any hyperplane containing x and separating $\{x\}$ from B.*

A convex closed set C, with the boundary set shown by ∂C, has at least one supporting hyperplane at any point of its boundary. Equivalently, a closed set with non-empty boundary is convex if it has at least one supporting hyperplane at any point of its boundary. The supporting hyperplanes can be used to classify the boundary points of convex sets. For instance, $x \in \partial C$ is a vertex if the intersection of all the supporting hyperplanes at x is an affine space of dimension zero, while ∂C is said to be smooth at x if it has only one supporting hyperplane at x. (Bonnesen and Fenchel, 1987, p.15) Note also that any convex closed set can have only a countable number of vertices.

1.4.1 Polyhedrons

Polyhedrons commonly arise in statistical inference of the extended exponential family and log-linear models in two specific forms; polytopes and polyhedral cones.

Definition 1.4 *A polyhedron is an intersection of a finitely many closed half-spaces, and a polytope is a bounded polyhedron.*

Equivalently, a polytope can be defined by the convex hull of a finite set of points. An important example of polytopes is a simplex which, for a given dimension, has the smallest number of vertices. A simplex is obtained by the convex hull of an affinely independent set of points. Recall that, a set of points are affinely independent if none of the points can be written as an affine combination of the rest of the points. The boundary of a polyhedron comprises vertices and higher-dimensional linear subspaces, which are all called faces of different orders, as characterized in the following definition.

Definition 1.5 *A face of a d-dimensional ($2 \leq d \leq dimA$) polytope P is defined as $P \cap \mathcal{H}$, where $\mathcal{H}$ is a hyperplane and P is contained in one of the closed half-spaces determined by $\mathcal{H}$. Specially, P is a face of itself called an improper face, vertices are 0-faces, and $(d-1)$-dim faces are called facets.*

An alternative criterion for characterizing a boundary point x of a polyhedron, is by the dimensionality of the polar cone at x. Specially, we can determine if x is vertex or a point in the interior of a face with the known dimensionality.

Definition 1.6 *A cone $\mathcal{C}$ is a subset of a vector space such that for any two non-negative real numbers a, b and vectors $v_1, v_2 \in \mathcal{C}$, we have $av_1 + bv_2 \in \mathcal{C}$. The cone of a finite vector set V, cone(V) –called a polyhedral cone– is defined by all the linear combinations with non-negative coefficients of vectors in V.*

Two important polyhedral cones, useful in describing the geometry of convex sets, are the projection cone and polar cone. The projection cone at any $x \in \partial C$ is $cone(V_x)$, where V_x is the set of all the rays emerging from x and containing another point of C. Polar cone of the projection cone at x, which is also called normal cone at x, is defined by $cone(V_{Nx})$ where V_{Nx} is the set of normal vectors to all the supporting hyper-planes at x. The dimensionality of the $cone(V_{Nx})$ at $x \in \partial C$ determines the type of x. Specifically, if C is a d-dimensional closed convex set, then x is a vertex, a point on a p-face, or a point on a facet if dimensionality of $cone(V_{Nx})$ is, respectively, d, $(d-p)$ or $(d-1)$.

1.5 Ruled Surfaces and Developable Surfaces

In this section, two particular surfaces, ruled surfaces and developable surfaces, are briefly introduced. In Chapter 3 we illustrate the role of these surfaces in approximating the boundary of convex sets with some special geometric structures. The technical definitions and results are taken from Do Carmo (1976) and Struik (1988). Preliminary concepts such as regular curves and surfaces are available explicitly in Do Carmo (1976, ch.1,2).

Intuitively, ruled surfaces are generated by a curve and a set of vectors, thus have more structure than a generic surface. A formal definition of ruled surfaces is as follows.

Definition 1.7 *A one-parameter family of lines $\{\alpha(x), \beta(x)\}$ is a correspondence that assigns to each $x \in I \subset \mathbb{R}$ a point $\alpha(x) \in \mathbb{R}^3$ and a vector $\beta(x) \in \mathbb{R}^3$, $\beta(x) \neq 0$. For each $x \in I$ the line L_x, parallel to $\beta(x)$ and passing through $\alpha(x)$, is called the line of the family at x. Given $\{\alpha(x), \beta(x)\}$, the parametrized surface*

$$\Gamma(x, \gamma) = \alpha(x) + \gamma \cdot \beta(x), \quad x \in I, \gamma \in \mathbb{R}.$$

is called the ruled surface generated by the family $\{\alpha(x), \beta(x)\}$. *Also, the lines* L_x *are called rulings and the curve* $\alpha(x)$ *is called the directrix of* Γ. *If, in addition,* $\beta(x)$, $\beta'(x)$ *and* $\alpha'(x)$ *are coplanar for all* $x \in I$, $\Gamma(x, \gamma)$ *is called a developable surface.*

Definition 1.7 can be generalized to higher dimensions by

$$\Gamma(x, \gamma) = \alpha(x) + \gamma \cdot \beta(x), \quad x \in I, \gamma \in \mathbb{R}^k \tag{1.5.1}$$

where $\alpha(x) : \mathbb{R} \to \mathbb{R}^k$, $\beta(x) \in \mathbb{R}^k$, $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_k)$ and $\gamma \cdot \beta(x) = \sum_{i=1}^{k} \gamma_i \beta_i(x)$. Clearly the geometric object defined by equation in (1.5.1) is not a surface or hyper-surface unless $\gamma_1 = \cdots = \gamma_k$; hence, we may call it a *ruled space*. Such spaces are simple examples of fiber bundles (Marriott, 2006).

Example 1.3 *Cylinders and Cones are the simplest ruled surfaces, and also developable surfaces. For a cylinder,* $\alpha(x)$ *is in a plane, say* P, *and the* $\beta(x)$*'s are parallel to a fixed direction. A cone however is obtained from an* $\alpha(x) \in P$ *with rulings passing through a point* $p \notin P$, *called the vertex of the cone (Figure 1.1).*

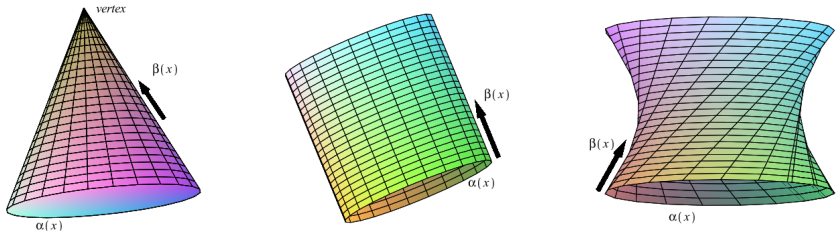

Figure 1.1: A cone, a cylinder and a ruled surface (right), which is not a developable surface, formed by $\alpha(x)$ and $\beta(x) = \alpha'(x) + e_3$, where $e_3 = (0, 0, 1)$

As illustrated in the definition, ruled surfaces consist of a curve and a set of straight lines attached to the curve, and developable surfaces are a specific form of ruled surfaces. Intuitively, a surface with vanishing Gaussian curvature, defined by determinant of the curvature matrix (Do Carmo, 1976, ch.3), at every point is a developable surface. Such surfaces are easily constructed by bending a plane without tearing and stretching (Sun and Fiume, 1996) and are widely used in many different areas of science and engineering.

1.5.1 Envelope of a Family of Planes

One way of constructing a developable surface is by finding the *envelope* of a one-parameter family of planes. This idea can also be generalized to the envelope of a one-parameter family of general surfaces (Struik, 1988, Sec.2-4 and 5-1). For convenience in writing, we use $\lambda = (\lambda_2, \lambda_3, \lambda_4)$ and $a(x) = (a_1(x), a_2(x), a_3(x))$.

Definition 1.8 *The family of planes,* $\mathcal{A} = \{\lambda |\ a(x) \cdot \lambda + d(x) = 0,\ x \in \mathbb{R}\}$*, in which* $a(x)$ *and* $d(x)$ *are differentiable, and each plane is determined by a value of the real parameter* x *, is called an infinite single parameter family of planes.*

Note that, Definition 1.8 can be generalized to a family of hyperplanes. Also, we exclude the family of parallel planes and the family forming a *pencil*, a family of planes passing through the same line.

To obtain the envelope of the family $\mathcal{A}$ and give a similar geometric structure as that of ruled surfaces, we need to find the corresponding directrix and rulings. For any x_1,$x_2 \in \mathbb{R}$ the corresponding planes in $\mathcal{A}$, under our mild regularity, intersect in a line called the *characteristic line*. If $x_2 \in (x_1 - \epsilon, x_1 + \epsilon)$ and $\epsilon \to 0$, the intersecting line is obtained by the following equations,

$$a(x_1) \cdot \lambda + d(x_1) = 0\,, \quad a'(x_1) \cdot \lambda + d'(x_1) = 0 \tag{1.5.2}$$

In a similar way for any x_1,x_2,$x_3 \in \mathbb{R}$ the planes in $\mathcal{A}$ intersect in a point known as the *characteristic point*. If also x_3, x_2 belong to an ϵ interval of x_1, and $\epsilon \to 0$, the characteristic point is determined by following equations,

$$a(x_1) \cdot \lambda + d(x_1) = 0\,, \quad a'(x_1) \cdot \lambda + d'(x_1) = 0\,, \quad a''(x_1) \cdot \lambda + d''(x_1) = 0 \tag{1.5.3}$$

where, in equations (1.5.2) and (1.5.3), prime and double prime represent first and second derivatives with respect to x, respectively.

Putting together the infinite number of characteristic points corresponding to the planes in the family $\mathcal{A}$, we obtain a curve, called the *edge of regression*. Moreover, all the characteristic lines together, side-by-side, construct a surface called the *envelope* of the family

$\mathcal{A}$, which is a developable surface (Figure 1.2). In addition, it can be shown that the characteristic lines are tangent to the edge of regression at their characteristic points, Struik (1988, p.67).

Example 1.4 *Another developable surface is tangential developable which is the envelope of a set of tangent planes to a space curve, L. Any plane P crossing L, intersects with the surface in a curve with a cusp on L; hence, edge of regression is also called* cuspidal edge. *The characteristic lines (rulings of the surface) are also tangent to L.*

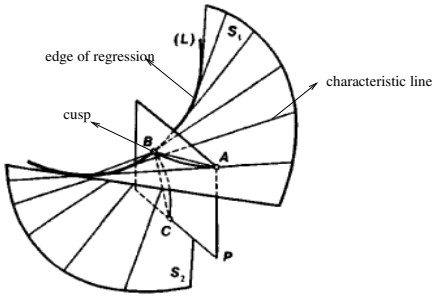

Figure 1.2: Tangential developable

1.6 Differential Geometry

Differential geometry methods are shown to be instrumental in statistical theory, as under some regularity conditions a statistical model can form a Riemannian manifold equipped with Fisher information metric or observed information metric (Amari, 1985; Barndorff-Nielsen and Cox, 1989; Murray and Rice, 1993). This section provides an abstract overview of differential geometry methods required for the statistical applications in the following chapters. As in Murray and Rice (1993), a statistical manifold is defined by embedding a family of smooth densities into the exponential affine space and its geometric components are obtained with respect to the embedding affine space. We use these tools in the latter chapters where the boundary of (the fiber at a point) a convex parameter subspace is characterized as a smooth manifold immersed in an Euclidean space, and covariant derivatives are exploited to design a gradient-based searching algorithm on the manifold.

1.6.1 Statistical Manifold

In differential geometry a manifold is defined in two different ways; as a space which is locally diffeomorphic to an open subspace of an Euclidean space at any point, or as a

non-linear space embedded into an affine space. Here we follow the second definition for two reasons. It is intuitive for statistical applications, as the set of density functions is a subset of the affine spaces in Section 1.3. Also, it is sufficient for our purposes throughout the thesis where we characterize the boundary of a convex parameter space as a manifold immersed in an Euclidean space. For formal definitions we follow Murray and Rice (1993) and Marriott and Salmon (2000) closely. Also, unless otherwise mentioned, for a family of densities $f(x;\theta)$, where $\theta = (\theta^1, \cdots, \theta^r)$, $x = (x_1, \cdots, x_d)$, and $\ell(\theta; x)$ is the loglikelihood function, we assume the following regularity conditions,
1- All members have common support.
2- The set of functions $\{\frac{\partial}{\partial\theta^i}\ell(\theta;x)|i = 1, \cdots, r\}$ are linearly independent and their moments exists up to a sufficient order.
3- Integration and partial derivative for all relevant functions to $f(x;\theta)$ are commutative.
The inverse function theorem (Dodson and Poston, 1979, p.221) and the implicit function theorem (Rudin, 1976, p.224) stated bellow, are also required for our definition of manifold to be concrete.

Theorem 1.2 *Suppose $h : A \to A'$, where A and A' are affine spaces and $h \in C^k$, set of k times differentiable function. Then the derivative mapping at x, $D_x h$, is an isomorphism, if and only if, there are neighborhoods N_x and $N_{h(x)}$ of x and $h(x)$ such that $h(N_x) = N_{h(x)}$ and h has a local C^k inverse $h^{-1} : N_{h(x)} \to N_x$.*

Theorem 1.3 *Suppose h is a continuously differentiable mapping from an open set $U \in \mathbb{R}^{m+n}$ into $\mathbb{R}^n$ such that, $h(a,b) = 0$ for some $(a,b) \in U$ and $D_x h$ is invertible. Then there exist open sets $U^* \subset \mathbb{R}^{m+n}$ and $W \subset \mathbb{R}^m$, with $(a,b) \in U^*$ and $b \in W$ such that, to every $y \in W$ corresponds a unique x where $(x,y) \in U^*$ and $h(x,y) = 0$. If x is defined by $g(y)$, then g is a continuously differentiable mapping from W into $\mathbb{R}^n$, $g(b) = a$ and $h(g(y), y) = 0$.*

Now we are at the position to define a manifold as the image of a smooth mapping into an affine space (Marriott and Salmon, 2000, p.17) as follows.

Definition 1.9 *For an open set $\Phi \in \mathbb{R}^r$ and affine space A, consider the map $\Upsilon : \Phi \to A$. Then the image $\Upsilon(\Phi)$ is an embedded manifold if,*

(i) the derivative of Υ *has full rank* r *for all points in* Φ*,*
(ii) the inverse image of any compact set is itself compact.

Condition (i) guarantees that Υ is invertible by Theorem 1.2, hence any point $\theta \in \Phi$ has a unique image in A. Restriction (ii) is required so that the differential map of Υ is one-to-one. Thus the image $\Upsilon(\Phi)$ is locally diffeomorphic to a subset of $\mathbb{R}^r$, so it has a structure of a manifold on its own right, and also it is mapped into A by a closed inclusion. Recall that an inclusion from $M(\subset A)$ into A maps any point $p \in M$ to the same point in A, and under a closed map the image of any closed set is a closed set. Although Definition 1.9 does not retain a global structure, this may not be an issue for applications in statistics as most of the statistical models only have one global coordinate; hence, they have one global differentiable chart (Amari, 1985, p.15).

Definition 1.9 together with Example 1.1 give a way of expressing a family of smooth densities as an embedded manifold inside the exponential affine space. Note first that the exponential affine space can be presented via its logarithm representation by $(\mathcal{M}_l, V_e)$ where $\mathcal{M}_l = \{\log(p) \,|\, p \in \mathcal{M}\}$, and the translation is

$$\log(p) \to \log(p) + f(x)$$

This representation is more intuitive to work with as it is common to represent statistical manifolds by logarithm of density functions, and since a loglikelihood function is defined only up to addition of a constant, the space of loglikelihood functions is a natural subspace of $\mathcal{M}_l$. Now consider a family of smooth density functions satisfying the above regularity conditions embedded in the exponential affine space. The set of loglikelihoods of this family lie in $\mathcal{M}_l$, since by regularity condition 1 they have the same support. Condition (i) is immediately implied by regularity condition (2). For (ii) to be satisfied it amounts for the mapping to have a continuous inverse.

1.6.2 Tangent Spaces

To apply geometric methods on a manifold, M, embedded in an affine space, A, a suitable geometric structure is required. The first step is to define the tangent space TM_p at any

point $p \in M$. Tangent spaces can be defined in two different ways; as the best first-order linear approximation of M about p, or as the space of directional derivatives which are smooth operators on M. The first definition is more intuitive since, based on our definition of manifold, TM_p is a linear subspace of A. Specifically, $TM_p \subset TA_p$ where TA_p is isomorphic to the translation vector space attached to A at p. The later definition, is useful for defining the rate of change of tangent vector fields which lie in tangent space as directional derivative operators. Hence, we briefly mention both definitions. Einstein's summation rule is used throughout (Amari, 1985, p.18).

According to the first approach, TM_p is defined as the space of tangent vectors to all the curves through p. Specifically, if $\rho(t) := \ell(\theta(t); x)$ defines a curve for $t \in (-\epsilon, \epsilon)$, where $t = 0$ represents p using θ parameterization, then by chain rule $\rho'(0) = \frac{\partial \ell}{\partial \theta^i}\frac{d\theta^i}{dt}$ which is a vector in A with origin at p. Thus, TM_p is a linear subspace of A spanned by $\{\partial_i,\, i = 1, \cdots, r\}$, where $\partial_i := \frac{\partial \ell}{\partial \theta^i}$, and by regularity condition 2 it is r-dimensional. The dual of TM_p, called cotangent space TM_p^*, is defined as the linear subspace spanned by $\{d\theta^1, \cdots, d\theta^r\}$, where $\partial_i(d\theta^j) = \delta_{ij}$ which is 1 if $i = j$ and 0 otherwise. Also, $TM = \{TM_p,\, p \in M\}$ and $TM^* = \{TM_p^*,\, p \in M\}$ are called tangent bundle and cotangent bundle, respectively.

The explicit definition of TM_p as above depends on the θ parametrization, however tangent spaces are geometric objects and invariant with respect to parameterization. Specially, if $\eta := \eta(\theta)$ is a new parameterization with basis $\{\partial_a,\, a = 1, \cdots, r\}$, using chain rule we obtain the base change formula $\partial_i = (\partial_i \eta^a)\partial_a$, and TM_p under both parameterizations is the same. For the dual space the base change formula is obtained similarly as $d\theta^i = \partial_a \theta^i d\eta^a$. This invariance property is critical in statistical theory where TM_p corresponds to the space of score vectors. For more about invariant methods in statistics, see McCullagh (1987).

Alternatively, a tangent vector can be seen as a smooth differential operator on M, which assigns directional derivative to smooth functions defined on M in a given direction (Amari, 1985, p.17). Then TM_p is the space of directional derivatives at p.

Definition 1.10 *A tangent vector at $p \in M$ is a mapping $X_p : C^\infty(M) \to \mathbb{R}$, which for all $f, g \in C^\infty(M)$ and $a, b \in \mathbb{R}$ satisfies*

$$X_p(af + bg) = aX_p(f) + bX_p(g), \qquad X_p(fg) = gX_p(f) + fX_p(g)$$

1.6.3 Metric Tensors

Metric tensors are the next component of geometric structure on manifolds, which allow us to calculate quantities such as length and angle on tangent spaces. Let $\chi(M)$ be the space of all tangent fields X, the deferential operators as in Definition 1.10. A metric tensor is defined as follows.

Definition 1.11 *A metric tensor is a smooth function defined by*

$$\begin{aligned} \langle\, ,\, \rangle : \chi(M) \times \chi(M) &\to C^{\infty}(M) \\ (X, Y) &\mapsto \langle X, Y \rangle \end{aligned}$$

which is bilinear, symmetric and positive definite.

Since $\langle\, ,\, \rangle$ is bilinear, for a parametrization θ, we have $\langle\, ,\, \rangle = g_{ij} d\theta^i d\theta^j$ where $g_{ij} = \langle\, \partial_i, \partial_j \,\rangle$ are called the components of the metric tensor. Under the new parameterization η, using the base change formula in the dual space the new metric components are obtained as $\tilde{g}_{ab} = \partial_a \theta^i \partial_b \theta^j g_{ij}$. This transformation rule guarantees that a metric tensor is also a geometric object, i.e, it is independent of coordinate systems, and consequently, ensures that the quantities such as lengths and angles defined by the metric are invariant under reparametrization. The common metric tensors for a statistical manifold are the Fisher information, observed information metrics and preferred-point metrics (Amari, 1985; Barndorff-Nielsen, 1986b; Critchley et al., 1993).

1.6.4 Affine connections

Concepts such as flatness, straightness and curvature are commonly used in differential geometry. For instance, affine spaces are flat and the minimum distance between any two points in an affine space is the straight line joining them. These properties do not hold for general manifolds, in which non-flatness raises the concept of curvature of different kinds, and straight lines are called geodesics which do not retain shortest distances in general. To formulate the above notions, one more geometric object is required, called an affine

connection or covariant derivative. Affine connections enable us to calculate the rate of change of tangent vector fields, which in turn gives a way of defining flatness, curvature and straightness.

An explicit definition of an affine connection, connection for short, as the operator $\nabla : \chi(M) \times \chi(M) \to \chi(M)$ is given in (Amari, 1985, p.35). In this section however, we use a more intuitive way of constructing connections which is consistent with Definition 1.2 of embedded manifold in an affine space, and it is sufficient for our uses throughout the thesis. Similar to (Murray and Rice, 1993, ch.4) we define a connection in two steps; (i) ordinary derivative in the embedding affine space, which is naturally defined, (ii) orthogonal projection into the tangent space of the embedded manifold. See also (Dodson and Poston, 1979, ch.8) for more details.

For an affine space A with translation vector space V, any vector $\nu \in V$ determines a tangent vector at each point $p \in A$, thus there is a linear isomorphism between V and T_pA. Then for a choice of basis $\nu_1, \cdots, \nu_r$, any tangent vector field X is written as $X^i\nu_i$ and a natural connection is defined as the vector field $\nabla X(\omega) = dX^i(\omega)\nu_i$ for any $\omega \in TA_p$. Since TM_p is a linear subspace of TA_p, this connection can be projected into TM_p by a linear map $\pi_p : TA_p \to TM_p$ at any point $p \in M$, where π_p is the orthogonal projection defined based on the imposed metric tensor (Murray and Rice, 1993, p.118). Hence, a connection on M is defined by $\bar{\nabla} X(\nu) = \pi_p(\nabla X(\nu))$, which can be shown that it satisfies the axioms of connections stated in Amari (1985, p.35). For a choice of coordinate θ a connection ∇ can be presented by its components Γ^k_{ij}, the k^{th} component of covariate derivative of $\frac{\partial}{\partial\theta^i}$ at $\frac{\partial}{\partial\theta^j}$,

$$\nabla \frac{\partial}{\partial\theta^i}(\frac{\partial}{\partial\theta^j}) = \Gamma^k_{ij}\frac{\partial}{\partial\theta^k}.$$

Example 1.5 uses this method to develop the two important connections, exponential (∇^{+1}) and mixture (∇^{-1}) connections defined in Amari (1985). Also, in the following chapters, we use this approach to obtain covariate derivative of a loglikelihood function restricted to the boundary surface of a parameter space which is shown to be a manifold.

Example 1.5 *Let M be the embedded manifold into the exponential affine space. The regular derivative of a tangent vector base ∂_i is given by $\partial_j\partial_i = \frac{\partial^2\ell}{\partial\theta^j\partial\theta^i}$. Orthogonal projection*

with respect to the metric $\langle\ ,\ \rangle = Cov_p(\ ,\ )$ *returns* ∇^{+1}. *Alternatively, if* M *is embedded into the mixture affine space similar projection defines* ∇^{-1} (Murray and Rice, 1993, p.119; Marriott, 2002).

As indicated in Section 1.2, different geometries can be defined on a statistical manifold by imposing different metrics and connections. For instance, Fisher information metric is used by Amari (1985) and observed Fisher information metric is exploited by Barndorff-Nielsen (1986a). A rather different and, in some sense more general, way to define a geometry on a statistical manifold is given by Eguchi (1983), called the minimum contrast geometry. In this method, the metric and connection are defined with respect to the choice of a divergence function. For a suitable contrast function ρ, which is a divergence function by definition, the minimum contrast geometry is defined by endowing a manifold with the following metric and connection,

$$g_{ij}^{(\rho)}(\theta) = -\frac{\partial}{\partial\theta_1^i}\frac{\partial}{\partial\theta_2^j}\,\rho(\theta_1,\theta_2)|_{\theta=\theta_1=\theta_2}, \quad \Gamma_{ijk}^{(\rho)}(\theta) = -\frac{\partial^3}{\partial\theta_1^i\partial\theta_1^j\partial\theta_2^k}\,\rho(\theta_1,\theta_2)|_{\theta=\theta_1=\theta_2} \quad (1.6.4)$$

By exploiting this geometry, the minimum contrast estimators are introduced and efficiency properties of this estimator for the curved exponential family are derived, see also Eguchi (1985, 1991), for more on this geometry. The expected geometry in Amari (1985), is obtained from this method by choosing KullbackLeibler divergence for the contrast function ρ.

Having defined connections ∇ and $\bar{\nabla}$ on A and M, respectively, Riemann curvature and embedding curvature can be introduced on M. Riemann curvature is defined using the components of $\bar{\nabla}$ for a choice of coordinate system and shows how M is curved as a disembodied manifold. Embedding curvature, the second fundamental form in Murray and Rice (1993), is defined as the orthogonal component of ∇ into $\bar{\nabla}$, showing how M is curved inside A. If Riemann curvature is zero at any $p \in M$ then M is said to be flat and there is a coordinate system for which the corresponding connection components are zero. Such a coordinate is called affine coordinate system for M.

The notion of straightness on a manifold is defined by geodesic curves. A curve is called geodesic with respect to a connection if the rate of change of its tangent vector field along

the curve is zero. As we mentioned earlier, in general these curves do not give the shortest path between two points on manifolds; however, for a special connection, called Levi-Civita, the two notions of straightness and minimum distance coincide. This connection is also known as metric connection and its components are called Christoffel symbols.

The two non-metric connections ∇^{+1} and ∇^{-1} are of great importance for theory of statistical manifolds. In Amari (1985), α-family of connections (∇^{α}) is defined by linear combinations of ∇^{+1} and ∇^{-1}, for any real value α. He also shows that there is a duality link between ∇^{α} and $\nabla^{-\alpha}$; hence, if a manifold is flat with respect to ∇^{α} then it is also flat with respect to $\nabla^{-\alpha}$. Essentially, two connections ∇ and $\nabla^{\star}$ are said to be dual if they satisfy

$$\langle A, B\rangle_{\theta} = \langle \Pi_{\rho} A, \Pi_{\rho}^{\star} B\rangle_{\theta'}$$

where θ and θ' are points on ρ corresponding to $p, p' \in M$, and $\Pi_{\rho} : T_p M \to T_{p'} M$ is the parallel mapping based on connection ∇ along ρ. According to this definition the Levi-Civita connection is self-dual (Amari, 1985, p.70).

1.7 Statistical Examples

1.7.1 Exponential Family

The family of continuous or discrete probability densities

$$f(x;\theta) = \exp\{\theta^i s_i - \psi(\theta)\} m(x) \tag{1.7.5}$$

with respect to some fixed measure ν, is called full exponential family, where $\theta \in \Theta \subset \mathbb{R}^r$, $m(x)$ is a non-negative function independent of θ, and $S(X) = (S_1(X), \cdots, S_r(X))$, a function of random variable X, is the vector of sufficient statistics for θ. Θ is called natural parameter space, and the family is said to be regular when Θ is open. For more details see Brown (1986).

This family, when parameterized by its natural parameter θ, satisfies the geometry of affine spaces as in Example 1.1. That is, the components of ∇^{+1} are zero, hence it is flat

with respect to ∇^{+1} and θ is its affine coordinate. Based on the duality theorem, this family is also flat with respect to ∇^{-1}, and the dual affine coordinate is the vector of expected parameters $E_\theta(S(X))$.

Inside the flat family of probability densities in 1.7.5, a d-dimensional ($d < r$) curved family is defined by a one-to-one and smooth mapping $\mathcal{B} : \Xi \to \Theta$ which assigns $\theta(\xi)$ to each $\xi \in \Xi$ and satisfies the following conditions,

(a) derivative of $\mathcal{B}$ has full rank at all $\xi \in \Xi$,

(b) if a sequence of points $\{\theta_j,\, j = 1, 2, \cdots\} \subseteq \mathcal{B}(\Xi)$ converges to $\theta_0 \in \mathcal{B}(\Xi)$ then $\{\mathcal{B}^{-1}(\theta_j),\, j = 1, 2, \cdots\}$ converges to $\mathcal{B}^{-1}(\theta_0) \in \Xi$,

and its probability density has the following form

$$f(x;\xi) = \exp\{\theta^i(\xi)s_i - \psi(\theta(\xi))\}m(x) \tag{1.7.6}$$

Conditions (a) and (b) together ensure that the curved family is an embedded manifold inside the affine space of the full exponential family, as in Definition 1.9. For more examples of curved exponential models such as Poisson regression and AR(1) models see Marriott and Salmon (2000) and Kass and Vos (1997).

This embedding structure is used in Amari (1985) for studying the asymptotic theory of inference in curved exponential models. Essentially, the MLE, $\hat{\xi}$, is defined as the point in the embedded manifold $M \subset A$ which is the -1-projection of $\bar{x} \in A$. Corresponding to $\hat{\xi}$ the ancillary space $B(\xi)$ is defined for any $\xi \in M$ and coordinatized by ω, then $\bar{x}$ is decomposed as $\bar{x} = (\hat{\xi}, \hat{\omega})$ where $\xi = (\xi^1, \cdots, \xi^r)$, $\omega = (\omega^{r+1}, \cdots, \omega^N)$ and $N = dim A$. Working with expected parameterization, it is shown that $\hat{\xi}$ is consistent, if and only if, any $\eta(\xi) \in M \subset A$ belongs to $B(\xi)$. For such a consistent estimator, considering the true parameter value ξ, the asymptotic distribution of $\tilde{\xi} = \sqrt{n}(\hat{\xi} - \xi)$ is shown to be normal with mean 0 and inverse asymptotic variance $g_{1ab} = g_{ab} - g_{ai}\, g_{bj}\, g^{ij}$, in which $a, b, \cdots$ represent quantities related to M and $i, j, \cdots$ represent those related to $B(\xi)$. Then clearly when $g_{bj} = 0$; that is, $B(\xi)$ is orthogonal to M, we have $g_{1ab} = g_{ab}$ and $\hat{\xi}$ is (first-order) efficient. Furthermore, the bias of this estimator is shown to be a combination of the components

of ∇^{-1} and the corresponding embedding curvature, and the bias corrected estimator is proved to be also second-order efficient. Finally, the third-order term in the expansion of the mean square error is shown to vanish, if and only if, $B(\xi)$ is -1-flat, giving third-order efficiency of the bias corrected MLE.

1.7.2 Extended exponential Family

Often, in discrete exponential families depending on the observed sample, the MLE is not attained, even though the likelihood function is bounded. Thus, the extended exponential family and extended MLE are defined. For a full exponential family $\mathcal{S}$ with respect to ν, let $\nu^{cl(F)}$ be the restriction of ν to the closure of F, where F is a face of the convex core of ν, $cc(\nu)$, defined as the intersection of all convex Borel subsets $B \subset \mathbb{R}^d$ for which $\nu(B) = \nu(\mathbb{R}^d)$. Then the exponential family with respect to $\nu^{cl(F)}$ is shown by $\mathcal{S}^F$ with natural parameter space Θ_F, and the extended exponential family of $\mathcal{S}$ is defined by the union of all the families $\mathcal{S}^F$ over all faces of $cc(\nu)$, (Csiszar and Matus, 2005; Malago and Pistone, 2010). An application of this extension is in maximum likelihood estimation in the log-linear models (Section 1.7.4).

Example 1.6 (Logistic regression) *Consider the logistic regression model for n binary responses, Y_i's, and an $n \times d$ design matrix X. The joint model of the vector $(Y_1, Y_2, \cdots, Y_n)$ lies in a $2^n - 1$ simplex,* (Critchley and Marriott, 2014; Anaya-Izquierdo et al., 2013a). *However, if Y_i's are independent and $p_i > 0$, $(i = 1, 2, \cdots, n)$ then the joint distribution is an n-dimensional full exponential family with natural parameters $x_i^T \beta$, where x_i^T is the i^{th} row of the design matrix. Now let*

$$\theta_i = \log\left(p_i/(1-p_i)\right) = x_i^T \beta, \quad \beta = (\beta_1, \beta_2, \cdots, \beta_d)^T, \tag{1.7.7}$$

defining a mapping $\mathbb{R}^d \to \mathbb{R}^n$, then the resulted model is a d-dimensional curved exponential family. Note however that, when $p_i \geq 0$ the family does not have manifold structure. In this case the joint model of independent Y_i's lies in the extended exponential family.

1.7.3 Mixture Family

Finite mixture model of k density functions $f_j(x)$, from the same family, is defined by

$$g(x;\eta) = \sum\nolimits_{j=1}^{k} \eta_j \, f_j(x), \quad \eta_j \geq 0, \quad \sum\nolimits_{j=1}^{k} \eta_j = 1 \tag{1.7.8}$$

where η is called the latent parameter vector. Continuous mixture models are also defined using a continuous latent distribution by integrating over the latent parameter (see Everitt and Hand, 1981; Mclachlan and Kaye, 1988; Lindsay, 1995). This family is flat with respect to ∇^{-1} connection, and consequently with respect to ∇^{+1} connection (Amari, 1985, p.41). To see the affine structure of this family in η parameterization, without loss of generality, let $k = 2$, so $\eta_2 = 1 - \eta_1$, then we have

$$g(x,\eta) = f_1(x) + \eta_1 \, [f_2(x) - f_2(x)] \tag{1.7.9}$$

This family satisfies the affine geometry as in Example 1.2 since $\int f_1(x)\, dx = 1$ and $\int [f_2(x) - f_1(x)]\, dx = 0$.

In general the geometry of mixture models is much more complicated than the geometry of parametric models. The space of all mixture models of family $\mathcal{F} = \{F_\theta\}$ is the smallest convex set containing $\mathcal{F}$, which is not a manifold because of the boundaries. Specifically, when $\eta_j > 0$ the model lies in the interior of a $(k-1)$-simplex, otherwise the model belongs to a lower dimensional simplex which would be a face of the $(k-1)$-simplex. Also Anaya-Izquierdo (2006, p.25-35) illustrates how the general mixture family can be defined as a convex subspace of the mixture affine space, and curved mixture models as curves inside the space of the general mixture models.

1.7.4 Local Mixture Models

Marriott (2002) introduced the family of local mixture models (LMM) by embedding a family of models into the mixture affine space, Example 1.2. For a density function $f(x;\mu)$, belonging to the exponential family, the LMM of order k is defined by

$$g(x;\mu,\lambda) = f(x;\mu) + \sum\nolimits_{j=1}^{k} \lambda_j \, f^{(j)}(x;\mu), \qquad \lambda \in \Lambda_\mu \subset \mathbb{R}^k \tag{1.7.10}$$

where $\lambda = (\lambda_1, \cdots, \lambda_k)$, $f^{(j)}(x;\mu) = \frac{\partial^j}{\partial \mu^j} f(x;\mu)$, and for any fixed and known μ,

$$\Lambda_\mu = \left\{ 1 + \sum\nolimits_{j=1}^{k} \lambda_j \, q_j(x;\mu) \geq 0, \;\; \text{for all} \;\; x \in S \right\},$$

is a subspace obtained from intersection of half-spaces, with $q_j(x;\mu) = \frac{f^{(j)}(x;\mu)}{f(x;\mu)}$, and S is the sample space of the density $f(x;\mu)$. Hence, Λ_μ is a non-empty convex set which has at least one supporting hyperplane at each boundary point. The boundary of Λ_μ is called the hard boundary and guarantees positivity.

The family can be seen as an example of the fiber bundle structure used by Amari (1985), where $f(x;\mu)$ is a curve inside the space of the full exponential family and $\sum_{j=1}^{k} \lambda_j \, f^{(j)}(x;\mu)$ is a -1-flat fiber attached to the curve at each μ. In fact, since λ coordinates of each fiber are restricted to the convex subspace Λ_μ, the family of LMMs is a union of convex subspaces inside the -1-affine space.

For identifiability purposes, Anaya-Izquierdo and Marriott (2007a) drop the first derivative from model (1.7.10) and study LMMs of the form

$$g(x;\mu,\lambda) = f(x;\mu) + \sum\nolimits_{j=2}^{k} \lambda_j \, f^{(j)}(x;\mu), \qquad \lambda \in \Lambda_\mu \subset \mathbb{R}^{k-1} \tag{1.7.11}$$

They show that this model is identifiable in all parameters, and (μ, λ) parameterizations are Fisher orthogonal at $\lambda = 0$. Also, for any fixed and known μ the loglikelihood function is concave.

LMMs can be applied to approximate continuous mixture models with small mixing variation, to an arbitrary order; however, they are richer than the general family of mixture models in some sense. That is, compared to parametric models with the same mean, LMMs unlike mixture models can also produce smaller dispersion. Hence, for a LMM to behave locally similar to a mixture model, it must be restricted to an additional boundary, called the soft boundary. This boundary will be explicitly defined and computed in Chapter 3. LMM's are also useful for modeling over-dispersion in binomial and Poisson regression models, frailty analysis in lifetime data analysis Anaya-Izquierdo and Marriott (2007b), measurement errors in covariates in regression models Marriott (2003), local influence analysis Critchley and Marriott (2004) and the analysis of predictive distributions Marriott (2002).

In the later chapters of this thesis we use LMMs for developing various statistical theories, methodologies and computational tools. In each chapter we explicitly define the LMM model which we will be using for the sake of completeness and preventing any confusion between the two models in equations (1.7.10) and (1.7.11). In Chapter 2 we use model (1.7.10) for extracting information from mixture data with unidentifiable components, and define a novel family of discrete mixture models which gives a tractable parametrization to the general space of mixture models. In Chapters 4 and 5 we exploit the same model for studying robustness in Bayesian analysis and survival data with unknown frailty distribution, respectively. In Chapter 3, however, we use the model in equation (1.7.11) for two reasons. First, it gives a suitable base for running profile likelihood estimation for the parameter μ. Second, it gives lower dimensional parameter subspaces and boundaries at each fiber which we can visualize.

1.7.5 Log-linear Models

For a set of discrete random variables $\{Y_1, Y_2, \cdots, Y_k\}$, the log-linear model with the set of generators $\mathscr{G} = \{\mathscr{G}_j;\ j = 1, \cdots, J;\ J < k\}$ is defined by

$$P(y) \propto \prod_{j=1}^{J} \phi_{\mathscr{G}_j}(y),$$

where $\mathscr{G}_j \subset \{Y_1, Y_2, \cdots, Y_k\}$, y is a realization of Y, and $\phi_{\mathscr{G}_j}$ is a function that depends on y only through the variables in $\mathscr{G}_i$ (Geiger et al., 2006). Log-linear models are a member of discrete exponential family and commonly used for studying association among categorical variables in contingency tables when one makes no distinction between response and explanatory variables (Agresti, 1990, ch.5). Natural parameter vector θ, or equivalently the average of counts μ are the parameters of interest, where $m = \log(\mu)$ lies in a linear space, $\mathcal{M}$ say.

In more details, let Y_j takes its value from $[d_j] = \{1, 2, \cdots, d_j\}$ and $\mathcal{I} = \bigotimes_{j=1}^{k} [d_j]$ be an index set such that any $i \in \mathcal{I}$ represents a cell in the corresponding cross-classified table of counts, and $I = |\mathcal{I}|$. Then the arrays of cell counts and expected cell counts can be represented by $n = (n_1, \cdots, n_I)^T$ and $\mu = (\mu_1, \cdots, \mu_I)^T \in \mathbb{R}^I$. Corresponding to a

log-linear model there is a 0/1-matrix $A_{I\times d}$, called the design matrix, such that $t = A^T n$ returns the vector of margins, a sufficient statistic for the natural parameter vector θ, and $\mathcal{M}$ is spanned by its column vectors. d is the dimension of t, and t is minimal if A is full rank. Also, the natural parameters are related to the log-expected parameters via $m = A\,\eta$.

The geometry of log-linear model, which is instrumental in characterizing existence of MLE, is tied with geometry of the space of all observable vector of margins t, shown by C_A. This space is the convex hull of all observable t vectors, which is a polyhedral cone, $cone(A)$ under poisson sampling scheme, or a polytope under multinomial sampling. For an observed table of counts the MLE exists if and only if t belongs to relative interior of C_A. Otherwise, the maximum point, called extended MLE, lies on the unique face C_A which contains t in its interior (Fienberg and Rinaldo, 2012; Eriksson et al., 2006).

Example 1.7 (Graphical models) *A specific form of a hierarchical log-linear model is an undirected graphical model defined by a graph with a set of vertexes, corresponding to a random vector, and edges showing dependence between the connected vertexes. A hierarchical log-linear model is a graphical model if its generators return the cliques, maximal complete subgraphs, of the corresponding graph and viz* (Edwards, 2000)*. Undirected graphical models are widely used in statistics, network models and biology. Two major inferential problems of interest related to them: finding marginal probabilities, and maximum a posteriori assignments, can be formulated as optimization of a nonlinear objective function in a polytope called the marginal polytope* (Sontag and Jaakkola, 2007)*. Also for more details on existed estimating methods refer to* Wainwright and Jordan (2006) *and* Peng et al. (2012)*.*

Example 1.8 (Contingency tables) *Consider a normalized two-by-two contingency table with probability vector* $(p_{11}, p_{12}, p_{21}, p_{22})$*. When* $p_{ij} > 0$ *for all* i, j*, this model has a multinomial distribution, lies in* 3*-dimensional full exponential family and can be embedded in the* 3*-simplex of the probability vectors. However, for* $p_{ij} \geq 0$ *the model is extended multinomial distribution in the extended exponential family* (Critchley and Marriott, 2014)*.* Fienberg and Gilbert (1970) *showed that the subset* C *of the points inside the tetrahedron, satisfying the row-column marginal independence, is a ruled surface called the surface of independence with explicit parametrization.*

$$(s,t) \to [st, s(1-t), (1-s)t, (1-s)(1-t)]\,, \qquad 0 \leq s,t \leq 1 \tag{1.7.12}$$

They also presented two different sets of rulings constructing this surface; as a result it is also called a doubly ruled surface. For a general $r \times c$ table the independence subspace is a manifold generated by non-intersecting linear spaces, inside the simplex (Fienberg, 1968)

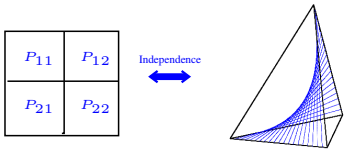

Figure 1.3: Ruled surface of a two-by-two contingency table inside 3D simplex

1.8 Summary and Contributions

This chapter covers the principal geometric tools, particularly in affine, convex and differential geometry, commonly applied in statistics, and illuminates geometric properties of a list of frequently used statistical models. Although the explicit proofs and technical derivations are not provided, it should be sufficient to follow the latter chapters which study new applications of these geometries in different areas of statistics, and describe the main contributions of this thesis.

In Chapter 2 we target the issues such as identifiability and estimability of general mixtures of the exponential family models. As a solution, we define the family of discrete mixtures of LMMs, which has the flexibility of general mixtures, yet always identifiable and estimable. It provides a novel parametrization to the family of mixture models and holds useful geometric properties which are fruitful in designing efficient estimation algorithms. We propose a type of Expectation-Maximization algorithm for estimating the parameters and give a new characterization to the number of components of a finite mixture model. These models are capable of approximating general mixture models, and give a way of fitting general mixture data without prior information about the mixing process and the number of components.

Chapter 3 exploits Sections 1.4 and 1.5 to characterize the parameter space of a LMM which is restricted by hard and soft boundaries. The boundaries are shown to have particular geometric structures that can be computed by polytopes, ruled and developable surfaces. In particular, we consider a normal and a Poisson model, revealing that the

boundaries can have both continuous and discrete aspects. The Appendix section gives some characterizations of the boundaries for LMMs of other distributions such as a binomial and an exponential distribution. We also propose an orthogonal parametrization for the computed boundary surfaces and show how the loglikelihood function can be restricted to these boundaries and maximized.

Chapter 4, exploits the general mixture model introduced in Chapter 2 for extending a base prior model and defining a perturbation space in Bayesian sensitive analysis. For assessing maximum sensitivity of inference or prediction, a perturbation space is defined which is natural, interpretable and flexible for incorporating modelers' prior knowledge. We aim at analyzing both local and global sensitivity to prior perturbation. The methodology leads to the problems of finding maximum local direction and maximum global sensitivity, both restricted to a convex space with a smooth boundary.

in Chapter 5 we target the identifiability issue in Cox's proportional hazard model with unobserved frailty for which no specific distribution is assumed. The likelihood has a mixture structure; hence, to overcome the identifiability problem we approximate that by the discrete mixture of LMMs introduced in Chapter 2, which is always identifiable and estimable. In the approximated likelihood frailty model is represented by a finite dimensional parameter vector lying inside a convex subspace of a linear space. We exploit these properties to design an efficient optimization algorithm for estimation of all the parameters.

Chapter 6 outlines few topics as future direction and possible extensions to the methodologies developed in this thesis.

Chapter 2

Mixture Models: Building a Parameter Space

2.1 Introduction

Mixtures of exponential family models have found application in almost all areas of statistics, see Lindsay (1995), Everitt (1996), Mclachlan and Peel (2000) and Schlattmann (2009). At their best they can achieve a balance between parsimony, flexibility and interpretability – the ideal of parametric statistical modelling. Despite their ubiquity there are fundamental open problems associated with inference on such models. Since the mixing mechanism is unobserved, a very wide choice of possibilities is always available to the modeller: discrete and finite with known or unknown support, discrete and infinite, continuous, or any plausible combination of these. This gives rise to the first open problem; what is a good way to define a suitable parameter space in this class of models? Other, related, problems include the difficulty of estimating the number of components, possible unboundedness and non-concavity of the log-likelihood function, non-finite Fisher information, and boundary problems giving rise to non-standard analysis. All these issues are described in more detail below. This chapter defines a new solution to first of these problems. We show how to construct a parameter space for general mixtures of exponential families, $\int f(x;\mu)dQ(\mu)$, where the parameters are identifiable, interpretable, and, due

to a tractable geometric structure, the space allows fast computational algorithms to be constructed.

2.1.1 Background

Let $f(x;\mu)$ be a member of the exponential family. It will be convenient, but not essential to any of the results of this chapter, to parameterize with the mean parameter μ. We will further assume that the dimension of μ is small enough to allow underlying Laplace expansions to be reasonable, Shun and Mccullagh (1995). A mixture over this family would have the form $\int f(x;\mu)dQ(\mu)$ where Q is the mixing distribution which, as stated above, can be very general. Since Q may lie in the set of all distributions the 'parameter space' of this set of models is infinite dimensional and very complex. It is tempting to restrict Q to be a finite discrete distribution indeed, as shown by Lindsay (1995), the non-parametric maximum likelihood estimate of Q lies in such a family. Despite this, as the following example clearly shows, this is too rich a class to be identified in a statistically meaningful way.

Example 2.1 *For this example let $f(x;\mu) = \phi(x;\mu,1)$, be a normal distribution with unit variance. The QQ plot in Figure 2.1 compares two data sets generated from two different finite mixture models with five and ten components respectively. The plot shows that data generated from each can have very similar empirical distributions – thus it would be very hard to differentiate between these models and hence estimate the number of components. In this example the components of the mixing distributions have been selected to be close to one another and to have the same lower order moment structure.*

Different methods have been proposed for determining the order of a finite mixture model, including graphical, Bayesian, penalized likelihood maximization, and likelihood ratio hypothesis testing (Mclachlan and Peel, 2000; Hall and Stewart, 2005; Li and Chen, 2010; Maciejowska, 2013). We question though if the order is, fundamentally, an estimable quantity:

(I) First, mixture components may be too close to one another to be resolved with a given set of data, as in Example 2.1.

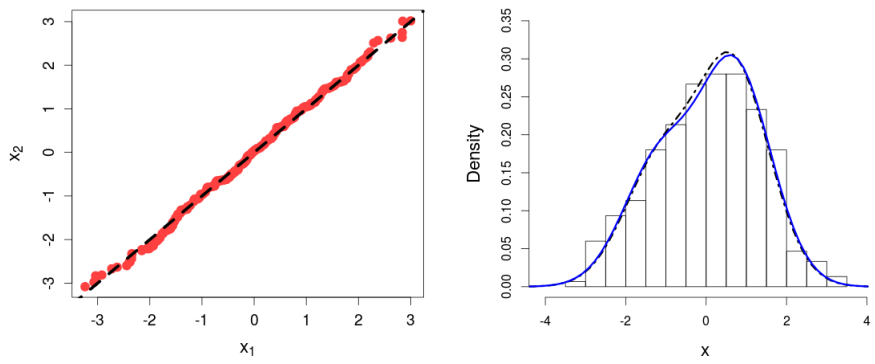

Figure 2.1: Left, the QQ-plot for samples generated from each finite mixture model. Right: the histogram of the sample with fitted local mixture density plots for each sample.

(II) Secondly, for any fixed sample size the mixing proportion for some components may be so small that contributions from these components may not be observed.

For instance, Culter and Windham (1994) show, using an extensive simulation, that when the sample size is small or the components are not well separated, likelihood based and penalized likelihood-based methods tend to overestimate or underestimate this parameter. Donoho (1988), studies the order as a functional of a mixture density and points out that, "near any distribution of interest, there are empirically indistinguishable distributions (indistinguishable at a given sample size) where the functional takes on arbitrarily large values". He adds, "untestable prior assumptions would be necessary", additional to the empirical data, for placing an upper bound. Celeux (2007) mentions that this problem is weakly identifiable from data as two mixture models with distinct number of components might not be distinguishable.

This fundamental identification issue has immediate consequences when we are trying to define a tractable parameter space. In particular its dimension is problematic: the space will have many dimensional singularities as component points merge or mixing distributions become singular. Identifiability with mixtures has been well studied of course, see Tallis (1969) and Lindsay (1993). The boundaries and singularities in the parameter space of a

finite mixture have been looked at in Leorox (1992), Chen and Kalbfleisch (1996) and Li et al. (2009) as have the corresponding effects on the shape of the log-likelihood function, for example see Gan and Jiang (1999).

2.1.2 The Local Mixture Approach

Examples where there is a single set of closely grouped components – or the much more general situation where Q is any small-variance distribution – motivated the design of the local mixture model (LMM), Marriott (2002), Anaya-Izquierdo and Marriott (2007a). This is constructed around parameters about which there is information in the data and can be justified by a Laplace, or Taylor, expansion argument.

Definition 2.1 *For a mean, μ, parametrized density $f(x;\mu)$ belonging to the regular exponential family, the local mixture of order k is defined as*

$$g_\mu(x;\lambda) := f(x;\mu) + \sum\nolimits_{j=1}^{k} \lambda_j f^{(j)}(x;\mu) \quad \lambda \in \Lambda_\mu \subset \mathbb{R}^k \tag{2.1.1}$$

where $\lambda = (\lambda_1, \cdots, \lambda_k)$ and $f^{(j)}(x;\mu) = \frac{\partial^j f}{\partial \mu^j}(x;\mu)$. We denote $q_j(x;\mu) := \frac{f^{(j)}(x;\mu)}{f(x;\mu)}$, then for common sample space S, and any fixed μ,

$$\Lambda_\mu = \left\{ \lambda | 1 + \sum\nolimits_{j=1}^{k} \lambda_j \, q_j(x;\mu) \geq 0, \forall x \in S \right\},$$

is a convex subspace obtained by intersection of half-spaces. The boundary of Λ_μ corresponds to a positivity condition on $g_\mu(x;\lambda)$.

Example 2.2 (2.1 revisited) *The right panel of Figure 2.1 shows the LMM fit to the two datasets considered above. We see that the model can successfully capture the shape of the data using only a small number of parameters about which the data is informative. This observation is formalized by Lemma 2.2 in Section 2.4.*

The local mixture approach is designed, using geometric principles, to generate an excellent inferential frame in the situation which motivated it. The 'cost' associated with

these properties is having to work explicitly with boundaries in the inference. We give more details of these properties and the tools associated with working with the boundaries in Section 2.2, and their explicit calculation in Chapter 3. Of course the major weakness of this approach is that it says nothing when the mixing is not 'local'. This chapter addresses this issue by looking at finite mixtures of local mixture models. This combines the nice properties of finite mixtures, for example the work of Lindsay (1995), while avoiding the fundamental trap of overidentifying the models as described in Section 2.1.1. We use this finite mixture of local mixtures to approximate the parameter space of all mixtures. In later sections estimation methods in this very rich model class are discussed, as is the problem of what a particular data set can tell us about the number of components examined in important classes of mixture models.

2.2 Local and Global Mixture Models

Let us consider a general mixture model of the form $\int_{\mu \in M} f(x;\mu) dQ(\mu)$ where we make the assumption that the support of Q, M, is compact. We can therefore partition M as $M = \cup_{i=1}^{L} M_i$ where $M_i \cap M_j = \emptyset$ for $i \neq j$, and each M_i is connected. Let us also select a set of 'grid points', $\mu_i \in M_i$, which will be fixed and known throughout.

The distribution Q can be written as a convex combination of distributions $Q = \sum_{i=1}^{L} \rho_i Q_i$, where (i) Q_i has support M_i, and (ii) for large enough L each Q_i is a localizing mixture in the sense required by Anaya-Izquierdo and Marriott (2007a), allowing each term $\int_{\mu \in M_i} f(x;\mu) dQ_i(\mu)$ to be well approximated by a LMM. In the form given in Definition 2.1 the mean of the LMM is $\mu + \lambda_1$, so there is one degree of ambiguity about the parametrization (μ, λ). In Anaya-Izquierdo and Marriott (2007a) this was resolved by always setting $\lambda_1 = 0$, see Section 1.7.4. In Definition 2.2 the mean ambiguity is resolved by fixing μ_i and having λ_1^i free.

Definition 2.2 *Let $g_{\mu_l}(x;\lambda^l)$ be the LMM from Definition 2.1, and $\lambda^l = (\lambda_1^l, \cdots, \lambda_k^l)$. A discrete mixture of LMMs is defined by*

$$h(x,\mu,\rho,\lambda) = \sum\nolimits_{l=1}^{L} \rho_l \, g_{\mu_l}(x;\lambda^l) \tag{2.2.2}$$

where $\lambda = (\lambda^1, \cdots, \lambda^L)$, $\mu = (\mu_1, \cdots, \mu_L)$ *is a fixed grid of support points,* $\rho = (\rho_1, \cdots, \rho_L)$ *such that* $0 \leq \rho_l \leq 1$ *and* $\sum_{l=1}^{L} \rho_l = 1$.

There are some points to consider in this definition. First, the choice of how to select the fixed grid points μ_i, in particular how far they should be separated, is clearly critical and discussed in Section 2.2.1. Second, throughout this chapter we only consider LMMs of order $k = 4$. Increasing this degree – while mathematically possible – only adds a small marginal improvement to the local modeling performance, (Marriott, 2006). Third, whenever $f(x;\mu)$ is a proper exponential family, the terms $q_j(x,\mu)$'s are polynomials of degree j, and the interior of the parameter space Λ_{μ_0} can be characterized by analyzing the roots of a quartic polynomial. Finally, we use throughout two illustrative examples: the normal and binomial.

Example 2.3 (Normal) *For the normal density function* $\phi(x;\mu,1)$*, with fixed variance at* $\sigma^2 = 1$*, the LMM at* $\mu = \mu_0$ *has the following form,*

$$\begin{aligned} g_{\mu_0}(x;\lambda) &= \phi(x;\mu_0,1)\{1 + \lambda_1(x-\mu_0) + \lambda_2[(x-\mu_0)^2 - 1] + \lambda_3[(x-\mu_0)^3 \\ &\quad -3(x-\mu_0)] + \lambda_4[(x-\mu_0)^4 - 6(x-\mu_0)^2 + 3]\} \end{aligned} \tag{2.2.3}$$

with, $\quad E(X) = \mu_0 + \lambda_1, \quad Var_g(X) = 1 + 2\lambda_2 - \lambda_1^2, \quad \mu_g^{(3)} = 6\lambda_3 + 2\lambda_1^3 - 6\lambda_1\lambda_2$

in which $\mu_g^{(3)}$ *is the third central moment. The expression for the first moment and an argument based on Fisher orthogonality of density derivatives (Morris, 1982) illustrate how identifiability is attained either by fixing* $\mu = \mu_0$ *or* $\lambda_1 = 0$.

Example 2.4 (Binomial) *The LMM for a binomial distribution, with mean* $\mu = \mu_0$ *and number of trials* n*, has a probability mass function of the form*

$$\begin{aligned} g_{\mu_0}(x;n,\lambda) &= \frac{n!\mu_0^x(n-\mu_0)^{n-x}}{x!(n-x)!n^n}\{1 + \lambda_1 p_1(x,\mu_0) + \lambda_2 p_2(x,\mu_0) + \lambda_3 p_3(x,\mu_0) \\ &\quad +\lambda_4 p_4(x,\mu_0)\} \end{aligned} \tag{2.2.4}$$

where $p_j(x,\mu_0)$ *is a polynomial with degree* j*. In this example there is extra boundary structure as* μ *is limited to the compact set* $[0,n]$.

Definition 2.3 *For fixed μ_0 the parameter space Λ_{μ_0} is a convex subset of $\mathbb{R}^4$ and its boundary, $\partial\Lambda_{\mu_0}$ is defined by the envelope of hyperplanes*

$$\Pi_x := \left\{\lambda | 1 + \sum\nolimits_{j=1}^{4} \lambda_j \, q_j(x;\mu) = 0\right\},$$

parametrized by $x \in S$, Struik (1988). The boundaries of LMMs are computed explicitly in Chapter 3.

2.2.1 Choosing the Support Points

In Definition 2.2 the set of support points, $\{\mu_1, \cdots, \mu_L\}$, is assumed fixed and the question remains: how to select it? Recall that the LMM gives a good approximation when the variance of the mixing distribution is small. This would imply that we want neighboring support points to be close, on the other hand the more support points the larger the value of L and hence the larger the dimension of the parameter space in Definition 2.2.

The following result follows from standard Taylor remainder results and formalizes the above discussion.

Lemma 2.1 *Suppose $g_{\mu_0}(x;\lambda)$ is the local mixture of the family of densities $f(x;\mu)$ and Q is a distribution. For any $\delta > 0$ there exist an interval $I = [\mu_0 - \epsilon_1(\delta), \mu_0 + \epsilon_2(\delta)]$ such that*

$$\left| \int_I f(x;\mu)\, dQ - g_{\mu_0}(x;\lambda) \right| < \delta,$$

for all x.

Example 2.5 (2.3 revisited) *By Taylor's theorem we have $f(x;\mu) - g_{\mu_0}(x;\lambda(\mu)) = \frac{(\mu-\mu_0)^5}{5!} f^{(5)}(x;m)$ where m is a value between μ and μ_0. For the normal family with standard deviation σ we have*

$$f^{(5)}(y,m) = \left(y^5 - 10\frac{y^3}{\sigma^2} + 15\frac{y}{\sigma^4}\right) \frac{e^{-\frac{\sigma^2 y^2}{2}}}{\sqrt{2\pi}\sigma},$$

where $y = \frac{(x-m)}{\sigma^2}$. *This function is obviously bounded, by* M *say, for all* $y \in \mathbb{R}$, *and the bound, which only depends on* σ, *can be numerically obtained. Letting* $\epsilon = \max\{\epsilon_1, \epsilon_2\}$ *gives,*

$$\begin{aligned} \left| \int_I f(x;\mu)\, dQ - \int_I g_{\mu_0}(x;\lambda(\mu))\, dQ \right| &\leq \int_I |f(x;\mu) - g_{\mu_0}(x;\lambda)|\, dQ \\ &\leq (\epsilon_1 + \epsilon_2)\, \frac{\epsilon^5}{5!}\, M \end{aligned} \tag{2.2.5}$$

The result follows since we can write $\int_I g_{\mu_0}(x;\lambda(\mu))\, dQ$ *as a LMM with* $\lambda_i := \int \lambda_i(\mu) dQ(\mu)$.

Example 2.6 (2.4 revisited) *For the binomial family, with probability function* $p(x;n,\mu)$, *again we want to bound the error by* $\frac{(\mu-\mu_0)^5}{5!}M$, *say. We have*

$$p^{(5)}(x;n,m) = p(x;n,m)\, q_5(x;n,m)$$

where $q_5(x;n,m)$ *is a polynomial of degree* 5 *of both* x *and* m, *which can be written as*

$$q_5(x;n,m) = \frac{1}{(n-m)^5} \sum_{j=0}^{5} \gamma(j) \binom{5}{j} (\frac{n}{m})^j (-1)^{5-j}$$

with $\gamma(j) = j!\,(5-j)!\binom{x}{j}\binom{n-j}{n-5}$. *It can be shown that uniformly for all* $x = 0, 1, \cdots, n$, $p(x;n,m) \leq p(x^\star;n,m)$, *where* $x^\star = \lfloor \frac{m(n+1)}{n} \rfloor$, *and*

$$L(n,m) < q_5(x;n,m) < U(n,m)$$

where, for all $m \in [0,n]$,

$$\begin{aligned} L(n,m) &= -\frac{\gamma(0)}{(n-m)^5 m^4}(5n^4 + 10n^2m^2 + m^4) \\ U(n,m) &= \frac{\gamma(0)}{(n-m)^5 m^5}(n^5 + 10n^3m^2 + 5nm^4 - m^5) \end{aligned} \tag{2.2.6}$$

Moreover, it can be shown that

$$\begin{cases} U(n,m) > |L(n,m)| & \text{if} \quad 0 \leq m \leq \frac{n}{2}; \\ U(n,m) < |L(n,m)| & \text{if} \quad \frac{n}{2} < m \leq n. \end{cases}$$

Therefore,

$$M = \max_{m \in I} q_5(x^\star;n,m)|L(n,m)| \quad \text{or} \quad M = \max_{m \in I} q_5(x^\star;n,m)U(n,m)$$

which depends on μ_0, ϵ_1 *and* ϵ_2.□

2.2.2 Estimation Methods

Estimation with a LMM is, in general, straightforward. The parameter space has nice convexity properties and the likelihood is log-concave, see Anaya-Izquierdo and Marriott (2007a). In Marriott (2002) Markov Chain Monte Carlo (MCMC) methods are used since boundaries in the parameter space can easily be accommodated by a simple rejection step whenever a parameter value is proposed that lies outside the boundary. Alternatively direct log-likelihood maximization can be done exploiting the convexity of the parameter space and the concavity of the objective function. See Section 5.2.2 for the explicit description of this algorithm.

Adopting these ideas to finite mixtures of LMMs, we can also easily use MCMC methods. However, here we define a new form of Expectation-Maximization (EM) algorithm, described below, and applied in Example 2.7. In this example we look at mixtures of normals, $\phi(x;\mu,\sigma_0^2)$, where grid-points for μ are selected as discussed in Section 2.2.1. To understand the selection of σ_0^2 by the modeler we return to point (II) of Section 2.1.1. This makes the case that we can only estimate clusters, and indeed features of such clusters, if there is the associated information in the data. One consequence of that is the well-known phenomenon that infinite likelihoods are attainable in the case where only a single observation has been associated with a normal cluster and the estimated variance is zero. In our approach we take issue (II) seriously and only put in a LMM component when there is enough data to support its inference. In particular we note that the variance of such a component is $\sigma_0^2+2\lambda_2-\lambda_1^2$, which will be bounded below, and vary from cluster to cluster. Hence the data can estimate the variance of each cluster as long as it is above our, modeler selected, threshold.

The Algorithm

Starting from initially selected grid points $\mu^{(0)} = (\mu_1^{(0)},\cdots,\mu_L^{(0)})$, proportions $\rho^{(0)} = (\rho_1^{(0)},\cdots,\rho_L^{(0)})$ and local mixture parameters $\lambda^{(0)} = (\lambda^{1,(0)},\cdots,\lambda^{L,(0)})$. Suppose, at step r, we have $\mu^{(r)}$ and $\rho^{(r)}$ and $\lambda^{(r)}$ and the number of components $L_r \leq L$. For obtaining the estimates at step $r+1$ run the following steps.

1. Calculate $\rho^{(r+1)} = \frac{n_l}{n}$, where $n_l = \sum_{i=1}^{n} w_{il}^{(r+1)}$ and

$$w_{il}^{(r+1)} = \frac{\rho_l^{(r)} g_{\mu_l}(x_i, \lambda^{l,(r)})}{\sum_{l=1}^{L_r} \rho_l^{(r)} g_{\mu_l}(x_i, \lambda^{l,(r)})}, \qquad x = 1, \cdots, n;\ l = 1, \cdots, L_r$$

2. Choose a positive value $0 < \gamma < 1$, and check if there is any l such that $\rho_l^{(r+1)} < \gamma$.

   (a) If yes: exclude the components corresponding to $\rho_l^{(r+1)} < \gamma$, update $L_r \to L_{r+1}$ and go back to step 1.

   (b) If no: go to step 3.

3. Classify the data set into $x^1, \cdots x^{L_{r+1}}$ by assigning each x_i to only one mixture component. For each $l = 1, \cdots, L_{r+1}$, update $\lambda^{l,(r)}$ by

$$\lambda^{l,(r+1)} = \arg\max_{\lambda \in \Lambda_{\mu_l}} l_{\mu_l}(x^l, \lambda),$$

   where $l_{\mu_l}(x^l, \cdot)$ is the log-likelihood function for the component l. This optimization step is implemented using our proposed algorithm in Section 5.2.2

Remark 2.1 *Step* 2 *restricts the number of required components for fitting a data set in a way that there is enough information necessary for running inference on each local mixture component. Its value has an influence on the final result of the algorithm in a similar way that an initial value affects the convergence of a general EM algorithm (Table 2.1).*

Example 2.7 (Acidity data) *The data includes acidity index measured in 155 lakes in north-central Wisconsin which is analyzed in Mclachlan and Peel (2000) and the references therein. Using likelihood ratio hypothesis testing, the bootstrap estimated p-value supports two or three components at the* 5% *or* 10% *level of significance, respectively. However, based on a Bayesian method all the values between two and six are equally supported, Richardson and Green, 1997.*

Here we select the grid-points $\mu^{(0)} = (3.6, 4.2, 4.8, 5.4, 6, 6.6, 7)$*, set* $\sigma_l = 0.5$ *and* $\gamma = 0.15$*, so that at least* 20 *observation is assigned to each cluster. The algorithm returns a*

two-component discrete mixture of LMMs with $\hat{\rho} = (0.676, 0.324)$, $\mu = (4.2, 6.6)$, *Figure 2.2 (left panel). The middle panel shows that if we give a set of slightly different initial grid points,* $\mu_6^{(0)} = 6.4$ *instead of* 6.6, *the algorithm returns the same order for the mixture, with* $m = (4.2, 6.4)$ *and* $\hat{\rho} = (0.651, 0.349)$, *(middle panel). In addition, if we let* σ_l*'s to take different values,* $\sigma_6 = 0.6$, *we get the same order with a slightly different fit (right panel).*

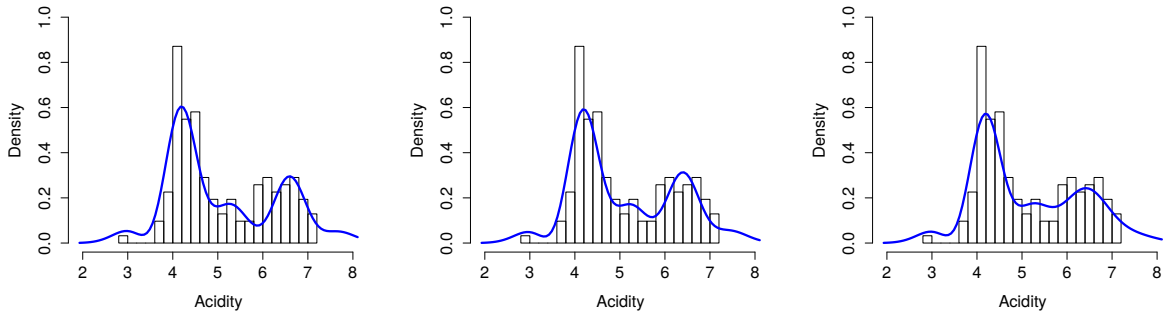

Figure 2.2: Discrete mixture of LMMs for Acidity data.

Further analysis of the data with different values of γ, *shows how the final results of the algorithm depend on* γ, *see Table 2.1.*

γ	μ	$\hat{\rho}$	Order
$0.13, 0.14, 0.15, 0.16, 0.17$	$(4.2, 6.6)$	$(0.67, 0.33)$	2
$0.1, 0.11, 0.12$	$(4.2, 4, 8, 6.6)$	$(0.57, 0.13, 0.3)$	3
$0.07, 0.08, 0.09$	$(4.2, 6, 6.6)$	$(0.63, 0.18, 0.19)$	3
0.06	$(4.2, 4.8, 6, 6.6)$	$(0.57, 0.08, 0.16, 0.19)$	4

Table 2.1: Further analysis for different values of γ

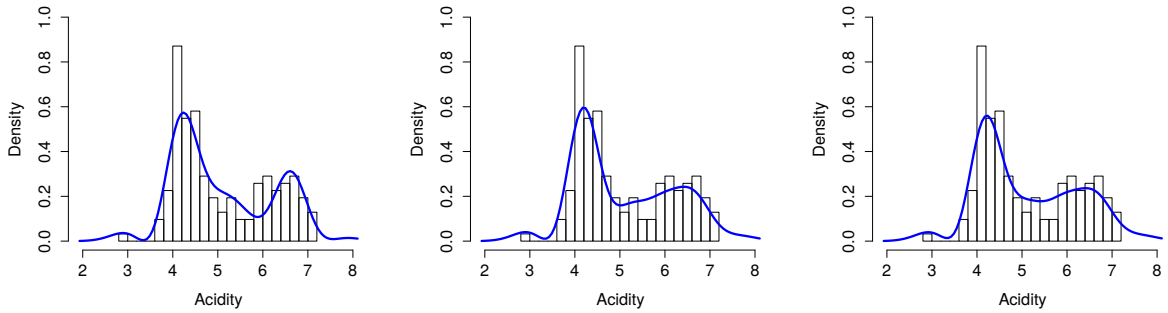

Figure 2.3: Left to right: three and four components fit corresponding to the last three rows in Table 2.1

2.3 Summary and Contributions

While finite mixtures of exponential families are very flexible they suffer from identification problems when support points cluster. This means estimating the order is a very hard problem with a fixed set of data. This chapter takes a new approach to this problem. We use a local mixture model to directly model each cluster in a very flexible way. This results in a finite mixture of LMMs. We propose counting these, now well-defined, components as the 'order' – which will now be statistically meaningful. In each of the component LMMs all the parameters are estimable with efficient algorithms where we have applied a principle that we do not considered models which are unestimable from the data at hand.

2.4 Supplementary materials and proofs

Lemma 2.2 *Let Q_1 and Q_2 be discrete distributions shrinking around a common mean μ,*

$$Q_1(m,\mu) = \sum_{d=1}^{D_1} \pi_d I\{m \geq \mu_d\}, \qquad Q_2(m,\mu) = \sum_{d=1}^{D_2} \pi'_d I\{m \geq \mu'_d\}$$

where I is the indicator function, $\sum_{d=1}^{D_2} \pi'_d \mu_d = \sum_{d=1}^{D_2} \pi_d \mu'_d = \mu$ and $\sum_{d=1}^{D_2} \pi'_d = \sum_{d=1}^{D_2} \pi_d = 1$. Then, if Q_1 and Q_2 share the first k central moments, the finite mixtures

$$g(x;Q_1) = \sum_{d=1}^{D_1} \pi_d f(x;\mu_d) \quad \text{and} \quad g(x;Q_2) = \sum_{d=1}^{D_2} \pi'_d f(x;\mu'_d)$$

have identical LMM approximation of order k.

Proof: Suppose f satisfies the conditions of Taylor's theorem. Then we have,

$$\begin{aligned} g(x;Q_1) &= f(x;\mu) + \sum_{j=1}^{k} \frac{f^{(j)}(x;\mu)}{j!} \left[\sum_{d=1}^{D_1} \pi_d \left(\mu_d - \mu\right)^j \right] + E_1 \\ &= f(x;\mu) + \sum_{j=1}^{k} \frac{f^{(j)}(x;\mu)}{j!} \mu_{Q_1}^{(j)} + E_1 \end{aligned} \tag{2.4.7}$$

and

$$\begin{aligned} g(x;Q_2) &= f(x;\mu) + \sum_{j=1}^{k} \frac{f^{(j)}(x;\mu)}{j!} \left[\sum_{d=1}^{D_2} \pi'_d \left(\mu'_d - \mu\right)^j \right] + E_2 \\ &= f(x;\mu) + \sum_{j=1}^{k} \frac{f^{(j)}(x;\mu)}{j!} \mu_{Q_1}^{(j)\prime} + E_1 \end{aligned} \tag{2.4.8}$$

where E_1 and E_2 are the error terms corresponded to dropping the higher order terms, and $\mu_{Q_1}^{(j)}$ and $\mu_{Q_1}^{(j)\prime}$ are the central moments of order j for Q_1 and Q_2, respectively. Hence, the two approximations coincide if $\mu_{Q_1}^{(j)} = \mu_{Q_1}^{(j)\prime}$ for $j = 2, \ldots, k$.

2.4.1 Orthogonal projection

To see how closely the LMM of density $f(x, \mu)$, at mean parameter value μ_0, approximates the discrete mixture $f_m(x) = (1-\alpha)f(x, \mu_0) + \alpha f(x, \mu)$, we find the orthogonal projection of f_m onto the family of LMMs with respect to Fisher information metric. Note that we can rewrite this mixture model as

$$f_m(x) = f(x, \mu_0) + \alpha[f(x, \mu) - f(x, \mu_0)]$$

where $[f(x, \mu) - f(x, \mu_0)]$ is a straight line in -1-geometry, and is a vector, as it integrates to zero for any fixed μ and μ_0. Considering the LMM of order $k = 4$, the coordinate of the projection at each direction (for $j = 1, 2, 3, 4$) is obtained by

$$\lambda_j^*(\mu, \mu_0) = \int \frac{1}{g_{jj}} (f(x;\mu) - f(x;\mu_0))\, q_j(x, \mu_0)\, dx$$

where $q_j(x,\mu) = \frac{f^{(j)}(x;\mu)}{f(x;\mu)}$, and g_{jj} is the Fisher norm of q_j. For instance, for normal family of models we have

$$\begin{aligned} \lambda_1^*(\mu,\mu_0) = \mu - \mu_0, \quad \lambda_2^*(\mu,\mu_0) &= \frac{1}{2}(\mu-\mu_0)^2, \quad \lambda_3^*(\mu,\mu_0) = \frac{1}{6}(\mu-\mu_0)^3 \\ \lambda_4^*(\mu,\mu_0) &= \frac{1}{24}(\mu-\mu_0)^4 \end{aligned} \tag{2.4.9}$$

and the projected models is

$$g_{\mu,\mu_0}(x) := f(x;\mu_0) + \sum_{j=1}^{4} \lambda_j^*(\mu,\mu_0) f^{(j)}(x;\mu_0) \tag{2.4.10}$$

Example 2.8 *This example shows that, any normal distribution on the line segment* $[\mu_0 \pm 0.6]$ *is well approximated by the LMM in equation (2.4.10). For* $\mu_0 = 0$ *the density, probability distribution function (pdf), and the difference between the two pdfs* $\phi(x,\mu)$ *and* $g_{\mu,\mu_0}(x)$ *are plotted in Figure 2.4.*

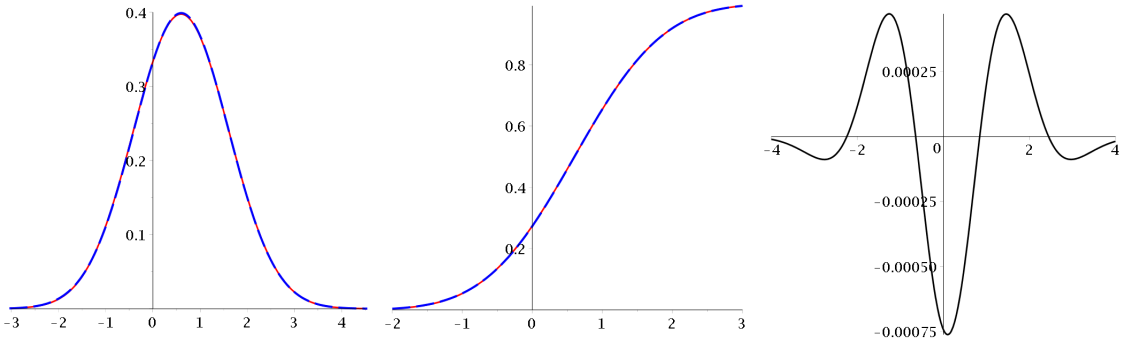

Figure 2.4: The density (left) and distribution function (middle) plots of $\phi(x,\mu,1)$ (blue dash line) and and $g_{\mu,\mu_0}(x)$ (red solid line), for $\mu_0 = 0$ and $\mu = 0.6$. Right panel; the difference between the two distribution functions.

Chapter 3

Computing Boundaries in Local Mixture Models

3.1 Introduction

Often, in statistical inference, the parameter space of a model includes a boundary, which can affect the maximum likelihood estimator (MLE) and its asymptotic properties. Important examples include the (extended) multinomial family, logistic regression models, contingency tables, graphical models and log-linear models, all of which are commonly used in statistical modelling, see Section 1.7 and Anaya-Izquierdo et al. (2013b). In Eriksson et al. (2006), Sontag and Jaakkola (2007) and Rinaldo et al. (2009), it is shown that the MLE in a log-linear model exists, if and only if, the observed sufficient statistic lies in the interior of the marginal polyhedron i.e. away from the boundary. The paper Geyer (2009) studies the influence of the non-existence of the MLE on asymptotic theory, confidence intervals and hypothesis testing for a binomial, a logistic regression model, and a contingency table. Further, in Anaya-Izquierdo et al. (2013b) a diagnostic criterion is provided for the MLE which defines how far it is required to be from the boundary so that first order asymptotics are adequate.

Boundary computation is, in general, a hard problem, Fukuda (2004) and Geyer (2009).

Although it is insufficiently explored in statistics, there are numerous mathematical and computational results in other literatures. Their focus are on (i) approximating a convex closed subspace by a polytope, see Dieker and Vempala (2014), Lopez and Reisner (2008), Boroczky and Fodor (2008) and Barvinok (2013) and (ii) approximating a polytope by a smooth manifold, see Ghomi (2001), Ghomi (2004) and Batyrev (1992). See also Section 3.5.1 for some details.

While the general problem of computing boundaries is difficult, in this chapter we show some new results about computing them for local mixture models (LMM), defined in Section 1.7.4 and studied further in Anaya-Izquierdo and Marriott (2007a). The parameter space of a LMM includes two forms of boundary: the hard and soft. Here we consider a continuous and a discrete LMM: based on the normal and Poisson distributions respectively. We show here that the boundary can have both discrete and smooth aspects, and provide novel geometric methods for computing the boundaries.

Section 2 is a brief review of LMM's and their geometry, while Section 3 introduces some explicit, and new, results on the structure of the fibre of a local mixture in important examples and uses the classical geometric notions of ruled surfaces in the computations. Section 4 concludes with discussion and future directions.

3.2 Local Mixture Models

The theory of local mixture models is motivated by a number of different statistical modeling situations which share a common structure. Suppose that there is a baseline statistical model which describes the majority of the observed variation, but there remains appreciable residual variation that is not consistent with the baseline model. These situations include over-dispersion in binomial and Poisson regression models, frailty analysis in lifetime data analysis Anaya-Izquierdo and Marriott (2007b) and measurement errors in covariates in regression models Marriott (2003). Other applications include local influence analysis Critchley and Marriott (2004) and the analysis of predictive distributions Marriott (2002).

The geometric complexity of the space of general mixture models means that undertaking inference in this class is a hard problem. It has issues of identification, singularity

and multi-modality in the likelihood function, interpretability problems and non-standard asymptotic expansions.

The key identification and multi-modality problem comes from the general observation that if a set of densities $f(x;\theta)$ lies uniformly close to a low-dimensional -1-affine space – as defined by Amari (1990)– then all mixtures of that model would also lie close to that space. Hence the space of mixtures is much lower dimensional than might be expected. The local mixture model is designed to have the 'correct' dimension by restricting the class of mixing distributions to so-called localizing distributions. This allows a much more tractable geometry and corresponding inference theory. The restriction often comes only at a small cost in modeling terms. The local mixture model is, in geometric terms, closely related to a fibre-bundle over the baseline model, and has the elegant information geometric properties, described formally in Theorem 3.1, that (i) inference on the 'interest parameters' of the baseline model only weakly depends on the values of the nuisance parameters of the fibres because of orthogonality, (ii) the log-likelihood on the fibre has only a single mode due to convexity (iii) the local mixture model is a higher order approximation to the actual mixture.

As defined in Section 1.7.4 a LMM is a union of -1-convex subsets of -1-affine subspaces of the set of densities, in the information geometry of Amari, Amari (1990). Here -1 refers to the $\alpha = -1$ or mixture connection.

Definition 3.1 *Let S be a common sample space. The local mixture, of order k, of a regular exponential family $f(x;\mu)$ in its mean parameterization, μ, is defined as*

$$g(x;\lambda,\mu) = f(x;\mu) + \lambda_2\, f^{(2)}(x;\mu) + \cdots + \lambda_k\, f^{(k)}(x;\mu), \qquad \lambda \in \Lambda_\mu \subset \mathbb{R}^{k-1} \tag{3.2.1}$$

where $\lambda = (\lambda_2, \cdots, \lambda_k)$ and $f^{(j)}(x;\mu) = \frac{\partial^j f}{\partial \mu^j}(x;\mu)$. Also $q_j(x;\mu) := \frac{f^{(j)}(x;\mu)}{f(x;\mu)}$, then for any fixed μ,

$$\Lambda_\mu = \left\{ \lambda | 1 + \sum\nolimits_{j=2}^{k} \lambda_j\, q_j(x;\mu) \geq 0, \forall x \in S \right\},$$

is a convex subspace obtained by intersection of half-spaces. Its boundary is called the hard boundary *and corresponds to a positivity condition on $g(x;\lambda,\mu)$.*

A local mixture model has a structure similar to that of a fibre bundle and for each fixed μ_0 the subfamily, $g(x;\lambda,\mu_0)$, is called a fibre – although more strictly it is a convex subset

of the full fibre. The paper Anaya-Izquierdo and Marriott (2007a) shows that LMMs have the following excellent statistical properties.

Theorem 3.1 *(i) The set* $\{g(x;\lambda,\mu_0) - f(x;\mu_0)\}$ *is* -1*-flat and Fisher orthogonal to the score of* $f(x;\mu)$ *at* μ_0*. Thus* μ *and* λ *are orthogonal parameters.*
(ii) On each fibre the log-likelihood function is concave - though not necessarily strictly concave.
(iii) A continuous mixture model $\int f(x;\mu)\,dQ(\mu)$ *can be approximated by a LMM to an arbitrary order if* Q *satisfies the properties of a localizing distribution defined in Marriott (2002).*

In such an approximation the parameter vector λ represents the mixing distribution Q through its moments; however, for some values of λ a LMM can have moments not attainable by a mixture model of the form $\int f(x;\mu)\,dQ(\mu)$. A true LMM, defined in Anaya-Izquierdo and Marriott (2007a), is a LMM which behaves similarly to a mixture model, in terms of a finite set of moments. For a true LMM, additional to hard boundary, there is another type of restricting boundary, called soft boundary, and characterized by following definition.

Definition 3.2 *For a density function* $f(x;\mu)$ *with* k *finite moments let,*

$$\mathcal{M}_k(f) := (E_f(X), E_f(X^2), \cdots, E_f(X^k)).$$

Then $g(x;\mu,\lambda)$*, defined in Definition 3.1, is called a* true local mixture*, if and only if, for each* μ *in a compact subset* I*,* $\mathcal{M}_k(g)$ *lies inside the convex hull of* $\{\mathcal{M}_k(f)|\mu \in I\}$*. The boundary of the convex hull is called the* soft boundary.

Inferentially Model (3.2.1) might be used for marginal inference about μ where λ is treated as a nuisance parameter in, for example, random effect or frailty models, see Marriott (2002) and Chapter 5. The properties of Theorem 3.1 on the (μ,λ)-parameterization guarantees asymptotic independence of $\hat{\mu}$ and $\hat{\lambda}$ and simplifies determination of $(\hat{\mu},\hat{\lambda})$, Cox and Reid (1987). Therefore, the profile likelihood method would be expected to be a promising approach for marginal inference about μ when λ is away from boundaries. This

intuition is confirmed by simulation exercises, see Section 3.5.2. To use such an approach in general it is necessary that the analyst can compute the inferential effect of the boundary. The rest of this chapter explores the geometric structure of the boundaries of LMMs and the computational consequences of such a structure.

3.3 Computing the boundaries

In this section we compute the hard and soft boundaries for LMMs of order $k = 4$, as lower order LMMs have trivial boundaries and typically LMMs with $k > 4$ do not add greatly to modelling performance, see Marriott (2006).

3.3.1 Hard Boundary for the LMM of Poisson distribution

Consider the following LMM of the Poisson probability mass function $p(x; \mu)$,

$$g(x; \mu, \lambda) = p(x; \mu) + \sum\nolimits_{j=2}^{4} \lambda_j \, p^{(j)}(x; \mu), \qquad \lambda \in \Lambda_\mu \subset \mathbb{R}^3. \tag{3.3.2}$$

It is straightforward to show that

$$E_g(X) = E_p(X) = \mu, \quad Var_g(X) = Var_p(X) + 2\,\lambda_2, \tag{3.3.3}$$

illustrating that the λ parametrization of LMMs is tractable and intuitive as the model in Equation (3.3.2) produces higher (lower) dispersion compared to $p(x; \mu)$. Furthermore, as shown in Anaya-Izquierdo and Marriott (2007a), the other parameters also have interpretable moment based meanings.

For model (3.3.2), the hard boundary is obtained by analyzing half spaces defined, for fixed μ, by

$$S_x = \{\lambda | A_2(x)\,\lambda_2 + A_3(x)\lambda_3 + A_4(x)\,\lambda_4 + 1 \geq 0\}\,, \;\; x \in \mathbb{Z}^+ \tag{3.3.4}$$

where $A_j(x)$'s are polynomials of x defined by Definition 3.1. The space Λ_μ, for fixed μ, will be the countable intersection of such half spaces over $x \in \{0, 1, \dots\}$ i.e., we can write $\Lambda_\mu = \bigcap_{x \in \mathbb{Z}^+} S_x$. In fact, as we show in Proposition 3.1, the space can be arbitrarily well approximated by a polytope.

Proposition 3.1 *For a LMM of a Poisson distribution, for each μ, the space Λ_μ can be arbitrarily well approximated, as measured by volume for example, by a finite polytope.*

Proof: See Section 3.5.3.

Figure 3.1 shows some issues related to this proposition. It shows two slices through the space Λ_μ by fixing a value of λ_2 (left panel) and λ_3 (right panel). The shaded polytope is a subset of Λ_μ in both cases. The lines are sets of the form

$$A_2(x)\,\lambda_2 + A_3(x)\lambda_3 + A_4(x)\,\lambda_4 + 1 = 0,$$

for different values of $x \in \{0, 1, 2, \cdots\}$, with solid lines being support lines and dashed lines representing redundant constraints. In $\mathbb{R}^3$ a finite number of such planes will define a polytope which is arbitrarily close to Λ_μ. A second feature, which is clear from Fig. 3.1, is that parts of the boundary look like they can be well approximated by a smooth curve (Section 3.5.1), which has the potential to simplify computational aspects of the problem.

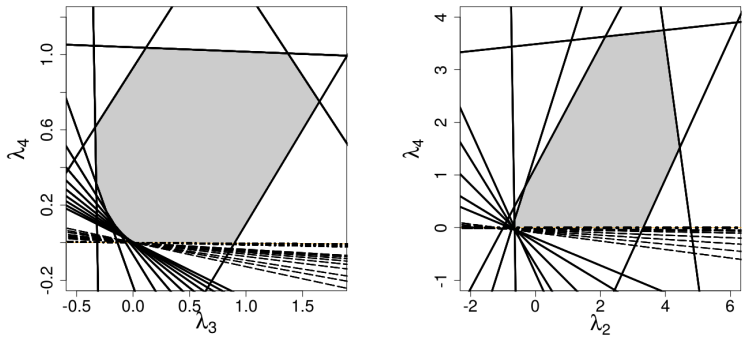

Figure 3.1: Left: slice through $\lambda_2 = -0.1$; Right: slice through $\lambda_3 = 0.3$. Solid lines represent active and dashed lines redundant constraints. For our model $\lambda_4 > 0$ is a necessary condition for positivity.

3.3.2 Hard Boundary for the LMM of Normal

In the previous example the boundary was defined by a countable intersection of half-spaces. Now we look at an example, the LMM of normal distributions, where we have an uncountably infinite intersection of half spaces and we observe a smooth, manifold like boundary. To compute this boundary we need the geometry of ruled and developable surfaces in Section 1.5; specifically, the one of the envelope of a infinite family of one-parameter planes, Definition 1.8.

Consider the LMM of a normal distribution $N(\mu, \sigma)$, where $\sigma > 0$ is fixed and known, and for which, without loss of generality, we assume $\sigma = 1$. Let $y = x - \mu$, then

$$\Lambda_\mu = \{\lambda \,|(y^2 - 1)\lambda_2 + (y^3 - 3y)\lambda_3 + (y^4 - 6y^2 + 3)\lambda_4 + 1 \geq 0, \forall y \in \mathbb{R}\} \tag{3.3.5}$$

is the intersection of infinite set of half-spaces in $\mathbb{R}^3$.

To understand the boundary of Λ_μ we first solve the equations in (1.5.2) to obtain the characteristic lines and consequently the envelope of a one parameter set of planes in $\mathbb{R}^3$. These planes, in λ-space, are parameterized by $y \in \mathbb{R}$, and are the solutions of

$$(y^2 - 1)\lambda_2 + (y^3 - 3y)\lambda_3 + (y^4 - 6y^2 + 3)\lambda_4 + 1 \ = 0.$$

The envelope of this family forms a ruled surface, and can be thought of as a self-intersecting surface in $\mathbb{R}^3$. The surface partitions $\mathbb{R}^3$ into disconnected regions and one of these – the one containing the origin $(0, 0, 0)$ – is the set Λ_μ. Figure 3.2 shows the self-intersecting surface and the shaded region is the subset which is the boundary of Λ_μ.

While the boundary of Λ_μ will have large regions which are smooth, it also has singular lines and points. These are the self-intersection points of the envelope and it is at these points where the boundary fails to be an embedded manifold, but is still locally smooth. The general structure of the boundary is a non-smooth union of a finite number of smooth components.

3.3.3 Soft Boundary calculations

The previous section looks at issues associated with the hard boundary calculations for LMMs. In this section we look at similar issues connected with computing soft boundaries,

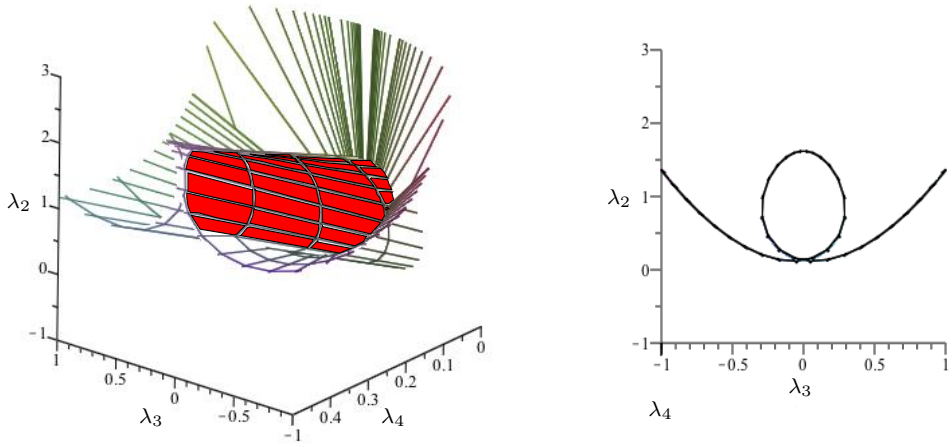

Figure 3.2: Left: The hard boundary for the normal LMM (shaded) as a subset of a self intersecting ruled surface (unshaded); Right: slice through $\lambda_4 = 0.2$.

Definition 3.2, in moment spaces for true LMMs.

For visualization purposes, consider $k = 3$ and we use the normal example from the previous section. The moment maps are given by

$$\begin{aligned} \mathcal{M}_3(f) &= (\mu, \mu^2 + \sigma^2, \mu^3 + 3\mu\sigma^2), \\ \mathcal{M}_3(g) &= (\mu, \mu^2 + \sigma^2 + 2\lambda_2, \mu^3 + 3\mu\sigma^2 + 6\mu\lambda_2 + 6\lambda_3). \end{aligned}$$

Suppose $I = [a, b]$, then $\mathcal{M}_3(f)$ defines a smooth space curve, $\varphi : [a, b] \to \mathbb{R}^3$. To construct the convex hull, denoted by $convh\{\mathcal{M}_3(f),\ \mu \in [a, b]\}$, all the lines between $\varphi(a)$ and $\varphi(\mu)$ and all the lines between $\varphi(\mu)$ and $\varphi(b)$, for $\mu \in [a, b]$, are required. Each of the two families of lines are attached to the curve and construct a surface in $\mathbb{R}^3$. Hence, we have two surfaces each formed by a smooth curve and a set of straight lines (Figure 3.3, right). Thus we have the following two ruled surfaces,

$$\begin{cases} \gamma_a(\mu, u) = \varphi(\mu) + u\, L_a(\mu), & \text{surface } a, \\ \gamma_b(\mu, u) = \varphi(\mu) + u\, L_b(\mu), & \text{surface } b, \end{cases}$$

where $u \in [0,1]$, and for each $\mu \in [a,b]$, $L_a(\mu)$ is the line connecting $\varphi(\mu)$ to $\varphi(a)$, and similarly $L_b(\mu)$ is the line between $\varphi(b)$ and $\varphi(\mu)$.

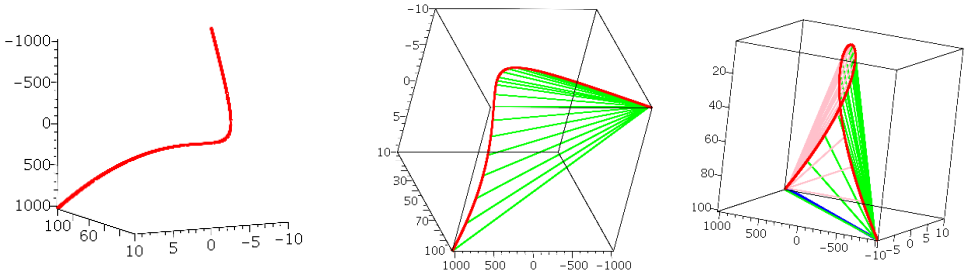

Figure 3.3: Left: the 3-D curve $\varphi(\mu)$; Middle: the bounding ruled surface $\gamma_a(\mu, u)$; Right: the convex subspace restricted to soft boundary.

The soft boundary of the Poisson model can be characterized similarly.

3.4 Summary and Contributions

This chapter gives an introduction to some of the issues associated with computing the boundaries of local mixture models. Understanding these boundaries is important if we want to exploit the nice statistical properties of LMM, given by Theorem 3.1. The 'cost' associated with these properties is that boundaries will potentially play a role in inference giving, typically, non-standard results. The boundaries described in this chapter have both discrete aspects, (i.e. the ability to be approximated by polytopes), and smooth aspects (i.e. regions where the boundaries are exactly or approximately smooth). We exploit the geometric properties of the parameter space to compute them for two working examples, normal and Poisson, by geometric objects such as polytopes, ruled and developable surfaces. Section 3.5.4 presents parallel analyses revealing similar structure for LMM of Binomial and exponential models to the working examples. It is an interesting and important open research area to develop computational information geometric tools which can efficiently deal with such geometric objects.

3.5 Supplementary materials and proofs

3.5.1 Approximating general boundaries

Smooth Convex Body

Consider the full dimensional polytope

$$P = \left\{ p \in \mathbb{R}^d \;\middle|\; \langle p, a^{(i)} \rangle \leq b_i, \;\; i = 1, \cdots, n \right\}. \tag{3.5.6}$$

If $0 \in P$, then $b_i \neq 0$; hence we can normalize the inequalities and

$$P = \left\{ p \in \mathbb{R}^d \;\middle|\; \langle p, \alpha^{(i)} \rangle \leq 1, \;\; i = 1, \cdots, n \right\}$$

where $\alpha^{(i)} = \frac{a^{(i)}}{b_i}$. Now let

$$F(p) = \frac{1}{2} \log \left(\sum_{i=1}^{n} e^{2\langle p, \alpha^{(i)} \rangle} \right)$$

and for a positive real number t define

$$Q_t = \{ p \in \mathbb{R}^d \;\big|\; F(tp) \leq t \}.$$

Finally, It can be shown that

1. $Q_t \subset P$ for all $t > 0$.
2. Q_t is a convex body with a smooth boundary.
3. $\lim_{t \to \infty} Q_t = P$.

(See Bonnesen and Fenchel (1987), Batyrev (1992) and Ghomi (2004))

Example 3.1 *Consider a full dimensional polygon with following vertices*

$$\alpha^{(1)} = (1, 1) \;,\; \alpha^{(2)} = (-1, 1) \;,\; \alpha^{(3)} = (-1, -1) \;,\; \alpha^{(4)} = (1, -1).$$

Then the smooth approximation of the boundary is obtained by following the mapping, Figure 3.4

$$F : \mathbb{R}^2 \to \mathbb{R} \qquad F(tp) = \frac{1}{2} \log \left(\sum_{i=1}^{4} e^{2\langle tp, \alpha^{(i)} \rangle} \right)$$

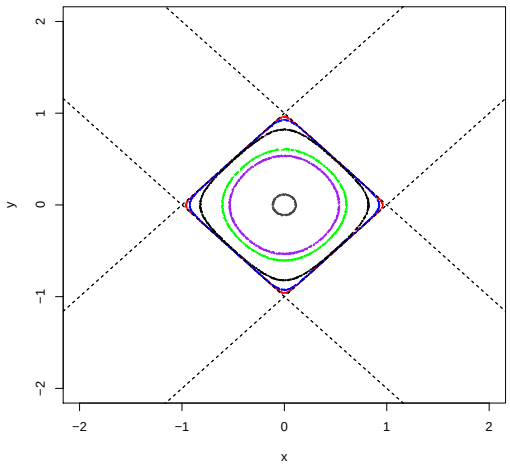

Figure 3.4: $t = 10, 5, 2, 1, 0.9, 0.7$ presented by red, blue, black, green, purple, gray.

Approximate by a Polytope

Another way of approximating boundaries is by convex polytopes constructed by the convex hull of randomly selected boundary points. There are various algorithms for sampling from a boundary. For instance, one way of taking a sample from a smooth surface, as the boundary of a convex body K, according to its Gauss Curvature, is to take a uniform sample form the surface of a sphere $\mathcal{S} \in K$, then blow them off from the origin. For different methods of uniform sampling from a d-dimensional sphere see Marsaglia (1972).

Alternatively, Narayanan and Niyogi (2008) provide and algorithm which generates an "approximately" uniform sample form boundary of $K \subset \mathbb{R}^d$ (∂K for short) given a uniform sample in K. Suppose, $p = (p_1, \cdots, p_n)$ is a uniform sample of K and choose $\epsilon > 0$, then

1. Estimate, with confidence $1-\epsilon$, the smallest eigenvalue τ of $A(K) = E[(p-\bar{p})(p-\bar{p})^T]$.
2. Set $\sqrt{t} = \frac{\epsilon\sqrt{\tau}}{32d}$.
3. (a) Generate a p uniformly from K.
   (b) Generate $q \sim Gaussian(p, 2tI)$.

(c) Let l be the segment between p and q.

(d) If $q \notin K$, output is $l \cap \partial K$, else output $\emptyset$.

4. If the output is $\emptyset$, go to 3. Else accept it.

3.5.2 Profile Likelihood Simulation

Figure 3.5 reveals the behavior of the profile likelihood function of a LMM. In the interior of the parameter space restricted by hard boundary, left panel, it seams to hold the normality property, while close to the hard boundary this property breaks down (right panel). In each example data are generated from a discrete mixture of normal densities.

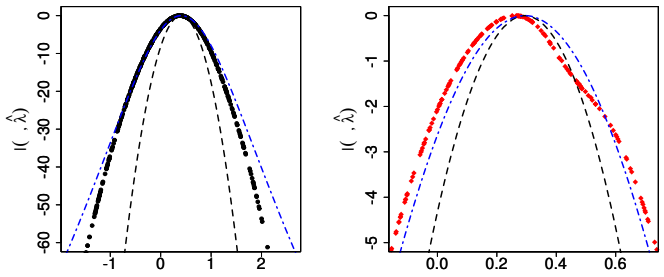

Figure 3.5: dashed (black) line presents $N(\mu, 1)$ and dash-dot (blue) the likelihood for $N(\mu, \hat{\sigma})$. For the panel on left $\hat{\lambda}$ is a interior point estimate, while in the panel on right it is boundary point estimate.

3.5.3 Proofs

Proposition 3.1 For each μ, the space Λ_μ is obtained $\Lambda_\mu = \bigcap_{x \in \mathbb{Z}^+} S_x$ (equation 3.3.4), where $A_j(x)$ is a polynomial of $x \in \aleph \cup \{0\}$, and $A(x) = (A_2(x), A_3(x), A_4(x))$ is the normal vector of the boundary plane of S_x, say H_x. Since the outward normal vector, $-A(x)$, for all $x \in \aleph \cup \{0\}$ do not point into a single half-space then the convex space Λ_μ is a bounded polytope (Alexandrov (2005), p.20). Now suppose $a_j \leq \lambda_j \leq b_j$ for real values $a_j < b_j$ and

$j = 1, 2, 3$. Suppose we approximate Λ_μ by intersection of a finite number of half-spaces

$$\Lambda_\mu \approx \bigcap_{x=0}^{x0} S_x \bigcap S_\infty$$

where $S_\infty = \{\lambda \mid \lambda_4 \geq 0\}$. That is, for some x_0 large enough all H_x's for $x > x_0$ are omitted form the intersection, eccept for the limiting suport plane. For a fixed value of λ_3 the difference between the area of the corresponding polygon (2-polytope) and the area of the approximated polytope is bounded by $(b_2 - a_2)h(\lambda_3, x_0, a_2)$ which is the area of a rectangle at the bottom of the polygon (Figure 3.6, left) and

$$h(\lambda_3, x_0, a_2) = -\frac{1}{A_4(x_0)}\{1 + A_3(x_0)\lambda_3 + A_2(x_0)a_2\}.$$

To obtain the difference in the volume we need to integrate this area over all values of λ_3; that is,

$$\begin{aligned} V_d(x_0) &= \int_{a_3}^{b_3} (b_2 - a_2)\{h(\lambda_3, x_0, a_2)\}\, d\lambda_3 \\ &= \frac{(b_2 - a_2)}{A_4(x_0)}\{(b_3 - a_3)(1 + A_2(x_0)a_2) + A_3(x_0)(b_3^2 - a_3^2)\} \end{aligned} \tag{3.5.7}$$

Since all a_j and b_j's are real values, $\frac{A_2(x_0)}{A_4(x_0)} \to 0$ and $\frac{A_3(x_0)}{A_4(x_0)} \to 0$ as x_0 increases, then $V_d(x_0)$ can get arbitrary small.

3.5.4 More on Hard Boundaries

Binomial Model

Consider a LMM of binomial distribution $Bin(n, \mu)$ with its mean parametrization. The sample space is $\{0, 1 \cdots, n\}$, hence Λ_μ is the intersection of $n + 1$ half-spaces determined by non-negative integers not larger than n. Figure 3.7 presents the 2-dimensional slices of Λ_μ through $\lambda_2 = -0.1$ for LMMs of $Bin(100, 3.5)$ (right) and $Bin(10, 3.5)$ (left), where each line represents a boundary plane projected onto $\lambda_2 = -0.1$. Figure 3.7 illustrates similar structure as of that for the LMM of a Poisson distribution.

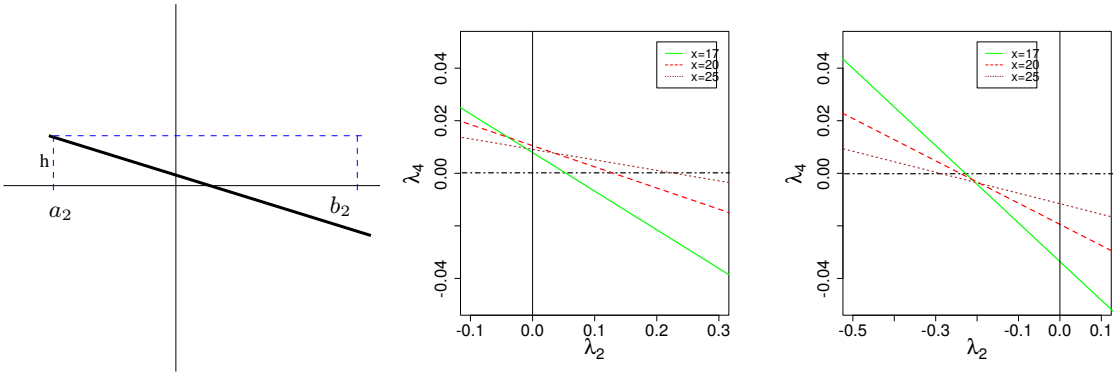

Figure 3.6: Left panel: gives a schematic graph for the proof. Middle and right panel present actual lines representing the boundary planes for a fixed $\lambda_3 < 0$ and a fixed $\lambda_3 > 0$, respectively.

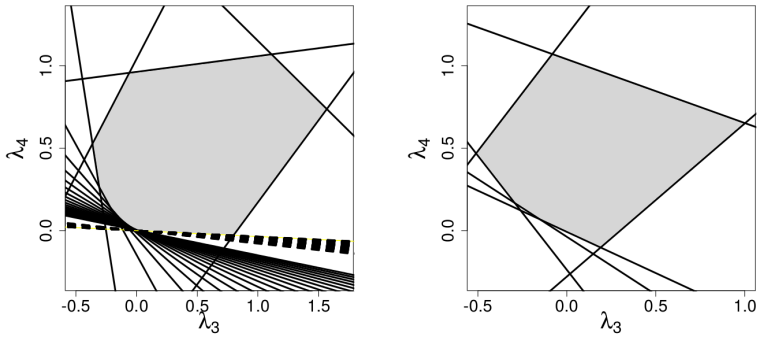

Figure 3.7: 2-dimensional slices of Λ_μ through $\lambda_2 = -0.1$. Left: LMM of $Bin(100, 3.5)$, the dashed lines represent the planes for $x \geq 50$. Right: $Bin(10, 3.5)$, and the planes for $x = 1, 2, 3, 4, 5$ are redundant in this slice.

Normal with Unknown σ

Figure 3.8 presents the boundary, plotted by characteristic lines, from three different angles. Black lines represent the characteristic lines and the active boundary, painted in red, presents the hard boundary. Also, the 2-dimensional intersections of the boundary are plotted in Figures 3.9 and 3.9.

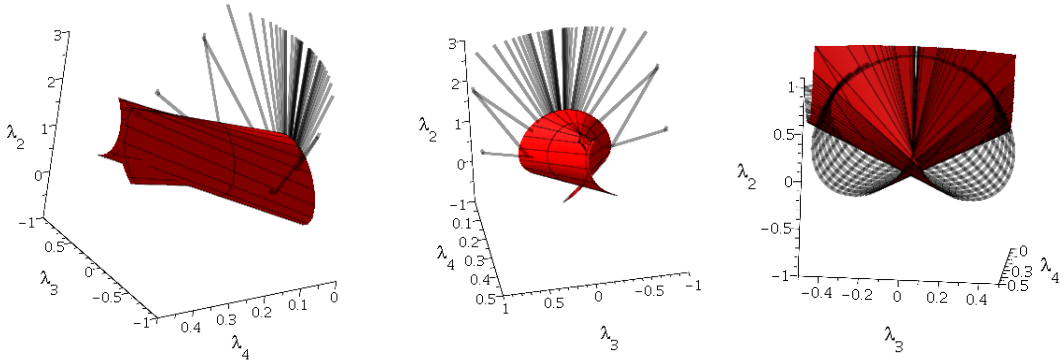

Figure 3.8: Different angles of Λ_μ

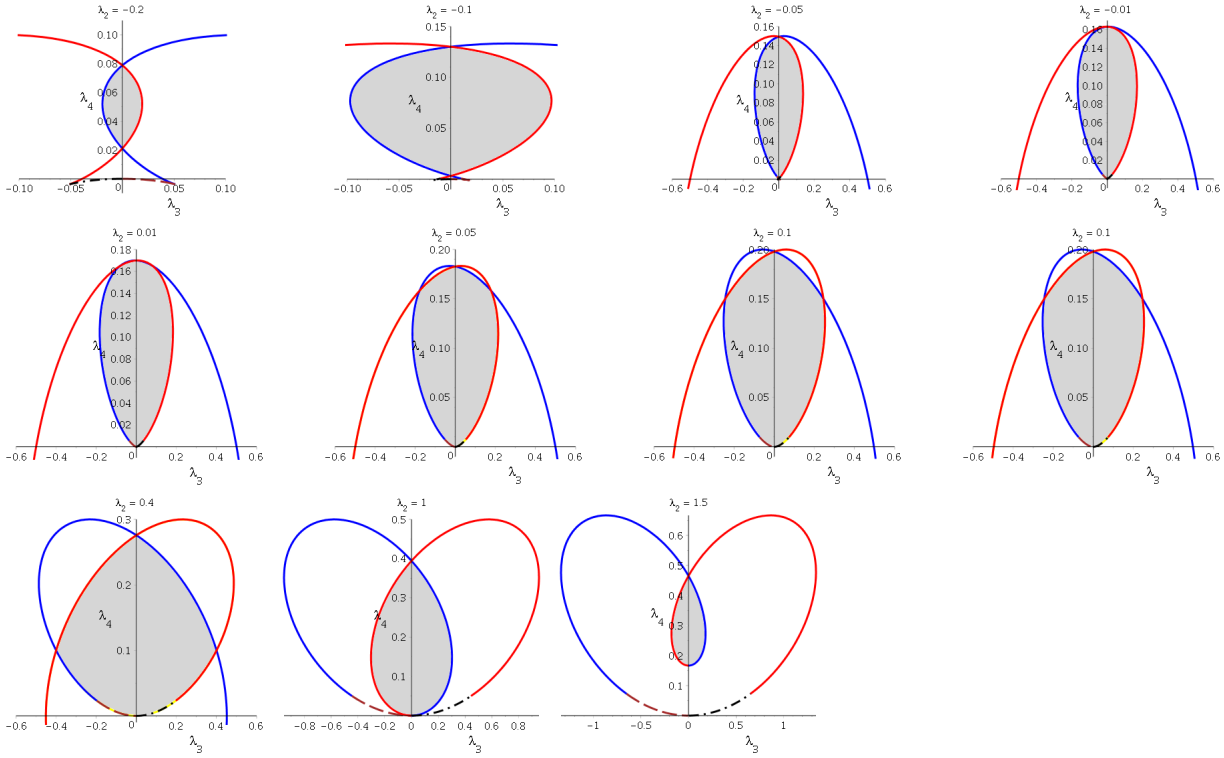

Figure 3.9: Different slices through λ_2

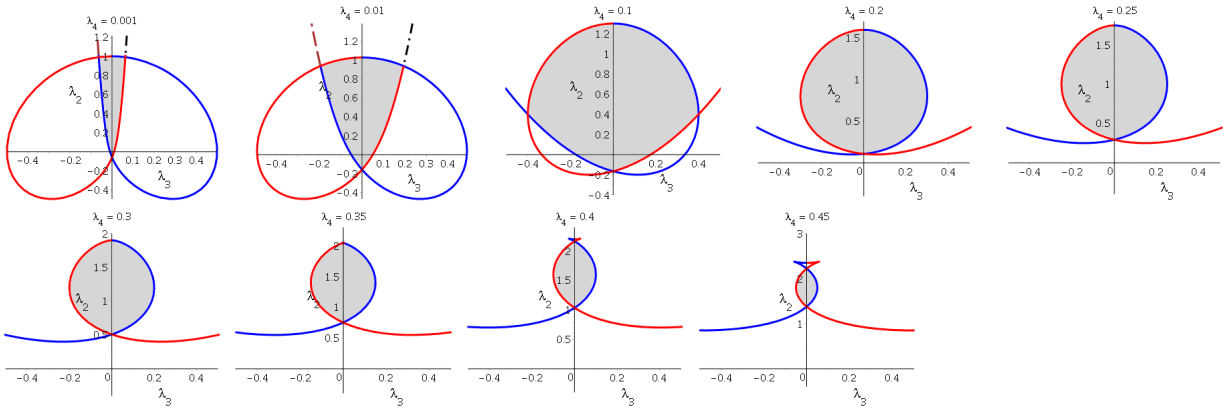

Figure 3.10: Different slices through λ_4

Geometry of Λ_μ of LMM of a normal distribution with unknown σ is also of interest, as the boundary planes clearly depend on the value of σ. Figure 3.11 illustrates that changing value of σ does not alter the geometry of the Λ_μ, but just acts as a scaler and inflate the subspace.

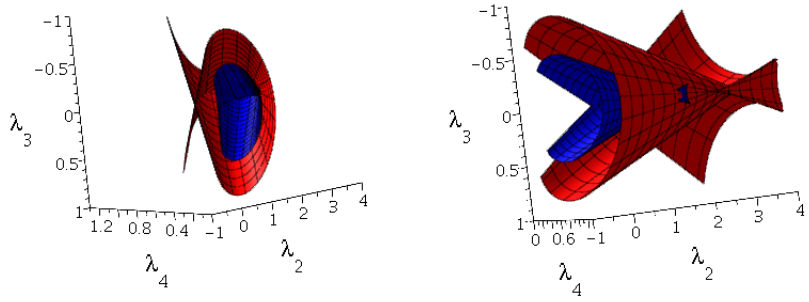

Figure 3.11: The hard boundary for $\sigma = 1$ (blue) vs $\sigma = 1.2$ (red)

Conjecture 3.1 *Suppose Λ_μ^σ is the parameter subspace of the LMM of the normal model $N(\mu, \sigma^2)$. If $\sigma_1 < \sigma_2$, then $\Lambda_\mu^{\sigma_1} \subset \Lambda_\mu^{\sigma_2}$.*

Since normal distribution with unknown σ has two parameters, its LMM is an over-parametrized model, which causes an identifiability problem. To see explicitly, consider a LMM of a normal model with $\sigma = 1$ and an unmixed normal distribution with unknown σ. For the LMM of $N(\mu, 1)$ we have

$$\begin{aligned} E\left[(X-\mu)^2\right] &= 1 + 2\lambda_2 \\ E\left[(X-\mu)^3\right] &= 6\lambda_3 \\ E\left[(X-\mu)^4\right] &= 12\lambda_2 + 24\lambda_4 + 3 \end{aligned} \tag{3.5.8}$$

the same central moments for the unmixed model $N(\mu, \sigma^2)$, unknown σ, are

$$E\left[(X-\mu)^2\right] = \sigma^2,\ E\left[(X-\mu)^3\right] = 0,\ E\left[(X-\mu)^4\right] = 3\sigma^4 \tag{3.5.9}$$

equating these two sets of central moments we get a curve as a function of σ^2 inside Λ_μ

$$\left[\lambda_2(\sigma^2) = \frac{\sigma^2 - 1}{2}, \lambda_3(\sigma^2) = 0, \lambda_4(\sigma^2) = 3\sigma^4 - 6\sigma^2 + 3\right]$$

That is up to forth moments we can represent the $N(\mu, \sigma^2)$ by the LMM of $N(\mu, 1)$ with characterized values of λ_j's.

Exponential Model

Similar to the LMM of a normal distribution, the boundary of Λ_μ for an exponential distribution is obtained by intersection of infinite number of half-spaces and can be constructed by their envelope. Illustrated in Figure 3.12, the boundary of Λ_μ is a locally smooth self-intersecting surface with singularity curves. The 2-dimensional slices of Λ_μ in Figure 3.13 illustrates the position of the singularities as points of the singularity curves. Inspection shows that these points also correspond to quartics with two double roots, as in the normal models.

3.5.5 Surface Parametrization and Optimization

Section 3.3.2 and 3.3.3 characterize the hard and soft boundaries of the LMM of a normal distribution as ruled surfaces. Specifically, any point $\lambda = (\lambda_2, \lambda_3, \lambda_4)$ restricted to these

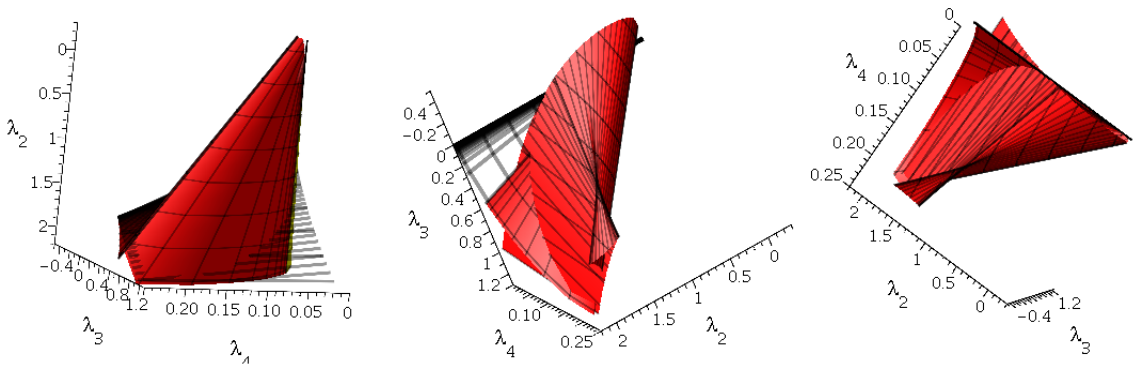

Figure 3.12: Different angles of Λ_μ for $f(x,\mu) = \mu\, e^{-\mu x}$

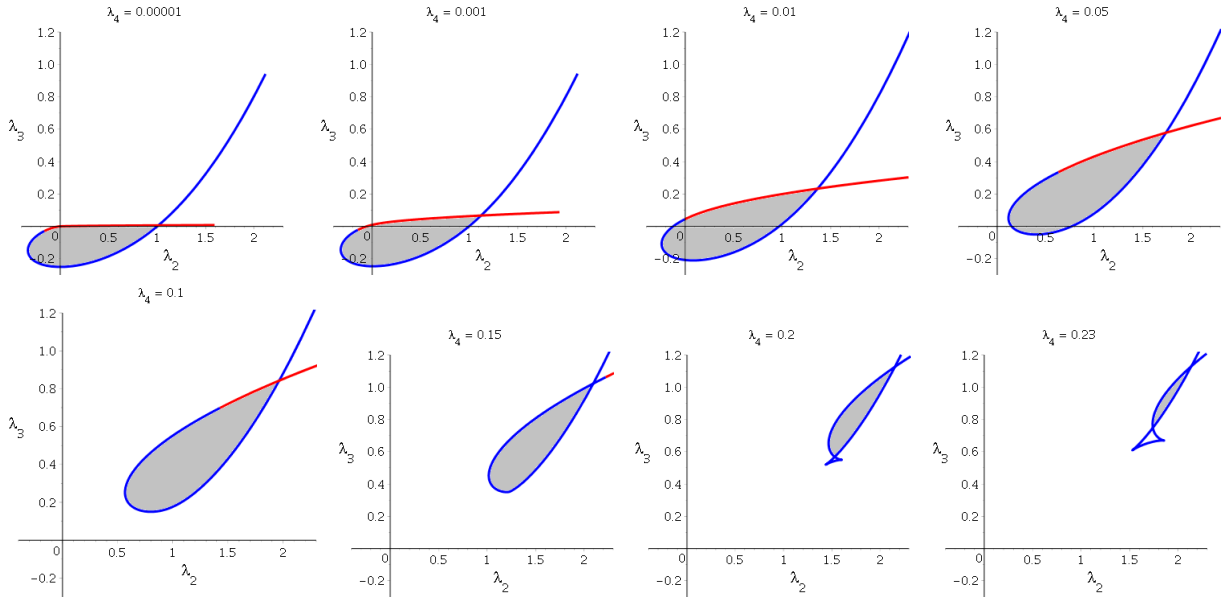

Figure 3.13: Different slices through λ_4, for $f(x,\mu) = \mu\, e^{-\mu x}$

surfaces is uniquely determined by a pair (x,u). In this section, we exploit this parameterization for finding the maximum likelihood estimate of λ on these boundaries. Let $\mathcal{D}$ be the envelope of the family of planes in Equation (3.3.5). For a fixed μ, the loglikelihood

function restricted to $\mathcal{D}$ is a function of (x, u) as follows

$$l_{\mathcal{D}}(x,u) = \sum_{i=1}^{n} \log\Big\{ f(x;\mu) + \sum_{j=2}^{4} \lambda_j(x,u)\, f^{(j)}(x_i;\mu) \Big\}. \tag{3.5.10}$$

However, a searching algorithm using the basis of coordinates (x, u) may not retain efficient steps, since they are not orthogonal. Hence, we need to find an alternative parameterization such that the corresponding basis are orthogonal. That is, instead of (x, u) parametrization, any point on the $\mathcal{D}$ is characterized by (x_1, x_0) implying orthogonal directions.

Orthogonal Parametrization

For finding an orthogonal parameterization on a ruled surface, we directly use the geometry of smooth space curves and their related geometric concepts such as, normal vectors, binormal vectors, tangent planes and normal planes (Struik, 1988, ch.1).

At any regular point $\lambda(x_1)$ on $\mathcal{C}(x)$ the normal plane π_{x_1} is orthogonal to $\mathcal{C}(x)$, and $t(x_1)$ is the normal vector to π_{x_1}. Also π_{x_1} is spanned by the normal vector $n(x_1)$ and binormal vectors $b(x_1)$ of $\mathcal{C}(x)$. In addition, for the normalized vectors $t(x_1)$, $n(x_1)$ and $b(x_1)$, we have the following equations,

$$t(x_1)\cdot b(x_1) = 0, \quad n(x_1)\cdot b(x_1) = 0, \quad b(x_1)\cdot b(x_1) = 1, \tag{3.5.11}$$

where "$\cdot$" represents the regular inner product in Euclidean space. $t(x_1)$ and $n(x_1)$ are obtained from the first and second derivatives of $\mathcal{C}(x)$ at x_1, and , $b(x_1)$ can be calculated from the equations in (3.5.11).

To obtain an orthogonal parametrization on $\mathcal{D}$, we characterize each point on $\mathcal{D}$ using the normal plane π_{x_1} to $\mathcal{C}$ at a suitable point, say $\lambda(x_1)$. In other words, for a $\lambda \in \mathcal{D}$, represented by (x_0, u_0), we want to find (x_1, v, w), where v and w are fully determined by x_0 and x_1 (see Figure 3.14). For any $\lambda \in \mathcal{D}$, there is a normal plane π_{x_1} which intersects $\mathcal{C}(x)$ at $\lambda(x_1)$ and includes λ, and can be represented by both implicit and explicit equations as follows,

$$\begin{aligned} \pi_{x_1} &= \Big\{ \lambda \in \mathbb{R}^3 \;\Big|\; \lambda = \lambda(x_1) + v\, n(x_1) + w\, b(x_1),\;\; v, w \in \mathbb{R} \Big\} \\ \pi_{x_1} &= \Big\{ \lambda \in \mathbb{R}^3 \;\Big|\; (\lambda - \lambda(x_1))\cdot t(x_1) = 0 \Big\}. \end{aligned} \tag{3.5.12}$$

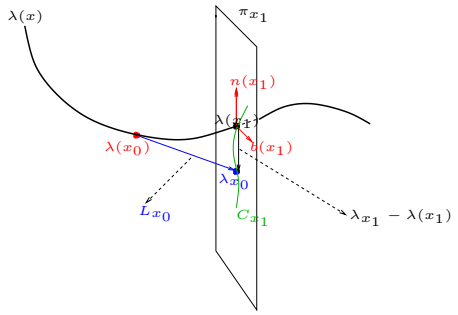

Figure 3.14: A schematic graph for visualizing the orthogonal reparametrization

The intersection of π_{x_1} and $\mathcal{D}$ is a plane curve on π_{x_1}, say C_{x_1}, such that it can be also spanned by the orthonormal bases $n(x_1)$ and $b(x_1)$, considering $\lambda(x_1)$ as the origin of π_{x_1}. On the other hand, according to ruled surface parameterization, λ is represented by a pair (x_0, u_0) as follows

$$\lambda = \lambda(x_0) + u_0\, t(x_0), \tag{3.5.13}$$

which is a point on the line

$$L_{x_0} = \lambda(x_0) + u\, t(x_0). \tag{3.5.14}$$

Therefore, λ can be determined as the intersecting point of π_{x_1} and L_{x_0}; that is, we need to find u_0 such that λ satisfies equation (3.5.13). For a fixed x_1, if the equation in (3.5.13) is substituted in the second equation in (3.5.12), after some algebra, u_0 is obtained as a function of x_0 as follows,

$$u(x_0) = \frac{(\lambda(x_1) - \lambda(x_0)) \cdot t(x_1)}{t(x_0) \cdot t(x_1)}. \tag{3.5.15}$$

Hence, λ can be written as

$$\lambda_{x_0} = \lambda(x_0) + u(x_0)\, t(x_0), \tag{3.5.16}$$

which is also spanned by $n(x_1)$ and $b(x_1)$, thus we can write

$$\lambda_{x_0} = \lambda(x_1) \; + v_{x_1}(x_0)\, n(x_1) \; + \; w_{x_1}(x_0)\, b(x_1), \tag{3.5.17}$$

where

$$v_{x_1}(x_0) = (\lambda_{x_0} - \lambda(x_1)) \cdot n(x_1), \quad w_{x_1}(x_0) = (\lambda_{x_0} - \lambda(x_1)) \cdot b(x_1).$$

Therefore, for any $\lambda \in \mathcal{D}$ there is a unique pair (x_1, x_0) where

$$\lambda(x_1, x_0) = \lambda(x_1) + v_{x_1}(x_0)\, n(x_1) + w_{x_1}(x_0)\, b(x_1), \qquad x_0, x_1 \in \mathbb{R} \tag{3.5.18}$$

According to this new orthogonal parametrization the loglikelihood function in Equation (3.5.10) can be written as follows,

$$l_{\mathcal{D}}(x_1, x_0) = \sum_{i=1}^{n} \log \left\{ f(x; \mu) + \sum_{j=2}^{4} \lambda_j(x_1, x_0)\, f^{(j)}(x_i; \mu) \right\} \tag{3.5.19}$$

Figure 3.15 shows $\mathcal{D}$ with the new parameterization and the contours of $l_{\mathcal{D}}(x_1, x_0)$ for a sample of size $n = 20$. The surface is in three pieces as the result of the singularity points in the edge of regression. However, each piece is quite well behaved on which any gradient based searching algorithm can be employed for finding the maximum point.

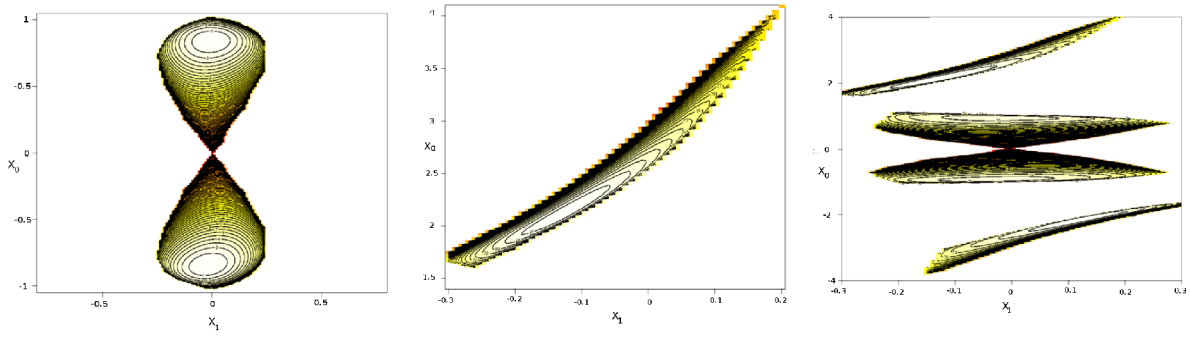

Figure 3.15: Contour plat of the loglikelihood function on the boundary surface $\mathcal{D}$. Left: for $-1 < x_0 < 1$ where $x_0 = 0$, represents the cusp in the edge of regression. Middle: for $x_0 > 1$. Right: the whole surface where it splits as the result of the two asymptotes at $x_0 = \pm 0.871$.

Chapter 4

Local and global robustness in conjugate Bayesian analysis

4.1 Introduction

Statistical analyses are often performed under certain assumptions which are not directly validated. Hence, there is always interest in investigating the degree to which a statistical inference is sensitive to perturbations of the model and data. Specifically, in a Bayesian analysis for which conjugate priors have been chosen, the sensitivity of the posterior to prior choice is an important issue. A rich literature on sensitivity to perturbations of data, prior and sampling distribution exists, see for example: Cook (1986), McCulloch (1989), Lavine (1991), Ruggeri and Wasserman (1993), Blyth (1994), Gustafson (1996), Critchley and Marriott (2004), Linde (2007), Copas and Eguchi (2001, 2010) and Zhu et al. (2011).

For instance, by maximizing a likelihood based divergence function, Cook (1986) introduces a version of influence analysis that finds a direction to which a putative base model is most sensitive; and, as an application, investigates the influence of case deletion in a standard linear regression model. Critchley and Marriott (2004) suggest a complementary method to Cook (1986) showing that data could be exploited for both selecting a suitable base model, rather than assuming it, and learning about the most effective perturbation.

In Gustafson (1996) local sensitivity of posterior expectations to a linear and nonlinear prior perturbation is studied. By adopting a mapping from the space of perturbations to the space of certain posterior expectation, the direction at which the posterior expectation has the maximum sensitivity to prior perturbation is obtained. In Linde (2007) a multiplicative term is used to perturb the base prior or likelihood model, and Kullback-Leibler divergence and χ^2-divergence functions are utilized for assessing local sensitivity. In this paper, local sensitivity is approximated by Fisher information of mixing parameter in an additive and a geometric mixing. Copas and Eguchi (2001, 2010) study robustness of likelihood inference with respect to model perturbation and departure from ignorability and randomness assumptions. They define the space of perturbation as a neighborhood around the base model with a small radius measured by Kullback-Leibler divergence. They show that inference under a misspecified model leads to a first order bias term.

In this chapter we consider both local and global sensitivity analyses with respect to perturbations of a conjugate base prior. We aim for four important properties for our method. Firstly, carefully selected perturbation spaces whose structure is such that it allows the analyst to select the generality of the perturbation in a clear way. Secondly, we want the space to be computationally tractable, hence we focus on convex sets inside linear spaces which have a clear geometric structure. Thirdly, in order to allow for meaningful comparisons, we want the spaces to be consistent with elicited prior knowledge. Thus if a subject matter expert indicates that a prior moment or quantile has a known value – or if a constraint such as symmetry is appropriate – then all perturbed priors should be consistent with respect to this information. Finally, we are going to base our perturbation spaces on mixtures over standard families. The motivation here is that the mixture allows us to explore if the analyst has been over-precise in the specification of the prior by allowing for unthought of heterogeneity. In general, spaces of mixture models are complex but we build on the work of Chapter 2 which shows how discrete mixtures of local mixture models can construct a very flexible but tractable space, see Section 4.2.1.

Sensitivity analysis with respect to a perturbation of the prior, which is the focus of this chapter, is commonly called robustness analysis (Insua and Ruggeri, 2000). In robustness analysis it is customary to choose a base prior model and a plausible class of perturbations. The influence of a perturbation is assessed either locally, or globally, by

measuring the divergence of certain features of the posterior distribution. For instance, Gustafson (1996) studies linear and non-linear model perturbations, and Weiss (1996) uses a multiplicative perturbation to the base prior and specifies the important perturbations using the posterior density of the parameter of interest. Common global measures of influence include divergence functions (Weiss, 1996) and relative sensitivity (Ruggeri and Sivaganesan, 2000).

In local analysis, the rate at which a posterior quantity changes, relative to the prior, quantifies sensitivity (Gustafson, 1996; Linde, 2007; Berger et al., 2000). Gustafson (1996), which we follow closely, obtains the direction in which a certain posterior expectation has the maximum sensitivity to prior perturbation by considering a mapping from the space of perturbations to the space of posterior expectations. In Linde (2007), the Kullback-Leibler and χ^2 divergence functions are utilized for assessing local sensitivity with respect to a multiplicative perturbation of the base prior or likelihood model. They approximate the local sensitivity using the Fisher information of the mixing parameter in additive and geometric mixing.

The approach of this chapter to defining the perturbation space extends the linear perturbations studied in Gustafson (1996) in a number of ways. We do not require the same positivity condition, rather using one which is more general and returns naturally normalized distributions. Further, our space is highly tractable, due to intrinsic linearity and convexity. Finally it is clear, with our formulation, how to remain consistent with prior information which may have been elicited from an expert. The cost associated with this generalization is the boundary defined by (4.2.1) in Section 4.2.1 and the methods we have developed to work with it. We can also compare our method with the geometric approach of Zhu et al. (2011) which uses a manifold based approach. Our, more linear, approach considerably improves interpretability and tractability while sharing an underlying geometric foundation.

In the examples of this chapter we work with our perturbation space in three ways. Similarly to Gustafson (1996) and Zhu et al. (2011) in Example 4.1 we look for the worst possible perturbation, both locally and globally. In Example 4.2 and 4.3 we add constraints to the perturbation space, representing prior knowledge, and again look for maximally bad local and global perturbations. In Example 4.4, we marginalize over the perturbation space

– rather than optimizing over it – as a way of dealing with the uncertainty of the prior. Finally, Example 4.5 applies the methodology on linear regression model using real data.

In Section 4.2, the perturbation space is introduced and its properties are studied. Sections 4.3 and 4.4 develop the theory of local and global sensitivity analysis. Section 4.5 describes the geometry of the perturbation parameter space and proposes possible algorithms for quantifying local and global sensitivity. Section 4.6 presents a number of simulated and real examples. The proofs are sketched in Section 4.8.2.

4.2 Perturbation Space

4.2.1 Theory and Geometry

We construct a perturbation space using the following definitions (Marriott, 2002; Marriott, 2006; Chapter 2). Also, see Chapter 1 for a review on convex and differential geometry.

Definition 4.1 *For the family of mean parameterized models $f(x;\theta)$ the perturbation space is defined by the family of models $f(x;\theta,\lambda)$ such that,*
(i) $f(x;\theta,0) = f(x;\theta)$ for all θ.
(ii) For fixed θ the $f(x;\theta_0,\lambda)$ space is closed under arbitrary mixing.

While other parameterizations can be used, we choose the mean parametrization because it leads to a clear interpretation as we will be working with functions of posterior expectations. A natural way to implement Definition 4.1 is to extend the family $f(x;\theta)$ by attaching to it, at each θ_0, the subfamily $f(x;\theta_0,\lambda)$, which is finite dimensional and spanned by a set of linearly independent functions $v_j(x;\theta_0)$, $j = 1,\cdots,k$. Thus, the subfamily $f(x;\theta_0,\lambda)$ can be defined as the linear space $f(x;\theta_0) + \sum \lambda_j v_j(x;\theta_0)$, where λ_j is a component of the vector λ. For $f(x;\theta_0,\lambda)$ to be a naturally normalized density, we need two further restrictions: (i) $\int v_j dx = 0$, and (ii) the λ parameters must be restricted such that each subfamily is non-negative for all x. This defines the parameter space as

$$\Lambda_{\theta_0} = \left\{ \lambda \mid f(x;\theta_0) + \sum\nolimits_{j=1}^{k} \lambda_j v_j(x;\theta_0) \geq 0, \text{ for all } x \right\}. \tag{4.2.1}$$

Note the space $\Lambda_{\theta_0} \subset \mathbb{R}^k$, is an intersection of half-spaces and consequently is convex.

Clearly, to construct such a perturbation space, the functions ν_j must be selected. A particular form of Definition 4.1 with naturally specified ν_j's is the family of local mixture models, Section 1.7.4.

Definition 4.2 *For a mean, θ, parametrized density $f(x;\theta)$ belonging to the regular exponential family, the local mixture of order k is defined as*

$$h(x;\lambda,\theta) = f(x;\theta) + \sum\nolimits_{j=1}^{k} \lambda_j \, f^{(j)}(x;\theta), \quad \lambda \in \Lambda_\theta \tag{4.2.2}$$

where $\lambda = (\lambda_1, \cdots, \lambda_k)$ and $f^{(j)}(x;\theta) = \frac{\partial^j}{\partial \theta^j} f(x;\theta)$. Also, Λ_θ, for any fixed and known θ, is a convex space defined by a set of supporting hyperplanes.

For regular exponential family $\int f^{(j)}(x;\theta_0)\, dx = 0$, and as shown in Morris (1982), for natural exponential family, the terms $f^{(j)}(x;\theta_0)$'s are all Fisher orthogonal; hence, this family is identifiable in all λ_j parameters when θ is fixed at some known θ_0.

While local mixtures have very attractive inferential properties – unlike general mixture models – they are restrictive in the sense that they are only 'local'. This restriction can be completely removed – while still keeping attractive inferential properties – by considering finite mixtures of local mixtures, Chapter 2.

Definition 4.3 *Let θ_i be a set of user selected, and suitably separated, grid-points as defined in Chapter 2, then a finite mixture of local mixtures is defined as the convex combination*

$$\sum_{i=1}^{K} \rho_i h(x;\lambda_i,\theta_i),$$

where $\lambda_i \in \Lambda_{\theta_i}$.

In this chapter, for simplicity, we mostly consider the single component case, but point out that the generalisation of Definition 4.3 is always possible.

4.2.2 Prior Perturbation

Suppose the base prior model is $\pi_0(\mu;\theta)$, the probability (density) function of a natural exponential family with the hyper-parameter θ.

Definition 4.4 *The perturbed prior model corresponding to $\pi_0(\mu;\theta)$ is defined by*

$$\begin{aligned}\pi(\mu,\lambda;\theta) &:= \pi_0(\mu;\theta) + \sum\nolimits_{j=1}^{k} \lambda_j\, \pi_0^{(j)}(\mu;\theta) \\ &= \pi_0(\mu;\theta)\left\{1 + \sum\nolimits_{j=1}^{k} \lambda_j\, q_j(\mu,\theta)\right\}, \qquad \lambda \in \Lambda_\theta \end{aligned} \tag{4.2.3}$$

where $\lambda = (\lambda_1, \cdots, \lambda_k)$ is the perturbation parameter vector, and $q_j(\mu,\theta) = \frac{\pi_0^{(j)}(\mu;\theta)}{\pi_0(\mu;\theta)}$ are polynomials of degree j.

In Definition (4.4), π_0 is perturbed linearly, similar to the linear perturbation

$$\tau(\cdot, \pi_0, u^*) = \pi_0(\cdot) + u^*(\cdot), \qquad u^*(\cdot) > 0 \tag{4.2.4}$$

studied in Gustafson (1996), but with a different positivity condition, and is, as we shall show, very interpretable for applications. Definition (4.4) can also be seen as the multiplicative perturbation model $\pi(\mu,\lambda;\theta) = \pi_0(\mu,;\theta)\, h^*(\mu;\lambda,\theta)$ studied in Linde (2007).

4.3 Local Sensitivity

In this section we study the influence of local perturbations, defined inside the perturbation space, on the posterior mean. Similar to Gustafson (1996) we obtain the direction of sensitivity using the Fréchet derivative of a mapping between two normed spaces. Throughout the rest of the chapter we denote the sampling density and base prior by $f(x;\mu)$ and $\pi_0(\mu;\theta)$, respectively, and $x = (x_1, \cdots, x_n)$ represents the vector of observations.

Lemma 4.1 *Under the prior perturbation (4.2.3), the perturbed posterior model is*

$$\pi_p(\mu,\lambda|x;\theta) = \frac{\pi_p^0(\mu|x,\theta)}{\xi(\lambda,\theta)}\left\{1 + \sum\nolimits_{j=1}^{k} \lambda_j\, q_j(\mu,\theta)\right\}, \tag{4.3.5}$$

with $\lambda \in \Lambda_\theta$, $\xi(\lambda,\theta) = 1 + \sum_{j=1}^{k} \lambda_j \, E_p^0[q_j(\mu,\theta)] > 0$, *where* $\pi_p^0(\mu|x,\theta)$ *and* $E_p^0(\cdot|x)$ *are the posterior density and posterior mean of the base model.*

The following lemma characterizes the l^{th} moment of the perturbed posterior model. Note that, throughout the rest of the chapter, for simplicity of exposition, we suppress the explicit dependence of ξ, q_j, π_p^0 and π_p on θ.

Lemma 4.2 *The moments of the perturbed posterior distribution are given by*

$$E_p(\mu^l|x,\lambda) = \frac{1}{\xi(\lambda)} \left\{ E_p^0(\mu^l) + \sum\nolimits_{j=1}^{k} \lambda_j \, A_j^l(x) \right\} \tag{4.3.6}$$

where $\lambda \in \Lambda_\theta$ *and* $A_j^l(x) = E_p^0(\mu^l \, q_j(\mu)|x)$.

To quantify the magnitude of perturbation we exploit the size function as defined in Gustafson (1996), i.e., the L^p norm of the ratio $\frac{u^*}{\pi_0}$, for $p < \infty$, with respect to the induced measure by π_0. Accordingly, the size function for $u(\mu) = \sum_{j=1}^{k} \lambda_j \, \pi_0^{(j)}(\mu;\theta)$ is

$$\text{size}(u) = \left[E_{\pi_0} \left(\left| \sum\nolimits_{j=1}^{k} \lambda_j \, q_j(\mu) \right| \right)^p \right]^{\frac{1}{p}},$$

which, (i) is a finite norm and (ii) is invariant with respect to change of the dominating measure and also with respect to any one-to-one transformation on the sample space. Clearly, size(u) is finite if the first $k+p$ moments of $\pi_0(\mu,\theta)$ exist. In addition, property (ii) holds by use of change of variable formula and the fact that for any one-to-one transformation $m = \nu(\mu)$ we have $\bar{\pi}_0^{(j)}(m,\theta)/\bar{\pi}_0(m,\theta) = \pi_0^{(j)}(\mu,\theta)/\pi_0(\mu,\theta)$.

For a mapping $T : \mathcal{U} \to \mathcal{V}$, where $\mathcal{U}$ and $\mathcal{V}$ are, respectively, the perturbations space normed with size($\cdot$), and the space of posterior expectations normed with absolute value, the Fréchet derivative at $u_0 \in \mathcal{U}$ is defined by the linear functional $\dot{T}(u_0) : \mathcal{U} \to \mathcal{V}$ satisfying

$$||T(u_0+u) - T(u_0) - \dot{T}(u_0)u||_{\mathcal{V}} = o(||u||_{\mathcal{U}}),$$

in which $\dot{T}(u_0)u$ is the rate of change of T at u_0 in direction u. Let $Cov_p^0(\cdot,\cdot)$ be the posterior covariance with respect to the base model, Theorem 4.1 expresses $\dot{T}(u_0)u$ as a linear function of λ, at $u_0 = 0$ which corresponds to the base prior model.

Theorem 4.1 *$\dot{T}(0)u$ is a linear function of λ as*

$$\varphi(\lambda) = \sum\nolimits_{j=1}^{k} \lambda_j \, Cov_p^0 \left(\mu, q_j(\mu)\right), \qquad \lambda \in \Lambda_\theta. \tag{4.3.7}$$

4.4 Global sensitivity

Here we use two commonly applied measures of sensitivity – the posterior mean difference and Kullback-Leibler divergence function – for assessing the global influence of prior perturbation on posterior mean and prediction, respectively. The following theorem characterizes the difference between the posterior mean of the base and perturbed models as a function of λ.

Theorem 4.2 *Let $\Psi(\lambda) = E_p(\mu|x,\lambda) - E_p^0(\mu|x)$ represents the difference between the posterior expectations, then*

$$\Psi(\lambda) = \frac{1}{\xi(\lambda)} \, \varphi(\lambda), \qquad \lambda \in \Lambda_\theta. \tag{4.4.8}$$

The function in (4.4.8) behaves in an intuitively natural way, for as $\lambda \to 0$ we have $\xi(\lambda) \to 1$, and consequently $\Psi(\lambda)$ behaves locally in a similar way to $\varphi(\lambda)$.

To assess the influence of the prior perturbation on prediction, we also quantify the change in the divergence in the posterior predictive distribution. As an illustrative example, suppose the sampling distribution and the base prior model are respectively $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(\theta, \sigma_0^2)$. The posterior predictive distribution for the base model is $\mathcal{N}(\mu_\pi, \sigma_\pi^2 + \sigma^2)$, where

$$\mu_\pi = \frac{\theta\sigma^2 + n\sigma_0^2\bar{x}}{n\sigma_0^2 + \sigma^2} \quad , \quad \sigma_\pi^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}.$$

Lemma 4.3 *The posterior predictive distribution for the perturbed model is*

$$g_p(y) = \frac{1}{\xi(\lambda)} \left\{ g_p^0(y) + \Gamma \sum\nolimits_{j=1}^{k} \lambda_j \, E^\star[q_j(\mu)] \right\} \tag{4.4.9}$$

in which, $g_p^0(y)$ is the posterior predictive density for the base model, Γ is a function of $(y, x, n, \theta_0, \sigma_0^2, \sigma^2)$ and $E^\star(\cdot)$ is the expectation with respect to a normal distribution.

For probability measures P_0 and P_1 with the same support space, S, and densities $g_p^0(\cdot)$ and $g_p(\cdot)$, respectively, the Kullback-Leibler divergence functional is defined by,

$$D_{KL}(P_0, P_1) = \int_S \log\left[g_p^0(y)/g_p(y)\right] g_p^0(y)\, dy \tag{4.4.10}$$

which satisfies the following conditions (see Amari (1990)),

1. $D_{KL}(P_0, P_1) \geq 0$, with equality if and only if $P_0 \equiv P_1$.
2. $D_{KL}(P_0, P_1)$ is invariant under any transformation of the sample space.

Theorem 4.3 *Kullback-Leibler divergence between $g_p^0(\cdot)$ and $g_p(\cdot)$ as a function of $\lambda \in \Lambda_\theta$ is*

$$\begin{aligned} D_{KL}(\lambda) &= \int_S \log\left[g_p^0(y)\right]\ g_p^0(y)\, dy + \log[\xi(\lambda)] \\ &\quad - \int_S \log\left(g_p^0(y) + \Gamma \sum\nolimits_{j=1}^{k} \lambda_j E^\star[q_j(\mu)]\right) g_p^0(y) dy, \end{aligned} \tag{4.4.11}$$

4.5 Estimating λ

Similar to the earlier chapters we fix $k = 4$. To obtain the values of λ which find the most sensitive local and global perturbations, as described in Section 4.1, we apply an optimization approach to the functions (4.3.7), (4.4.8) and (4.4.11). $\varphi(\lambda)$ is a linear function of λ on the space Λ_θ which presents the directional derivative of the mapping T at $\lambda = 0$. Thus, for obtaining the maximum direction of sensitivity, called the worst local sensitivity direction in Gustafson (1996), we need to maximize $\varphi(\lambda)$ over all the possible directions at $\lambda = 0$ restricted by the boundary of Λ_θ. However, $\Psi(\lambda)$ and $D_{KL}(\lambda)$ are smooth objective functions on the convex space Λ_θ, for which we propose a suitable gradient based constraint optimization method that exploits the geometry of the parameter space. By Definition 4.2, for a fixed known θ, the space Λ_θ is a non-empty convex subspace in $\mathbb{R}^k$ with its boundary obtained by the following infinite set of hyperplanes

$$\mathcal{H} = \left\{\lambda \;\middle|\; 1 + \sum\nolimits_{j=1}^{k} \lambda_j\, q_j(\mu) = 0\ ;\ \mu \in \mathbb{R}\right\}.$$

Specifically, for the normal example with order $k = 4$, $\mathcal{H}$ is the infinite set of planes of the form

$$P_\lambda(z) = z\lambda_1 + \left(z^2 - \frac{1}{\sigma_0^2}\right)\lambda_2 + \left(z^3 - \frac{3z}{\sigma_0^2}\right)\lambda_3 + \left(z^4 - \frac{6z^2}{\sigma_0^2} + \frac{3}{\sigma_0^4}\right)\lambda_4 + 1. \quad (4.5.12)$$

where $z = \frac{\mu-\theta}{\sigma_0^2}$. Lemma 4.4 describes the boundary of Λ_θ as a smooth immersed manifold which can have self intersections.

Lemma 4.4 *The boundary of Λ_θ is a manifold immersed in $\mathbb{R}^4$ Euclidean space.*

In addition, the interior of Λ_θ, which guarantees positivity of $\pi(\mu, \lambda; \theta)$ for all $\mu \in \mathbb{R}$, can be characterized by the necessary and sufficient positivity conditions on general polynomials of degree four. Comprehensive necessary and sufficient conditions are given in Barnard and Child (1936) and Bandy (1966), see Section 4.8.1.

Lemma 4.5 *The function $\varphi(\lambda)$ attains its maximum value at the gradient direction $\nabla\varphi$ if it is feasible; otherwise, the maximum direction is the direction of the orthogonal projection of $\nabla\varphi$ onto the boundary plane corresponding to $\lambda_4 = 0$.*

$D_{KL}(\lambda)$ and $\Psi(\lambda)$ are smooth functions which can achieve their maximum either in the interior or on the smooth boundary of Λ_θ. Therefore, optimization shall be implemented in two steps: searching the interior using regular Newton-Raphson algorithm, and then searching the boundary using a generalized form of Newton-Raphson algorithm on smooth manifolds, see Section 5.2.2.

4.6 Examples

We consider five examples, where the first three study local and global sensitivity in the normal conjugate model using the optimization approaches developed earlier. In Example 4.4, we address sensitivity analysis in finite mixture models with independent conjugate prior models for all parameters of interest. Rather than using an optimization approach,

for this example, a Markov Chain Monte Carlo method is used and sensitivity of the posterior distribution of each parameter is assessed. The last example studies the effect of prior perturbation in a hierarchical linear regression model based on real data. For demonstrating the effect of the perturbation we compare the posterior distributions before and after perturbation and also use the relative difference between the Bayes estimates defined by

$$d = \frac{|E_p^0(\mu) - E_p^{\hat{\lambda}}(\mu)|}{std_p^0(\mu)}$$

in which $E_p^0(\mu)$ and $E_p^{\hat{\lambda}}(\mu)$ are the Bayes estimates with respect to the base and perturbed models, respectively, and $std_p^0(\mu)$ is the posterior standard deviation under the base model.

Example 4.1 *A sample of size* $n = 15$ *is taken from* $\mathcal{N}(1, 1)$*, and the base prior is* $\mathcal{N}(2, 1)$*. The estimate* $\hat{\lambda}_\Psi = (-0.323, 1.44, -0.218, 0.441)$ *is obtained from minimizing* $\Psi(\lambda)$*, and the corresponding relative discrepancy in Bayes estimate is* $d = 1.19$*; that is, the resulted change in posterior expectation is* 119% *of the posterior standard deviation of the base model. The corresponding density plots of both models are given in Figure 4.1. For a local analysis, we obtained the unit vector* $\hat{\lambda}_\varphi$ *which maximizes the directional derivative* $\varphi(\lambda)$*. Figure 4.2 shows the posterior density displacement corresponding to the perturbation parameter* $\lambda_\alpha = \alpha\hat{\lambda}_\varphi$ *for different values of* $\alpha > 0$*, as well as the boundary point* λ_b *in direction of* $\hat{\lambda}_\varphi$*. The corresponding relative differences in posterior expectation are* $d = 0.09, 0.15, 0.3, 0.55$*. Hence, additionally to obtaining the worst direction, these values suggest that how far one can perturb the base prior along the worst direction so that relative discrepancy in posterior mean estimation is less than, say* 50%*. These results imply that although conjugate priors are convenient in applications, they might cause significant bias relative to other plausible priors.*

Example 4.2 *The central moments of the perturbed prior model, in Definition (4.4), are linearly related to the perturbation parameter* λ*. Specifically, for the normal model the mean, second and third central moments are*

$$\begin{aligned} \bar{\mu}_\pi &= \theta + \lambda_1, \qquad \bar{\mu}_\pi^{(2)} = \sigma^2 + 2\lambda_2 - \lambda_1^2, \\ \bar{\mu}_\pi^{(3)} &= 6\lambda_3 + 2\lambda_1^3 - 6\lambda_1\lambda_2 \end{aligned} \tag{4.6.13}$$

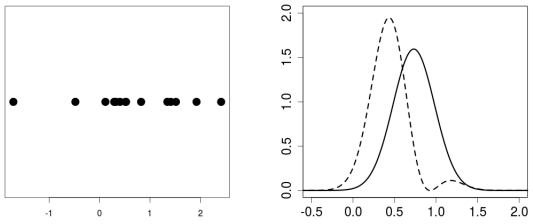

Figure 4.1: respectively, plots for sample, and posterior densities of the based (solid) and perturbed (dashed) model corresponding to $\hat{\lambda}_{\Psi}$.

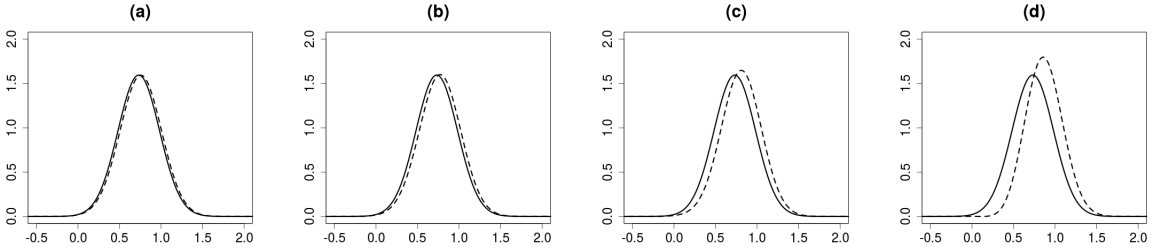

Figure 4.2: Posterior density displacement corresponding to $\lambda = \alpha\hat{\lambda}_{\varphi}$ for $\alpha = 0.1, 0.15, 0.25$ and the boundary point at the maximum direction.

Clearly, λ_1 modifies the mean value, (λ_1, λ_2) adjust variance, and $(\lambda_1, \lambda_2, \lambda_3)$ add skewness to the normal base model. Assuming $\lambda_1 = 0$, guarantees the perturbed model with its mean unchanged, and restricting $\lambda_1 = \lambda_3 = 0$ returns a symmetric perturbed model with same mean as the base prior model.

In this example we fix $\lambda_1 = 0$ and find the most effective local and global perturbations for the similar data in Example 4.1. The estimate $\hat{\lambda}_D = (1.821, -0.011, 0.482)$ and $\hat{\lambda}_{\Psi} = (1.836, 0.016, 0.481)$ are obtained from maximizing $D_{KL}(\lambda)$ and minimizing $\Psi(\lambda)$, respectively. The corresponding relative discrepancies in the Bayes estimate are respectively $d = 1.19, 1.2$; that is, the resultant changes in posterior expectation are respectively 119% and 120% of the posterior standard deviation of the base model. Also, the corresponding posterior distributions are plotted in Figure 4.3. Considering the fact that we constructed the perturbation space as a subfamily of a local mixture model which are 'close' to the base

prior model, these maximum global perturbations are obtained by searching over a reasonably small space of prior distributions which are only different from the base prior by their tail behaviour.

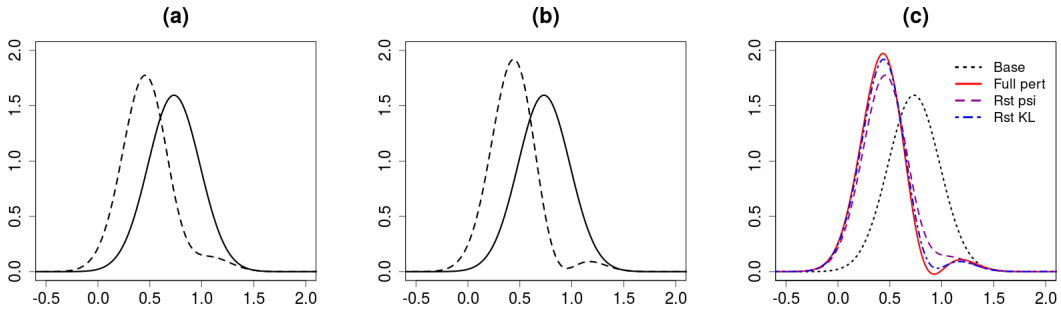

Figure 4.3: (a),(b) correspond to $\hat{\lambda}_\Psi$ and $\hat{\lambda}_D$, under $\lambda_1 = 0$, respectively, including the base (solid) and perturbed posterior (dashed). (c) presents posterior densities of based model (Base), and perturbed models for $\hat{\lambda}_\Psi$ (Rst psi) and $\hat{\lambda}_D$ (Rst KL) under $\lambda_1 = 0$, and the full perturbed posterior model (Full pert) from Example 4.1.

For local analysis, we obtained the unit vector $\hat{\lambda}_\varphi$ which maximizes the directional derivative $\varphi(\lambda)$. Figure 4.4 shows the posterior density displacement by perturbation parameter $\lambda_\alpha = \alpha\hat{\lambda}_\varphi$ for different values of $\alpha > 0$, as well as the boundary point λ_b in direction of $\hat{\lambda}_\varphi$. The corresponding relative differences in posterior expectation are $d = 0.1$, 0.16, 0.25, 0.38, 0.49, 0.56.

Example 4.3 *Suppose that elicited prior knowledge requires a symmetric prior after perturbation, then the perturbation space must be modified by the extra restriction $\lambda_3 = 0$, which gives zero skewness. Consequently, we should be exploring the restricted space, say Λ_θ^0, instead of Λ_θ, for the worst direction and maximum global perturbation. Λ_θ^0 is a 2-dimensional cross section obtained from intersection of Λ_θ with the plane defined by $\lambda_1 = \lambda_3 = 0$. Hence the boundary properties are preserved. For the same data in Example 4.2, sensitivity in the worst direction returns* $d = 0.1$, 0.16, 0.26, 0.42, 0.57, 0.64 *(Figure 4.5). Also, minimizing $\Psi(\lambda)|_{\lambda_1=\lambda_3=0}$ returns $\hat{\lambda}_\Psi^0 = (1.837, 0.494)$.*

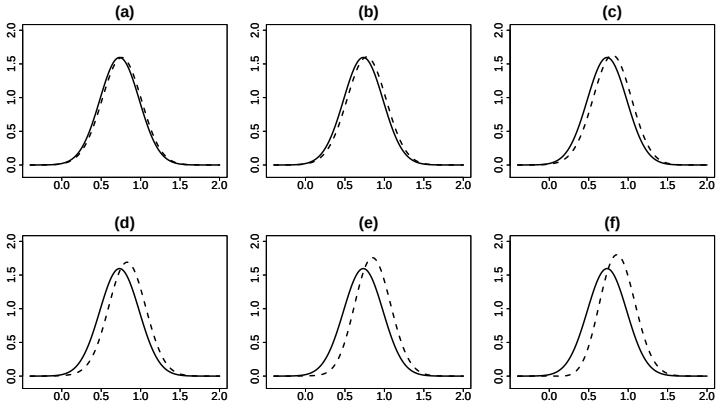

Figure 4.4: (a)-(e) posterior densities of based models and perturbed model (dashed) corresponding to $\lambda = \alpha\hat{\lambda}_\varphi$ where $\alpha = 0.05, 0.07, 0.1, 0.13, 0.15$ and (f) for boundary point in direction of $\hat{\lambda}_\varphi$.

Two observations can be made from these results. First, as in Example 4.2, although we have restricted the perturbation space further, there are still noticeable discrepancies in posterior densities caused by perturbation along the worst direction. Second, the results in Example 4.3 agree with that in Example 4.2, where the estimate of λ_3 does not seem to be significantly different from zero, and the rest of two parameter estimates are quite close in both examples.

Figure 4.6 presents the behavior of the base prior and perturbed prior models corresponding to Examples 4.1-4.3, demonstrating smaller maximal perturbation in the prior in Examples 4.2 and 4.3, where the perturbation parameter is restricted, compared to Example 4.1.

Example 4.4 (Finite Mixture) *Using a missing value formulation, the likelihood function of the mixture model $\rho\mathcal{N}(x;\mu_1,\sigma_1) + (1-\rho)\mathcal{N}(x;\mu_2,\sigma_2)$ can be written as follows*

$$L = \prod_{j=1}^{2} \rho^{n_j} \prod_{i\in A_j} \phi(x_i;\mu_j,\sigma_j),$$

where $A_j = \{i|w_i = j\}$, and w_i is the latent missing variable for x_i such that $p(w_i = 1) = \rho$, and $p(w_i = 2) = 1-\rho$. The marginal conjugate base prior models are $\mu_j \sim \mathcal{N}(\theta_j, \sigma_{0j})$, $\sigma_j^{-2} \sim \Gamma(k_j, \tau_j)$, and $\rho \sim Beta(\alpha,\beta)$, $(j = 1, 2)$.

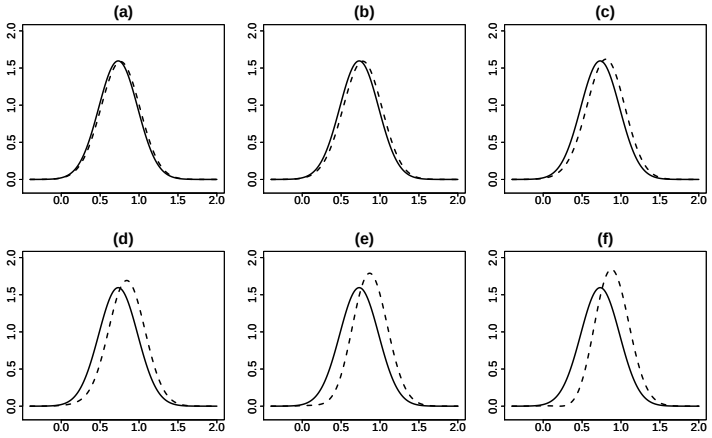

Figure 4.5: (a)-(e) posterior densities of based models and perturbed model (dashed) corresponding to $\lambda = \alpha\hat{\lambda}_\varphi$ where $\alpha = 0.05, 0.07, 0.1, 0.13, 0.15$ and (f) for boundary point in direction of $\hat{\lambda}_\varphi$.

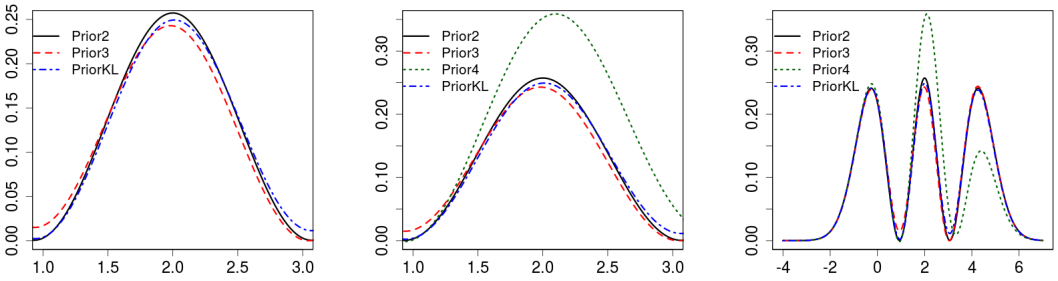

Figure 4.6: Presents all the perturbed prior models in Example 4.1 (Prior4), Example 4.2 (Prior3 and PriorKL for $\hat{\lambda}_\Psi$ and $\hat{\lambda}_D$) and Example 4.3 (Prior2)

In this example the base prior model can be split into five independent components and, correspondingly, five independent perturbation spaces are naturally defined. Unlike previous examples, where we find the maximum local and global perturbations, here we explore each marginalized perturbation space by generating perturbation parameters and observing their influence on the posterior of parameters of interest. We keep $\lambda_1 = 0$ *for all perturbed priors, because of the identifiability issue discussed above.*

Specifically, we use Markov Chain Monte Carlo sampling for estimating the marginal posterior distribution of all parameters of interest corresponding to the base and perturbed models. Each perturbation parameter is generated, independently from the rest, through a Metropolis algorithm with a uniform proposal distribution. Figure 4.7 shows the histograms of generated sample for an observed data set of size $n = 15$ *from* $0.4\mathcal{N}(x; -1, 1) + 0.6\mathcal{N}(x; 1, 1)$*, and the hyper-parameters are set to be* $\theta_1 = -1.5$*,* $\theta_2 = 0.5$*,* $\tau_1 = \tau_2 = 1$*,* $k_1 = k_2 = 2$ *and* $\alpha = \beta = 1$*. Comparing the two histograms for each parameter, the posterior models for* ρ*,* μ_1 *and* μ_2 *show significant differences between the base and perturbed models, as they are extremely skewed for* μ_1 *and* μ_2*. The marginal relative differences are* $d = 0.87, 0.44, 0.60, 0.48, 0.45$*, respectively for* $(\rho, \mu_1, \mu_2, \sigma_1, \sigma_2)$*. These differences, however, are not as significant as those in the previous examples since they do not correspond to maximum perturbations; instead, they return the average influences over all generated perturbation parameter values.*

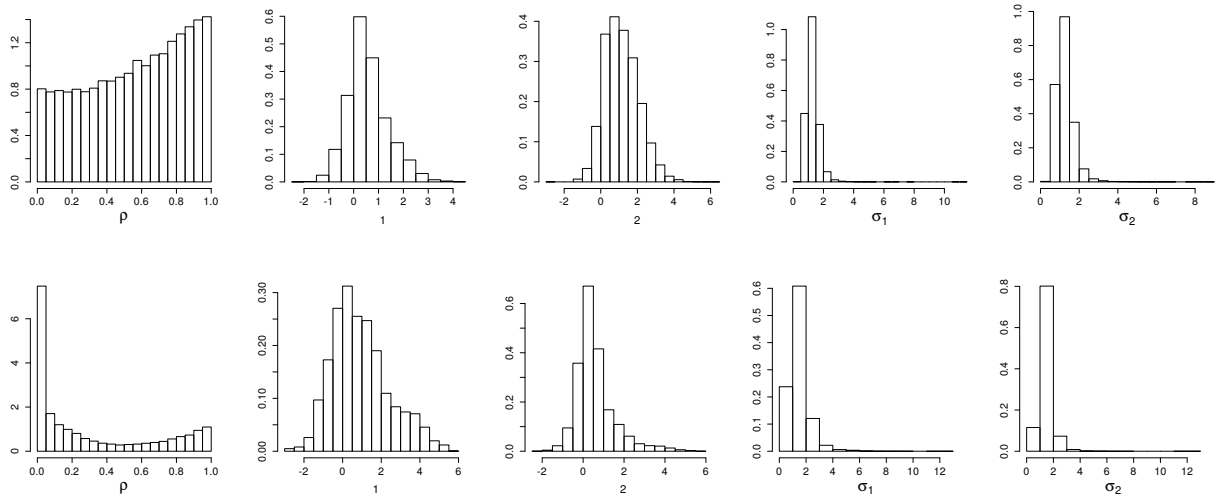

Figure 4.7: First row: estimates from the base model; second row: estimates form the perturbed model

Previous examples have explored the perturbation space in three ways using simulated data. In Example 4.1 we looked for the worst possible perturbation, both locally and globally. In Examples 4.2 and 4.3 we added constraints to the perturbation space, representing

prior knowledge, and again looked for maximally bad local and global perturbations. Finally, in Example 4.4, we marginalize over the perturbation space – rather than optimizing over it – as a way of dealing with the uncertainty of the prior.

The following example studies influence of prior perturbation in hierarchical linear regression model.

Example 4.5 (linear regression) *The full data looks at 919 households in Minnesota. For each, log-radon (y), the logarithm of the amount of radioactive radon in the house, and floor, an 0-1 indicator variable showing whether the house has a basement or not, are recorded (Gelman et al., 2007). Since with a large sample size the data dominates the role of any prior information we take a sample of size* 15 *from households, for which a simple linear regression with model* $y = \alpha + \beta x$ *returns* $\hat{\alpha} = 1.4, (sd = 0.15)$ *and* $\hat{\beta} = -1.27, (sd = 0.34)$. *We want to investigate the influence of perturbation of the base prior* $\beta \sim N(-1, 1)$ *on the Bayes estimate of the linear model. Simple posterior likelihood maximization for the base model gives* $\hat{\alpha}_b = 1.06$ *and* $\hat{\beta}_b = -0.92$. *Maximum perturbation of the prior is obtained via maximizing the* Ψ *function, and the estimates based on maximization of the perturbed posterior loglikelihood are* $\hat{\alpha}_p = 0.66$ *and* $\hat{\beta}_p = -0.39$. *The absolute differences* $|\hat{\alpha}_p - \hat{\alpha}_b| = 0.4$ *and* $|\hat{\beta}_p - \hat{\beta}_b| = 0.53$, *which are bigger than the reported standard deviation for the ordinary regression model, may imply that the perturbation has significant effect on the estimation of the parameters.*

4.7 Summary and Contributions

This chapter uses the model space introduced in Chapter 2 for extending a prior model and defining a perturbation space in the Bayesian sensitivity analysis. This perturbation space is well-defined, tractable, and consistent with the elicited prior knowledge, the three properties that improve the methodology in Gustafson (1996). The perturbation neighborhood defined in this chapter can be considered as the -1 counterpart of the $+1$ perturbation neighborhood in Copas and Eguchi (2001, 2010). Here, however, the "radius" for the neighborhood is defined naturally by the hard boundary, and the discrete mixture of the LMM components takes one step further to defining a much bigger perturbation

space. We study both local and global sensitivity in conjugate Bayesian models. In the local analysis the worst direction of sensitivity is obtained by maximizing the directional derivative of a functional between the perturbation space and the space of posterior expectations. For finding the maximum global sensitivity, however, two criteria are used; the divergence between posterior predictive distributions and the difference between posterior expectations. Both local and global analyses lead to optimization problems with a smooth boundary restriction

4.8 Supplementary materials and proofs

4.8.1 Positivity Conditions

Consider the following quartic

$$p(x) = a\,x^4 + 4b\,x^3 + 6c\,x^2 + 4d\,x + e$$

Barnard and Child (1936) show that the necessary and sufficient positivity conditions for $p(x)$ with $x \in \mathbb{R}$ are

$$\begin{cases} \Delta > 0, \;\; e > 0, \;\; a > 0 \\ H \geq 0, \;\; or \;\; 12H^2 < I^2 \end{cases}$$

where

$$\Delta = I^3 - 27J^2, \quad H = ac - b^2, \quad I = ae - 4bd + 3c^2, \quad J = ace + 2bcd - ad^2 - c^3 - eb^2.$$

Bandy (1966) modified these conditions as

$$\begin{cases} I > 0, \;\; e > 0, \;\; a > 0 \\ I\sqrt{I} + 3\sqrt{3}J > 0 \\ H + a\sqrt{\frac{I}{12}} > 0. \end{cases} \tag{4.8.14}$$

4.8.2 Proofs

Lemma 4.1

$$\pi_p(\mu|x,\lambda) = \frac{\pi(\mu,\lambda) f(x;\mu)}{g(x,\lambda)} \tag{4.8.15}$$

where

$$\begin{aligned}
g(x,\lambda) &= \int \pi(\mu,\lambda;\theta) f(x;\mu)\, d\mu \\
&= \int f(x;\mu)\, \pi_0(\mu;\theta) \left\{1 + \sum\nolimits_{j=1}^{k} \lambda_j\, q_j(\mu,\theta)\right\} d\mu \qquad (4.8.16) \\
&= \int f(x;\mu)\pi_0(\mu;\theta)\, d\mu \\
&\quad + \sum\nolimits_{j=1}^{k} \lambda_j \int q_j(\mu,\theta) f(x;\mu)\pi_0(\mu;\theta)\, d\mu \\
&= g(x) + g(x) \sum\nolimits_{j=1}^{k} \lambda_j \int q_j(\mu,\theta)\pi_p^0(\mu|x,\theta)\, d\mu \qquad (4.8.17) \\
&= g(x)\left\{1 + \sum\nolimits_{j=1}^{k} \lambda_j\, E_p^0[q_j(\mu,\theta)]\right\} \qquad (4.8.18)
\end{aligned}$$

Since $f(x;\mu)\pi_0(\mu;\theta) = g(x)\pi_p^0(\mu|x,\theta)$ and $g(x) = \int f(x;\mu)\pi_0\ (\mu;\theta)\, d\mu$

$$f(x;\mu)\pi_0(\mu;\theta) = g(x)\pi_p^0(\mu|x,\theta) \quad , \quad g(x) = \int f(x;\mu)\pi_0(\mu;\theta)\, d\mu.$$

where, $g(x)$ is the marginal density of sample in the base model. Hence,

$$\begin{aligned}
\pi_p(\mu,\lambda|x;\theta) &= \frac{f(x;\mu)\pi_0(\mu;\theta)\left\{1 + \sum_{j=1}^{k} \lambda_j\, q_j(\mu,\theta)\right\}}{g(x)\left\{1 + \sum_{j=1}^{k} \lambda_j\, E_p^0[q_j(\mu,\theta)]\right\}} \\
&= \frac{g(x)\, \pi_p^0(\mu|x,\theta)\left\{1 + \sum_{j=1}^{k} \lambda_j\, q_j(\mu,\theta)\right\}}{g(x)\left\{1 + \sum_{j=2}^{k} \lambda_j\, E_p^0[q_j(\mu,\theta)]\right\}} \\
&= \frac{\pi_p^0(\mu|x,\theta)}{\xi(\lambda,\theta)}\left\{1 + \sum\nolimits_{j=1}^{k} \lambda_j\, q_j(\mu,\theta)\right\}, \lambda \in \Lambda_\theta
\end{aligned}$$

with $\xi(\lambda,\theta) = 1 + \sum_{j=1}^{k} \lambda_j\, E_p^0[q_j(\mu,\theta)]$.

$$\xi(\lambda,\theta) = 1 + \sum\nolimits_{j=1}^{k} \lambda_j\, E_p^0[q_j(\mu,\theta)]$$

Also $\xi(\lambda,\theta) > 0$, since $1 + \sum_{j=1}^{k} \lambda_j \, q_j(\mu,\theta) > 0$, for all $\mu \in \mathbb{R}$ and $\lambda \in \Lambda_\theta$, and $\xi(\lambda,\theta) = E_p^0[1 + \sum_{j=1}^{k} \lambda_j \, q_j(\mu,\theta)]$.

Lemma 4.2 Result follows by direct calculation and using the fact that,

$$A_j^l(x) := \int \mu^l q_j(\mu) \, \pi_{post}^0(\mu|x) \, d\mu = E_p^0[\mu^l q_j(\mu)] \tag{4.8.19}$$

then for the normal model and $l = 1, 2$ we have

$$A_2^1(x) = \frac{1}{\sigma_0^4}\left[\mu_\pi^{(3)} - 2\theta\mu_\pi^{(2)} + (\theta^2 - \sigma_0^2)\mu_\pi\right]$$

$$A_2^2(x) = \frac{1}{\sigma_0^4}\left[\mu_\pi^{(4)} - 2\theta\mu_\pi^{(3)} + (\theta^2 - \sigma_0^2)\mu_\pi^{(2)}\right]$$

$$A_3^1(x) = \frac{1}{\sigma_0^6}\left[-\mu_\pi^{(4)} + 3\theta\mu_\pi^{(3)} + 3(\sigma_0^2 - \theta^2)\mu_\pi^{(2)} + (-3\theta\sigma_0^2 + \theta^3)\mu_\pi\right]$$

$$A_3^2(x) = \frac{1}{\sigma_0^6}\left[-\mu_\pi^{(5)} + 3\theta\mu_\pi^{(4)} + 3(\sigma_0^2 - \theta^2)\mu_\pi^{(3)} + (-3\theta\sigma_0^2 + \theta^3)\mu_\pi^{(2)}\right]$$

$$A_4^1(x) = \frac{1}{\sigma_0^8}\left[\mu_\pi^{(5)} - 4\theta\mu_\pi^{(4)} + 6(\theta^2 - \sigma_0^2)\mu_\pi^{(3)} + (12\theta\sigma_0^2 - 4\theta^3)\mu_\pi^{(2)} + (3\sigma_0^4 - 6\sigma_0^2\theta^2 + \theta^4)\mu_\pi\right]$$

$$A_4^2(x) = \frac{1}{\sigma_0^8}\left[\mu_\pi^{(6)} - 4\theta\mu_\pi^{(5)} + 6(\theta^2 - \sigma_0^2)\mu_\pi^{(4)} + (12\theta\sigma_0^2 - 4\theta^3)\mu_\pi^{(3)} + (3\sigma_0^4 - 6\sigma_0^2\theta^2 + \theta^4)\mu_\pi^{(2)}\right]$$

since for $j = 2, 3, 4$ we have

$$q_2(\mu) = \frac{(\mu-\theta)^2}{\sigma_0^4} - \frac{1}{\sigma_0^2} \qquad q_3(\mu) = -\frac{(\mu-\theta)^3}{\sigma_0^6} + \frac{3(\mu-\theta)}{\sigma_0^4}$$

$$q_3(\mu,\theta) = -\frac{(\mu-\theta)^3}{\sigma_0^6} + \frac{3(\mu-\theta)}{\sigma_0^4}$$

$$q_4(\mu) = \frac{(\mu-\theta)^4}{\sigma_0^8} - \frac{6(\mu-\theta)^2}{\sigma_0^6} + \frac{3}{\sigma_0^4} \tag{4.8.20}$$

where

$$\mu_\pi^{(2)} = \mu_\pi^2 + \sigma_\pi^2, \quad \mu_\pi^{(3)} = \mu_\pi^3 + 3\sigma_\pi^2\mu_\pi, \quad \mu_\pi^{(4)} = \mu_\pi^4 + 6\sigma_\pi^2\mu_\pi^2 + 3\sigma_\pi^4$$

$$\mu_\pi^{(5)} = \mu_\pi^5 + 10\sigma_\pi^2\mu_\pi^3 + 15\sigma_\pi^4\mu_\pi, \quad \mu_\pi^{(6)} = \mu_\pi^6 + 15\sigma_\pi^2\mu_\pi^4 + 45\sigma_\pi^4\mu_\pi^2 + 15\sigma_\pi^6$$

Theorem 4.1 (Gustafson, 1996, Result 8)
For any bounded $h(\cdot)$ let

$$I_h u = \int h(\mu)\, w(\cdot, \pi, u)\, d\mu$$

then, for a function $r(\mu)$, bounded likelihood model l and perturbed prior $w(\cdot, \pi, u) = \pi(\cdot) + u(\cdot)$ the posterior expectation mapping is defined by

$$T_l^r u = \frac{\int r(\mu)\, l\, w(\cdot, \pi, u)\, d\mu}{\int l\, w(\cdot, \pi, u)\, d\mu} = \frac{I_{rl} u}{I_l u}.$$

According to the definition of Gateaux derivative at u_0 in the direction of u we have

$$\begin{aligned}
\dot{I}_h(u_0)u &= \lim_{\epsilon \to 0} = \frac{\int h(\mu)\, w(\mu, \pi, u_0 + \epsilon u)\, d\mu - \int h(\mu)\, w(\mu, \pi, u_0)\, d\mu}{\epsilon} \\
&= \lim_{\epsilon \to 0} = \frac{\int h(\mu)\, [\pi + u_0 + \epsilon u]\, d\mu - \int h(\mu)\, [\pi + u_0]\, d\mu}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\epsilon \int h(\mu) u\, d\mu}{\epsilon} = \int h(\mu) u\, d\mu \qquad (4.8.21)
\end{aligned}$$

Now the derivative of the mapping T at $u_0 = 0$ is obtained as follows,

$$\begin{aligned}
\dot{T}_l^r(0)u &= \frac{\dot{I}_r l(0) u}{I_l 0} - \frac{I_r l0}{I_l 0} \cdot \frac{\dot{I}_l(0) u}{I_l 0} \\
&= \frac{\int r\, l\, u\, d\mu}{\int l\, \pi\, d\mu} - \frac{\int r\, l\, \pi\, d\mu}{\int l\, \pi\, d\mu} \cdot \frac{\int l\, u\, d\mu}{\int l\, \pi\, d\mu} \\
&= \frac{\int r \frac{u}{\pi}\, l\, \pi\, d\mu}{\int l\, \pi\, d\mu} - \frac{\int r\, l\, \pi\, d\mu}{\int l\, \pi\, d\mu} \cdot \frac{\int \frac{u}{\pi}\, l\, \pi\, d\mu}{\int l\, \pi\, d\mu} \\
&= E_p^0(r\frac{u}{\pi}) - E_p^0(r) \cdot E_p^0(\frac{u}{\pi}) \\
&= Cov_p^0(r, \frac{u}{\pi}) \qquad (4.8.22)
\end{aligned}$$

Theorem 4.2 By direct calculation and use of equation (4.8.19)

Lemma 4.3

$$g_p(y) = \int f(y; \mu) \pi_p(\mu, \lambda | x)\, d\mu \qquad (4.8.23)$$

is the convolution of $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(\mu_\pi, \sigma_\pi^2)$. Since,

$$\frac{(y-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_\pi)^2}{\sigma_\pi^2} = \frac{\left(\mu - \frac{\sigma_\pi^2 y + \sigma^2 \mu_\pi}{\sigma^2+\sigma_\pi^2}\right)^2}{\frac{\sigma^2\sigma_\pi^2}{\sigma^2+\sigma_\pi^2}} + \frac{(y-\mu_\pi)^2}{\sigma^2+\sigma_\pi^2}$$

hence, the posterior predictive distribution for base model is $\mathcal{N}(\mu_\pi, \sigma_\pi^2 + \sigma^2)$ and (4.4.9) is obtained by direct calculation, where,

$$\Gamma = \frac{1}{\sqrt{2\pi(\sigma_\pi^2+\sigma^2)}} \exp\left\{-\frac{(y-\mu_\pi)^2}{2(\sigma_\pi^2+\sigma^2)}\right\}$$

and $E^\star(\cdot)$ is expectation with respect to μ according to the following normal distribution

$$\mathcal{N}\left(\frac{\sigma_\pi^2 y + \sigma^2\mu_\pi}{\sigma_\pi^2+\sigma^2}, \frac{\sigma_\pi^2\sigma^2}{\sigma_\pi^2+\sigma^2}\right)$$

Theorem 4.3 Use of Lemma 4.3 and direct calculation finishes the proof.

Lemma 4.4 Implied by direct application of the implicit function theorem, Rudin, 1976, p.225. Let $\sigma_0 = 1$ in equation (4.5.12) for convenience. From solving $P_\lambda(z) = 0$ and $P'_\lambda(z) = 0$, simultaneously for λ_2 and λ_3, we get a smooth parametrization for the boundary as follows

$$\begin{aligned} \lambda_2(z, \lambda_1, \lambda_4) &= \frac{(z^6 - 3z^4 + 9z^2 + 9)\lambda_4 - 2z^3\lambda_1 - 3z^2 + 3}{z^4+3} \\ \lambda_3(z, \lambda_1, \lambda_4) &= -\frac{(-z^2-1)\lambda_1 + (2z^5 - 4z^3 + 6z)\lambda_4 - 2z}{z^4+3} \end{aligned} \tag{4.8.24}$$

Hence, by implicit function theorem (Rudin, 1976, p.225) the boundary of Λ_{θ_0} is a smooth surface (Manifold) embedded in $\mathbb{R}^4$ by

$$\begin{aligned} \mathcal{S}_1 &: \mathbb{R} \times U \times \to \mathbb{R}^4 \\ \mathcal{S}_1(z, \lambda_1, \lambda_4) &= [\lambda_1, \lambda_2(z, \lambda_1, \lambda_4), \lambda_3(z, \lambda_1, \lambda_4), \lambda_4] \end{aligned} \tag{4.8.25}$$

Lemma 4.5 $\nabla\varphi = (a_1, a_2, a_3, a_4)$, is a vector originated at $\lambda = 0$, where $a_j = Cov_p^0(\mu, q_j(\mu))$. If it is feasible then it clearly gives the maximum direction. However, if it is not feasible then $a_4 \leq 0$ since the condition $a_4 > 0$ is necessary for feasibility. Thus, the direction of the orthogonal projection of $\nabla\varphi$ onto the boundary plane corresponding to $\lambda_4 = 0$ is the closest we get to a maximum and feasible direction.

Chapter 5

Generalizing the Frailty Assumptions in Survival Analysis

5.1 Introduction

Frailty models are important for analyzing survival time data and have been studied by many researchers; for example, Klein (1992), Hougaard (1986), Clayton (1978) and Gorfine et al. (2006). One way of deriving frailty survival models, which we do not follow here, is to formulate the frailty factor as a single parameter, θ, presenting the association between time-to-event data of two correlated events, in which $\theta = 1$ is interpreted as being no correlation while $\theta > 1$ and $\theta < 1$ demonstrate positive and negative association, respectively (Hu et al., 2011; Nan et al., 2006; Clayton and Cuzick, 1985;Clayton, 1978; Oakes, 1982).

An alternative approach for modeling heterogeneity as unobserved covariate, is to add the frailty variable as a multiplicative factor to the baseline hazard function. In Hougaard (1986) a positive stable family is assumed for the frailty variable and the marginal survival time is assumed to have an exponential or Weibull distribution or be unspecified. Various hazard functions, including Cox's regression model, have been generalized by assuming a gamma frailty variable with mean equal to 1 and variance η, see Nielsen et al. (1992) and Klein (1992). They utilize the Expectation-Maximization algorithm for es-

timation of parametric and nonparametric accumulated hazard function and regression coefficients. Further, Gorfine et al. (2006) proposed a different approach for estimation in non-parametric frailty survival models, which is applicable for any parametric model with finite mean on the frailty variable.

Although, different models have been assumed for the multiplicative frailty variable, (Gorfine et al., 2006; Hougaard, 1984; Hougaard, 1986), one of the most frequently used distributions is the gamma distribution, because of its tractable properties (Klein, 1992; Nielsen et al., 1992; Vaupel et al., 1979). For example Martinussen et al. (2011) used gamma frailty in the Aalen additive model, and Zeng et al. (2009) studied transformation models with gamma frailty for multivariate survival analysis, in which $\eta = 0$ (no frailty) is also allowed. In addition, Abbring and Van Den Berg (2007) establish the fact that conditional frailty among survivors is always gamma distributed if and only if the frailty distribution is regularly varying at zero.

In this chapter, we consider Cox's regression model (Cox, 1972) with a multiplicative frailty factor on which no specific model is imposed, as biased estimators might be obtained if the frailty model is misspecified (Abbring and Van Den Berg, 2007; Hougaard, 1984). Similar to a general mixture model problem, the frailty survival models with unknown frailty distribution, suffers from identification issues. Although, when all the covariates variables are continuous with continuous distribution, Eleber and Ridder, 1982 shows that given the distribution of the time duration variable, all the three multiplicative factors are identified, his theoretical result does not solve the identifiability issue in the general sense. For instance, when there is a discrete covariate then identifiability requires the corresponding regression coefficient to be limited to a known compact set (Horowitz, 2010, ch.2). Consequently, the estimation method developed using unknown transformation models in Horowitz (1999), although useful for econometric models, has the same limitation.

In this chapter, however, we use the discrete mixture of local mixture models introduced in Definition 2.2, as they are always identifiable and estimable, and their geometric and inferential properties allow for fast and efficient estimation algorithms. They give the right parametrization and the tool for learning the frailty model without any further information about its structure.

Notation, motivation and the main result of the chapter, including our proposed method, the local mixture method, are presented in Section 5.2 for a fixed hazard frailty model. We estimate the regression coefficients through a two step optimization process, the first of which is implemented using our proposed algorithm. The algorithm comprises using a gradient search method on smooth manifolds embedded in finite dimensional Euclidean spaces. The methodology is generalized to non-parametric baseline hazard in Section 5.3. Section 5.4 is devoted to simulation studies, illustrating that the local mixture method returns similar bias, but larger standard deviation for the estimates compared to the to the existed Expected-Maximization method in Klein (1992) when a gamma frailty is assumed. In Section 5.5, rhDNAse data is analyzed and the results are compared, for both treatment and placebo group, with the method in Klein (1992).

5.2 Methodology

Throughout this section, we follow the notation and definitions in Lawless (1981) and Gorfine et al. (2006). Let (T_i^0, C_i), for $i = 1, \cdots, n$, be the failure time and censoring time of the ith individual, and also let X be the $n \times p$ design matrix of the covariate vectors. Define $T_i = \min(T_i^0, C_i)$ and $\delta_i = I(T_i^0 < C_i)$, where $I(\cdot)$ is an indicator function. In addition, associated with the ith individual, an unobservable covariate θ_i, the frailty, is assumed, where θ_i's follow some distribution, Q.

Suppose, at least initially, that the marginal lifetime distribution given frailty is an exponential model with rate λ_0. Then the baseline hazard function is $\lambda_0(t) = \lambda_0$. Adapting the regression model in Cox (1972), the hazard function for the ith individual conditional on the frailty θ_i takes the following form,

$$\lambda_i = \theta_i \, \lambda_0 \exp\{X_i \beta\}, \tag{5.2.1}$$

where X_i is the ith row of X and $\beta = (\beta_0, \cdots, \beta_{p-1})$ is a p-vector parameter. For the ith individual with the hazard function in Equation (5.2.1), the cumulative hazard function and survival function are, respectively, defined as

$$\Lambda_i(t) = \int_0^t \lambda_i(u) \, du, \qquad S_i(t) = \exp\{-\Lambda_i(t)\}. \tag{5.2.2}$$

Following the arguments in Gorfine et al. (2006), we assume that the frailty θ is independent of X, and further that, given X and θ, censoring is independent and noninformative for θ and (λ_0, β). Then, for the exponential failure time, the full likelihood function for the parameter vector (λ_0, β) is written as

$$L(\lambda_0, \beta) = \prod_{i=1}^{n} \int \left[\left(\theta\, \lambda_0\, e^{X_i\beta}\right)^{\delta_i} \exp\left\{-\theta\, T_i \lambda_0\, e^{X_i\beta}\right\} \right] dQ(\theta), \tag{5.2.3}$$

and the log likelihood function is

$$l(\lambda_0, \beta) = \sum_{i=1}^{n} \delta_i [\log \lambda_0 + X_i\beta] \; + \sum_{i=1}^{n} \log \int \theta^{\delta_i} \; \exp\{-\theta\lambda_0\, T_i\, e^{X_i\beta}\}\, dQ(\theta) \tag{5.2.4}$$

5.2.1 Local Mixture Method

As mentioned in Section 1, it is common in the literature to assume a gamma model, with $E_Q(\theta) = 1$ and variance η, for θ and apply the Expectation-Maximization algorithm for maximizing the loglikelihood function in Equation (5.2.4). However, since frailty model misspecification causes biased coefficient estimation, we relax the gamma restriction and assume a more general family of distributions for the frailty. Essentially, we use Definition 2.2 and substitute the integral term in equation (5.2.4) by a discrete mixture of LMMs on a set of grid points $\vartheta_1, \cdots \vartheta_L$. For consistency with the assumption $E_Q(\theta) = 1$ and also for simplicity, in this chapter we only use one component at $\vartheta = 1$; but generalization is always possible. For instance the set of grid points can be selected in a ways that their average is 1.

Let

$$f(T_i, \beta, \theta) = \theta^{\delta_i} \; \exp\{-\theta\lambda_0\, T_i\, e^{X_i\beta}\},$$

then the integral term in Equation (5.2.4) can be written as

$$\int_{\Theta} f(T_i, \beta, \theta)\, dQ(\theta) = f(T_i, \beta, \vartheta) + \sum_{j=1}^{k} \lambda_j\, f^{(j)}(T_i, \beta, \vartheta) + O\left(\epsilon^{\lfloor \frac{k+1}{2} \rfloor}\right), \tag{5.2.5}$$

where $f^{(j)}(T_i, \beta, \vartheta) = \frac{\partial}{\partial\theta^j} \mid_{\theta=\vartheta} f(T_i, \beta, \theta)$, and $\lambda = (\lambda_1, \cdots, \lambda_k)$ is a parameter vector, the last term characterizes the error of approximation, and $\epsilon > 0$ represents the variation of Q (Marriott, 2002).

The model in Equation (5.2.5) is similar to the local mixture models, equation (1.7.10) in Section 1.7.4. For any fixed ϑ, the finite dimensional parameter vectors λ represent the frailty distribution through its central moments. Local mixing extends a parametric model to a larger and more flexible space of densities which holds nice geometric and inferential properties. Identifiability is achieved by fixing $\vartheta = 1$ and Fisher orthogonality of the higher derivatives as discussed in Chapter 2.

Using only one component LMM for learning all the information in a general frailty distribution seems to be a bit restrictive. However, as seen in Example 2.1 a single component LMM is quite flexible for modeling contaminated data, where the latent variation is obvious. In fact, our simulation studies show that LMMs can even produce multimodal models. In addition, as mentioned earlier, generalization is always possible by adding more LMM components if the results are not satisfactory for one data set.

Substituting Equation (5.2.5) in Equation (5.2.4) we obtain

$$\begin{aligned} l(\lambda_0, \beta, \lambda) &= \sum_{i=1}^{n} \left(\delta_i[\log \lambda_0 + X_i\beta] + \log f(T_i, \beta, \vartheta)\right) \\ &\quad + \sum_{i=1}^{n} \log\left(1 + \sum\nolimits_{j=2}^{k} \lambda_j \, A_j(\delta_i, y_i)\right), \qquad \lambda \in \Lambda_\vartheta \end{aligned} \tag{5.2.6}$$

in which $A_j(\delta_i, y_i) = \frac{f^{(j)}(T_i,\beta,\vartheta)}{f(T_i,\beta,\vartheta)}$, and $y_i = \lambda_0 T_i \exp\{X_i\beta\}$ is positive. Assuming $\vartheta = 1$, we maximize Equation (5.2.6) when estimating β, where λ_0 and λ are considered as nuisance parameters which are required to be obtained in advance. Thus, a profile likelihood optimization method is employed. That is, we first maximize for λ over Λ_ϑ to obtain $\hat{\lambda}$ and then maximize $l_p(\beta) = l(\hat{\lambda}_0, \beta, \hat{\lambda})$ to estimate β. $\hat{\lambda}_0$ is imputed into the loglikelihood function at each iteration. A method for computing $\hat{\lambda}_0$ is described in Section 5.3, for a more general situation.

5.2.2 Maximum Likelihood Estimator for λ

The λ parameter space, Λ_ϑ, is characterized as the space of all λ's such that, for all $y > 0$

$$1 + \sum\nolimits_{j=1}^{k} \lambda_j \, A_j(\delta_i, y) > 0, \quad \delta_i = 0, 1 \tag{5.2.7}$$

where, $A_j(\delta_i, y)$, as a function of $y > 0$, is a polynomial of degree j. For $k = 4$, the inequality in Equation (5.2.7) is equivalent to the simultaneous positivity conditions of the following two quartics,

$$\begin{aligned} p(y) &= \lambda_4 y^4 - \lambda_3 y^3 + \lambda_2 y^2 - \lambda_1 y + 1, \\ q(y) &= \lambda_4 y^4 - (4\lambda_4 + \lambda_3) y^3 + (3\lambda_3 + \lambda_2) y^2 - (2\lambda_2 + \lambda_1) y + \lambda_1 + 1. \end{aligned} \tag{5.2.8}$$

for which we can prove the following result.

Lemma 5.1 *If Λ^1 and Λ^2 are the space of all $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ such that $p(y)$ and $q(y)$ are positive on $y > 0$, respectively, then $\Lambda^2 \subset \Lambda^1$.*

Proof: First note that $\lambda_4 > 0$ is a necessary condition hence $P(y)$ has a minimum. Also $q(y) = p(y) - p'(y)$, $p(0) = 1$ and $p'(0) = 0$. For all $y > 0$, If $q(y) > 0$, then $p(y) > p'(y)$. Since $p(y)$ attains its minimum value at some y_1 for which $p'(y_1) = 0$, therefore, $p(y) \geq p(y_1) > p'(y_1) = 0$. If $y_1 = 0$, we have $p(y) > p(0) = 1$.

Lemma 5.1 implies that Λ_ϑ can be characterized just by investigating the positivity domain of $q(y)$, for which the following theorem is required (see Ulrich and Watson (1994)).

Theorem 5.1 *For the quartic polynomial $p(x) = ax^4 + bx^3 + cx^2 + dx + e$, with $a > 0$ and $e > 0$, define $\alpha = b\, a^{-3/4} e^{-1/4}$, $\beta = c\, a^{-1/2} e^{-1/2}$, $\gamma = d\, a^{-1/4} e^{-3/4}$,*

$$\Delta = 4[\beta^2 - 3\alpha\gamma + 12]^3 - [72\beta + 9\alpha\beta\gamma - 2\beta^3 - 27\alpha^2 - 27\gamma^2]^2$$

$$L_1 = (\alpha - \gamma)^2 - 16(\alpha + \beta + \gamma + 2)$$

$$L_2 = (\alpha - \gamma)^2 - \tfrac{4(\beta+2)}{\sqrt{\beta-2}} \left(\alpha + \gamma + 4\sqrt{\beta - 2}\right).$$

Then, $p(x) \geq 0$ for all $x > 0$ if and only if

- $\beta < -2 \quad , \quad \Delta \leq 0 \quad , \quad \alpha + \gamma > 0$
- $-2 \leq \beta \leq 6 \ , \ (\Delta \leq 0 \ , \ \alpha + \gamma > 0)$ *or* $(\Delta \geq 0 \ , \ L_1 \leq 0)$

- $6 < \beta$, $(\Delta \leq 0\ ,\ \alpha + \gamma > 0)$ *or* $(\alpha > 0\ ,\ \gamma > 0)$ *or* $(\Delta \geq 0\ ,\ L_2 \leq 0)$

Due to the existence of hard boundaries, obtaining $\hat{\lambda}$ is a nonstandard inference problem. A suitable maximization algorithm should be flexible enough to converge to a turning point $\hat{\lambda}$ in the interior if $\hat{\lambda} \in \Lambda_\vartheta$; otherwise, it must converge to the unique boundary point with the highest likelihood, say $\hat{\lambda}_b$ (Berger, 1987, p.337). In the rest of this section, we propose a gradient based optimization algorithm, utilizing the geometry of Λ_ϑ and concavity of the local mixture term in Equation (5.2.6) for finding the global maximum value $\hat{\lambda}$ or $\hat{\lambda}_b$ in two major steps. The following lemma reveals the geometry of the boundary surface of Λ_ϑ, as a manifold immersed in $\mathbb{R}^k$, where k is the order of the corresponding local mixture model.

Lemma 5.2 *The boundary of the parameter space Λ_ϑ, shown by Λ_ϑ^b, is an immersed manifold in $\mathbb{R}^4$.*

Proof: see Section 5.7.2.

Algorithm

0: Start with an initial value $\lambda^{(0)} \in \Lambda_\vartheta$.

1: Run Newton-Raphson algorithm, until either algorithm converges to $\hat{\lambda} \in \Lambda_\vartheta$ (then stop) or the first update $\lambda^{(j)} \notin \Lambda_\vartheta$ is obtained (go to step 2).

2: Find the boundary point $\lambda^\star$ on the line segment between $\lambda^{(j-1)}$ and $\lambda^{(j)}$, let $\lambda^{(j-1)} = \lambda^\star$ and run the following steps.

  2a: Find the gradient g_j and the supporting plane t_j at $\lambda^{(j-1)}$.

  2b: Update $\lambda^{(j)} = \lambda^{(j-1)} + (\Pi_j H_j^{-1})(\Pi_j g_j)$, (Figure 5.1, middle panel).

  2c: Update $\lambda^\star$ by finding the boundary point on the line segment in the direction of N_j, the normal vector of t_j, passing through $\lambda^{(j)}$ (Figure 5.1, right panel).

  2d: Let $\lambda^{(j-1)} = \lambda^\star$ and repeat $(2a)$-$(2c)$, until convergence; that is $||P_{t_j}(g_j)|| \to 0$.

Step 1, obviously applies the well understood Newton-Raphson algorithm on the interior of Λ_ϑ as a subspace of $\mathbb{R}^k$. In Step 2, however, a generalization of Newton's method on smooth manifolds is exploited. Applying Lemma 2 and using the technical details in Section 5.7.1 we can prove the following result.

Theorem 5.2 *The algorithm either converges to $\hat{\lambda}$ quadratically in step(1), or there is an open neighborhood $V \subset \Lambda_\vartheta^b$ of $\hat{\lambda}_b$, that for any $\lambda^\star \in V$ it converges to $\hat{\lambda}_b$ in quadratic order, in step(2).*

Proof: see Section 5.7.2.

5.3 Non-parametric Hazard

Although our working example in Section 2 has a fixed hazard rate and exponential lifetime model, the methodology can be generalized to other parametric marginal lifetime distributions with known hazard function up to a finite dimensional parameter vector. Furthermore, identifiability property of local mixture models allows the methodology to be generalized for nonparametric hazard function. When the baseline hazard function is an unknown time-dependent function $\lambda_0(t)$, the hazard function for ith individual takes the form

$$\lambda_i(t) = \theta_i \, \lambda_0(t) \exp\{X_i \beta^T\}. \tag{5.3.9}$$

The log likelihood function in Equation (5.2.4) has the following form

$$l(\beta, Q) = \sum_{i=1}^{n} \delta_i [\log \lambda_0(T_i) + X_i \beta] \; + \sum_{i=1}^{n} \log \int \theta^{\delta_i} \, \exp\{-\theta \Lambda_0(T_i) \, e^{X_i \beta}\} \, dQ(\theta)$$

and after approximating the integral using a local mixture we obtain

$$\begin{aligned} l(\beta, \lambda) \;\; &= \;\; \sum\nolimits_{i=1}^{n} \left(\delta_i [\log \lambda_0(T_i) + X_i \beta] + \log f(T_i, \beta, \vartheta)\right) \\ &\quad + \sum\nolimits_{i=1}^{n} \log \left(1 + \sum\nolimits_{j=1}^{k} \lambda_j \, A_j(\delta_i, y_i)\right), \qquad \lambda \in \Lambda_\vartheta \end{aligned} \tag{5.3.10}$$

where $y_i = \Lambda_0(T_i)\exp\{X_i\beta\}$. Therefore, the geometric and inferential properties of the model stays the same and we can proceed as in previous section.

To impute $\lambda_0(t)$ and $\Lambda_0(t)$ we can use the same argument in Gorfine et al. (2006) to provide a recursive estimate of the cumulative hazard function using the fact that for two consecutive failure times $T_{(i)}$ and $T_{(i+1)}$ we have $\Lambda_0(T_{(i+1)}) = \Lambda_0(T_{(i)}) + \Delta\Lambda_i$. Substituting this recursive equation in the log likelihood function in (5.3.10), considering the conventions in Breslow (1972) and taking partial derivative with respect to $\Delta\Lambda_i$ we obtain

$$\frac{\partial l}{\partial \Delta\Lambda_i} = \frac{1}{\Delta\Lambda_i} - \sum_{\ell=i}^{n} e^{X_\ell\beta} + \frac{P'(e^{X_i\beta}[\Lambda_0(T_{(i)}) + \Delta\Lambda_i])}{P(e^{X_i\beta}[\Lambda_0(T_{(i)}) + \Delta\Lambda_i])} \tag{5.3.11}$$

which is a function of just $\Delta\Lambda_i$ when $\hat{\Lambda}_0(T_{(i)})$ is given at time $T_{(i+1)}$, where $P(\cdot)$ is a polynomial of degree four with its coefficients as linear functions of $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ and $P'(\cdot)$ is its derivative with respect to $\Delta\Lambda_i$. When the denominator is not zero, equation (5.3.11) is a polynomial of degree five which can be solved numerically for $\Delta\Lambda_i$. Note that when there is no frailty factor; that is, $\lambda = (0, 0, 0, 0)$ then the last term in equation (5.3.11) is zero, and the estimate of the cumulative hazard function reduces to the form in Johansen (1983) which is the estimate in Klein (1992) with $\hat{\omega} = 1$.

5.4 Simulation Study

In this section a simulation study is conducted to compare the local mixture method with the method in Klein (1992), which assumes a gamma model with mean 1 and variance η, for the frailty and applies the Expectation-Maximization algorithm. The Expectation-Maximization method is a repetitive optimization method while in local mixture method the optimization is performed in just two steps; hence, it is faster.

We let $C = 0.01$, $\tau = 4.6$ and follow a similar set-up as found in Hsu et al. (2004). For each individual the event time is $T = [-\log(1-U)\{\theta\exp\{\beta X\}\}^{-1}]^{-1/\tau}C^{-1}$, where $X \sim N(0,1)$, $U \sim$ uniform$[0,1]$. The censoring distribution is $N(100, 15)$, and frailty is assumed to follow a gamma distribution with mean 1 and variance η. Table 5.1 shows the bias and standard deviation for the estimates of the regression coefficient using both

methods for three different values of η. It is illustrated that the LMM method which does not use any information about the frailty model returns almost similar bias as the EM method. However, the standard deviation for the estimates in LMM method are almost twice as big as that for the EM method.

			EM		LMM	
n	η	β	*bias*	*std*	*bias*	*std*
200	0.5	$\log 3$	-0.057	0.18	-0.048	0.43
200	0.7	$\log 3$	-0.056	0.21	-0.038	0.40
200	1	$\log 3$	-0.117	0.22	-0.094	0.41

Table 5.1: Bias and standard deviation of coefficient estimates, when frailty is generated from $\Gamma(\frac{1}{\eta}, \eta)$.

Although the results produced by the LMM method are promising in terms of the bias of estimator, the following two reasons may cause the bigger standard deviations compared to the EM method. First, the LMM method clearly uses no information about the shape of the frailty distributions, and it just exploits its flexibility to extract that form the data. Second, we obtain λ and β only once in the LMM method, while the EM method iterates between η and β until convergence.

Table 5.2 shows a similar result for $\eta = 0.5$, where we let the LMM method run more than one iteration between λ and β, similar to the EM method, until convergence. The results do not show any significant changes in the variance of estimation. This observation may be explained by the Fisher orthogonality of parametrization in LMMs, meaning that asymptotically the parameters do not affect each other.

The simulation study in this chapter is rather unfinished, and more works needs to be done. Adding more components to the LMM approximation might be one way of returning better estimates. In addition, we want to run similar simulation studies for misspecified frailty models and compare the results between the two methodologies.

			EM		LMM	
n	η	β	$bias$	std	$bias$	std
200	0.5	$\log 3$	-0.06	0.18	-0.02	0.38

Table 5.2: Bias and standard deviation of coefficient estimates, when frailty is generated from $\Gamma(\frac{1}{\eta}, \eta)$. λ and β are iterated until convergence.

5.5 Example

The data was reported based on a clinical trial for assessing the influence of rhDNase on the occurrence of respiratory exacerbations among patients with cystic fibrosis (Fuchs et al., 1994). Among the 645 patients, 324 were assigned to a placebo group using a double-blind randomized design. For both treatment and placebo group, we study the time to the first occurrence of respiratory exacerbation with two baseline measurements of forced expository volume, FEV_1 and FEV_2 as covariates.

As illustrated in both Tables 5.3 and 5.4 the estimate of the coefficients do not show significant differences between the two methods.

Method	$\hat{\beta}_1$	$\hat{\beta}_2$
EM	0.114	-0.142
LMM	0.083	-0.104

Table 5.3: Coefficient estimates for placebo group of rhDNAse data using both methods with unspecified hazard function are obtained.

Method	$\hat{\beta}_1$	$\hat{\beta}_2$
EM	0.039	-0.061
LMM	-0.003	-0.092

Table 5.4: Coefficient estimates for the treatment group of rhDNAse data.

5.6 Summary and Contributions

This chapter studies Cox's proportional hazard model with unobserved frailty for which no specific distribution is assumed. The likelihood function, which has a mixture structure with an unknown mixing distribution, is approximated by a local mixture model, which is always identifiable and estimable. The nuisance parameters in the approximating model, which represent the frailty distribution through its moments, lie in a convex space with a smooth boundary characterized as a smooth manifold. Using differential geometric tools, a new algorithm is proposed for maximizing the likelihood function restricted by the smooth yet non-trivial boundary. The regression coefficients, the parameters of interest, are estimated in a two step optimization process, unlike the existing methodology in Klein (1992) which assumes a gamma assumption and uses the Expectation-Maximization approach. An early stage simulation study shows that, the LMM method gives as good estimation bias as that in EM method when frailty factor is generated form a gamma model; however, it returns a bigger variance for the estimators. The Simulation study and data example in this section is not complete and some extra works need to be done.

5.7 Supplementary materials and proofs

5.7.1 The Algorithm Description

To clarify the technical background and convergence proof of the algorithm, the following paragraphs are in order. For convenience we present the local mixture term in (5.2.6) by $l_\vartheta(\lambda)$.
In step (2a), t_j is tangent to Λ_ϑ^b at $\lambda^{(j-1)} = \lambda^\star$ and can be obtained as follows. If we collect the quartic $q(y)$ in (5.2.8) with respect to $\lambda_1^\star$, $\lambda_2^\star$, $\lambda_3^\star$ and $\lambda_4^\star$, we obtain the supporting plane with the normal vector $(1 - y^\star, y^{\star 2} - 2y^\star, -y^{\star 3} + 3y^{\star 2}, y^{\star 4} - 4y^{\star 3})$, where $y^\star$ is the real multiple root of $q(y)$.

In step (2b), $\Pi_j = I - N_j N_j^T$ presents the matrix of orthogonal projection onto tangent plane t_j, with respect to Euclidean inner product, in which I is the identity matrix.

Therefore, for g_j and H_j the gradient vector and hessian matrix of $l_\vartheta(\lambda)$ at $\lambda^{(j-1)}$, the first and second covariant derivatives are $\Pi_j g_j$ and $\Pi_j H_j$, respectively.

Step (2c), describes the so called exponential -also called retraction- mapping $\mathbb{R} : T\Lambda_\vartheta^b \to \Lambda_\vartheta^b$, where $T\Lambda_\vartheta^b$ represents the tangent bundle of Λ_ϑ^b, the disjoint union of t_j's (See Shub 1986). Let R_j be the restriction of R to t_j, then R_j is a one-to-one mapping that maps the vector $(\Pi_j H_j^{-1})(\Pi_j g_j) \in t_j$ to a curve between $\lambda^{(j-1)}$ and $R_j(\lambda^{(j)})$ on Λ_ϑ^b and holds the following assumptions,

1. R_j is defined in an open interval $U_{r_j}(0_j) \in t_j$, about 0_j of radius $r_j > 0$, where 0_j is the representation of $\lambda^{(j-1)}$ in t_j.
2. $R_j(\dot{\lambda}) = \lambda$ if and only if $\dot{\lambda} = 0_j$.
3. R is smooth and $DR_j(0_j) = id_{t_j}$, since $R_j(\lambda^{(j-1)}) = \lambda^{(j-1)}$.

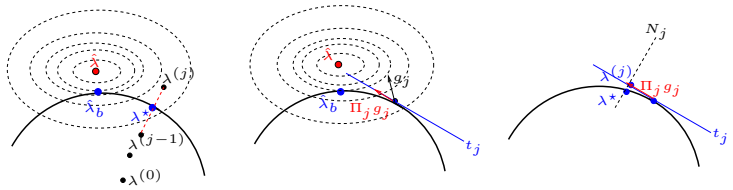

Figure 5.1: Schematic visualization of the algorithm steps

5.7.2 Proofs

Proof: (Lemma 5.2) For $k = 4$ (without loss of generality), $\Lambda_b(\vartheta)$ can be parametrized using the solution set of equations $q(y) = 0$, $q'(y) = 0$, as functions of λ_1, λ_2, λ_3 and λ_4 for all $y > 0$, which retain the locus of the intersecting line for any two consecutive supporting planes (Do Carmo 1976 ch.2). Direct calculation shows that $\Lambda_b(\vartheta)$ can be obtained by the following smooth mapping,

$$\mathcal{C} : (0, \infty) \times U \times V \to \mathbb{R}^4, \qquad (y, \lambda_1, \lambda_4) = [\lambda_1, \lambda_2(y, \lambda_1, \lambda_4), \lambda_3(y, \lambda_1, \lambda_4), \lambda_4]$$

where, $U, V \subset \mathbb{R}$ is an open interval and

$$\lambda_2(y, \lambda_1, \lambda_4) = \frac{(y^5 - 6y^4 + 12y^3)\lambda_4 + (2y^2 - 6y + 6)\lambda_1 - 3y + 6}{y(y^2 - 4y + 6)} \quad (5.7.12)$$

$$\lambda_3(y, \lambda_1, \lambda_4) = \frac{(2y^5 - 10y^4 + 16y^3)\lambda_4 + (y^2 - 2y + 2)\lambda_1 - 2y + 2}{y^2(y^2 - 4y + 6)} \quad (5.7.13)$$

Therefore, the implicit function theorem (Rudin, 1976, p.224) implies that $\Lambda_b(\vartheta)$ is a smooth 3-manifold.

In general, the boundary of parameter space of local mixture models may not be smooth manifolds; hence, in those situations either the possible singularity points must be characterized or other optimization approaches must be applied for finding maximum on boundary.

Proof: (Theorem 5.2) Consider the following two cases,
(I) $\hat{\lambda} \in \Lambda_\vartheta$
Since $l_\vartheta(\lambda)$, for any fixed ϑ, is concave and satisfy the second-order sufficient conditions, then step (1) of the algorithm converges to the unique global maximum $\hat{\lambda}$ in quadratic order, for any initial point $\lambda^{(0)}$ inside the interior of Λ_ϑ (see Nocedal & Wright 2006 p.45).
(II) $\hat{\lambda} \notin \Lambda_\vartheta$
Since Λ_ϑ is closed and convex in a finite dimensional vector space, there is a unique $\hat{\lambda}_b \in \Lambda_\vartheta$ with minimum distance from $\hat{\lambda}$, and consequently $l_\vartheta(\hat{\lambda}_b) \geq l_\vartheta(\lambda)$ for all $\lambda \in \Lambda_\vartheta$, since $l_\vartheta(\lambda)$ is concave. Moreover, the vector $\hat{\lambda}_b\hat{\lambda}$ is orthogonal to the supporting plane t_b, tangent to Λ_ϑ at $\hat{\lambda}_b$; hence, $(\Pi_b g_b)$ is a zero vector in the tangent vector space t_b.
In addition, according to Lemma 2, $\Lambda_b(\vartheta)$, is an Immersed manifold in $\mathbb{R}^k$. According to notations in Shub (1986), step 2 can be presented by the following mapping

$$\begin{aligned} \mathcal{S} &: \Lambda_b(\vartheta) \to \Lambda_b(\vartheta) \\ \lambda^{(j-1)} &\to R_j\left(\lambda^{(j-1)}, (\Pi_j H_j^{-1})(\Pi_j g_j)\right) \end{aligned} \quad (5.7.14)$$

where, by condition (3), $\mathcal{S}$ is smooth. Also, if $(\Pi_j H_j^{-1})$ exists then by conditions (1) and (2), the fixed points of $\mathcal{S}$ (i.e, $\mathcal{S}(\lambda) = \lambda$) are the zeros of the covariant gradient, and at fixed points the derivative of $\mathcal{S}$ vanishes.

Chapter 6

Discussion and Future Work

6.1 Discussion

The earlier chapters of this thesis exploit geometric tools along with the useful geometric and inferential properties of LMMs to develop novel methodologies for dealing with non-standard statistical problems. Flexibility, identifiability and interpretability of LMMs allow for building large and flexible classes of models which are naturally defined for a wide range of statistical models. Such flexible models are applied to build a perturbation space in Chapter 4 which includes the base prior model as a member, and to model lifetime data with an unobservable frailty variable in Chapter 5, where the the hidden information is extracted by learning the parameters of a LMM. In addition, because of the geometric properties, such as linearity and convexity, fast and efficient algorithms are designed for estimating the parameters.

The major cost to all these incredible properties is the fact that the parameter space includes two types of boundaries, which affect existence of MLE or its nice asymptotic properties. We show in Chapter 3 how the geometric properties and underlying structure assist in computing these types of boundaries by smooth or non-smooth objects.

Chapters 2-5 represent the four distinct papers as contributions of the thesis. However, there are still possible generalizations and extensions to some of the papers. For instance,

two possible computational problems, equally important for application purposes, are extending the computational methodology in Chapters 4 and 5 to the discrete mixture of LMMs with more than one component. Essentially, such computational tools can be developed by combining the theories and computational algorithms in Chapter 2 with the methodology developed in Chapters 4 and 5.

6.2 Future Work

6.2.1 Hypothesis Testing for Mixture of LMMs

The discrete mixture of local mixture models (DMLMM), introduced in Chapter 2, is shown to be identifiable, estimable, flexible, and having well-defined number of components. We also characterized their boundaries in Chapter 3 and proposed a fast numerical algorithm for their parameter estimation in Chapter 2. One major difficulty in working with mixture models is related to their lake of identifiability which influences any inference and hypothesis testing on this family, particularly for determining the number of components of a mixture model. One consequence is that the distribution of the test statistic usually has a complicated form, if a closed form is attainable (Hall and Stewart, 2005; Li and Chen, 2010; Maciejowska, 2013).

The geometric and inferential properties of DMLMM, as well as their ability to approximate the general family of mixture models, motivate alternative approaches for solving these problems. The goal is to provide methodologies for hypothesis testing, inference, and modeling mixture data which are significantly less complicated both mathematically and computationally. We have seen in Chapter 2 that a fast and efficient version of EM algorithm does the computational task. Regarding to the theory, the motivation lies within the geometric properties similar to those of the exponential family, affine property and convexity. The exponential family is flat with respect to $+1$-geometry and has a convex natural parameter space; consequently it has a clean asymptotic theory. Similarly, a local mixture model is a union of convex subspaces of a -1-affine space, flat with respect to -1-geometry (Section 1.6).

Another possible direction can be exploiting DMLMMs for clustering. Finite mixture models are popular tools for clustering. Essentially, in this method a dataset, which is a mixture of many subclasses of data, is fitted by a mixture model, which requires estimating the mixing proportions and the component parameters.Then the individuals are assigned to suitable clusters according to the estimated parameters. Finite mixture of normal models and also finite mixture of t-student distributions are commonly used models, and the parameters can be estimated via likelihood maximization or using a Bayesian methodology. Since, each component of a DMLMM is a naturally defined flexible model, and because of the identifiability and estimability of this model it can be a promising tool for this purpose. Also, as it has clear geometry such as linearity and convexity, both likelihood maximization and Bayesian methods, such as MCMC, are significantly straightforward and efficient.

6.2.2 Inference on Log-linear and Graphical Models

Log-linear models, 1.7, are the most popular statistical models for analyzing categorical data. They have applications in a wide range of areas of science and engineering, including, social and biological sciences, data mining, manufacturing, image and language precessing. Specifically, they have been increasingly applied to analyzing sparse contingency tables in the last two decades. Also, graphical models, as their special cases, are widely applied to network analysis, machine learning, and latent analysis.

Despite their popularity, inference on these models is extremely challenging, except for their special forms; for example, decomposable log-linear models. Their inference problems are essentially constrained optimization problems where the difficulty is due to existence of high-dimensional non-trivial boundaries. As a result, the maximum likelihood estimate (MLE) for the natural parameters are not attainable. To work around this issue it is common to estimate the expected parameters which are related to the natural parameter by the design matrix. The space of these parameters, called marginal polytope (or cone), is either a convex polytope, a convex cone, or a combination of both, depending on the sampling scheme. Therefore, the inference problem is reduced to obtaining the MEL, which exists when the observed sufficient statistics lie in the interior of the marginal polytope, or obtaining the boundary point which maximizes the loglikelihood function. Because

of existing possible sampling zeros in contingency tables, the boundaries are extremely difficult to approximate, since they have redundant faces, and maximization algorithms are computationally challenging.

In Chapter 3 we develop two methods of approximating boundaries, by smooth surfaces or by simpler polytopes. We are aiming to extend our methods and use differential and convex geometric tools for approximating the boundaries of marginal polytopes. The initial difficulty is developing a methodology to eliminate the redundant faces which do not contribute to the boundary, and then finding the full dimensional representation of the polytope. The next step will be employing smoothing, or polytope approximation methods. Finding a smooth approximation for a full dimensional polytope, which is a pure geometrical problem, is not hard; however, since marginal polytopes are of high dimensions we want to find the one on which we can apply efficient searching algorithms. We have designed such an algorithm for finding MLE in models called local mixture models, in our earlier works. Also, one possible approach for approximating a polytope with a simpler one, is to sample form its boundary and construct the simpler polytope by convex hull of the selected points. This approach is expected to simplify the computational difficulty of inference in log-linear models to a great extent

6.2.3 Over-dispersion in Count data

The Poisson distribution is commonly used for modeling count data and analyzing contingency tables. However, count data rarely have equal mean and variance; hence, the Poisson assumption is doubtful. They often have larger variance (over-dispersed) due to unobservable variations, and possibly – but less likely – have smaller variation (under-dispersed) as a result of clustering, rounding or censoring. A common approach for dealing with this problem is to use a mixture of Poisson distributions with a latent variable which is responsible for the extra structure. However, there are two issues with this approach. First, it is only capable of modeling over-dispersion; second, the latent distribution is unknown, and is required be estimated or postulated.

We propose local mixture models (LMM) of Poisson distribution which are naturally defined and have flexible moment structure. Under LMM, the problem of estimation of

the latent distribution is reduced to inference about a finite set of parameters, which represent the latent model via its moments. In earlier chapters we have shown that LMMs are sufficiently flexible in modeling complex data and they are, in some sense, richer than mixture models. The goal is to design a hypothesis test for assessing over(under)-dispersion in count data. The advantages of this test are two folded. First, the null hypothesis parameter subspace lies in the interior of the parameter space. Second, we conjecture that, compared to the existing approaches, deriving the asymptotic distribution of the test statistic is less complicated, since LMMs have similar geometric properties to those of the exponential family.

References

Abbring, J. H. and G. J. Van Den Berg (2007). The unobserved hetreogeneity distribution in duration analysis. *Biometrika 94*(1), 87–99.

Agresti, A. (1990). *Categorical Data Analysis* (Second ed.). John Wiley & Sons.

Alexandrov, A. D. (2005). *Convex polyhedra*. Springer Science and Business Media.

Amari, S. I. (1985). *Differential-Geometrical Methods in Statistics: Lecture Notes in Statistics*. New York: Springer-Verlag.

Amari, S. I. (1990). *Differential-Geometrical Methods in Statisitics: Lecture Notes in Statisitics* (Second ed.). New York: Springer-Verlag.

Anaya-Izquierdo, K. (2006). *Statistical and Geometrical Analysis of Local Mixture Models and a Proposal of Some New Tests of Fit with Censored Data*. Ph. D. thesis, Facultad de Ciencias Universidad Nacional Autnoma de Mexico.

Anaya-Izquierdo, K., F. Critchley, and P. Marriott (2013a). Logistic regression geometry. *arXiv preprint arXiv:1304.1720*.

Anaya-Izquierdo, K., F. Critchley, and P. Marriott (2013b). When are first order asymptotics adequate? a diagnostic. *Stat 3*(1), 17–22.

Anaya-Izquierdo, K. and P. Marriott (2007a). Local mixture models of exponential families. *Bernoulli 13*, 623–640.

Anaya-Izquierdo, K. and P. Marriott (2007b). local mixture of the exponential distribution. *Annals of the Insti. Math. Statist. 59*(1), 111–134.

Bandy, M. L. (1966). A theorem on positive quartic forms. *The American Mathematical Monthly 73*(8), 864–866.

Barnard, S. and J. M. Child (1936). *Higher Algebra*. Macimillian and co.

Barndorff-Nielsen, O. E. (1986a). Likelihood and observed geometries. *The Annals of Statistics 14*(3), 856–873.

Barndorff-Nielsen, O. E. (1986b). Strings, tensorial combinants, and Bartlett adjustments. *Proceedings of the Royal Society of London. Series A 406*, 127–137.

Barndorff-Nielsen, O. E. (1987a). Differential and integral geometry in statistical inference. *Institute of Mathematical Statistics, Hayward, Ca.*, 95–162.

Barndorff-Nielsen, O. E. (1987b). Differential geometry and statisitics: Some mathematical aspects. *Indian Journal of Math 29*, 335–350.

Barndorff-Nielsen, O. E. (1988). *Parametric Statistical Models and Likelihood*. Springer, London.

Barndorff-Nielsen, O. E. and D. R. Cox (1989). *Asymptotic Techniques for use in Statistics*. Chapman and Hall.

Barndorff-Nielsen, O. E., D. R. Cox, and N. Reid (1986). The role of differential geometry in statistical theory. *International Statistical Review 54*(1), 83–96.

Barvinok, A. (2013). Thrifty approximations of convex bodies by polytopes. *International Mathematics Research Notices rnt078*.

Batyrev, V. V. (1992). Toric varieties and smooth convex approximations of a polytope. *RIMS Kokyuroku 776*, 20.

Berger, J. O., D. Rios Insua, and F. Ruggeri (2000). Bayesian robustness. *Robust Bayesian Analysis*, 1–32.

Berger, M. (1987). *Geometry I.* Springer.

Blyth, S. (1994). Local divergence and association. *Biometrika 81*(3), 579–584.

Bonnesen, T. and W. Fenchel (1987). *Theory of Convex Bodies.* BCS Associatee, Moscow, Idaho.

Boroczky, K. and F. Fodor (2008). Approximating 3-dimensional convex bodies by polytopes with a restricted number of edges. *Contributions to Algebra and Geometry 49*(1), 177–193.

Breslow, N. (1972). Contribution to the discussion on the paper of Cox (1972). *Biom 34*, 216–217.

Brown, L. D. (1986). *Fundamentals of Statistical Esponential Families: with Application in Statistical Decision Theory.* Institute of Mathematical Statistics.

Celeux, G. (2007). Mixture models for classification. *Advances in Data Analysis Springer Berlin Heidellberg*, 3–14.

Chen, J. and J. D. Kalbfleisch (1996). Penalized minimum-distance estimates in finite mixture models. *The Canadian Journal of Statistics 24*(2), 167–175.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika 65*(1), 141–151.

Clayton, D. G. and J. Cuzick (1985). Multivariate generalizations of the proportional hazard model. *Journal of the Royal Statistical Society A 148*(2), 82–117.

Cook, D. R. (1986). Assessment of local influence. *Journal of the Royal Statistical Society* (2), 133–169.

Copas, J. and S. Eguchi (2001). Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(4), 871–895.

Copas, J. and S. Eguchi (2010). Likelihood for statistically equivalent models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(2), 193–217.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society B 34*, 187–200.

Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society 49*(1), 1–39.

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

Critchley, F. and P. Marriott (2014). Computational information geometry: Theory and practice. *Entropy 16*, 2454–2471.

Critchley, F., P. Marriott, and M. Salmon (1993). Preferred point geometry and statistical manifolds. *The Annals of Statistics 21*(3), 1197–1224.

Critchley, F., P. Marriott, and M. Salmon (1994). Preferred point geometry and the local differential geometry of the kullback-leibler divergence. *The Annals of Statistics 22*(3), 1587–1602.

Csiszar, I. and F. Matus (2005). Closure of exponential families. *The Annals of Probability 33*(2), 582–600.

Culter, A. and Windham (1994). Information-based validity functionals for mixture analysis. *In Proceedings of the First US/Japan Conference on the frontires of Statistical Modeling in Informational Approach Amsterdam: Kluwer*, 149–170.

Dieker, A. B. and S. Vempala (2014). Stochastic billiards for sampling form boundary of a convex set. *arXiv:1410.5775*.

Do Carmo, M. P. (1976). *Differential Geometry of Curves and Surfaces*, Volume 2. Prentice-Hall Englewood Cliffs, NJ.

Dodson, C. T. J. and T. Poston (1979). *Tensor Geometry: the geometric viewpoint and its uses*, Volume 130. Springer.

Donoho, D. L. (1988). One-sided inference about functionals of a density. *Annals of statistics 16*, 1390–1420.

Edwards, D. (2000). *Introduction to Graphical Modeling* (second ed.). Springer.

Efron, B. (1975). Defining the curvature of a statistical problem (with application to second order efficiency). *The Annals of Statistics 3*(6), 1189–1242.

Eguchi, S. (1983). Second order efficiency of minimum contrast estimatores in a curved exponential family. *The Annals of Statistics 11*, 793–803.

Eguchi, S. (1985). A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Math. Part A 38*, 385–398.

Eguchi, S. (1991). A geometric look at nuisance parameter effects of local powers in testing hypothesis. *Annals of the Institute of statistical mathematics 43*, 245–260.

Eleber, C. and G. Ridder (1982). True and spurious duration dependence: The identifiability of the proportional hazard model. *Review of Economic Studies 49*, 403–409.

Eriksson, N., S. E. Fienberg, A. Rinaldo, and S. Sullivant (2006). Polyhedral conditions for the nonexistence of the mle for hierarchical log-linear models. *Journal of Symbolic Computation 41*, 222–233.

Everitt, B. S. (1996). An introduction to finite mixture distributions. *Statistical Methods in Medical Research 5*(2), 107–127.

Everitt, B. S. and D. J. Hand (1981). *Finite Mixture Distributions*. Chapman and Hall.

Fienberg, S. E. (1968). The geometry of an r by c contingency table. *The Annals Mathematical Statistics 39*, 1186–1190.

Fienberg, S. E. and J. P. Gilbert (1970). The geometry of a two by two contingency table. *Journal of the American Statistical Association*, 694–701.

Fienberg, S. E. and A. Rinaldo (2012). Maximum likelihood estimation in log-linear models. *The Annals of Statistics 40*(2), 996–1023.

Fuchs, H. J., B. D. S., D. H. Christiansen, E. M. Morris, M. L. Nash, B. W. Ramsey, B. J. Rosenstein, A. L. Smith, and M. E. Wohl (1994). Effect of aerosolized recombinant human dnase on exacerbations of respiratory symptoms on pulmonary function in paitients with cystic fibrosis. *New Eng. J. Med 331*, 637–643.

Fukuda, K. (2004). From the zonotope construction to the minkowski addition of convex polytopes. *Journal of Symbolic Computation 38*(4), 1261–1272.

Gan, L. and J. Jiang (1999). A test for global maximum. *Journal of the American Statistical Assiciation 94*(447), 847–854.

Geiger, D., C. Meek, and B. Sturmfels (2006). On the toric algebra of graphical models. *The Annals of Statistics 34*(3), 1463–1492.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2007). *Replication data for: Bayesian Data Analysis* (Second ed.).

Geyer, C. J. (2009). Likelihood inference in exponential familes and direction of recession. *Electronic Journal of Statistics 3*, 259–289.

Ghomi, M. (2001). Strictly convex submanifolds and hypersurfaces of positive curvature. *Journal of Differential Geometry 57*(2), 239–271.

Ghomi, M. (2004). Optimal smoothing for convex polytopes. *Bulletin of the London Mathematical Society 36*(4), 483–492.

Gorfine, M., D. M. Zucher, and L. Hsu (2006). Prospective survival analysis with general semiparametric shared frialty model: A pseudo full likelihood approach. *Biometrika 93*(3), 735–741.

Gustafson, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics 24*(1), 174–195.

Hall, P. and M. Stewart (2005). Theoretical analysis of power in a two-componet normal mixture model. *Journal of Statistical Planning and Inference 134*, 158–179.

Horowitz, J. L. (1999). Semiparametric estimation of a proportional hazard model with unobserved hetrogeneity. *Econometrica 67*(5), 1001–1028.

Horowitz, J. L. (2010). *Semiparametric and Nonparametric Methods in Econometrics*. Springer.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika 71*(1), 75–83.

Hougaard, P. (1986). Survival models for heterogeneus population derived from stable distributions. *Biometrika 73*(2), 387–396.

Hsu, L., L. Chen, M. Gorfine, and K. Malone (2004). Semiparametric estimation of marginal hazard function from case-control family studies. *Biometrics 60*, 936–944.

Hu, T., B. Nan, X. Lin, and J. M. Robbins (2011). Time-dependent cross ratio estimation for bivariate failure times. *Biometrika 98*(2), 341–354.

Insua, D. R. and F. Ruggeri (2000). *Robust Bayesian Analysis*. New York: Springer.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences 186*, 453–461.

Johansen, S. (1983). An extenssion of Cox's regression model. *International Statistical Review 51*, 258–262.

Kass, R. E. and P. W. Vos (1997). *Geometrical Foundations of Asymptotic Inference*. John Wiley and sons.

Klein, J. P. (1992). Semiparametric estimation of random effects using cox model based on the em algorithm. *Biometrics 48*, 795–806.

Lavine, M. (1991). The prior and the likelihood. *Journal of the American Statistical Assiciation 86*, 396–399.

Lawless, J. F. (1981). *Statistical Models and Methods for Lifetime Data*. Wiley.

Leorox, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics 20*(3), 1350–1360.

Li, P. and J. Chen (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association 105:491*, 1084–1092.

Li, P., J. Chen, and P. Marriott (2009). Non-finite fisher information and homogeneity: an EM approach. *Biometrika*, 1–16.

Linde, V. D. A. (2007). Local influence on posterior distributions under multiplicative models of perturbation. *Bayesian Analysis 2*(2), 319–332.

Lindsay, B. G. (1993). Uniqueness of estimation and identifiability in mixture models. *Canadian Journal of Statistics 21*(2), 139–147.

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

Lopez, M. and S. Reisner (2008). Hausdorff approximation of 3d convex polytopes. *Information Processing Letters 107*(2), 76–82.

Maciejowska, K. (2013). Assessing the number of components of a normal mixture: an alternative appraoch. *University Library of Munich* (No. 50303).

Malago, L. and G. Pistone (2010). A note on the border of an exponential family. *arXiv preprint arXiv:1012.0637*.

Marriott, P. (1989). *Applications of Differential Geometry to Statistics.* Ph. D. thesis, University of Warwick.

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

Marriott, P. (2003). On the geometry of measurement error models. *Biometrika 90*(3), 567–576.

Marriott, P. (2006). Extending local mixture models. *AISM 59*, 95–110.

Marriott, P. and M. Salmon (2000). *Applications of Differential Geometry to Econometrics*. Cambridge University Press.

Marriott, P. and P. W. Vos (2004). On the global geometry of parametric models and information recovery. *Bernoulli 10*(4), 639–649.

Marsaglia, G. (1972). Choosing a point from the surface of a sphere. *The Annals of Mathematical Statistics 43*(2), 645–646.

Martinussen, T., T. H. Scheike, and D. M. Zucker (2011). The Aalen additive gamma frailty hazard model. *Biometrika 98*(4), 831–843.

Matousek, J. (2002). *Lectures on Discrete Geometry*, Volume 212. Springer.

McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall.

McCulloch, R. E. (1989). Local model influence. *Journal of the American Statistical Association 84*(406), 473–478.

Mclachlan, G. and B. Kaye (1988). *Mixture Models: inference and applications to clustering*. New York, N.Y. :M.Dekker.

Mclachlan, G. and D. Peel (2000). *Finite Mixture Models*. John Wiley and sons.

Morris, C. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics 10*(1), 65–80.

Murray, M. K. and J. W. Rice (1993). *Differential Geometry and Statistics*. Chapman and Hall.

Nan, B., X. Lin, L. D. Lisabeth, and S. Harlow (2006). Piecewise constant cross ratio estimation for association of age at a marker event age at a menopause. *Journal of the American Statistical Association 101*, 65–77.

Narayanan, H. and P. Niyogi (2008). Sampling hypersurfaces through diffusion. *Approximation, Randomization and Combinatorial Optimization Algorithms and Techniques*, 535–548.

Nielsen, G. G., R. D. Gill, P. K. Anderson, and T. I. A. Sorensen (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics 19*(1), 25–43.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society B 44*(3), 414–422.

Peng, J., T. Hazan, N. Srebro, and J. Xu (2012). Approximate inference by intersecting semidefinite bound and local polytope. *Proceedings of the 15th International Conference at Artificial Intelligence and Statistics. Canaries Spanish: JMLR W & CP*, 868–876.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc. 37*, 81–91.

Richardson, S. and P. J. Green (1997). On bayesian analysis of mixtures with an unknown number of components (with discuassion). *Journal of the Royal Statistical Society B 59*, 731–792.

Rinaldo, A., S. E. Fienberg, and Y. Zhou (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics 3*, 446–484.

Rudin, W. (1976). *Principles of Mathematical Analysis* (3rd ed.). New York: McGraw-Hill.

Ruggeri, F. and S. Sivaganesan (2000). On a global sensiticity measure for bayesian inference. *Sankhya 62*, 110–127.

Ruggeri, F. and L. Wasserman (1993). Infinitesimal sensitivity of posterior distributions. *The Canadian Journal of Statistics 21*(2), 195–203.

Schlattmann, P. (2009). *Medical Applications of Finite Mixture models*. Springer.

Shun, Z. and P. Mccullagh (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal statistical Society. Series B (Methodological) 57*(4), 749–760.

Sontag, D. and T. Jaakkola (2007). New outer bounds on the marginal polytopes. *Advances in Natural Information Processing 20*, 1393–1400.

Struik, D. J. (1988). *Lectures on Classical Differential Geometry*. Dover Publications.

Sun, M. and E. Fiume (1996). A technique for constructing developable surfaces. *Graphics Interface*, 176–185.

Tallis, G. (1969). The identifiability of mixtures of distributions. *Journal of Applied Probability 6*(2), 389–398.

Ulrich, G. and L. T. Watson (1994). Positivity conditions for quartic polynomials. *AIAM J. Sci. Comput. 15*, 528–544.

Vaupel, J. W., K. G. Manton, and E. Stallard (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography 16*, 539–454.

Wainwright, M. J. and M. I. Jordan (2006). Log-determinat relaxation for approximate inference in discrete Markov random fields. *IEEE Transactions on Signal Processing 54*, 2099–2109.

Weiss, R. (1996). An approach to Bayesian sensitivity analysis. *Journal of the Royal Statistical Society 58*(4), 739–770.

Zeng, D., Q. Chen, and J. Ibrahim (2009). Gamma frailty transformation models for multivariate survival times. *Biometrika 96*(2), 277–291.

Zhu, H., J. G. Ibrahim, and N. Tang (2011). Bayesian influence analysis: a geometric approach. *Biometrika 98*(2), 307–323.