ABSTRACT

Rates of Convergence in Semiparametric Mixture Models

Hemant Ishwaran
Yale University
1993

The semiparametric mixture model has density

$$f(x \mid \theta, Q) = \int f(x \mid \theta, \eta)\, dQ(\eta),$$

where θ is a real vector, sometimes referred to as the structural parameter, and Q is an unknown distribution commonly referred to as the mixing distribution. The thesis considers the problem of estimation for θ with Q playing the role of a nuisance parameter.

The problem of estimation for θ in the semiparametric mixture model has generally focused on the question of efficiency. Van der Vaart (1988) and Pfanzagl (1990) show that when the type of mixing behavior is constrained it is possible to state efficiency results as local asymptotic minimax (LAM) theorems and convolution theorems. These results implicitly presuppose the existence of a $O_p\left(n^{-1/2}\right)$ estimator for θ.

However, it is not always clear that the structural parameter is estimable at a $O_p\left(n^{-1/2}\right)$ rate. Carroll and Hall (1988) and Zhang (1990) show that when θ is known, the mixing distribution in a location mixture model is typically estimable only at very slow rates. It is possible, therefore, for the difficulty in estimating the mixing distribution to create problems, as well, in the estimation for θ.

Rates of convergence for estimators of θ are defined in a locally uniform sense to incorporate a minimax approach to estimation. Le Cam (1973) and Donoho and Liu (1987) describe a general approach for determining lower bounds for uniform rates of convergence which is adapted to the mixture setting.

The thesis presents a general class of mixture models for which the structural parameter can only be estimated at rates slower than $O_p\left(n^{-1/2}\right)$. A new Fourier technique is used to determine explicit lower bounds for rates of convergence in location-mixture models, where estimation is for an unknown scale parameter. The theory is illustrated by application to two well known models: the mixture model formed by constrained mixing over the mean of a normal density with unknown variance, and the Weibull mixture model as studied by Heckman and Singer (1984).

Rates of Convergence in Semiparametric Mixture Models

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Hemant Ishwaran

May, 1993

Acknowledgements

I wish to extend my humblest thanks to David Pollard for his generous gifts of time (especially time spent listening to my arguments on the little blackboard), for his encouragement, and for his many subtle comments that have greatly influenced and shaped this research. His advice on matters of writing and his careful reading of many early drafts have led to a much more presentable and readable piece of work. I am also grateful for his technical assistance concerning TEX matters and for allowing me to use the wonderful "Pollard.sty" output routine. It made the mathematical typsetting for the thesis so simple and so nice.

I am grateful to Kathryn Roeder who served as my advisor for one term and for her role as a reader. Both Kathryn and reader Bin Yu helped with several useful suggestions. Thanks are also extended to Andrew Barron, who pointed out an interesting fact about Fourier transforms which helped to make some of the work in Chapters 4 and 5 easier.

Richard Savage gave me my first glimpse of real statistics. He also gave me important advice that helped guide me through difficulties which I encountered during my first two years. His genuine concern was greatly appreciated and not forgotten. Another person who could never be forgotten is John Hartigan. His bizarre sense of humour and charisma made the classes he taught entertaining as well as educational.

I also thank all the faculty, student, staff, and family who helped to make my experiences at 24 Hillhouse so rich and rewarding. I cannot thank everyone by name, but special mention must be given to: Barbara Amato, Joseph Chang, Surya Mohanty, and Robert Sherman.

Finally, I thank my father and mother for the emotional support and the encouragement they have given me throughout the course of this research. They gave me strength when I badly needed it, and I am so grateful to them.

This research was parially funded by NSF Grant DMS 9102286.

Table of Contents

Notation

The linear functional notation for expectation is usually employed. For example, the expected value of a function g with respect to a probability measure P is written as Pg rather than the usual convention of $\int g(x)\,dP(x)$. One exception is that we write $\int g(x)\,dx$ for the integral of g with respect to Lebesgue measure.

The function $f(x \mid \theta, \eta)$ is used to describe a real valued density in x, indexed by two real parameters θ and η. The parameter set for θ is denoted by Θ, and may be k-dimensional, while the parameter set for η is always one-dimensional and denoted by $\mathcal{N}$. We assume that $f(x \mid \theta, \eta)$ is measurable as a function of (x, η), so that the mixture

$$f(x \mid \theta, Q) = \int f(x \mid \theta, \eta)\, dQ(\eta)$$

formed by the mixing distribution, Q, is well defined (we adopt the conventional form for expectation when describing mixtures). Denote the corresponding probability measure by $P_{\theta,Q}$, which is sometimes referred to as the mixed distribution.

Symbols which are assignable to letters (followed by the convention with which they are usually employed):

$\mathcal{B}(\mathbb{R})$	Borel σ-algebra of the real line
(D, d)	metric space D with metric d
δ_n	sequence converging to zero
η	nuisance parameter
f_0	true mixed density
f_τ	perturbed mixed density
$f(x \mid \theta, \eta)$	real valued density with real parameters (θ, η)
$f(x \mid \theta, Q)$	mixed density with real parameter θ and mixing distribution Q
$\mathcal{F}(\Theta, \mathcal{N})$	parametric family of densities
Γ	gamma function
$H(P_1, P_2)$	Hellinger distance between probability measures P_1 and P_2
i	$\sqrt{-1}$
$\kappa(P)$	functional mapping a probability measure P onto a metric space (D, d)
$L_0^p(P)$	equivalence class of of p-integrable functions with zero expectation with respect to probability measure P
$\mathcal{L}_0^\infty(P)$	class of essentially bounded functions that have zero P-expectation
μ	Lebesgue measure

$N(0,1)$	standard normal distribution
$\mathcal{N}$	parameter set for η
$\mathcal{N}(\theta)$	natural parameter space at θ
P	probability measure
P^n	n-fold product measure for P
P_0	true probability measure
P_τ	perturbed probability measure
$\mathcal{U}$	topology on $\mathcal{P}$ induced by the total variation distance
$\mathcal{U}(P_0, \alpha_n)$	$O(\alpha_n)$ shrinking neighborhood of a probability measure P_0
$P_{\theta,Q}$	mixed distribution with real parameter θ and mixing distribution Q
$\mathcal{P}$	class of probability measures
$\mathcal{P}(\Theta, \mathcal{Q})$	class of mixed distributions
q	mixing density
$q^{(m)}$	m^{th} derivative for the mixing density q
q_0	true mixing density
$q(\cdot, \tau)$	perturbed mixing density
Q	mixing distribution
Q_0	true mixing distribution
Q_τ	perturbed mixing distribution
$\mathcal{Q}$	class of mixing distributions
$\mathbb{R}^k$	k-dimensional real space
τ	small real number
θ	structural parameter
θ_0	true structural parameter
Θ	parameter space for θ
ν	σ-finite measure dominated by Lebesgue measure
$v(P_1, P_2)$	total variation distance between two probability measures P_1 and P_2
$(\mathcal{X}, \mathcal{A})$	measurable space with set $\mathcal{X}$ and σ-algebra $\mathcal{A}$

Symbols not assignable to letters (followed by the convention with which they are usually employed):

$*$	$g * h$ denotes the convolution of two Lebesgue integrable functions g and h
$\ll$	measure R dominates measure P: $P \ll R$

$\in$	element of
$\widehat{\ }$	$\widehat{g}$ denotes the Fourier transform for Lebesgue integrable g
$\wedge$	$a \wedge b =$ minumum (a, b)
$\subset$	proper subset
$\subseteq$	subset

Chapter 1

Introduction

1. Semiparametric Mixture Models The semiparametric mixture model has density

<1.1> $$f(x \mid \theta, Q) = \int f(x \mid \theta, \eta)\, dQ(\eta),$$

where θ is a real vector, sometimes referred to as the structural parameter, and Q is an unknown distribution commonly referred to as the mixing distribution.

When θ is known, the model <1.1> is completely nonparametric and interest focuses on estimation for the unknown mixing distribution, Q.

Carroll and Hall (1988) address the problem of determining rates of convergence in deconvolving densities. Their results, therefore, pertain to the problem of estimating the mixing density in location mixture models. What they find is that the smoother the known location density, the slower the optimal rate of pointwise estimation for the unknown mixing density. For example, in the normal mean-mixture model they find optimal rates which are logarithmic, while in the double-exponential mixture model they find geometric optimal rates.

Zhang (1990) also considers the problem of estimation for the mixing density (and distribution) in the location mixture model. He gives lower bounds for rates of convergence for the mixing density in the form of pointwise results and integrated mean square error and shows that the rates are related to the tails of the characteristic function of the known location density, with slower rates being found for more rapidly decreasing tails. By considering mixing densities whose Fourier transforms decrease on the order of $O\left(1/t^2\right)$, he finds a mean square error lower bound in the normal mean-mixture model of $(\log n)^{-1/2}$, while in the double exponential mixture model the bound is $n^{-1/6}$.

Therefore, estimating the mixing distribution can sometimes be difficult.

The question that we will pursue in this thesis is whether the difficulty in estimation for the mixing distribution creates problems for estimation of the structural parameter. Therefore, we will consider the problem of estimation for the structural parameter, with the mixing distribution acting as nuisance. The problem is motivated by the following examples.

<1.2> **Example (normal mean-mixture model):** The semiparametric normal mean-mixture model can be written in the form <1.1> by integrating over the mean of a normal density with unknown standard deviation. The model can also be more conveniently

expressed as the convolution $\theta Z + Y$, where Z has a standard $N(0,1)$ distribution, θ is the unknown standard deviation, and Y has unknown distribution Q independent of Z. Determining θ in this model corresponds to estimating the common standard deviation amongst various normal populations which are heterogeneous in their means.

Roeder (1990) uses such a model to investigate a clustering hypothesis in astronomy. The data consist of velocities at which galaxies in the Corona Borealis region are receding from us. According to the Big Bang theory of cosmology, the further the galaxy is from us, the faster it is receding. If galaxies are clustered, the data should be multimodal generated with modes corresponding to clusters of galaxies and could be be modeled as realizations from a normal mean-mixture. Under this hypothesis, the unknown standard deviation corresponds to the tightness of the clustering in the galaxies and is of cosmological interest. □

<1.3> **Example (Weibull mixture model):** Heckman and Singer (1984) study economic theories concerning continuous durations of occupancy of states. To test such theories and to estimate structural parameters, they propose the use of a semiparametric mixture model to account for population heterogeneity in unobserved variables. They assume a Weibull functional form for the hazard function

$$h(x \mid \mathbf{z}, \theta, \boldsymbol{\alpha}, \eta) = \exp(\mathbf{z}\boldsymbol{\alpha})\, \theta x^{\theta-1} \eta,$$

where x is the observed positive duration time, $\mathbf{z}$ is a vector of time invariant observed covariates independent of the positive heterogeneity component η, and $(\theta, \boldsymbol{\alpha})$ is a vector of real parameters to be estimated. The authors model the unobserved η as a random variable with unknown distribution, Q, and propose the Weibull mixture model

<1.4> $$f(x \mid \mathbf{z}, \theta, \boldsymbol{\alpha}, Q) = \int \theta x^{\theta-1} \eta \exp\left(\mathbf{z}\boldsymbol{\alpha} - \eta x^{\theta} \exp(\mathbf{z}\boldsymbol{\alpha})\right)\, dQ(\eta)$$

as a device for modeling duration data. The hazard function for the mixture <1.4> takes the form

$$h(x \mid \mathbf{z}, \theta, \boldsymbol{\alpha}, Q) = \exp(\mathbf{z}\boldsymbol{\alpha})\, \theta x^{\theta-1} \psi(x, \mathbf{z} \mid, \theta, \boldsymbol{\alpha}, Q),$$

where

$$\psi(x \mid \mathbf{z}, \theta, \boldsymbol{\alpha}, Q) = \frac{\int \eta \exp\left(-\eta x^{\theta} \exp(\mathbf{z}\boldsymbol{\alpha})\right)\, dQ(\eta)}{\int \exp\left(-\eta x^{\theta} \exp(\mathbf{z}\boldsymbol{\alpha})\right)\, dQ(\eta)}.$$

Notice that the hazard function is not a proportional hazard as discussed in Cox (1972), so that the usual conditioning argument employed to estimate $\boldsymbol{\alpha}$ will not work here.

The authors use a nonparametric maximum likelihood estimator (NPMLE) as a means for estimating $(\theta, \boldsymbol{\alpha}, Q)$ and verify conditions that Kiefer and Wolfowitz (1956) prove to be sufficient for establishing the consistency of the maximum likelihood estimator in general

semiparametric mixture models (cf. Wald, 1949). Heckman and Singer also state results from simulations indicating the difficulty in estimation for the unknown distribution, Q. They remark:

> "A limited set of Monte Carlo experiments was conducted to evaluate the performance of the estimator. The NPMLE estimated the parameters of the structural duration model very well for samples as small as 500. Estimation of the distribution of the unobservables was less successful."

They note that the same phenomena is observed for both finite and continuous mixing distributions. □

2. Estimation for the Structural Parameter Let us first consider the parametric problem created when the mixing distribution is known and interest is in estimation for the unknown structural parameter. Let P_{θ_0} equal the true distribution with structural parameter θ_0 and known mixing distribution Q_0. Under standard regularity conditions we should expect the maximum likelihood estimator for θ_0 to be $\sqrt{n}$-consistent. Typically it should be asymptotically normal with expectation θ_0 and with variance equal to the reciprocal of the Fisher information for θ_0, where the Fisher information is defined as

<1.5> $$I(\theta_0, Q_0) = P_{\theta_0}\left(\frac{\partial}{\partial\theta}\log f(x \mid \theta_0, Q_0)\right)^2.$$

By the Cramér-Rao inequality, this implies that the asymptotic variance of the maximum likelihood estimator equals the smallest variance possible amongst unbiased estimators for θ_0. One might naively be led to believe that $I(\theta_0, Q_0)^{-1}$ provides a lower bound for the asymptotic variance for all $O_p(n^{-1/2})$ estimators for θ_0. The assertion, as is well known, is false and disproved by superefficient estimators.

To eliminate the superefficiency problem and to rescue the concept of efficiency, the modern fix is to recast the definition in a minimax framework. A treatment of efficiency can be found in Le Cam (1972), Hájek (1972) and Millar (1981, Chapter 7). We describe a more recent treatment of the problem by Le Cam and Yang (1990, Chapter 5.6, Theorem 1) and apply it to the parametric mixture setting. Let

$$\Theta(\theta_0, C, n) = \{\theta : |\theta - \theta_0| \le Cn^{-1/2}\}$$

be a $O\left(n^{-1/2}\right)$ shrinking neighborhood of θ_0. Let P_θ^n be the n-fold product distribution for P_θ and assume that $\{P^n_{\theta_0 + t/\sqrt{n}} : t \in I\!R\}$ is locally asymptotically normal (Le Cam and Yang, 1990, Chapter 5.7). Then if $\widehat{\theta}_n$ is an estimator for θ_0,

<1.6> $$\lim_{B\to\infty}\lim_{C\to\infty}\liminf_n \sup_{\theta\in\Theta(\theta_0,C,n)} P_\theta^n\left[B \wedge \left(n^{1/2}(\widehat{\theta}_n - \theta)\right)^2\right] \ge I(\theta_0, Q_0)^{-1}.$$

The minimax statement <1.6> asserts a lower bound for the risk of an estimator over $n^{-1/2}$ shrinking neighborhoods of the true parameter, θ_0. The statement implicitly assumes the existence of a $O_p\left(n^{-1/2}\right)$ estimator for θ which performs uniformly well under a whole class of models, for if such an estimator were not to exist the left-hand side of <1.6> would be infinite for each $\widehat{\theta}_n$ and the minimax expression trivial. However, under fairly mild regularity conditions, it is possible to prove the existence of theoretical minimum distance estimators which attain the $n^{-1/2}$ rate (Millar, 1981, Chapter 10). In fact it is possible to use these preliminary $O_p\left(n^{-1/2}\right)$ estimators to construct new estimators which have asymptotic minimax risk equal to the asserted lower bound (Le Cam and Yang, 1990, Chapter 5.3; Millar, 1981, Chapter 10). Such estimators are said to be efficient.

When the mixing distribution Q is unknown and the model semiparametric, the problem of estimation for θ has generally focused on the question of efficiency. The approach taken to this problem follows a line of argument described by Stein (1956). He argues that a nonparametric problem is at least as difficult as any parametric sub-problem and states:

> "it frequently happens that ... there is, through each state of nature, a one-dimensional problem which is, for large samples, at least as difficult (to a first approximation) as any other finite-dimensional problem at that point. If a procedure does essentially as well, for large samples, as one could do for each such one-dimensional problem, one is justified in considering the procedure efficient for large samples."

Koshevnik and Levit (1976) and more recently Begun, Hall, Huang, and Wellner (1983) expand upon this idea in the semiparametric setting. The general approach is to find the smooth path (indexed by a one-dimensional real parameter) through a space of probability models which makes estimation for a finite dimensional parameter as difficult as possible. This worst one-dimensional problem, for large sample sizes, determines the minimax efficiency for estimators of the finite dimensional parameter.

Begun et al. consider paths which are smooth in a Hellinger differentiable sense. Their argument specialized to the semiparametric mixture model is as follows. They work with a specified class of mixing densities $\mathcal{Q}$ taken with respect to a σ-finite measure λ. The semiparametric mixture densities are taken with respect to a σ-finite measure ν. Let $\|\cdot\|_\lambda$ and $\|\cdot\|_\nu$ denote the $L^2(\lambda)$ and $L^2(\nu)$ norms respectively. Thus if f is a semiparametric mixture density and q a mixing density, $\sqrt{f} \in L^2(\nu)$ and $\sqrt{q} \in L^2(\lambda)$.

The semiparametric density $f(x \mid \theta, q)$ is Hellinger differentiable at a fixed structural parameter, θ_0, and mixing density, q_0, if

<1.7> $$\sqrt{f}(x \mid \theta, q) = \sqrt{f}(x \mid \theta_0, q_0) + (\theta - \theta_0)\rho(\theta_0, q_0) + A(\theta_0, q_0)(\sqrt{q} - \sqrt{q_0}) + r(x, \theta, q)$$

with

$$\|r(x,\theta,q)\|_\nu = o\left(|\theta-\theta_0| + \|\sqrt{q}-\sqrt{q_0}\,\|_\lambda\right),$$

and $A(\theta_0,q_0) : L^2(\lambda) \to L^2(\nu)$ is a bounded linear operator.

They assume that <1.7> holds for θ sequences

<1.8> $$\theta_0 + Kn^{-1/2},$$

for real K, and sequences $q_n \in \mathcal{Q}$ such that

<1.9> $$\sqrt{q_n} = \sqrt{q_0} + n^{-1/2}\Delta + R_n,$$

where $\Delta \in L^2(\lambda)$ and $\|R_n\|_\lambda = o\left(n^{-1/2}\right)$. By doing so, they assume a sufficient condition for local asymptotic normality (cf. Le Cam and Yang, 1990, Chapter 5.7). Let $\mathcal{T}$ equal the set of all tangent scores, Δ, obtained from sequences <1.9>. By assuming that $A(\theta_0,q_0)\mathcal{T}$ is a closed space, the authors prove a minimax theorem similar to <1.6> over sequences in θ and q of the form <1.8> and <1.9>. They show that semiparametric minimax risk equals the reciprocal of

<1.10> $$4\left\|\rho(\theta_0,Q_0) - \Pi\,\rho(\theta_0,Q_0)\right\|_\nu^2,$$

where Π is the L^2-projection of $\rho(\theta_0,Q_0)$ onto $A(\theta_0,q_0)\mathcal{T}$. Thus for example, when the mixing distribution is known, the minimax risk equals the reciprocal of the Fisher information <1.5> from the parametric problem (the factor of four results from the use of square roots). However, when the mixing distribution is unknown, the information for θ_0 is smaller and consequently the minimax risk increases.

The information <1.10> can also be calculated as

$$4\left\|\rho(\theta_0,Q_0) - A(\theta_0,q_0)\Delta^*\right\|_\nu^2,$$

for a unique $\Delta^* \in \mathcal{T}$. In the context of Stein's paper, the path of mixing densities $\sqrt{f}(x \mid \theta_0+\tau, q_\tau)$ with $\sqrt{q_\tau} = \sqrt{q_0} - \tau\Delta^*$ represents, as τ converges to zero, the worst possible one-dimensional approach through $\sqrt{f}(x \mid \theta_0, q_0)$ for the problem of estimating θ_0.

Pfanzagl (1990) and van der Vaart (1988) show that in certain semiparametric mixture models, the information <1.10> can be calculated explicitly and used to state efficiency results.

Pfanzagl considers exponential mixture models

$$f(x \mid \theta_0, Q_0) = \int \exp(-\eta S(x,\theta_0) + b(\eta))\, dQ_0(\eta),$$

which belong to the class considered by Lindsay (1983). The models which are studied have the interesting property that no loss in information for θ_0 results due to the presence of the

nuisance mixing. The information in these models equals the ordinary Fisher information for θ_0 and can be computed as the squared $L^2(P_0)$-norm of

$$<1.11> \qquad -\frac{\partial}{\partial\theta}S(x,\theta_0)\frac{\int \eta \exp(-\eta S(x,\theta_0)+b(\eta))\,dQ_0(\eta)}{\int \exp(-\eta S(x,\theta_0)+b(\eta))\,dQ_0(\eta)},$$

where P_0 equals the true model.

Pfanzagl describes a procedure (Chapter 5, Theorem 5.6) for constructing an efficient estimator for θ_0 which is based on a preliminary $O_p\left(n^{-1/2}\right)$ estimator for θ_0 and a consistent estimator for the P_0-expectation of <1.11>. The procedure is implemented in several examples and simulation results indicate that even when Q_0 is crudely estimated, the variance for the new estimator for θ_0 is substantially smaller than the variance for the preliminary estimator and close to the theoretical asymptotic lower bound.

Van der Vaart assumes that the information <1.10> is positive and states under certain regularity conditions a local asymptotic minimax theorem and a convolution theorem. These results also implicitly presuppose the existence of an $n^{-1/2}$ estimator for θ_0. He describes a method for constructing such estimators.

In certain cases, it is easy to see why the structural parameter can be estimated at the classical $n^{-1/2}$ rate. Pfanzagl (1990, Example 1, page 67) and van der Vaart (1988, Example 5.2) both consider the paired exponential mixture model with density

$$f(x_1,x_2 \mid \theta,Q)=\int \eta\exp(-\eta x_1)\theta\eta\exp(-\theta\eta x_2)\,dQ(\eta),$$

with respect to Lebesgue measure for x_1, x_2 positive. This model corresponds to observing the random variables $X_1 = Y^{-1}Z_1$ and $X_2 = (\theta Y)^{-1}Z_2$, where Z_1 and Z_2 are independent standard exponentials, independent of the random variable Y with unknown distribution Q. The model becomes parametric when the data is transformed by taking the ratio X_1/X_2. Consequently, the question of whether the structural parameter can be estimated at classical rates is not an issue. The only problem that remains is the determination of lower bounds for efficiency and the construction of $O_p\left(n^{-1/2}\right)$ efficient estimators for the structural parameter θ.

3. Objective and Layout of Thesis It is not always clear whether $n^{-1/2}$ is the correct rate of convergence for uniform estimators of the structural parameter. Nor is it clear how rates of convergence are related to constraints on the mixing behavior, especially when this behavior is not restricted to be smooth.

The main objective of the thesis will be to determine lower bounds for rates of convergence for estimators of θ, and to identify how these rates are related to the type of mixing

behavior. Because of the semiparametric nature of the model, rates of convergence will be defined in a locally uniform sense to incorporate a minimax approach to estimation.

Chapter 2 formally introduces the idea of locally uniform rates of convergence and adapts, to the mixture setting, a general approach by Le Cam (1973) and Donoho and Liu (1987) for determining lower bounds for rates of convergence. The chapter collects together standard ideas and inequalities which make it possible to relate distances between probability measures to lower bounds for rates of convergence. Some simple motivating examples illustrate the ideas and mechanics applied to the mixture problem.

Chapter 3 presents techniques for determining lower bounds for rates. The first step is to identify mixture models for which it is not possible to estimate θ at a $O_p\left(n^{-1/2}\right)$ rate. Section 2 states sufficient conditions for identifying mixture models whose information, in the sense of <1.10>, can be made arbitrarily close to zero. Consequently, the section makes it possible to identify models whose structural parameters cannot be estimated at classical rates. The proof of the main theorem revolves around showing that a linear operator much like the one in <1.7> has a dense range, so that the information <1.10> is nearly zero. The result is limited in that it asserts that estimation rates for θ must be slower than $n^{-1/2}$ but does not give explicit rates. The second step is to explicitly determine lower bounds. Section 3 describes a Fourier technique for constructing explicit examples that establish lower bounds for rates of convergence. The technique is applicable to location mixture models with unknown scale parameter.

Chapter 4 is devoted solely to the normal mean-mixture model and the problem of determining lower bounds for rates of estimators for the unknown standard deviation. Both techniques of Chapter 3 are used here. The Fourier technique shows that rates of convergence depend upon the smoothness of the mixing. We find that the smoother the mixing is allowed to be, the slower the rates of convergence.

Chapter 5 considers the question of how well the shape parameter, θ, can be estimated in the Weibull mixture model <1.4> without covariates. By applying a simple transformation, the model is recast as a location-mixture with unknown scale parameter. The Fourier technique of Chapter 3.3 is used to show that the scale parameter in the transformed model cannot be estimated at a rate faster than $O_p\left(n^{-1/4}\right)$. This will indirectly establish $O_p\left(n^{-1/4}\right)$ as a lower bound for estimation in the original problem.

4. What Remains The main accomplishment of the thesis is to show that the structural parameter in the semiparametric mixture model is in general not estimable at classical $O_p\left(n^{-1/2}\right)$ rates. We do so by exhibiting a general class of mixture models for which the structural parameter can only be estimated at rates slower than $O_p\left(n^{-1/2}\right)$, and by

establishing explicit lower bounds in two well known models.

This research, therefore, has pushed the knowledge about rates of convergence in one direction. The next logical step is to pursue the problem of determining upper bounds for rates. In Chapters 4 and 5, we describe constructions which establish explicit lower bounds in the normal mean-mixture model and the Weibull mixture model. The constructions are chosen to push the lower bounds as far as possible, but it is unclear whether the rates given there are optimal or what the upper bound for rates are.

It is left for future research to determine the answers to these questions.

Chapter 2

Regular Estimators

1. Introduction A sensible requirement for a good estimator to satisfy should be that the estimator perform well under a fairly wide class of models. As discussed in Chapter 1, such a requirement is quite common in the modern semiparametric literature where the requirement that estimators possess good uniformity properties is motivated by the existence of super-efficient estimators. We adopt this uniformity approach to the mixture problem, where our interest in estimation will be from a rates of convergence perspective rather than an efficiency standpoint.

For simplicity assume that we are interested in uniformly estimating a real parameter θ_0 in a real set Θ. Informally, uniformity (or what we will later refer to as regularity) of an estimator $\widehat{\theta}_n$ for θ_0 will mean for a small fixed $\epsilon > 0$ there exists a small $\tau > 0$, such that for a large enough sample size

<2.1> $$P_\theta^n\{\mid \widehat{\theta}_n - \theta \mid > \tau\} < \epsilon,$$

for all θ in some neighborhood of θ_0 (here P_θ^n denotes the n-fold product measure of P_θ). By allowing both τ and the neighborhood to depend upon the sample size, expression <2.1> can be made into a precise statement for locally uniform rates of convergence (Definition <2.3>). Such a definition depends explicitly on the choice of neighborhoods, so that rates of convergence defined in this fashion should be interpreted in the context of the implied topology.

The chapter collects together many standard results and illustrates how they can be applied to the mixture problem. One key idea, introduced by Le Cam (1973), is described in Lemma <2.5> which asserts that if an estimator satisfies a local uniformity condition, then a "good" test for discrimination exists. By running the argument in reverse, the lemma provides a method for determining lower bounds for rates of convergence. Section 3 presents a motivating example that illustrate how the method can be used to determine lower bounds in the mixture problem.

A story book rendition of the Le Cam argument is as follows. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a class of probability models. The parameter space Θ can be arbitrary, but for simplicity take it to be real. You are the statistician interested in estimating θ for a fixed large sample size. Someone tries to sell you an estimator for θ which is reputed to have good properties. In fact the seller tells you that he can estimate θ with high precision and with P_θ-probability of at least 0.99 for a class of models of your choice, $\mathcal{C} \subseteq \mathcal{P}$. You ask for the

precision of the estimator and are told an outrageously small number, such as 10^{-50}. You doubt that such an estimator can exist, so you perform a simple calculation. If such an estimator were to exist, then it would be possible to construct a test between two models in $\mathcal{C}$ whose θ parameters are separated by more than 2×10^{-50}, such that under either model the test has a probability of misclassification of only 1%. This in turn means that the two sample distributions are "distinguishable". Therefore, to show that such an estimator cannot exist, you need only find a pair of models in $\mathcal{C}$ with θ parameters differing by at least 2×10^{-50} but whose sample distributions are "close". If such a construction were possible, then you would know not to waste your money.

Story book aside, the mathematical argument follows much along those lines and provides a means for determining lower bounds for rates of convergence. The key to the argument involves being able to gauge whether two sample distributions (product measures) are close in some sense. This is done through the use of the *total variation distance*, a metric which measures distances between probability measures. It is defined as:

<2.2> **Definition:** *Let P, Q be probability measures on $(\mathcal{X}, \mathcal{A})$. The total variation distance between P and Q is defined as,*

$$v(P,Q) = \sup_{A \epsilon \mathcal{A}} \mid PA - QA \mid .$$

If P and Q have densities p and q with respect to a σ-finite measure λ on $(\mathcal{X}, \mathcal{A})$, then

$$v(P,Q) = \frac{1}{2} \lambda \mid p - q \mid .$$

The total variation distance is a metric with a range of values that lie between zero and one. A value of zero between two probability measures implies that the measures are identical while a value of one implies that the two measures are singular and therefore can be perfectly discriminated between.

In the context of hypothesis testing, the metric has the following interpretation. If $\mathcal{F}$ is the set of all measurable functions over $(\mathcal{X}, \mathcal{A})$ that are bounded between zero and one, then

$$v(P,Q) = 1 - \inf_{f \in \mathcal{F}} [P(1-f) + Qf] .$$

That is, the total variation distance between two probability measures equals one minus the minimum sum of the type-1 and type-2 error amongst all zero-one tests between the two models. It is this relationship between testing and distance which makes it possible to translate the problem of determining rates of convergence into one concerning hypothesis testing.

2. Locally Uniform Rates of Convergence Let $\mathcal{P}$ be a family of probability measures on a common measurable space $(\mathcal{X}, \mathcal{A})$. Assume that we observe a sequence of n independent realizations from a distribution $P_0 \in \mathcal{P}$. Our aim is to uniformly estimate $\kappa(P_0)$, where κ is a functional mapping $\mathcal{P}$ onto a metric space (D, d).

Uniform estimation by an estimator T_n for $\kappa(P_0)$ will mean that under sampling from P_n, the estimator T_n gets close to $\kappa(P_n)$ for sequences P_n which converge to P_0 in the total variation sense. In terms of rates of convergence, this idea is made more precise as follows. Let $\mathcal{U}$ equal the topology on $\mathcal{P}$ induced by the total variation metric. Let α_n be a positive sequence converging to zero, and define

$$\mathcal{U}(P_0, \alpha_n) = \Big\{ P \in \mathcal{P} : v(P_0, P) \leq \alpha_n \Big\}$$

to be a sequence of $O(\alpha_n)$ shrinking neighborhoods of P_0 in $\mathcal{U}$. With this notation, we adopt the following as our definition for uniform rates of convergence:

<2.3> **Definition:** *Estimators T_n for $\kappa(P)$ are said to be regular at P_0 for the rate of convergence δ_n if for each $\epsilon > 0$ there exists a finite positive constant $K(\epsilon)$ such that for each sequence α_n decreasing to zero*

$$\limsup_n \sup_{P \in \mathcal{U}(P_0, \alpha_n)} P^n\{d(T_n, \kappa(P)) \geq K(\epsilon)\delta_n\} < \epsilon,$$

where P^n denotes the n-fold product measure for P.

Definition <2.3> asserts that if an estimator T_n is regular with rate of convergence δ_n, then T_n estimates $\kappa(P)$ to a precision of $O(\delta_n)$ with high P-probability over a class of $O(\alpha_n)$ shrinking neighborhoods of P_0. It also asserts that this hold for each sequence α_n converging to zero.

One argument for insisting that uniform estimation hold for different α_n sequences can be made in terms of confidence intervals. If an estimator T_n is regular with rate of convergence δ_n, then Definition <2.3> implies that for each $\epsilon > 0$ and each sequence α_n decreasing to zero there exists an integer $n_0 = n_0(\epsilon, \alpha_n)$ such that

$$\sup_{P \in \mathcal{U}(P_0, \alpha_n)} P^n\{d(T_n, \kappa(P)) \geq K(\epsilon)\delta_n\} < \epsilon, \qquad \text{for } n \geq n_0.$$

By inverting this probability statement, we can construct for each $n \geq n_0$ a confidence interval for $\kappa(P)$ with size at least $(1 - \epsilon) \times 100\%$

$$\{\theta \in D : d(T_n, \theta) < K(\epsilon)\delta_n\},$$

that holds for all $P \in \mathcal{U}(P_0, \alpha_n)$. Of course the confidence statement is only useful if the data is sampled from a model which lies in the $O(\delta_n)$ neighborhood $\mathcal{U}(P_0, \alpha_n)$ of P_0. However, since P_0 is unknown, it may be impossible to discern whether the marginal sampling

distribution falls in the required neighborhood. Thus, to make the confidence statement practicable, it is desirable to require uniform estimation to hold over neighboorhoods which shrink at different rates.

If T_n is a regular estimator with rate of convergence δ_n, then it is also regular with rate δ'_n for each sequence $\delta'_n \geq \delta_n$. Thus, Definition <2.3> is a statement concerning achievability of rates and does not address the issue of whether a regular rate is optimal in the sense that it is the fastest achievable rate possible (see Stone, 1980 for another definition of optimality in the context of rates of convergence). We do not pursue the issue of optimality in the thesis, instead our goal will be to determine lower bounds for rates of convergence. A problem which amounts to finding a rate $\widetilde{\delta_n}$ such that each estimator for the functional of interest fails to satisfy the probability inequality in Definition <2.3>.

For example, assume that for each small $\epsilon > 0$, we can show that a regular estimator for $\kappa(P_0)$ cannot achieve the rate $\widetilde{\delta_n} = n^{-1/(d-\epsilon)}$. Then this establishes $O_p\left(n^{-1/(d-\epsilon)}\right)$ as a strict lower bound for each $\epsilon > 0$, and implies that a regular estimator cannot convergence faster then the lower bound $O_p\left(n^{-1/d}\right)$. Of course to be able to establish that this is in fact the best lower bound, or more precisely that $O_p\left(n^{-1/d}\right)$ is the optimal rate of convergence, we would need to show that $O_p\left(n^{-1/d}\right)$ is an achievable rate of convergence (for example, by constructing an estimator that is regular with the asserted rate).

Because we only consider the problem for lower bounds of rates of convergence, we hereafter omit the distinction between achievability and optimality except when a clarification is necessary.

To see how Definition <2.3> applies to the mixture problem, first introduce the following terminology and convenient notation.

<2.4> **Definition:** *Let $\mathcal{F}(\Theta, \mathcal{N}) = \{f(\cdot \mid \theta, \eta) : (\theta, \eta) \in \Theta \times \mathcal{N} \subseteq \mathbb{R}^k \times \mathbb{R}\}$ be a parametric family of probability density functions with respect to a σ-finite measure $\nu \ll \mu$, where μ is Lebesgue measure. We assume that $f(x \mid \theta, \eta)$ is measurable as a function of (x, η). Form the mixed distribution $P_{\theta,Q}$ over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with ν-density*

$$f(x \mid \theta, Q) = \int f(x \mid \theta, \eta)\, dQ(\eta),$$

and form the class of all such distributions $\mathcal{P}(\Theta, \mathcal{Q})$ as θ ranges over Θ and Q ranges over a class of mixing distributions $\mathcal{Q}$ with support on $\mathcal{N}$.

In the context of the semiparametric mixture problem, take the family of probability measures $\mathcal{P}$ of Definition <2.3> to be the class of mixed distributions $\mathcal{P}(\Theta, \mathcal{Q})$ described in Definition <2.4>. Our interest is in estimation for the structural parameter so that the functional of interest, κ, is defined by $\kappa(P_{\theta,Q}) = \theta$ for $P_{\theta,Q} \in \mathcal{P}$.

The general approach in the semiparametric mixture literature, at least for the purpose of calculating information bounds, has been to consider estimation along sequences generated from smooth paths. Rates of convergence for the structural parameter are implicity set at a $O_p\left(n^{-1/2}\right)$ rate and uniform estimators are required to perform well over $O\left(n^{-1/2}\right)$ shrinking Hellinger neighborhoods where the mixing densities satisfy a Hellinger differentiability property (Begun, Hall, Huang, and Wellner, 1984). We will not require that the mixing densities be Hellinger differentiable (we will not even require that the mixing distribution have a density), rather we investigate rates of convergence over $O\left(\alpha_n\right)$ shrinking neighborhoods of P_0 without any additional constraints on the mixing distribution other than those implied by our choice for $\mathcal{Q}$.

The main differences, therefore, in this approach to uniform estimation is firstly that we do not presume that rates of convergence are set at a $O_p\left(n^{-1/2}\right)$ rate, and secondly that we require uniform estimation to hold over classes of $O\left(\alpha_n\right)$ shrinking neighborhoods, for all α_n sequences which decrease to zero. Another argument in favour of using different shrinking neighborhoods can be made in the mixture problem, as follows. If $\mathcal{F}(\Theta, \mathcal{N})$ is a parameteric family which is smooth in θ, then we would expect that $v(P_0, P_n) = O\left(\delta_n\right)$ for sequences $P_n \in \mathcal{P}$ which have parameters $\left(\theta_0 + O\left(\delta_n\right), Q_0\right)$ converging to the parameter (θ_0, Q_0) for the distribution P_0. If estimation for the structural parameter in such a family is at a $O\left(\delta_n\right)$ rate, then it seems reasonable to require that the rate take into account shrinking neighborhoods that contain such P_n sequences. Our definition takes into account such neighborhoods.

By working with the total variation distance, the next lemma shows that the uniformity of an estimator implies the existence of a good test for discriminating between models. Indirectly, the existence of such a test will provide a method for determining lower bounds for the rate of convergence of regular estimators. The lemma is a reformulation from Le Cam (1973, Lemma 1). The same idea is also used effectively in Donoho and Liu (1987, 1991).

<2.5> **Lemma:** *Let $\mathcal{P}, \mathcal{U}$, κ, and α_n be as in Definition <2.3>. If estimators T_n for $\kappa(P)$ are regular at $P_0 \in \mathcal{P}$ with rate of convergence δ_n, then for each $\epsilon > 0$ there exists a finite constant $K(\epsilon)$ such that for each sequence $P_n \in \mathcal{U}(P_0, \alpha_n)$ with $d(\kappa(P_0), \kappa(P_n)) \geq 2K(\epsilon)\delta_n$,*

$$v(P_0^n, P_n^n) \geq 1 - 2\epsilon, \qquad \text{for } n \geq n_0,$$

where $n_0 = n_0(\epsilon, \alpha_n)$ is an integer which depends upon both ϵ and the sequence α_n.

The proof of this lemma is remarkably simple and is included for completeness.

Proof of Lemma <2.5>: By the regularity of T_n it is possible to choose, for a fixed

$\epsilon > 0$, a constant $K(\epsilon)$ such that for all $P \in \mathcal{U}(P_0, \alpha_n)$

$$P^n\{d(T_n, \kappa(P)) < K(\epsilon)\delta_n\} \geq 1 - \epsilon, \qquad \text{for } n \geq n_0,$$

where $n_0 = n_0(\epsilon, \alpha_n)$. In particular, because $P_0 \in \mathcal{U}(P_0, \alpha_n)$ we have $P_0^n A_n \geq 1 - \epsilon$, for $n \geq n_0$, where

$$A_n = \{d(T_n, \kappa(P_0)) < K(\epsilon)\delta_n\}.$$

If $d(\kappa(P_0), \kappa(P_n)) \geq 2K(\epsilon)\delta_n$, then

$$P_n^n A_n \leq P_n^n\{d(T_n, \kappa(P_n)) \geq K(\epsilon)\delta_n\}.$$

Therefore if $P_n \in \mathcal{U}(P_0, \alpha_n)$, then the right-hand side is less than ϵ for $n \geq n_0$.

We have exhibited a set A_n so that $P_0^n A_n \geq 1 - \epsilon$ and $P_n^n A_n < \epsilon$ for $n \geq n_0$. This implies the lower bound

$$v(P_0^n, P_n^n) \geq 1 - 2\epsilon, \qquad \text{for } n \geq n_0. \qquad \square$$

Informally, Lemma <2.5> asserts that if an estimator T_n is regular with rate of convergence δ_n, then there exists a test based on T_n which can discriminate between P_0 and a sequence $P_n \in \mathcal{P}$ with total variation distance $v(P_0, P_n) = O(\alpha_n)$ and whose functionals are larger than $2K\delta_n$ away from $\kappa(P_0)$, for a large positive K.

We can exploit the lemma to establish lower bounds for rates of convergence by the following contrapositive argument. Let δ be a fixed small positive number. For a given suggested rate of convergence δ_n exhibit for each finite $K > 0$ a sequence $P_n \in \mathcal{P}$ such that $v(P_0, P_n) = O(\alpha_n)$ and

$$|\kappa(P_0) - \kappa(P_n)| \geq 2K\delta_n,$$

but with product total variation distance bounded *away* from one:

$$v(P_0^n, P_n^n) < \delta.$$

By Lemma <2.5> this construction shows that a regular estimator for $\kappa(P_0)$ cannot have rate of convergence δ_n and establishes δ_n as a lower bound for the rate of convergence.

To be able to implement the Le Cam approach, it is necessary to calculate L^1-distances between product measures. To circumvent the difficulty of such a calculation, the usual practise is to work with the *Hellinger distance* because of the convenient manner in which the distance factorizes for product measures.

<2.6> **Definition:** *Let P, Q be probability measures over $(\mathcal{X}, \mathcal{A})$. The Hellinger distance between P and Q is defined as*

$$H(P, Q) = \sqrt{\lambda\,(\sqrt{p} - \sqrt{q}\,)^2}$$

where p and q represent the density of P and Q with respect to a σ-finite measure λ on $(\mathcal{X}, \mathcal{A})$.

We work with densities

$$f_\tau(x) = \int f(x \mid \theta + \tau, \eta)\, dQ_\tau(\eta),$$

and frequently will need to calculate the L^1-distance between the product measure for f_τ, for $\tau > 0$, and the product measure for f_0, when $\tau = 0$. In this setting it might be inappropriate to work with square roots of densities and the Hellinger distance. Instead we will find it more convenient to work with the squared L^2-distance:

$$\int f_0 \left(\frac{f_\tau}{f_0} - 1 \right)^2.$$

(The observant reader will recognize that in the discrete case this is the Pearson goodness of fit value.)

The following lemma makes explicit the relationship between the "Pearson distance" and the total variation distance. It is a slightly weaker result than the one used by Donoho and Liu (1987, 1991) who work with the Hellinger distance.

<2.7> **Lemma:** *Let $Q \ll P$ be probability measures over $(\mathcal{X}, \mathcal{A})$. Then for each $\delta > 0$ there exists a $\gamma > 0$ such that,*

$$v(P^n, Q^n) < \delta \quad \textit{if} \quad P \left(\frac{dQ}{dP} - 1 \right)^2 < \frac{\gamma}{n}.$$

Proof of Lemma <2.7>: Let $\lambda = P + Q$ and denote its n-fold product measure by λ^n. Let $p = dP/d\lambda$, $q = dQ/d\lambda$, and denote their n-fold densities by p^n and q^n respectively.

From the Cauchy-Schwarz inequality:

$$\begin{aligned} v(P, Q) &= \frac{1}{2} \lambda \mid \sqrt{p} - \sqrt{q} \mid\mid \sqrt{p} + \sqrt{q} \mid \\ &\leq \frac{1}{2} H(P, Q) \sqrt{4 - H(P, Q)^2}. \end{aligned}$$

Rewrite this as,

$$v(P, Q)^2 \leq 1 - (1 - \frac{1}{2} H(P, Q)^2)^2.$$

Use this inequality and the definition for the Hellinger distance to show:

$$\begin{aligned} v(P^n, Q^n)^2 &\le 1 - \left[1 - \frac{1}{2} H(P^n, Q^n)^2\right]^2 \\ &= 1 - \left[\lambda^n \sqrt{p^n q^n}\,\right]^2 \\ &= 1 - \left[\lambda \sqrt{pq}\,\right]^{2n} \\ &= 1 - \left[1 - \frac{1}{2} H(P, Q)^2\right]^{2n}. \end{aligned} \tag{<2.8>}$$

Use the fact that $Q \ll P$ to bound the Hellinger distance (see for example: Le Cam, 1986, Chapter 4):

$$\begin{aligned} P\left(\frac{dQ}{dP} - 1\right)^2 &\ge \lambda \left[\left(\sqrt{p} - \sqrt{q}\right)^2 \left(\sqrt{p} + \sqrt{q}\right)^2 \frac{1}{p} \{p \ne 0\}\right] \\ &\ge \lambda \left[\left(\sqrt{p} - \sqrt{q}\right)^2 \{p \ne 0\}\right] \\ &= H(P, Q)^2. \end{aligned} \tag{<2.9>}$$

Use this inequality in <2.8> to show that $v(P^n, Q^n) < \delta$ when

$$P\left(\frac{dQ}{dP} - 1\right)^2 < \frac{\gamma}{n},$$

for a suitably choosen $\gamma > 0$. □

The next lemma puts together some of the previous ideas and states sufficient conditions to infer lower bounds for rates of convergence. We refer to it frequently throughout the thesis.

<2.10> **Lemma:** *Suppose there exists a path $P_\tau \in \mathcal{P}$ through P_0 indexed by $\tau \ge 0$, such that for a $\beta > 0$:*

(i) $|\kappa(P_\tau) - \kappa(P_0)| \ge \beta\tau$, *for τ near zero*
(ii) $P_0 \left(dP_\tau/dP_0 - 1\right)^2 = O\left(\tau^d\right)$.

Then the rate of convergence, in the sense of Definition <2.3>, cannot be faster than $O\left(n^{-1/d}\right)$.

Proof of Lemma <2.10>: Suppose $\delta_n = o\left(n^{-1/d}\right)$. Write P_n for P_τ at $\tau = 2K\delta_n/\beta$, so that $|\kappa(P_n) - \kappa(P_0)| \ge 2K\delta_n$ and $P_0 \left(dP_n/dP_0 - 1\right)^2 = o\left(n^{-1}\right)$. By Lemma <2.7>,

$$v(P_0^n, P_n^n) \to 0.$$

No matter how large K is chosen, we cannot keep the total variation distance bounded away from zero. Deduce the asserted lower bound by Lemma <2.5>. □

In certain cases, we will not be able to establish the conditions of Lemma <2.10> which are sufficient to show explicit lower bounds for rates of convergence. However, it still may be possible to obtain a lower bound. The next lemma describes how.

<2.11> **Lemma:** *Suppose the conditions of Lemma <2.10> are replaced by: for each $\epsilon > 0$ there is a path satisfying (i), where $\beta > 0$ is independent of ϵ, such that*

$$P_0 \left(\frac{dP_\tau}{dP_0} - 1 \right)^2 < \epsilon \tau^2$$

for τ near zero. Then the rate of convergence, in the sense of Definition <2.3>, must be slower than $O\left(n^{-1/2}\right)$.

Proof of Lemma <2.11>: Fix $K > 0$ and let $\delta > 0$ be a fixed small number. Choose a path P_τ so that $4K^2\epsilon/(n\beta^2)$ is less than the γ in Lemma <2.7>. Let P_n denote P_τ at $\tau = 2K/(\sqrt{n}\beta)$, so that $|\kappa(P_n) - \kappa(P_0)| \geq 2K/\sqrt{n}$ and

$$P_0 \left(\frac{dP_\tau}{dP_0} - 1 \right)^2 < \frac{4K^2\epsilon}{n\beta^2}, \qquad \text{eventually.}$$

Then by Lemma <2.7>

$$v(P_0^n, P_n^n) < \delta, \qquad \text{eventually.}$$

Because this holds for each $K > 0$, Lemma <2.5> shows that a regular estimator must have rate of convergence slower than $O\left(n^{-1/2}\right)$. □

3. Two Motivating Examples This section presents motivating examples that indicate how the inequalities and ideas of the previous section can be used to describe rates of convergence in the mixture problem. Before proceeding we first need to consider the problem of identifiability, for without some type of identifiability constraints it would be futile to pursue the problem of determining rates of convergence.

For example, consider the normal mean-mixture model presented in Chapter 1. This model is formed by integrating over the mean of a normal density with unknown standard deviation. As previously observed the model has the convenient representation as a convolution $\theta Z + Y$, where Z has a $N(0,1)$ distribution, θ is the unknown positive standard deviation, and Y has an unknown distribution independent of Z. Without any constraints on the types of distributions that Y can take, the model as it stands is unidentifiable. For example, there is no way to tell the mixture model (slightly abusing notation)

$$2N(0,1) + 3N(0,1)$$

from the mixture model

$$3N(0,1) + 2N(0,1),$$

even though the structural parameters are different. Clearly then, it is necessary to require identifiability in the model before pursuing the estimation problem. That is, we need to assume that if $P_{\theta_1,Q_1} = P_{\theta_2,Q_2}$, then $\theta_1 = \theta_2$ and $Q_1 = Q_2$.

Assuming that we have identifiablity, consider the problem of estimating an unknown location parameter in the mixture model formed by mixing over the scale of a scale-location density. This model is described by the random variable

<2.12> $$X = YZ + \theta,$$

where Z has a known density f, and Y is noise with an unknown distribution Q independent of Z. The problem is to estimate the unknown location parameter θ.

If Z has zero median, then the median for X equals the parameter of interest, θ. Therefore, if the mixture model is identifiable and the mixing distribution Q suitably restricted, we should expect the sample median to be a regular estimator for θ with *achievable* rate of convergence $O_p\left(n^{-1/2}\right)$. The next example shows this to be true.

<2.13> **Example (regularity of the median):** Let F be the distribution function for the random variable YZ of <2.12>. Let $\mathcal{Q}$ be a class of mixing distributions composed of distributions which satisfy

<2.14> $$|F(t) - F(0)| \geq \gamma|t|,$$

for all t in a fixed neighborhood of zero, where $\gamma > 0$ is a fixed constant. (Notice that because the median for Z is zero, $F(0) = 1/2$.) Let $\mathcal{P}(\Theta, \mathcal{Q})$ be the class of mixed distributions of the form <2.12>, where the unknown location parameter, θ, takes values in $\Theta = \mathbb{R}$. We will show that the median is a regular estimator for the structural parameter of each $P \in \mathcal{P}(\Theta, \mathcal{Q})$, and has achievable rate of convergence $O_p\left(n^{-1/2}\right)$.

Let M_n be the median of the sample obtained from n independent realizations of a $P \in \mathcal{P}$ with structural parameter θ. We will show that by choosing K to be suitably large,

<2.15> $$P^n\{|M_n - \theta| \geq Kn^{-1/2}\}$$

can be made arbitrarily small for large n, independent of the sampling scheme P. This will be more than enough to demonstrate our assertion concerning the regularity of the median.

For convenience assume that the sample size is even and that some method is used to determine the sample median in the case of ties. Consider one side of the bound in <2.15>:

<2.16> $$P^n\{M_n - \theta \geq Kn^{-1/2}\} = P^n\{M_n \geq \theta + Kn^{-1/2}\}.$$

If $M_n \geq C$, then the number of observations which are greater than or equal to C must be at least $n/2$. Thus,

$$\begin{aligned} P^n\{M_n \geq \theta + Kn^{-1/2}\} &= P^n\Big\{\sum_{i=1}^{n}\{X_i \geq \theta + Kn^{-1/2}\} \geq \frac{n}{2}\Big\} \\ &= P^n\Big\{\text{Bin}\,(n, 1 - F(Kn^{-1/2})) \geq \frac{n}{2}\Big\} \\ <2.17> \qquad &\leq P^n\Big\{\text{Bin}\,(n, \frac{1}{2} - \gamma Kn^{-1/2}) \geq \frac{n}{2}\Big\}, \end{aligned}$$

where the last inequality is obtained from the constraint <2.14>, and $\text{Bin}\,(n,p)$ is used to denote a random variable with a binomial (n,p) distribution.

Use the inequality,

$$\begin{aligned} P^n\{\text{Bin}\,(n,p) - np \geq C\sqrt{n}\} &\leq \frac{P^n|\text{Bin}\,(n,p) - np|^2}{nC^2}, \qquad \text{for } C > 0 \\ &\leq \frac{1}{C^2}, \end{aligned}$$

in <2.17> with $p = 1/2 - \gamma Kn^{-1/2}$ and $C = \gamma K$, to show that

$$P^n\{M_n \geq \theta + Kn^{-1/2}\} \leq \frac{1}{(\gamma K)^2}.$$

Deduce that the inequality <2.17> can be made arbitrarily small by choosing K large enough, independent of the sequence P. This takes care of one half of the inequality in <2.15>. The other half is dealt with in the same fashion.

<2.18> **Remarks:** It is not hard to construct a class $\mathcal{Q}$ which satisfies condition <2.14>. For example, assume that $\mathcal{Q}$ equals the class of distributions that have support on the set $[A, \infty)$, where A is a fixed positive constant. Then, by interchanging the order of integration, we can express $F(t)$ as

$$F(t) = \int\int_{-\infty}^{t} f\left(\frac{x}{\eta}\right)\frac{1}{\eta}\,dx\,dQ(\eta).$$

A change of variables and the fact that the support of Q is a subset of $[A, \infty)$ allows us to bound $F(t)$ by

$$\int_{-\infty}^{-|t|/A} f(x)\,dx \;\leq F(t) \leq \int_{-\infty}^{|t|/A} f(x)\,dx.$$

Express each integral as $F(0)$ plus a contribution over a range depending upon t to show that <2.14> holds for all $Q \in \mathcal{Q}$, for values of t in a small neigborhood of zero.

□

On the other hand, there are cases where mixing can pathologically affect rates of convergence. Consider the normal mean-mixture model once again. As we previously

observed, the model is unidentifiable if the mixing distribution is allowed to contain normal components. Consequently no rate of convergence is attainable in the model with normal components. We would like to find out if this is still the case even when the class of mixing distributions is constrained to ensure identifiability.

The next example begins to answer this question by showing that the standard deviation might not be estimable at any guaranteed rate of convergence if the model is only required to be identifiable. The example shows how to construct two normal mixture models whose standard deviations are of some distance apart, and yet whose densities can be made to be as close as possible by an appropriate choice for the mixing distribution. The construction takes advantage of the near lack of identifiability in the model created by allowing too large a class of mixing distributions, and because of this the result is not of any practical consequence. Rather we introduce it at this time to illustrate how the various inequalities and ideas of the previous sections can be used in the mixture setting to derive rates of convergence.

<2.19> **Example (normal mean-mixture model):** Let $\mathcal{P}(\Theta, \mathcal{Q})$ equal the class of normal mean-mixtures, where the unknown standard deviation takes values in $\Theta = (0, \infty)$, and the class of mixing distributions $\mathcal{Q}$ is composed of distributions that contain no normal component. That is, no Q in $\mathcal{Q}$ can be expressed as a convolution of a nondegenerate normal with another distribution.

We will show that $\mathcal{P}(\Theta, \mathcal{Q})$ is identifiable and yet a regular estimator for the standard deviation cannot have rate of convergence faster than $O_p\left(n^{-1/d}\right)$, for each $d > 0$.

Establishing the identifiability of $\mathcal{P}(\Theta, \mathcal{Q})$ involves a characteristic function argument which utilizes the fact that the normal mean-mixture model can be expressed as a convolution. Let $\gamma_i = (\theta_i, Q_i)$ where $\theta_i \in \Theta$ and $Q_i \in \mathcal{Q}$ for $i = 1, 2$. Let P_{γ_i} be the normal mixture model corresponding to $\theta_i Z + Y_i$, where Z is normally distributed, and Y_i has distribution Q_i independent of Z.

If P_{γ_1} and P_{γ_2} are equal, then their characteristic functions must also be equal. Therefore, if we denote the characteristic function for Q_i by ψ_{Q_i}

$$\exp(-\tfrac{1}{2}(t\theta_1)^2)\psi_{Q_1}(t) = \exp(-\tfrac{1}{2}(t\theta_2)^2)\psi_{Q_2}(t),$$

which implies that

$$\psi_{Q_1}(t) = \exp(-\tfrac{1}{2}t^2(\theta_2^2 - \theta_1^2))\psi_{Q_2}(t).$$

Therefore, Q_1 is the convolution of Q_2 and a $N(0, \theta_2^2 - \theta_1^2)$ distribution if $\theta_2^2 > \theta_1^2$. Consequently, $\mathcal{P}(\Theta, \mathcal{Q})$ is identifiable if $\mathcal{Q}$ consists of distributions that contain no normal components.

Let P_0 be the assumed true mixed distribution with standard deviation θ_0 and mixing distribution Q_0 corresponding to the atomic distribution which puts all its mass at zero. Without loss of generality assume that $\theta_0 = 1$. Then P_0 is the standard $N(0,1)$ distribution.

Let $\delta_n = n^{-1/d}$ be a fixed sequence decreasing to zero and let $K > 0$ be a large fixed constant. Let Z_n denote the random variable with distribution $N(0, 1 - K\delta_n)$, and let Y_n be independent of Z_n with distribution $N(0, K\delta_n)$. Let $\widetilde{Y}_n = Y_n\{|Y_n| \le M_n\}$, where M_n is a positive sequence converging to infinity. Denote the distribution for Y_n and $\widetilde{Y}_n$ by Q_n and $\widetilde{Q}_n$ respectively.

Define the sequence of normal mean-mixture models, P_n, by the random variables

$$Z_n + \widetilde{Y}_n.$$

That is, $P_n \in \mathcal{P}$ with standard deviation $(1 - K\delta_n)^{1/2}$ and mixing distribution $\widetilde{Q}_n$. We will show that although the standard deviations for P_n and P_0 are separated by more than δ_n, the total variation distance, $v(P_0^n, P_n^n)$, will be bounded away from one for each $K > 0$. By Lemma <2.5> this will establish δ_n as a lower bound for the rate of convergence.

Notice that P_0 can also be represented as

$$Z_n + Y_n.$$

Therefore, the problem of discriminating between a sample from P_0 and P_n should be as difficult as discriminating between a sample from Q_n and $\widetilde{Q}_n$. In particular,

$$v(P_0^n, P_n^n) \le v(Q_n^n, \widetilde{Q}_n^n).$$

A proof of this fact follows by working with the product densities. Let h_n denote the density for Z_n. By interchanging the order of integration:

$$\begin{aligned} v(P_0^n, P_n^n) &= \frac{1}{2} \int \left| \int h_n(x_1 - y_1) \cdots h_n(x_n - y_n) \left(dQ_n^n(\mathbf{y}) - d\widetilde{Q}_n^n(\mathbf{y}) \right) \right| d\mathbf{x} \\ &\le \int h_n(x_1 - y_1) \cdots h_n(x_n - y_n)\, d\mathbf{x} \;\; \frac{1}{2} \int \left| \, dQ_n^n(\mathbf{y}) - d\widetilde{Q}_n^n(\mathbf{y}) \, \right| \\ &= v(Q_n^n, \widetilde{Q}_n^n). \end{aligned}$$

By the definition for $\widetilde{Y}_n$,

$$\frac{d\widetilde{Q}_n(y)}{dQ_n(y)} - 1 = \{|y| \le M_n\} \left(Q_n\{|Y| \le M_n\}\right)^{-1} - 1.$$

Expand the quadratic to obtain,

$$Q_n \left(\frac{d\widetilde{Q}_n}{dQ_n} - 1 \right)^2 = \left(Q_n\{|Y| \le M_n\}\right)^{-1} - 1.$$

By choosing M_n to converge to infinity fast enough, the right-hand side can be made less than γ/n for small $\gamma > 0$. Deduce by Lemma <2.7> that $v(Q_n^n, \widetilde{Q}_n^n)$ can be made arbitrarily small for each $K > 0$, and therefore conclude that the standard deviation is not estimable at the rate of convergence δ_n. Because δ_n is arbitrary, this shows that the standard deviation is not estimable at any rate. □

<2.20> **Remarks:** We return to this example in Chapter 4 where we discuss different methods for constraining the class of mixing distribution and show how lower bounds for rates of convergence are related to these constraints.

Chapter 3

Techniques for Determining Rates of Convergence

1. Introduction Consider the parametric family of densities

<3.1> $$\mathcal{F}(\Theta,\mathcal{N}) = \{f(\cdot \mid \theta,\eta) : (\theta,\eta) \in \Theta \times \mathcal{N} \subseteq I\!R^k \times I\!R\},$$

where the underlying densities $f(x \mid \theta,\eta)$ are taken with respect to a σ-finite measure, ν, dominated by Lebesgue measure. As in Definition <2.4>, we assume that $f(x \mid \theta,\eta)$ is measurable as a function of (x,η) and form the class of mixed distributions, $\mathcal{P}(\Theta,\mathcal{Q})$, to consist of distributions with ν-densities

$$f(x \mid \theta, Q) = \int f(x \mid \theta,\eta)\, dQ(\eta),$$

where $\mathcal{Q}$ is a class of mixing distributions with support on $\mathcal{N}$.

Our objective is to determine lower bounds for rates of convergence for estimators of θ assuming that Θ and $\mathcal{Q}$ are chosen so as to make $\mathcal{P}(\Theta,\mathcal{Q})$ identifiable. Determining a lower bound for a particular component of the vector θ is at least as difficult as determining the rate when the remaining $k-1$ components are known. Therefore, for simplicity we will take Θ to be a one-dimensional parameter space.

Let $P_0 \in \mathcal{P}(\Theta,\mathcal{Q})$ have parameter (θ_0, Q_0). Let $P_\tau \in \mathcal{P}(\Theta,\mathcal{Q})$ be a path through P_0 formed by perturbing θ_0 by τ and by perturbing Q_0 in some fashion. If $\mathcal{F}(\Theta,\mathcal{N})$ satisfies mild regularity conditions and if the perturbation in Q_0 is smooth enough, then we would expect the likelihood ratio to be expressible as

<3.2> $$\frac{dP_\tau}{dP_0} = 1 + \tau\Delta + R(\cdot,\tau),$$

where Δ is an $L^2(P_0)$-function, and $R(\cdot,\tau)$ has $L^2(P_0)$-norm of order $o(\tau)$ (see Pfanzagl, 1990 for a different example where a similar form for the likelihood is assumed).

The layout of the chapter is as follows. Section 2 presents sufficient conditions (Regularity Conditions <3.6>) which ensure that the likelihood ratio for smooth P_τ paths can be expressed in a form similar to <3.2>. There we consider families $\mathcal{F}(\Theta,\mathcal{N})$ which have a specific exponential form, and show that when the class of mixing distributions is rich enough, it is possible to construct for any positive ϵ, a smooth path $P_\tau \in \mathcal{P}(\Theta,\mathcal{Q})$ through $P_0 \in \mathcal{P}(\Theta,\mathcal{Q})$ with likelihood ratio expressible as <3.2> such that $P_0\Delta^2 < \epsilon$. That is, the path P_τ is constructed so that the component in Δ attributable to the perturbation in the mixing distribution nearly cancels the component due to the perturbation in the structural parameter. This will imply that the structural parameter cannot be estimated at the usual $O_p\left(n^{-1/2}\right)$ rate.

Section 3.3 describes a Fourier technique for constructing paths, P_τ, whose likelihood ratio can be expressed in a form similar to <3.2>, but whose linear coefficient Δ is an $L^2(P_0)$-function depending upon τ. The rate of decrease of the $L^2(P_0)$-norm of these tangents determines explicit lower bounds for rates of convergence. The technique is applicable to location-mixture models with unknown scale parameter.

2. Checking for Zero Information Now to formalize the discussion of the previous section. Let $\mathcal{F}(\Theta, \mathcal{N})$ be a parametric family of densities of the form <3.1> which can be written as,

<3.3> $$f(x \mid \theta, \eta) = \exp\left(\eta s(x,\theta) + t(x,\theta) + b(\theta,\eta)\right).$$

For fixed θ it defines an exponential family indexed by η, with natural parameter space

$$\mathcal{N}(\theta) = \Big\{\eta : \int \exp\left(\eta s(x,\theta) + t(x,\theta)\right)\, dx < \infty\Big\}.$$

We assume that $\mathcal{N}(\theta) = \mathcal{N}$ for each $\theta \in \Theta$ and that $(0,\infty) \subseteq \mathcal{N}$.

<3.4> **Example (Weibull mixture model):** Let $\mathcal{F}(\Theta, \mathcal{N})$ be the parametric family of Weibull densities

$$f(x \mid \theta, \eta) = \theta x^{\theta-1} \eta \exp(-\eta x^\theta),$$

with respect to Lebesgue measure on $(0,\infty)$, where $\Theta = (0,\infty)$, and $\mathcal{N} = (0,\infty)$. The densities in this family conform to the parametric requirement <3.3> with $s(x,\theta) = -x^\theta$, $t(x,\theta) = (\theta - 1)\log x$ and $b(\theta,\eta) = \log(\theta\eta)$. Notice that for each θ, the natural parameter space is $\mathcal{N} = (0,\infty)$. □

Let $Q_0 \in \mathcal{Q}$ and define $\mathcal{L}_0^\infty(Q_0)$ to be the set of functions which are bounded a.e. $[Q_0]$ and which have zero Q_0-expectation. For $h \in \mathcal{L}_0^\infty(Q_0)$, define the perturbed mixing distribution $Q_{\tau,h}$ by

<3.5> $$dQ_{\tau,h}(\eta) = dQ_0(\eta)\left(1 + \tau h(\eta)\right),$$

for small τ. In addition to assuming that $\mathcal{Q}$ is constrained to ensure that $\mathcal{P}(\Theta, \mathcal{Q})$ is identifiable, we also assume that $\mathcal{Q}$ contains all distributions $Q_{\tau,h}$ for τ in a neighborhood of zero depending upon h and Q_0.

Let $P_0 \in \mathcal{P}(\Theta, \mathcal{Q})$ be the assumed true distribution with structural parameter $\theta_0 \in \Theta$ and mixing distribution $Q_0 \in \mathcal{Q}$. Let $f_{\tau,h}$ denote the perturbed mixed density

$$\begin{aligned} f_{\tau,h}(x) &= f(x \mid \theta_0 + \tau, Q_{\tau,h}) \\ &= \int f(x \mid \theta_0 + \tau, , \eta)\, dQ_{\tau,h}(\eta). \end{aligned}$$

Denote the density for P_0 by f_0. It is obtained by evaluating $f_{\tau,h}$ at $\tau = 0$.

Using paths of the form <3.5>, we will show that if $\mathcal{Q}$ is a large enough class and if $\mathcal{F}(\Theta,\mathcal{N})$ satisfies regularity conditions, then it is possible to find a $Q_0 \in \mathcal{Q}$ and a $h \in \mathcal{L}_0^\infty(Q_0)$ so that $f_{\tau,h} \approx f_0$. This in turn will show that the structural parameter cannot be estimated at a $O_p\left(n^{-1/2}\right)$ rate.

The heuristic for the argument is as follows. Assuming that it is possible to differentiate densities:

$$\begin{aligned} f_{\tau,h}(x) &= \int f(x \mid \theta_0+\tau,,\eta)[1+\tau h(\eta)]\, dQ_{\tau,h}(\eta) \\ &\approx \int \left[f(x \mid \theta_0,\eta) + \tau \frac{\partial}{\partial\theta} f(x \mid \theta_0,\eta) + \cdots \right] [1+\tau h(\eta)]\, dQ_{\tau,h}(\eta) \\ &= f_0(x) + \tau \int \left[h(\eta) f(x \mid \theta_0,\eta) + \frac{\partial}{\partial\theta} f(x \mid \theta_0,\eta) \right] dQ_0(\eta) + \cdots . \end{aligned}$$

By dividing throughout by f_0, obtain the approximation

$$\frac{f_{\tau,h}(x)}{f_0(x)} \approx 1 + \tau\left[A(Q_0,h)(x) + \rho(x)\right],$$

where

$$A(Q_0,h)(x) = \frac{1}{f_0(x)} \int h(\eta) f(x \mid \theta_0,\eta)\, dQ_0(\eta),$$

and

$$\rho(x) = \frac{1}{f_0(x)} \int \frac{\partial}{\partial\theta} f(x \mid \theta_0,\eta)\, dQ_0(\eta).$$

To make $f_{\tau,h} \approx f_0$, the aim will be to find a Q_0 and an h such that

$$\rho(x) \approx -A(Q_0,h)(x).$$

Here are the conditions that justify the formal differentiation:

<3.6> **Regularity Conditions:** *Suppose $P_0 \in \mathcal{P}(\Theta,\mathcal{Q})$ with parameter (θ_0,Q_0). Say that P_0 satisfies regularity conditions, if for small τ*

$$\Delta(x,\eta,\tau) = \frac{\partial}{\partial\theta} f(x \mid \theta_0+\tau,\eta)$$

exists for a.a. $x\,[\nu]$ and a.a. $\eta\,[Q_0]$, such that:

(i) $\Delta(x,\eta,\tau)$ is continuous in τ,

(ii) there exists a dominating function M such that

<3.7>
$$\frac{\Delta(x,\eta,\tau)^2}{f(x \mid \theta_0,\eta)} \le M(x,\eta),$$

and $Q_0 M(x,\cdot) \in L^1(\nu)$.

Define $L_0^2(P_0)$ to be the equivalence class of P_0 square integrable functions which have zero P_0-expectation. Under the previous conditions it is possible to show:

<3.8> **Lemma:** *Assume that Regularity Conditions <3.6> hold for $P_0 \in \mathcal{P}(\Theta, \mathcal{Q})$ with parameter (θ_0, Q_0). Then for each $h \in \mathcal{L}_0^\infty(Q_0)$*

<3.9> $$P_0\left(\frac{f_{\tau,h}}{f_0} - 1\right)^2 \leq \tau^2 P_0(A(Q_0, h) \ + \rho)^2,$$

for small τ, where $\rho \in L_0^2(P_0)$ and $A(Q_0, \cdot)$ is the linear map from $\mathcal{L}_0^\infty(Q_0)$ into $L_0^2(P_0)$.

The next lemma asserts the existence of a Q_0 and an $h \in \mathcal{L}_0^\infty(Q_0)$ which makes the right-hand side of <3.9> small.

<3.10> **Lemma:** *Suppose $Q_0 \in \mathcal{Q}$ is a discrete distribution with countably infinite many atoms on $\mathcal{N}^* \subset \mathcal{N}$. Then the range of the linear operator $A(Q_0, \cdot)$ is dense in $L_0^2(P_0)$.*

Lemma <3.8> and Lemma <3.10> will establish the following theorem.

<3.11> **Theorem:** *Assume that Regularity Conditions <3.6> hold for $P_0 \in \mathcal{P}(\Theta, \mathcal{Q})$ with parameter (θ_0, Q_0), and that Q_0 satisfies the conditions of Lemma <3.10>. Then a regular estimator for θ_0 at P_0 must have rate of convergence slower than $O_p\left(n^{-1/2}\right)$.*

Proof of Theorem <3.11>: The operator $A(Q_0, \cdot)$ is dense by Lemma <3.10>. Therefore, for each $\epsilon > 0$ there exists a function $h_0 \in \mathcal{L}_0^\infty(Q_0)$ satisfying $P_0(A(Q_0, h_0) + \rho)^2 < \epsilon$. By Lemma <3.8>,

$$P_0\left(\frac{f_{\tau,h_0}}{f_0} - 1\right)^2 < \epsilon\tau^2, \qquad \text{eventually.}$$

The fact that ϵ is arbitrary proves the theorem by Lemma <2.11>. □

<3.12> **Example (Weibull mixture model, continued):** Let $\mathcal{Q}$ be the class of mixing distributions with support on the set $\mathcal{N}^* = (\eta_0, \eta_1)$, where $0 < \eta_0 < \eta_1 < \infty$. Because $\mathcal{N}^*$ is bounded, each $Q \in \mathcal{Q}$ must have first moment bounded by η_1. Chapter 5.1 shows that this is a sufficient condition to ensure that $\mathcal{P}(\Theta, \mathcal{Q})$ is identifiable. Notice, as well, that if $Q_0 \in \mathcal{Q}$ and $h \in \mathcal{L}_0^\infty(Q_0)$, then

$$\begin{aligned}\int \eta \, dQ_{\tau,h}(\eta) &= \int \eta \, dQ_0(\eta) + \tau \int h(\eta) \, dQ_0(\eta) \\ &< \eta_1 \left(1 + |\tau| \, \|h\|_\infty\right),\end{aligned}$$

where $\|h\|_\infty$ is the sup-norm for h. This implies that the left-hand side of the expression must be less than or equal to η_1 eventually, and therefore, that $\mathcal{Q}$ is rich enough to contain all distributions of the form <3.5>, for small enough τ.

We show below that Regularity Conditions <3.6> holds for each $P \in \mathcal{P}(\Theta, \mathcal{Q})$. Therefore if $P_0 \in \mathcal{P}(\Theta, \mathcal{Q})$ with parameter (θ_0, Q_0), then Theorem <3.11> shows that it is not

possible to estimate θ_0 at a $O_p\left(n^{-1/2}\right)$ rate when Q_0 is discrete with countably infinite many atoms.

Let us verify the regularity conditions. Differentiation with respect to θ gives

$$\frac{\partial}{\partial\theta} f(x \mid \theta, \eta) = g(x, \theta, \eta) f(x \mid \theta, \eta),$$

where

$$g(x, \theta, \eta) = \frac{1}{\theta} + \log x - \eta x^{\theta} \log x.$$

To establish the regularity conditions we must show that the right-hand side of the inequality

$$\begin{aligned} \frac{\Delta(x, \eta, \tau)^2}{f(x \mid \theta_0, \eta)} &= g(x, \theta_0 + \tau, \eta)^2 \frac{f(x \mid \theta_0 + \tau, \eta)^2}{f(x \mid \theta_0, \eta)} \\ &\leq C g(x, \theta_0 + \tau, \eta)^2 x^{(\theta_0 + 2\tau - 1)} \exp\left[-\eta x^{\theta_0} \left(2x^{\tau} - 1\right)\right], \end{aligned} \qquad \text{<3.13>}$$

can be bounded by a function independent of τ satisfying <3.7> (here C is a finite constant depending upon θ_0 for small τ).

By restricting τ to be positive, construct a dominating integrable function for <3.13> over the separate regions $0 \leq x \leq 1$ and $x > 1$ (for example, use the fact $x^s |\log x|^t$ is integrable over $0 \leq x \leq 1$, for $s, t > -1$). Use the boundedness of $\mathcal{N}^*$ to construct the function independent of η to show that <3.7> holds. This verifies the regularity conditions. □

Now to prove the two lemmas.

Proof of Lemma <3.8>: Use the mean value theorem to expand the perturbed density as

$$f(x \mid \theta_0 + \tau, \eta) = f(x \mid \theta_0, \eta) + \tau \Delta(x, \eta, 0) + r(x, \eta, \tau), \qquad \text{<3.14>}$$

where the remainder term is defined to be

$$r(x, \eta, \tau) = \tau \left[\Delta(x, \eta, \tau^*) - \Delta(x, \eta, 0)\right], \qquad \text{<3.15>}$$

and $\tau^* = \tau^*(x, \eta, \tau)$ is bounded by $|\tau|$.

Square both sides of <3.15>, divide throughout by $\tau^2 f(\cdot \mid \theta_0, \eta)$, and integrate with respect to ν to write

$$\frac{1}{\tau^2} \int \frac{r(x, \eta, \tau)^2}{f(x \mid \theta_0, \eta)} \, d\nu(x) = \int \frac{1}{f(x \mid \theta_0, \eta)} \left[\Delta(x, \eta, \tau^*) - \Delta(x, \eta, 0)\right]^2 d\nu(x).$$

Take expectations with respect to Q_0 on both sides of the expression. Use the dominated convergence theorem with dominating function

$$4 \int M(x, \eta)\, d\nu(x)$$

(which must be Q_0 integrable by regularity condition <3.7> and Fubini's Theorem), to show by the continuity of $\Delta(x, \eta, \tau)$ in τ that:

<3.16> $$Q_0 \int \frac{r(x, \eta, \tau)^2}{f(x \mid \theta_0, \eta)}\, d\nu(x) = o\left(\tau^2\right).$$

Divide <3.14> throughout by f_0 on the set where it is nonzero (which has P_0 measure one), and take expectations with respect to $Q_{\tau,h}$ to write

$$\frac{f_{\tau,h}(x)}{f_0(x)} = \frac{1}{f_0(x)} \int \left[f(x \mid \theta_0, \eta) + \tau\Delta(x, \eta, 0) + r(x, \eta, \tau)\right] \left[1 + \tau h(\eta)\right]\, dQ_0(\eta).$$

Assume that a term by term expansion on the right side is justifiable, yielding

<3.17> $$\begin{aligned} &1 + \frac{\tau}{f_0(x)} \left[\int h(\eta) f(x \mid \theta_0, \eta)\, dQ_0(\eta) + \int \Delta(x, \eta, 0)\, dQ_0(\eta)\right] \\ &\qquad + \frac{\tau^2}{f_0(x)} \int h(\eta)\Delta(x, \eta, 0)\, dQ_0(\eta) + \frac{1}{f_0(x)} \int r(x, \eta, \tau)\, dQ_{\tau,h}(\eta). \end{aligned}$$

Recognize that the coefficient of τ equals $A(Q_0, h) + \rho$. Collect the remainder terms, to write the likelihood ratio as

<3.18> $$\frac{f_{\tau,h}(x)}{f_0(x)} - 1 = \tau\Big(A(Q_0, h)(x) + \rho(x)\Big) + R(x, h, \tau),$$

for a.a. $x\,[P_0]$.

The lemma will be established by showing that $A(Q_0, h)$ and ρ are $L_0^2(P_0)$-functions, and that $R(\cdot, h, \tau)$ has squared $L^2(P_0)$-norm of order $o\left(\tau^2\right)$. Let us start with the last term in <3.17>. Because we can bound h by its sup-norm, it is sufficient to consider

<3.19> $$\begin{aligned} &P_0\left(\frac{1}{f_0} \int r(\cdot, \eta, \tau)\, dQ_0(\eta)\right)^2 \\ &\qquad = \int \frac{1}{f_0(x)} \left[\int \frac{r(x, \eta, \tau)}{f(x \mid \theta_0, \eta)^{1/2}} f(x \mid \theta_0, \eta)^{1/2}\, dQ_0(\eta)\right]^2 \{f_0(x) \neq 0\}\, d\nu(x). \end{aligned}$$

From the Cauchy-Schwarz inequality, bound <3.19> by

$$\iint \frac{r(x, \eta, \tau)^2}{f(x \mid \theta_0, \eta)}\, dQ_0(\eta)\, d\nu(x).$$

Interchange the order of integration to deduce by <3.16> that this term is of order $o\left(\tau^2\right)$. This takes care of the last term in <3.17>. The other terms are dealt with in much the same fashion. Notice also, that if T is a function such that $Q_0 T(x, \eta) \in L^2(P_0)$,

then $T(x,\cdot) \in L^1(Q_0)$ a.a. $x\,[P_0]$. Therefore, the term by term expansion leading to the expression <3.17> is valid for a.a. $x\,[P_0]$.

Finally, choose h to equal zero. Take expectations of <3.18> with respect to P_0, to show that ρ has zero P_0-expectation. Consequently, $A(Q_0,h)$ has zero P_0-expectation. □

To prove Lemma <3.10> we will have need of the following Fourier result.

<3.20> **Lemma:** *Let λ be a σ-finite measure dominated by Lebesgue measure. If there exists a sequence of distinct real numbers z_n converging to $z_0 \in (\alpha,\beta)$ such that*

$$\int |\gamma(x)| \exp(zx)\, d\lambda(x) < \infty \qquad \text{for} \quad \alpha < z < \beta,$$

and

$$\int \gamma(x) \exp(z_n x)\, d\lambda(x) = 0,$$

then

$$\gamma = 0 \qquad \text{a.e.}\,[\lambda].$$

Proof of Lemma <3.20>: The integrability condition implies that

$$g(z) = \int \gamma(x) \exp(zx)\, d\lambda(x)$$

is analytic on $z \in \Lambda(\alpha,\beta) = \{z : \alpha < \mathrm{Re}(z) < \beta\}$ (Lehmann, 1986: Theorem 9, Chapter 2.7). Because every neighborhood of z_0 contains points of the sequence z_n, the analytic nature of g implies that it must equal zero in some neighborhood of z_0. Therefore by analytic continuation, g equals zero on $\Lambda(\alpha,\beta)$. Appeal to the uniqueness of the Fourier transform (Rudin, 1987: Chapter 9) to establish the result. □

Proof of Lemma <3.10>: Now to prove that the range of $A(Q_0,\cdot)$ is dense in $L_0^2(P_0)$. To do so, we will show that if $\psi \in L_0^2(P_0)$ satisfies $P_0\psi A(Q_0,h) = 0$ for all $h \in \mathcal{L}_0^\infty(Q_0)$, then $\psi = 0$ a.e. $[P_0]$.

Suppose $\psi \in L_0^2(P_0)$ such that $P_0\psi A(Q_0,h) = 0$ for all h. Because both ψ and $A(Q_0,h)$ are elements of $L^2(P_0)$, we can interchange the order of integration to obtain

<3.21> $$\iint h(\eta)\psi(x)\frac{f(x \mid \theta_0,\eta)}{f_0(x)}\, dP_0(x)\, dQ_0(\eta) = 0.$$

Therefore, $Q_0 hT = 0$ for all $h \in \mathcal{L}_0^\infty(Q_0)$, where

$$T(\eta) = \int \psi(x) f(x \mid \theta_0,\eta)\, d\nu(x).$$

Define $h_0 = T - \xi$, where $\xi = Q_0 T$. If T is bounded on $\mathcal{N}^*$ (which contains the support of Q_0), then $h_0 \in \mathcal{L}_0^\infty(Q_0)$. For the moment assume that this is the case. From <3.21>, we

have that $Q_0\, h_0 T = 0$. Because h_0 has zero Q_0-expectation, $Q_0\, h_0(T-\xi) = 0$. This implies that $Q_0\, h_0^2 = 0$, and hence that T must equal ξ for all η in the support of $[Q_0]$ (recall that Q_0 is discrete). To determine what ξ equals, interchange the order of integration to show

$$\xi = \int \psi(x) \int f(x \mid \theta_0, \eta)\, dQ_0(\eta)\, d\nu(x),$$

which equals $P_0\psi$ and is therefore zero.

Therefore, we have shown that

<3.22> $$\int \psi(x) f(x \mid \theta_0, \eta)\, d\nu(x) = 0,$$

for all η in the support of Q_0.

Standard results about exponential families show that $b(\theta_0, \cdot)$ is continuous on $\mathcal{N}(\theta_0)$. Because $\mathcal{N}^*$ is a proper subset of $\mathcal{N} = \mathcal{N}(\theta_0)$ we can find a compact subset $[\eta_0, \eta_1] \subset \mathcal{N}$ which contains it. Thus, the continuity of $b(\theta_0, \cdot)$ implies

<3.23> $$\sup_{\eta_0 \le \eta \le \eta_1} \exp(-b(\theta_0, \eta)) < \infty.$$

This allows us to rewrite <3.22> as

<3.24> $$\int \psi(x) \exp(\eta s(x, \theta_0))\, d\nu^*(x) = 0,$$

for all η in the support of Q_0, where $d\nu^*(x) = \exp(t(x, \theta_0)) dv(x)$.

By assumption, Q_0 puts positive mass on a sequence of atoms $a_n \in \mathcal{N}^*$ converging to an interior point of $\mathcal{N}^*$. Thus, the integral in <3.24> equals zero when η equals a_n. Therefore, if we can show that

<3.25> $$\int |\psi(x)| \exp(\eta s(x, \theta_0))\, d\nu(x)^* < \infty, \qquad \text{for } \eta_0 < \eta < \eta_1,$$

then a change of variables and an application of Lemma <3.20> will establish the desired result.

The same argument which shows T to be bounded, will also show that <3.25> holds. Let us first show that T is bounded. Write T as

<3.26> $$T(\eta) = \int \left(\psi(x) f_0(x)^{1/2}\right) \left(\frac{f(x \mid \theta_0, \eta)}{f_0(x)^{1/2}}\right) d\nu(x).$$

If we can show that each expression in parenthesis is square integrable, uniformly in η, then an application of the Cauchy-Schwarz inequality will establish the result. The first expression is easily dealt with by our assumption that $\psi \in L^2(P_0)$. To deal with the second term, consider the inequality

$$f(x \mid \theta_0, \eta) \ge f(x \mid \theta_0, \widetilde{\eta})\, \{\eta = \widetilde{\eta}\}\,,$$

where $\widetilde{\eta}$ is an atom of Q_0. Take expectations with respect to Q_0 on both sides of the inequality to show that $1/f_0 \leq C/f(\cdot \mid \theta_0, \widetilde{\eta})$, where C is a finite constant depending upon $\widetilde{\eta}$. Therefore,

<3.27> $$\int \frac{f(x \mid \theta_0, \eta)^2}{f_0(x)}\, d\nu(x) \leq C \int \frac{f(x \mid \theta_0, \eta)^2}{f(x \mid \theta_0, \widetilde{\eta})}\, d\nu(x).$$

The integrand on the right side can be written as

<3.28> $$\exp\Big[(2\eta - \widetilde{\eta})s(x, \theta_0) + t(x, \theta_0) + 2b(\theta_0, \eta) - b(\theta_0, \widetilde{\eta})\Big].$$

Use the continuity of $b(\theta_0, \cdot)$ on $\mathcal{N}^* \subset [\eta_0, \eta_1]$ to bound the terms involving $b(\theta_0, \cdot)$ by a finite constant, B_0. By considering whether $s(\cdot, \theta_0)$ is positive or negative, bound <3.28> by

$$B_0 \exp(t(x, \theta_0))\Big[\exp[(2\eta_1 - \eta_0)s(x, \theta_0)] + \exp[\eta_0 s(x, \theta_0)]\Big].$$

Because $2\eta_1 - \eta_0$ and η_0 are interior points of $\mathcal{N} = \mathcal{N}(\theta_0)$, deduce that the expression <3.28> must be ν-integrable . Apply the Cauchy-Schwarz inequality in <3.26> to deduce that

<3.29> $$|T(\eta)|^2 \leq \int \psi(x)^2 f_0(x)\, d\nu(x) \int \frac{f(x \mid \theta_0, \eta)^2}{f_0(x)}\, d\nu(x)$$

is bounded uniformly for $\eta \in \mathcal{N}^*$.

To show that <3.25> holds, multiply the left side of the expression by $\exp(b(\theta_0, \eta))$ and use the previous argument to establish that $\psi f(\cdot \mid \theta_0, \eta)$ is integrable for $\eta_0 < \eta < \eta_1$. Now use <3.23> to conclude the proof. □

<3.30> **Remarks:** It is worth pointing out an interesting connection to the paper of Begun, Hall, Huang, and Wellner (1984) discussed in Chapter 1. By the inequality <2.9>, Theorem <3.11> shows that it is possible to find a model P_0, such that

$$\begin{aligned}\int \left(\sqrt{f_{\tau,h}} - \sqrt{f_0}\right)^2 &\leq \tau^2 P_0(A(Q_0, h) \ + \rho)^2 \\ &< \epsilon\tau^2,\end{aligned}$$

for arbitrarily small ϵ. In the context of the Begun paper, this implies that the model has zero information even for fairly smooth paths <3.5>.

<3.31> **Remarks:** In the proof of Lemma <3.8>, we show using Regularity conditions <3.6> that $r(\cdot, \tau)^2/f(\cdot \mid \theta_0, \eta)$ has $L^1(\nu)$-norm of order $o\left(\tau^2\right)$, for a.a. $\eta\,[Q_0]$. The same argument also implies that $\Delta(\cdot, \eta, 0)^2/f(\cdot \mid \theta_0, \eta)$ is a ν-integrable function, for a.a.

$\eta\,[Q_0]$. Consequently, the likelihood ratio

$$\frac{f(\cdot \mid \theta_0 + \tau, \eta)}{f(\cdot \mid \theta_0, \eta)}$$

is $L^2(f(\cdot \mid \theta_0, \eta))$-differentiable with first derivative $\Delta(\cdot, \eta, 0)/f(\cdot \mid \theta_0, \eta)$, for a.a. $\eta\,[Q_0]$ in the following sense:

<3.32> **Definition:** *Let $g(\cdot, \tau)$ be a density, where τ is real. Define $g_0 = g(\cdot, 0)$. Say that $g(\cdot, \tau)/g_0$ is $L^2(g_0)$-differentiable, with first derivative g_1, if*

$$\frac{g(x, \tau)}{g_0(x)} = 1 + \tau g_1(x) + r(x, \tau),$$

where $g_1 \in L^2(g_0)$ and $r(\cdot, \tau)$ has $L^2(g_0)$-norm of order $o\,(\tau)$.

This type of differentiabilty will become useful in the next section when we come to work with L^2-distances of convolutions (Lemma <3.38>).

3. A Fourier Technique for Scale Location-Mixtures Let us introduce a new technique for determining rates of convergence. Consider a random variable

$$X = \theta Z + Y,$$

where the random variable Z has a known density h_0, the parameter θ is real, and Y has an unknown distribution Q, independent of Z. The random variable X describes a location-mixture model with unknown scale parameter θ. Let Θ be a real parameter space and $\mathcal{Q}$ a class of mixing distributions that are absolutely continuous with respect to Lebesgue measure. Take $\mathcal{P}(\Theta, \mathcal{Q})$ to be the class of mixed distributions, induced by Θ and $\mathcal{Q}$, which can be described as convolutions like X. The problem will be to estimate the structural parameter θ, assuming that enough constraints have been placed on Θ and $\mathcal{Q}$ to make $\mathcal{P}(\Theta, \mathcal{Q})$ identifiable.

Because the distribution for X is a convolution, we can write its characteristic function as the product of the characteristic functions of the distributions for θZ and Y. The interaction between the structural parameter and the mixing distribution in the real domain becomes in the complex domain a much simpler relationship involving products of characteristic functions. The following heuristic argument takes advantage of this fact and utilizes Fourier analysis for determining rates of convergence.

First some notation. Denote the Fourier transform of $f \in L^1(\mu)$ by $\widehat{f}$, where

$$\widehat{f}(t) = \int_{-\infty}^{+\infty} \exp(it\eta) f(\eta)\, d\eta, \qquad t \in I\!R.$$

Define the convolution between f and $g \in L^1(\mu)$ as

$$[f * g](x) = \int_{-\infty}^{+\infty} f(x-y)g(y)\,dy, \qquad x \in I\!R.$$

For convenience let us assume that $\theta_0 = 1$ is an element of Θ. Let $P_0 \in P(\Theta, \mathcal{Q})$ equal the true mixed distribution with density f_0, structural parameter θ_0, and mixing density q_0. All densities are taken with respect to Lebesgue measure. Let $P_\tau \in \mathcal{P}(\Theta, \mathcal{Q})$ be a perturbed distribution with density $f(\cdot, \tau)$, structural parameter $1 + \tau$, and mixing distribution Q_τ. Assume that the density $h(\cdot, \tau)$ for the random variable $(1 + \tau)Z$ can be expressed as the Taylor series expansion

<3.33> $$h(x, \tau) = h_0(x) + \tau h_1(x) + \tau^2 h_2(x) + \cdots,$$

where each term in the expansion is Lebesgue integrable.

Assume that Q_τ has a density which can be expressed as

<3.34> $$q(y, \tau) = q_0(y) + \tau q_1(y) + \tau^2 q_2(y) + \cdots,$$

where, again, all terms in the expansion are integrable. Then, the density for P_τ is

$$\begin{aligned} f(x, \tau) &= [h(\cdot, \tau) * q(\cdot, \tau)](x) \\ &= \left[\left(h_0 + \tau h_1 + \tau^2 h_2 + \cdots\right) * \left(q_0 + \tau q_1 + \tau^2 q_2 \cdots\right)\right](x). \end{aligned}$$

Expand by collecting coefficients in powers of τ. Recognize that $f_0 = h_0 * q_0$ to show

<3.35> $$\begin{aligned} f(x, \tau) - f_0(x) = \tau&\Big([h_0 * q_1](x) + [h_1 * q_0](x)\Big) \\ &+ \tau^2\Big([h_0 * q_2](x) + [h_1 * q_1](x) + [h_2 * q_0](x)\Big) + \cdots. \end{aligned}$$

To make discrimination between P_0 and P_τ difficult, we would like to make $f(\cdot, \tau)$ as close as possible to f_0. Expression <3.35> indicates the first place of attack should be the linear term and suggests we choose Q_τ such that

<3.36> $$h_0 * q_1 + h_1 * q_0 = 0,$$

or so that the left-hand side can be made as close to zero as possible. It will be convenient to recast this problem into one involving products of Fourier transforms. Shortly we will see how to do this.

Let us assume that the left-hand side of <3.36> can be made to be exactly zero. Divide <3.35> by f_0 to obtain

<3.37> $$\frac{f(x, \tau)}{f_0(x)} - 1 = \tau^2\left(\frac{1}{f_0(x)}[h_0 * q_2](x) + \frac{1}{f_0(x)}[h_1 * q_1](x) + \frac{1}{f_0(x)}[h_2 * q_0](x)\right) + \cdots.$$

By Lemma <2.10>, to be able to infer a lower bound of $n^{-1/4}$ for the rate of convergence, it is sufficient to show

$$P_0\left(\frac{f(\cdot,\tau)}{f_0}-1\right)^2 = O\left(\tau^4\right).$$

This can be established by showing that the right-hand side of <3.37> has squared $L^2(P_0)$-norm of order $O\left(\tau^4\right)$. The following lemma will be helpful in showing this.

<3.38> **Lemma:** *For functions F_1, F_2, G_1, and G_2 in $L^2(\mu)$*

<3.39>
$$\int \frac{\left[(F_1G_1)*(F_2G_2)\right]^2}{F_1^2*F_2^2} \le \int G_1^2 \int G_2^2.$$

Proof of Lemma <3.38>: By the Cauchy-Schwarz inequality

$$\begin{aligned}\left[(F_1G_1)*(F_2G_2)\right]^2 &= \left(\int F_1(x-y)G_1(x-y)F_2(y)G_2(y)\,dy\right)^2 \\ &\le \int F_1(x-y)^2F_2(y)^2\,dy \int G_1(x-y)^2G_2(y)^2\,dy.\end{aligned}$$

The first integral on the right-hand side equals $F_1^2*F_2^2$, which is finite almost everywhere, being a convolution of two integrable functions. When this factor is finite (and nonzero), we deduce that the integrand on the left-hand side of <3.39> is bounded by

<3.40>
$$\int G_1(x-y)^2G_2(y)^2\,dy.$$

(When $F_1^2*F_2^2$ equals zero, the left-hand side of <3.39> equals zero, so that the same bound still holds.)

The integral over x of the expression <3.40> factorizes into the product on the right-hand side of <3.39>. □

For example, to show that the term $\tau^2\left[h_1*q_1\right]/f_0$ has squared $L^2(P_0)$-norm of order $O\left(\tau^4\right)$, we need to show that $\left[h_1*q_1\right]^2/f_0$ is Lebesgue integrable. This can be established by applying the previous lemma with $F_1=\sqrt{h_0}$, $F_2=\sqrt{q_0}$, $G_1=h_1/\sqrt{h_0}$ and $G_2=q_1/\sqrt{q_0}$:

$$\begin{aligned}\int \frac{\left[h_1*q_1\right]^2}{f_0} &= \int \frac{\left[\left(\sqrt{h_0}\;\; h_1/\sqrt{h_0}\right)*\left(\sqrt{q_0}\;\; q_1/\sqrt{q_0}\right)\right]^2}{\left(\sqrt{h_0}\right)^2*\left(\sqrt{q_0}\right)^2} \\ &\le \int \frac{h_1^2}{h_0}\int \frac{q_1^2}{q_0}.\end{aligned}$$

The integrability of $\left[h_1*q_1\right]^2/f_0$ can be established by showing that that both integrals on the right-hand side are finite. The first integral can be made finite by assuming that that the likelihood ratio $h(\cdot,\tau)/h_0$ is $L^2(h_0)$-differentiable, while the second integral can be dealt with by imposing constraints on the the mixing distribution Q_τ.

Now, let us describe a Fourier technique for solving <3.36>. The Fourier analog of <3.36> is

<3.41> $$\widehat{h}_0(t)\widehat{q}_1(t) + \widehat{h}_1(t)\widehat{q}_0(t) = 0.$$

This problem can be formulated more simply by assuming that $\widehat{h}_0(t)$ is expressible as $\exp(l(t))$. Assume, as well, that $l(t)$ is smooth enough that a Taylor series expansion of $\widehat{h}(t,\tau) = \exp\bigl[l(t+\tau t)\bigr]$ about $\tau = 0$ yields

<3.42> $$\exp(l(t)) + \tau \exp(l(t))tl'(t) + \cdots .$$

The Fourier transform of $h(\cdot,\tau)$ should have an expansion analogous to <3.33>

$$\widehat{h}(t,\tau) = \widehat{h}_0(t) + \tau\widehat{h}_1(t) + \tau^2\widehat{h}_2(t) + \cdots .$$

Comparing this with <3.42>, gives

$$\widehat{h}_1(t) = \widehat{h}_0(t)tl'(t).$$

Take out the common factor $\widehat{h}_0$ so as to reformulate the Fourier problem <3.41> more simply as,

<3.43> $$\widehat{q}_1(t) + tl'(t)\widehat{q}_0(t) = 0.$$

The advantage of trying to satisfy the Fourier expression <3.43> over the analogous real expression <3.36> is that the problem shifts from one involving convolutions of functions to a simpler one involving products of functions. In particular, solving the Fourier problem will involve choosing Q_0 so that the density q_0 and its Fourier transform $\widehat{q}_0$ are easy to work with, while at the same time choosing an appropriate Fourier perturbation $\widehat{q}_1$.

It will not always be possible to choose q_0 and q_1 to satisfy all the constraints and to make the left-hand side of <3.43> exactly zero, but it is possible to use <3.43> as a starting point to choose q_0 and q_1 in such a way that the leading term on the right-hand side of

$$\frac{f(x,\tau)}{f_0(x)} - 1 = \frac{\tau}{f_0(x)}\Bigl([h_0 * q_1](x) + [h_1 * q_0](x)\Bigr) + \frac{\tau^2}{f_0(x)}\Bigl([h_0 * q_2](x) + [h_1 * q_1](x) + [h_2 * q_0](x)\Bigr) + \cdots$$

contributes only a term with squared $L^2(P_0)$-norm of order $o\left(\tau^2\right)$ to the left-hand side. A careful analysis will then involve consideration of the remainder terms, to ensure that they not undo the work that goes into making the coefficient of τ small.

Chapter 4

Normal Mean-Mixture Model

1. Introduction The normal mean-mixture model is formed by mixing over the mean of a normal density with unknown standard deviation. Interest in this chapter will focus on the problem of estimation for the standard deviation in the presence of the nuisance mixing distribution. More precisely, we have independent observations from $\theta Z + Y$, where Z has a $N(0,1)$ distribution, θ is a unknown parameter in $\Theta = (0,\infty)$, and Y has an unknown distribution Q, independent of Z. The problem is to estimate the standard deviation, θ.

We showed earlier in Example <2.19> of Chapter 2, that the class of mixed distributions $\mathcal{P}(\Theta,\mathcal{Q})$ induced by Θ and a class of mixing distributions $\mathcal{Q}$ will not be identifiable if $\mathcal{Q}$ contains distributions which have normal components. For example, if τ is a small real number, then there is no way to discriminate between a $N(0,1-\tau)$ distribution convolved with a $N(0,1+\tau)$ distribution, and a $N(0,1)$ distribution convolved with a $N(0,1)$ distribution. Thus, for the estimation problem to be of any practical significance, we need to place constraints on Θ and $\mathcal{Q}$ so as to make the model identifiable.

Example <2.19> makes it clear that the standard deviation might not be estimable at any rate of convergence if the model is only required to be identifiable. The example takes advantage of the near lack of identifiablity by considering mixing distributions which are nearly normal. One wonders then, if the only examples which establish slow rates are those which take advantage of the near lack of identifiability. We show that this is not the case in sections 2 and 3 of the chapter.

Section 2 verifies the regularity conditions for Theorem <3.11> and shows that the model has zero information even when the mixing distribution is constrained to be discrete with finite support. Furthermore, the proof that the model has zero information follows from considering smooth paths through the mixing space $\mathcal{Q}$.

Section 3 uses the Fourier technique of Chapter 3.3 to deduce lower bounds for rates of convergence. There $\mathcal{Q}$ is constrained in the frequency domain and rates are directly related to the manner of constraint.

2. Zero Information Let $\mathcal{F}(\Theta,\mathcal{N})$ be the parametric family of normal densities of the form

$$f(x \mid \theta,\eta) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\left[-\frac{1}{2\theta^2}(x-\eta)^2\right],$$

where θ takes values in $\Theta = (0,\infty)$, and η ranges over $\mathcal{N} = I\!R$. Let $\mathcal{Q}$ be the class of mixing distributions with support contained within the set $\mathcal{N}^* \subset \mathcal{N}$. Form the class of

identifiable normal-mean mixture models $\mathcal{P}(\Theta, \mathcal{Q})$ by mixing over the densities in $\mathcal{F}(\Theta, \mathcal{N})$ by distributions in $\mathcal{Q}$.

We show that the model has zero information according to Theorem <3.11> of Chapter 3.2, by verifying Regularity Conditions <3.6> of the same chapter.

First verify that the densities satisfy the parametric form <3.3>. Check that this holds with $s(x,\theta) = x/\theta^2$, $t(x,\theta) = -x^2/(2\theta^2)$ and $b(\theta,\eta) = -\eta^2/(2\theta^2) - \log(2\pi\theta^2)/2$. Now verify the remaining conditions. Observe that for small τ,

$$\frac{f(x \mid \theta_0+\tau,\eta)^2}{f(x \mid \theta_0,\eta)} \le C_1 \exp\left[-\frac{1}{4\theta_0^2}(x-\eta)^2\right], \tag{4.1}$$

where C_1 is a fixed positive constant.

Differentiation yields

$$\frac{\partial}{\partial\theta} f(x \mid \theta,\eta) = g(x,\theta,\eta) f(x \mid \theta,\eta),$$

where

$$g(x,\theta,\eta) = -\frac{1}{\theta} + \frac{1}{\theta^3}(x-\eta)^2.$$

Use <4.1> to obtain the bound

$$\frac{[\partial f(x \mid \theta_0+\tau,\eta)/\partial\theta]^2}{f(x \mid \theta_0,\eta)} \le C_2 \left[1+(x-\eta)^2\right]^2 \exp\left[-\frac{1}{4\theta_0^2}(x-\eta)^2\right], \tag{4.2}$$

for small τ, where C_2 is a fixed constant. Integrate <4.2> over x to obtain a function independent of η to show that <3.7> holds and consequently that Regularity Conditions <3.6> hold.

Therefore, if $P_0 \in \mathcal{P}(\Theta, \mathcal{Q})$ with parameter (θ_0, Q_0), then Theorem <3.11> shows that the rate of convergence for estimators of the standard deviation must be slower than $O_p\left(n^{-1/2}\right)$ when Q_0 is discrete with countably infinite many atoms. In particular the theorem shows, for such Q_0, the existence of an $h \in \mathcal{L}_0^\infty(Q_0)$ which makes discrimination difficult between P_0 and the model P_τ with structural parameter $\theta_0 + \tau$ and mixing distribution

$$dQ_{\tau,h} = dQ_0(1+\tau h). \tag{4.3}$$

The next section extends this result. By working with a constrained class $\mathcal{Q}$, we show how to construct mixing paths, analogous to <4.3>, of the form

$$q(y,\tau) = q_0(y) + \tau q_1(y) + \cdots + \tau^d q_d(y),$$

to derive explicit lower bounds for rates of convergence which depend upon d and the manner in which $\mathcal{Q}$ is constrained.

3. Rates of Convergence Using Fourier Analysis In this section we rigorously apply the Fourier argument of Chapter 3.3 to the problem of estimation for the standard deviation in the normal mean-mixture model. Because we work with Fourier transforms, the most natural way to ensure identifiability in the model will be to place Fourier constraints on the class of mixing distributions, $\mathcal{Q}$.

Assume that $\mathcal{Q}$ is absolutely continuous with respect to Lebesgue measure. To ensure that $\mathcal{Q}$ contain no normal distributions, require that each $Q \in \mathcal{Q}$ have density q (with respect to Lebesgue measure) such that

<4.4> $$\int \widehat{q}\,(t) t^{2d+1}\, dt = \infty,$$

where d is a fixed positive integer.

By constraining the tail behavior for the Fourier transform, condition <4.4> limits the amount of smoothness a distribution in $\mathcal{Q}$ might have. For example, consider the case when $d = 1$. Let Q be the distribution for the convolution of four uniform distributions on $[-1, +1]$. Then Q has Fourier transform $(\sin t)^4/t^4$ which satisfies <4.4>. Notice, however, that the distribution for five uniform $[-1, +1]$ distributions would be too smooth in this case. Its Fourier transform, $(\sin t)^5/t^5$, decreases too rapidly to satisfy <4.4> when $d = 1$.

Let $\mathcal{P}(\Theta, \mathcal{Q})$ be the class of identifiable normal-mean mixture models induced by $\mathcal{Q}$ and $\Theta = (0, \infty)$. We will show for this class of models, a lower bound of $O_p\left(n^{-1/(2d+2)}\right)$ for the rate of convergence. The fact that the result depends upon d shows that rates of convergence depend directly upon the amount of smoothness allowed in the mixing densities: the smoother the mixing density, the slower the rates of convergence.

Let $P_0 \in \mathcal{P}(\Theta, \mathcal{Q})$ be the true model with structural parameter θ_0 and mixing distribution Q_0 with density q_0. Without loss of generality, let $\theta_0 = 1$. Define P_τ as the perturbed mixed distribution with structural parameter $(1+\tau)^{1/2}$ and mixing distribution Q_τ. Then P_τ corresponds to the random variable $(1+\tau)^{1/2} Z + Y$, where Z is normally distributed, and Y has distribution Q_τ independent of Z. (Notice that the standard deviations for P_0 and P_τ are separated rougly by τ. Therefore, the convenient parameterization that we employ here will not affect the argument given in Chapter 3.3.)

Let h_0 equal the standard normal density. Expand the density $h(\cdot, \tau)$ for the random variable $(1+\tau)^{1/2} Z$ to $(d+2)$-terms in a Taylor series expansion about $\tau = 0$

<4.5> $$h(x, \tau) = h_0(x) + \tau h_1(x) + \cdots + \tau^d h_d(x) + \tau^{d+1} h_{d+1}(x, \tau),$$

where $h_m(x) = h_0(x) P_m(x)$ and $h_{d+1}(x, \tau) = h(x, \tau) P_{d+1}(x, \tau)$, and where $P_m(x)$ is a polynomial in x of degree $2m$, for $m = 1, 2, \cdots, d$ and $P_{d+1}(x, \tau)$ is a polynomialy in x of degree $2d + 2$.

Let Q_τ have density which can be expressed as the sum of $(d+1)$-integrable functions

$$q(y,\tau) = q_0(y) + \tau q_1(y) + \cdots + \tau^d q_d(y).$$

We will construct Q_τ by choosing q_0 and the perturbations $q_1, \cdots, q_d$ so as make discrimination between P_0 and P_τ difficult.

The density for $(1+\tau)^{1/2}Z + Y$ equals,

$$\begin{aligned} f(x,\tau) &= [h(\cdot,\tau) * q(\cdot,\tau)](x) \\ &= \left[\left(h_0 + \tau h_1 + \cdots + \tau^d h_d + \tau^{d+1} h_{d+1}(x,\tau)\right) * \left(q_0 + \tau q_1 + \cdots + \tau^d q_d\right)\right](x). \end{aligned}$$

The density for P_0 can be written as $f_0 = h_0 * q_0$. Expand the previous expression by collecting coefficients in powers of τ and use this identity to obtain

$$\begin{aligned} f(x,\tau) - f_0(x) = \tau&\Big([h_0 * q_1](x) + [h_1 * q_0](x)\Big) \\ &+ \cdots + \tau^d\Big([h_0 * q_d](x) + [h_1 * q_{d-1}](x) + \cdots + [h_d * q_0](x)\Big) \\ &+ \cdots + \tau^{2d+1}[h_{d+1}(\cdot,\tau) * q_d](x). \end{aligned} \qquad \text{<4.6>}$$

As discussed in the heuristic, estimation for θ is made difficult by making $f(\cdot,\tau)$ as close to f_0 as possible, for a suitable choice $q(\cdot,\tau)$. To establish the asserted lower bound for the rate of convergence, we will construct $q(\cdot,\tau)$ so that the first d coefficients in the expansion <4.6> are zero. Each coefficient is made zero by requiring that q_0 is smooth: the higher the order of the coefficient, the more smoothness required to eliminate it. Eventually the Fourier constraint imposed on $\mathcal{Q}$ hinders the construction from affecting coefficients of higher order than d. This gives the required lower bound.

Let us start with the first coefficient on the right of <4.6>. Therefore, try to solve for q_1 so that

$$h_0 * q_1 + h_1 * q_0 = 0,$$

or equivalently, solve the Fourier analog

$$\widehat{h}_0\widehat{q}_1 + \widehat{h}_1\widehat{q}_0 = 0. \qquad \text{<4.7>}$$

Expand the Fourier transform for $h(\cdot,\tau)$ as

$$\begin{aligned} \widehat{h}(t,\tau) &= \exp\left(-\tfrac{1}{2}t^2(1+\tau)\right) \\ &= \widehat{h}_0(t) + c_1\tau t^2\widehat{h}_0(t) + \cdots + c_d\tau^d t^{2d}\widehat{h}_0(t) + \cdots, \end{aligned}$$

where $\widehat{h}_0(t) = \exp(-t^2/2)$ and $c_m = (-1)^m/(2^m m!)$, for $m \geq 1$. A careful analysis shows that by comparing the Fourier expansion of <4.5> with the previous expansion,

$$\widehat{h}_m(t) = c_m t^{2m}\widehat{h}_0(t). \qquad \text{<4.8>}$$

In particular <4.8> asserts that $\widehat{h}_1(t) = c_1 t^2 \widehat{h}_0(t)$. Factor out the common term, $\widehat{h}_0$, and reexpress <4.7> as

<4.9> $$\widehat{q}_1(t) = -c_1 t^2 \widehat{q}_0(t).$$

Solving this equation is simple by an application of the following standard result (see for example, Feller Vol. II, 1971, Chapter XV.4)

<4.10> **Lemma:** *Suppose a density q has integrable derivatives up to l^{th}-order, and a Fourier transform for which $t^l \widehat{q}(t)$ is integrable. Then the Fourier transform for the m^{th} derivative, $q^{(m)}$, equals*

$$\widehat{q}^{\,(m)}(t) = (-it)^m \widehat{q}(t), \qquad \text{for } m = 1, 2, \cdots, l.$$

Therefore, if we assume that q_0 has integrable second derivative, then Lemma <4.10> asserts that

$$q_1(y) = c_1 q_0^{(2)}(y)$$

is a solution to <4.9>.

Let us require q_0 to be even smoother. Assume that q_0 has $2d$ integrable derivatives and a Fourier transform such that $t^{2d}\widehat{q}_0(t)$ is integrable. The recursive argument below shows that not only can we eliminate the first coefficient in the expansion <4.6> of f_τ, but we can eliminate each of the next $d-1$ coefficients as well.

The d^{th} coefficient in the expansion <4.6> equals

<4.11> $$h_0 * q_d + h_1 * q_{d-1} + \cdots + h_d * q_0.$$

To make the expression zero, we will show that its Fourier transform equals zero:

<4.12> $$\widehat{h}_0 \widehat{q}_d + \widehat{h}_1 \widehat{q}_{d-1} + \cdots + \widehat{h}_d \widehat{q}_0 = 0.$$

By identity <4.8>, we can factor out the common term, $\widehat{h}_0$, to reexpress this d^{th} Fourier problem as

<4.13> $$\widehat{q}_d(t) + c_1 t^2 \widehat{q}_{d-1}(t) + \cdots + c_d t^{2d} \widehat{q}_0(t) = 0.$$

Assume that the first $d-1$ Fourier problems are of the form

<4.14> $$\widehat{q}_m(t) = \gamma_m t^{2m} \widehat{q}_0(t),$$

for constants γ_m, and that each problem has the solution

<4.15> $$q_m(y) = (-1)^m \gamma_m q_0^{(2m)}(y),$$

where $m \leq d-1$. We have already shown this is true for $m = 1$ with $\gamma_1 = -c_1$. By recursion we will show that the Fourier problem and its solution are of this form for all $m \leq d$.

The d^{th} Fourier problem <4.13> upon substitution of <4.14> becomes

$$\begin{aligned}\widehat{q}_d(t) &= -c_1 t^2 \widehat{q}_{d-1}(t) - \cdots - c_{d-1} t^{2d-2} \widehat{q}_1(t) - c_d t^{2d} \widehat{q}_0(t) \\ &= \left(-c_1\gamma_{d-1} - \cdots - c_{d-1}\gamma_1 - c_d\right) t^{2d} \widehat{q}_0(t) \\ &= \gamma_d t^{2d} \widehat{q}_0(t).\end{aligned}$$

By Lemma <4.10> the solution to this problem is of the form <4.15> when $m = d$. This establishes our assertion as to the form and solution of each of the first d Fourier problems.

Reintroduce the factor $\widehat{h}_0$ to deduce that the d^{th} term <4.12> and the preceding $d-1$ terms all equal zero. By the uniqueness of the Fourier transform conclude that the the d^{th} coefficient <4.11> as well as the preceding $d-1$ coefficients are all zero.

However, it is not enough to simply show that these coefficients are zero. We must also show that the solutions <4.15> for q_m satisfy the constraints required for Q_τ to be a member of $\mathcal{Q}$. This amounts to showing that one can find a density q_0 with $2d$ integrable derivatives such that $t^{2d}\widehat{q}_0(t)$ is integrable (but so that $\widehat{q}_0$ satisfies the Fourier constraint <4.4>) and such that the expression

<4.16> $$q(y,\tau) = q_0(y) - \gamma_1\tau q_0^{(2)}(y) + \gamma_2\tau^2 q_0^{(4)}(y) + \cdots + (-1)^d\gamma_d\tau^d q_0^{(2d)}(y)$$

is a density.

This is fairly easy to do. One such choice being the distribution Q_0 formed by the random variables

$$(E_1 - E_2) + (E_3 - E_4) + \cdots + (E_{2d+1} - E_{2d+2}),$$

where E_m are independent standard exponentials, for $m = 1, 2, \cdots, 2d+2$.

The Fourier transform for q_0 equals

$$\widehat{q}_0(t) = (1+t^2)^{-(d+1)},$$

so that $\widehat{q}_0(t)$ has tails of order $O\left(t^{-(2d+2)}\right)$. A later observation (Remark <4.19>) shows that the density can be expressed as

<4.17> $$q_0(y) = \exp(-|y|)P(|y|, d),$$

where $P(y,d)$ is a polynomial in y of order d. Deduce, therefore, that the derivatives $q_0^{(2m)}$ for $m = 1, \cdots, d$ exist and can be expressed in a form similar to <4.17> (in fact the polynomials will also be of order d). Hence, q_0 has $2d$ integrable derivatives.

To show that <4.16> is a density, we need to verify that the expression is non-negative and integrates to one. The smoothness of q_0 will imply that $\int q_0^{(2m)} = 0$, for $m = 1, \cdots, d$. Thus, we need only establish the non-negativity of the function. This amounts to showing:

$$\frac{q(\cdot,\tau)}{q_0} = 1 - \gamma_1\tau\frac{q_0^{(2)}}{q_0} + \gamma_2\tau^2\frac{q_0^{(4)}}{q_0} + \cdots + (-1)^d\gamma_d\tau^d\frac{q_0^{(2d)}}{q_0} \geq 0.$$

Remark <4.19> also shows that $q_0(y)$ is bounded away from zero for finite values of y. Therefore, because the derivatives $q_0^{(2m)}$ are expressible in the same form as q_0 deduce that $q_0^{(m)}/q_0$ is bounded. This verifies the non-negativity requirement.

Now let us formally establish that our choice has led to the desired rate of convergence.

We have already shown that the first d coefficients in the expansion <4.6> equal zero for our choice <4.16> for $q(\cdot,\tau)$. Therefore, dividing throughout by f_0 we are left with

$$\begin{aligned}\frac{f(x,\tau)}{f_0(x)} - 1 = {} & \frac{\tau^{d+1}}{f_0(x)}\Big([h_1 * q_d](x) + \cdots + [h_d * q_1](x) + [h_{d+1}(\cdot,\tau) * q_0](x)\Big) \\ & + \frac{\tau^{d+2}}{f_0(x)}\Big([h_2 * q_d](x) + \cdots + [h_d * q_2](x) + [h_{d+1}(\cdot,\tau) * q_1](x)\Big) \\ & + \cdots + \frac{\tau^{2d+1}}{f_0(x)}[h_{d+1}(\cdot,\tau) * q_d](x). \end{aligned} \qquad <4.18>$$

We will show that

$$P_0\left(\frac{f(\cdot,\tau)}{f_0} - 1\right)^2 = O\left(\tau^{2d+2}\right)$$

by using Lemma <3.38> to show that each of the terms on the right-hand side of <4.18> have squared $L^2(P_0)$-norms of the same order.

For example, a typical term on the right-hand side is of the form

$$\frac{h_l * q_m}{f_0} = (-1)^m\gamma_m\frac{h_l * q_0^{(2m)}}{h_0 * q_0},$$

where $1 \leq l, m \leq d$.

Use Lemma <3.38> with $F_1 = \sqrt{h_0}$, $F_2 = \sqrt{q_0}$, $G_1 = h_l/\sqrt{h_0}$ and $G_2 = q_0^{(2m)}/\sqrt{q_0}$, to bound the squared $L^2(P_0)$-norm of this term by:

$$\begin{aligned}\gamma_m^2\int\frac{[h_l * q_0^{(2m)}]^2}{h_0 * q_0} &= \gamma_m^2\int\frac{\left[\left(\sqrt{h_0}\;\; h_l/\sqrt{h_0}\right) * \left(\sqrt{q_0}\;\; q_0^{(2m)}/\sqrt{q_0}\right)\right]^2}{\left(\sqrt{h_0}\right)^2 * \left(\sqrt{q_0}\right)^2} \\ &\leq \gamma_m^2\int\frac{h_l^2}{h_0}\int\frac{\left(q_0^{(2m)}\right)^2}{q_0}.\end{aligned}$$

We observed earlier that $h_l(x) = h_0(x)P_l(x)$, where $P_l(x)$ is a polynomial in x of degree $2l$. This shows that the first integral on the right is finite (this is the same phenomenon

that shows $h(\cdot,\tau)/h_0$ to be $L^2(h_0)$-differentiable in the sense of Definition <3.32>). Our previous observation that $q_0^{(2m)}$ has an expression similar to <4.17> shows that the second term is also bounded.

This takes care of the majority of the terms on the right of <4.18>. The terms that contain $h_{d+1}(\cdot,\tau)$ are dealt with in a similar fashion.

Therefore, all the terms on the right side of <4.18> have squared $L^2(P_0)$-norms of order $O\left(\tau^{2d+2}\right)$. Hence,

$$P_0\left(\frac{f(\cdot,\tau)}{f_0}-1\right)^2 = O\left(\tau^{2d+2}\right).$$

Appeal to Lemma <2.10> to infer that a regular estimator for the standard deviation has rate of convergence no faster than $O_p\left(n^{-1/(2d+2)}\right)$.

<4.19> **Remarks:** One way to derive an explicit representation for q_0 is to use a contour integration argument. We know that q_0 has Fourier transform

$$\widehat{q}_0(t) = (1+t^2)^{-(d+1)}.$$

This transform is integrable so that q_0 can be expressed as the inversion of its transform

$$q_0(y) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \exp(-ity)(1+t^2)^{-(d+1)}\,dt.$$

Define

$$v(z) = \exp\left(-zy\right)(1+z)^{-(d+1)}(1-z)^{-(d+1)}.$$

Observe that the integrand in the expression for q_0 equals the function $v(it)$. The function v is analytic expect for the $(d+1)$-poles at $z=1$ and $z=-1$. Let $\mathcal{C}$ equal the positively oriented semi-circle contour, with large radius, centered at zero and which lies in $\{z : \mathrm{Re}(z) \geq 0\}$. Integrate v over $\mathcal{C}$ for $y \geq 0$. Apply the Cauchy Residue theorem (Cartan, 1973, III.5.2) and let the radius of the contours go off to infinity, to obtain for $y \geq 0$

$$q_0(y) = \frac{1}{d!}\lim_{z\to -1}\frac{\partial^d}{\partial z^d}\exp\left(-zy\right)(1-z)^{-(d+1)}.$$

This function can be written as $\exp(-y)$ multiplied by a polynomial in y of order d. The representation for $y<0$ is obtained by the symmetry of q_0 about zero, which proves our earlier assertion that

$$q_0(y) = \exp(-|y|)P(|y|,d),$$

where $P(y,d)$ is a polynomial in y of order d.

Finally, to show that $q_0(y)$ is bounded away from zero for finite y, recognize that it can be written as $d+1$ convolutions of the double exponential density, $\exp(-|y|)/2$. Because each of these densities are strictly positive for finite y, deduce that q_0 can be zero only for infinite y values.

Chapter 5

Weibull Mixture Model

1. Introduction The Weibull density with unknown shape and scale parameter can be written as

$$f(x \mid \theta, \eta) = \theta x^{\theta-1} \eta \exp(-\eta x^{\theta}), \qquad <5.1>$$

where the density is taken with respect to Lebesgue measure on $(0, \infty)$. The shape parameter θ is assumed to lie in the set $\Theta = (0, \infty)$, while the scale parameter η lies in the set $\mathcal{N} = (0, \infty)$. In Chapter 1, we presented the semiparametric Weibull mixture model studied by Heckman and Singer (1984). A special case of this model is formed by integrating over the scale parameter of the density <5.1> with an unknown mixing distribution. The chapter studies this model and investigates the problem of estimation for the shape parameter, θ.

Let $\mathcal{Q}$ be the class of mixing distributions consisting of those distributions which have support on $\mathcal{N}$ and whose first moment is less than a fixed positive constant. Heckman and Singer (1984) show that for this $\mathcal{Q}$ the class of Weibull mixture models, $\mathcal{P}(\Theta, \mathcal{Q})$, is identifiable. Their proof, expressed in our notation, is as follows.

Let $\gamma_i = (\theta_i, Q_i)$, where $\theta_i \in \Theta$ and $Q_i \in \mathcal{Q}$ for $i = 1, 2$. If the two distributions $P_{\gamma_1}, P_{\gamma_2}$ are equal, then the continuity of the density <5.1> implies that $f_{\gamma_1}(x) = f_{\gamma_2}(x)$ for all positive x. Therefore,

$$\begin{aligned} 1 &= \frac{f_{\gamma_1}(x)}{f_{\gamma_2}(x)} \\ &= \frac{\theta_1}{\theta_2} x^{\theta_1 - \theta_2} \frac{\int \eta \exp(-\eta x^{\theta_1})\, dQ_1(\eta)}{\int \eta \exp(-\eta x^{\theta_2})\, dQ_2(\eta)}, \qquad x > 0. \end{aligned}$$

The monotone convergence theorem shows that the ratio of integrals tends to the finite positive constant $\int \eta\, dQ_1(\eta) / \int \eta\, dQ_2(\eta)$ as $x \to 0$. Therefore, the right-hand side would converge to either 0 or $+\infty$ if $\theta_1 \neq \theta_2$. That leaves

$$\int \eta \exp(-\eta x^{\theta_1})\, dQ_1(\eta) = \int \eta \exp(-\eta x^{\theta_1})\, dQ_2(\eta),$$

for all $x > 0$. The uniqueness theorem for Laplace transforms (Feller Vol. II, 1971, Chapter XIII.1) gives equality of Q_1 and Q_2. Therefore, the two parameters γ_1 and γ_2 must be equal. Conclude that $\mathcal{P}(\Theta, \mathcal{Q})$ is identifiable.

This chapter will consider the question of how well the shape parameter θ can be estimated from a sample of independent realizations of an identifiable Weibull mixture

model. We have already seen, by Example <3.12> of Chapter 3, that the model has zero information and consequently that the shape parameter can only be estimated at a rate slower than $O_p\left(n^{-1/2}\right)$. By applying a simple transformation, the model can be recast as a location-mixture with unknown scale parameter. Section 5.2 takes advantage of this and extends the result of Example <3.12> by showing an explicit lower bound of $O_p\left(n^{-1/4}\right)$ for the rate of convergence.

2. Rates of Convergence Using Fourier Analysis The Weibull density <5.1> with shape parameter $1/\theta$ and scale parameter η is that of the random variable $(E/\eta)^\theta$, where E has a standard exponential distribution. Form the Weibull mixture model by taking η to be a random variable independent of E, with unknown distribution concentrated on $(0,\infty)$. The problem is to determine the rate at which $1/\theta$, or equivalently θ, can be estimated at given an independent sample of realizations from the model.

Transform the data by taking logs. The new data have the form

$$X = \theta Z + Y, \qquad \text{<5.2>}$$

where $Z = \log E$, and $Y = -\theta \log \eta$ has unknown distribution Q. By observing that Z is independent of Y recognize that the transformation describes a location mixture model with unknown shape parameter θ.

The following argument shows that it is difficult to discriminate between a mixture model of the form <5.2> with structural parameter θ_0 compared against models with structural parameters smaller than θ_0. Without loss of generality we take the true structural parameter to be $\theta_0 = 1$, and the parameter space as $\Theta = (1-\epsilon, \infty)$ for a fixed small $\epsilon > 0$. We show by using the Fourier technique of Chapter 3.3 that θ_0 can be estimated at a rate no faster than $O_p\left(n^{-1/4}\right)$. The result is obtained as a consequence of working with the total variation distance. Therefore, because the log transformation is invertible and measurable, the result can be readily translated back into the Weibull setting. Thus, we prove that the shape parameter in a Weibull mixture model can be estimated at a rate no faster than $O_p\left(n^{-1/4}\right)$.

We first need to introduce constraints which ensure identifiability of the model. Let $\mathcal{P}(\Theta, \mathcal{Q})$ be the collection of all distributions of the form <5.2> as θ ranges over the parameter space Θ and Q ranges over a class of distributions $\mathcal{Q}$. As discussed in the introduction, identifiability in the Weibull model is ensured by assuming that the unknown mixing distribution has finite moment bounded by a fixed constant, M. By the invertibility of the transformation this requirement becomes $Q\exp(-Y/\theta) < M$. If $0 < \epsilon < 1/2$, then $1/\theta < 1/(1-\epsilon) < 1 + 2\epsilon$ for each $\theta \in \Theta$. Thus, a sufficient condition to satisfy the

integrability constraint becomes

<5.3> $$\int \exp\Big((1+2\epsilon)|y|\Big)\, dQ(y) < M.$$

Let $\mathcal{Q}$ be the class of mixing distributions that have support on $\mathbb{R}$ and which satisfy condition <5.3>. Then $\mathcal{P}(\Theta, \mathcal{Q})$ is identifiable.

Let $P_0 \in \mathcal{P}(\Theta, \mathcal{Q})$ equal the true mixed distribution with structural parameter $\theta_0 = 1$ and mixing distribution $Q_0 \in \mathcal{Q}$. Let Q_0 have density q_0 with respect to Lebesgue measure. Let $P_\tau \in \mathcal{P}(\Theta, \mathcal{Q})$ equal the perturbed mixed distribution with structural parameter $1-\tau$ and mixing distribution $Q_\tau \in \mathcal{Q}$, for small $\tau \geq 0$.

Let h_0 equal the density of Z and $h(\cdot, \tau)$ the density for the random variable $(1-\tau)Z$. The density h_0 is smooth enough to allow a Taylor series expansion of $h(\cdot, \tau)$ about $\tau = 0$ as the sum of integrable functions:

$$\begin{aligned} h(x,\tau) &= \frac{1}{1-\tau}\exp\Big[-\exp\left(\frac{x}{1-\tau}\right) + \frac{x}{1-\tau}\Big] \\ &= h_0(x) + \tau h_1(x) + \tau^2 h_2(x) + \cdots, \end{aligned} \qquad <5.4>$$

where, for example,

$$\begin{aligned} h_1(x) &= \frac{\partial}{\partial \tau} h(x,0) \\ &= h_0(x)\left[1 - \exp(x) + x\right]. \end{aligned}$$

Some calculus shows that when $\tau \geq 0$, the ratio $h(\cdot,\tau)/h_0$ is $L^2(h_0)$-differentiable (in the sense of Definition <3.32>) with first derivative h_1/h_0. This fact will become useful later.

Let Q_τ have density

$$q(y,\tau) = q_0(y) + \tau q_1(y).$$

To make estimation for the structural parameter difficult, we will try to construct Q_τ so that

<5.5> $$h_0 * q_1 + h_1 * q_0 = 0,$$

or at least so that the expression on the left-hand side is made close to zero.

The argument of Chapter 3.3 suggests it may easier to work with the Fourier analog of <5.5>

<5.6> $$\widehat{h}_0\widehat{q}_1 + \widehat{h}_1\widehat{q}_0 = 0.$$

The same argument also indicates that it may be easier to satisfy the expression <5.6> if $\widehat{h}_0(t)$ is expressible as $\exp(l(t))$ for a smooth function l.

Let us show that $\widehat{h}_0$ can be expressed in such a form. The Fourier transform for h_0 equals:

$$\begin{aligned}\widehat{h}_0(t) &= P\exp(it\log E)\\ &= \int_0^{+\infty}\exp(-x)\exp(it\log x)\,dx\\ &= \int_0^{+\infty}\exp(-x)x^{(it+1)-1}\,dx\\ &= \Gamma(it+1),\end{aligned}$$

where Γ is the Gamma function. On the domain $\{z : \mathrm{Re}(z) > 0\}$, where it is analytic, the Gamma function has an infinite product expansion (Ahlfors, 1979, Chapter 5.2.4)

$$\Gamma(z) = \frac{1}{z}\exp(-\gamma z)\prod_{j\geq 1}\exp\left(\frac{z}{j}\right)\left(1+\frac{z}{j}\right)^{-1},$$

where γ is Euler's constant with approximate value 0.57722. This expansion and the property $\Gamma(z+1) = z\Gamma(z)$ enables us to express $\widehat{h}_0$ as

<5.7> $$\widehat{h}_0(t) = \exp(-\gamma it)\prod_{j\geq 1}\exp\left(\frac{it}{j}\right)\left(1+\frac{it}{j}\right)^{-1},$$

which can also be written as $\widehat{h}_0(t) = \exp(l(t))$, where $l(t) = -\gamma it + S(it)$ and S is the analytic function on $\{z : \mathrm{Re}(z) > -1\}$ defined by

$$S(z) = \sum_{j\geq 1}\left(\frac{z}{j} - \log\left(1+\frac{z}{j}\right)\right).$$

(For definiteness, we take the principal branch for each log in the expression for S.)

The analytic nature of l enables us to write

$$\begin{aligned}\widehat{h}(t,\tau) &= \exp\bigl(l(t-\tau t)\bigr)\\ &= \widehat{h}_0(t) - \tau\widehat{h}_0(t)tl'(t) + \cdots,\end{aligned}$$ <5.8>

where

$$tl'(t) = -\gamma it - t^2\sum_{j\geq 1}\frac{1}{j^2}\,\widehat{m}_j(t),$$

and

$$\widehat{m}_j(t) = \left(1+\frac{it}{j}\right)^{-1}.$$

Notice that $\widehat{m}_j$ is the Fourier transform for the distribution of the random variable $-E/j$, where E has a standard exponential distribution. It has density $j\exp(-j|y|)\{y \leq 0\}$.

By comparing the expansion <5.8> for $\widehat{h}(\cdot,\tau)$ with the corresponding Fourier expansion for <5.4>, we show later (Remark <5.26>) that

<5.9> $$\widehat{h}_1(t) = -\widehat{h}_0(t)tl'(t).$$

Factor out the common term $\widehat{h}_0$ in the Fourier expression <5.6>. We should try to solve for q_1 in

<5.10> $$\begin{aligned}\widehat{q}_1(t) &= tl'(t)\widehat{q}_0(t)\\ &= -\gamma it\widehat{q}_0(t) - t^2\sum_{j\geq 1}\frac{1}{j^2}\,\widehat{m}_j(t)\,\widehat{q}_0(t),\end{aligned}$$

subject to the constraints of the model.

The next lemma shows that if q_0 is smooth enough, then an exact solution to the unconstrained problem exists. (First some notation: let $L_0^1(\mu)$ be the set of Lebesgue integrable functions which integrate to zero.)

<5.11> **Lemma:** *Assume that the density q_0 has first two derivatives $q_0^{(1)}, q_0^{(2)} \in L_0^1(\mu)$, and a Fourier transform for which $t^2\widehat{q}_0(t)$ is integrable. Then,*

<5.12> $$q_1(y) = \gamma q_0^{(1)}(y) + \sum_{j\geq 1}\frac{1}{j^2}\,\left[m_j * q_0^{(2)}\right](y)$$

is an element of $L_0^1(\mu)$ with Fourier transform

$$-\gamma it\widehat{q}_0(t) - t^2\sum_{j\geq 1}\frac{1}{j^2}\,\widehat{m}_j(t)\,\widehat{q}_0(t).$$

Proof of Lemma <5.11>: Let us first prove that q_1 is integrable. Interchange the order of integration, by the assumption that $q_0^{(2)}$ is integrable, to show

$$\iint m_j(z-y)|q_0^{(2)}(y)|\,dy\,dz = \int |q_0^{(2)}(y)|\,dy.$$

The integrability of q_1 follows by:

$$\begin{aligned}\int |q_1(y)|\,dy &\leq \gamma\int |q_0^{(1)}(y)|\,dy + \sum_{j=1}^{\infty}\frac{1}{j^2}\int\left|\left[m_j * q_0^{(2)}\right](y)\right|dy\\ &\leq \gamma\int |q_0^{(1)}(y)|\,dy + \int |q_0^{(2)}(y)|\,dy\ \sum_{j\geq 1}\frac{1}{j^2}\\ &< \infty.\end{aligned}$$

<5.13>

Use the fact that $q_0^{(2)}$ integrates to zero to infer that $\int m_j * q_0^{(2)} = 0$. Because $q_0^{(1)}$ also integrates to zero, deduce that $q_1 \in L_0^1(\mu)$.

The integrability of $t^2\widehat{q}_0(t)$ and the integrability of the first two derivatives for q_0 shows by Lemma <4.10>, that

$$\widehat{q}_0^{(1)}(t) = -it\widehat{q}_0(t) \qquad \text{<5.14>}$$

and

$$\widehat{q}_0^{(2)}(t) = -t^2\widehat{q}_0(t). \qquad \text{<5.15>}$$

Expression <5.13> shows that q_1 can be bounded by an integrable function. Use this dominating function coupled with identities <5.14> and <5.15> to deduce by the dominated convergence theorem that q_1 has the desired Fourier transform. □

Therefore to solve the unconstrained problem <5.10>, choose Q_0 to be a distribution satisfying the conditions of the lemma and take q_1 to be the function <5.12> defined by Q_0.

However, to be able to solve the constrained problem, we also need to show that

$$q(y,\tau) = q_0(y) + \tau q_1(y)$$

satisfies, at least for τ near zero, the constraints necessary for $q(\cdot,\tau)$ to be a density

(C1) $\int q(y,\tau)\,dy = 1$,

(C2) $q(y,\tau) \geq 0$,

and the constraint

(C3) $\int \exp\left((1+2\epsilon)|y|\right) q(y,\tau)\,dy < M$, for Q_τ to be a member of $\mathcal{Q}$.

To satisfy the constraints we will perturb the solution $q_0(y)+\tau q_1(y)$ to $q_0(y)+\tau q_1(y,\tau)$, with $q_1(y,\tau) \approx q_1(y)$. We will then no longer have an exact solution to <5.10>, but it will still be possible to prove the asserted $O_p\left(n^{-1/4}\right)$ lower bound for the rate of convergence.

First let us consider the effect that each constraint has on the possible solutions <5.12> for q_1. Condition (C1) will not present an obstacle, for we know that q_1 must integrate to zero by Lemma <5.11>. For condition (C2) to hold, we would need to show for small values of τ

$$\begin{aligned}\frac{1}{q_0}(q_0+\tau q_1) &= 1 + \gamma\tau\frac{q_0^{(1)}}{q_0} + \tau\sum_{j\geq 1}\frac{\left[m_j * q_0^{(2)}\right]}{j^2 q_0} \\ &\geq 0. \qquad \text{<5.16>}\end{aligned}$$

If q_0 were convex in the tails, the second derivative, $q_0^{(2)}(y)$, would be positive for large absolute values of y. In addition, if $q_0(y)$ were bounded away from zero for finite y, then

for small positive τ the expression involving $m_j * q_0^{(2)}$ in <5.16> should be positive. If in addition $q_0^{(1)}$ decreases more rapidly than q_0, then the left-side of <5.16> should be positive, and condition (C2) should hold.

The constraint (C3) is the most difficult to satisfy. The density

$$m_j(y) = j \exp(-j|y|)\{y \le 0\}$$

has tails decreasing faster than $\exp(-(1+2\epsilon)|y|)$ except when $j = 1$. Therefore, we seek a q_0 for which $q_0^{(1)}$ and $m_j * q_0^{(2)}$ decrease faster than $\exp(-(1+2\epsilon)|y|)$, for $j \ge 1$.

Here is one way to construct an approximate solution to <5.10>. Let Q_0 be the distribution for the random variable $\beta^{-1}(L_1 + L_2)$, where L_j are independent double exponential random variables, for $j = 1, 2$, and β is chosen larger than $1 + 2\epsilon$ so that the density

$$q_0(y) = \frac{\beta}{4} \exp\bigl(-\beta|y|\bigr)\bigl(1 + \beta|y|\bigr)$$

satisfies the integrability constraint <5.3> needed to ensure that $Q_0 \in \mathcal{Q}$.

The density for Q_0 has first derivative

<5.17> $$q_0^{(1)}(y) = \frac{-\beta^3}{4} \exp(-\beta|y|)y,$$

second derivative

<5.18> $$q_0^{(2)}(y) = \frac{\beta^3}{4} \exp(-\beta|y|)(\beta|y| - 1),$$

and Fourier transform

$$\widehat{q}_0(t) = \left(1 + \left(\frac{t}{\beta}\right)^2\right)^{-2}.$$

The integrability of $t^2\widehat{q}_0(t)$ and the smoothness of q_0 shows that Q_0 satisfies the conditions of Lemma <5.11> (it is fairly easy to check that the derivatives have zero expectation).

To be able to satisfy the constraint (C3), we need that both $q_0^{(1)}$ and $m_j * q_0^{(2)}$ decrease faster than $\exp(-(1+2\epsilon)|y|)$, for $j \ge 1$. Use the expression <5.17> for the first derivative $q_0^{(1)}$ to see that the presence of this term will not violate the constraint. A little bit of work shows

<5.19>
$$\begin{aligned}\left[m_j * q_0^{(2)}\right](y) = \{y < 0\}&\left(C_1(j,\beta)\exp(-j|y|) + C_2(j,\beta,y)\exp(-\beta|y|)\right)\\ &+ \{y \ge 0\}\, C_3(j,\beta,y)\exp(-\beta|y|),\end{aligned}$$

where $C_1(j,\beta)$ is a constant which is uniformly bounded in j and $C_2(j,\beta,y)$, $C_3(j,\beta,y)$ are functions which are uniformly bounded in j by a function of y which is of order $O(|y|)$, for large y.

Therefore, the tails for <5.19> are of order $o\left(\exp(-(1+2\epsilon)|y|)\right)$ when $j \geq 2$. To ensure that condition (C3) not be violated when $j = 1$, define Q_τ to be the distribution with density

$$q(y,\tau) = q_0(y) + \tau q_1(y,\tau),$$

where $q_1(y,\tau)$ is an approximate solution for <5.10> defined by:

$$\begin{aligned} q_1(y,\tau) =& q_1(y) + \left[\left(m_1(y,\tau) - m_1\right) * q_0^{(2)}\right](y) \\ <5.20> \qquad =& \gamma q_0^{(1)}(y) + \left[m_1(y,\tau) * q_0^{(2)}\right](y) + \sum_{j\geq 2} \frac{1}{j^2}\left[m_j * q_0^{(2)}\right](y), \end{aligned}$$

where the truncated function $m(y,\tau)$ is defined as

$$\begin{aligned} m_1(y,\tau) &= m_1(y)\left\{\frac{\delta}{2\epsilon}\log\tau \leq y\right\} \\ &= \exp(y)\left\{\frac{\delta}{2\epsilon}\log\tau \leq y \leq 0\right\}, \end{aligned}$$

for a fixed small $0 < \delta < 1$.

Condition (C3) is now no longer a problem, for

$$\left|\tau \int \exp\Big((1+2\epsilon)|y|\Big)\left[m_1(\cdot,\tau) * q_0^{(2)}\right](y)\,dy\right| = O\left(\tau^{1-\delta}\right)$$

will be small enough, eventually, to guarantee that the integrability constraint <5.3> holds. This combined with our previoius observations concerning the tail behavior of $m_j * q_0^{(2)}$, for $j \geq 2$, and the tail behavior of $q_0^{(1)}$ shows that (C2) holds.

Condition (C1) must also hold for this choice, for $q_1(\cdot,\tau)$ must integrate to one by Lemma <5.11> and the fact that

$$\iint m_1(z,\tau) q_0^{(2)}(y-z)\,dy\,dz = 0.$$

To verify the nonnegativity constraint implied by condition (C2), it is sufficient to observe that the ratio $q_0^{(1)}/q_0$ is bounded and that the second derivative $q_0^{(2)}(y)$ is positive for large absolute values of y. Therefore, deduce that $Q_\tau \in \mathcal{Q}$.

Notice that we can choose δ and ϵ as we wish subject to the constraints of their range. Later we will need $\delta/2\epsilon \geq 2$; therefore assume that the two values are chosen accordingly.

Now let us rigorously show that our choice for Q_τ, and its density, have led to the desired rate of convergence. Expand $h(\cdot,\tau)$ to three terms in the Taylor series expansion

$$h(x,\tau) = h_0(x) + \tau h_1(x) + \tau^2 h_2(x,\tau).$$

Let f_0 and $f(\cdot,\tau)$ be the mixed densities for P_0 and P_τ. Then

$$f(x,\tau) = \left[\Big(h_0 + \tau h_1 + \tau^2 h_2(\cdot,\tau)\Big) * \Big(q_0 + \tau q_1(\cdot,\tau)\Big)\right](x).$$

We can expand this expression term by term, since all terms are integrable. Collect coefficients in powers of τ, and use the identity $f_0 = h_0 * q_0$ to find by dividing throughout by f_0:

$$\frac{f(x,\tau)}{f_0(x)} - 1 = \frac{\tau}{f_0(x)}\Big([h_0 * q_1(\cdot,\tau)](x) + [h_1 * q_0](x)\Big) + \frac{\tau^2}{f_0(x)}[h_1 * q_1(\cdot,\tau)](x)$$

<5.21> $$+ \frac{\tau^2}{f_0(x)}[h_2(\cdot,\tau) * q_0](x) + \frac{\tau^3}{f_0(x)}[h_2(\cdot,\tau) * q_1(\cdot,\tau)](x).$$

We will show that

$$P_0\left(\frac{f(\cdot,\tau)}{f_0} - 1\right)^2 = O\left(\tau^4\right)$$

by using Lemma <3.38> to show that each of the four terms on the right side of <5.21> have squared $L^2(P_0)$-norms of order $O\left(\tau^4\right)$.

Let us start with the first term. We have already established that Q_0 satisfies the conditions of Lemma <5.11>. Therefore by the conclusion of the same lemma and the identity <5.9>, deduce that the Fourier transform for $h_0 * q_1(\cdot,\tau)$ equals

$$-\widehat{h}_1(t)\widehat{q}_0(t) + \widehat{h}_0(t)\Big(\widehat{m}_1(t,\tau) - \widehat{m}_1(t)\Big)\widehat{q}_0^{(2)}(t).$$

Remark <5.26> shows that $\widehat{h}_0$ and $\widehat{h}_1$ are integrable. Thus, by the uniqueness of the Fourier transform deduce that the first term on the right side of <5.21> equals

$$\frac{\tau}{f_0(x)}\left[h_0 * \big(m_1(\cdot,\tau) - m_1\big) * q_0^{(2)}\right](x).$$

Use Lemma <3.38> with $F_1 = \sqrt{h_0}$, $F_2 = \sqrt{q_0}$, $G_1 = h_0 * (m_1(\cdot,\tau) - m_1)/\sqrt{h_0}$ and $G_2 = q_0^{(2)}/\sqrt{q_0}$, to bound the squared $L^2(P_0)$-norm of this term by:

$$\begin{aligned}
&\tau^2 \int \frac{\left[h_0 * (m_1(\cdot,\tau) - m_1) * q_0^{(2)}\right]^2}{h_0 * q_0} \\
&\qquad = \tau^2 \int \frac{\left[\left(\sqrt{h_0}\ \ h_0 * (m_1(\cdot,\tau) - m_1)/\sqrt{h_0}\right) * \left(\sqrt{q_0}\ \ q_0^{(2)}/\sqrt{q_0}\right)\right]^2}{\left(\sqrt{h_0}\right)^2 * \left(\sqrt{q_0}\right)^2} \\
&\qquad \le \tau^2 \int \frac{[h_0 * (m_1(\cdot,\tau) - m_1)]^2}{h_0} \int \frac{(q_0^{(2)})^2}{q_0}.
\end{aligned}$$

Deduce from the expression <5.18> that the second integral on the right is finite. To deal with the first integral, observe that the density

$$h_0(x) = \exp(-\exp(x) + x)$$

has tails of the order $\exp(x)$ for large negative x and tails of the order $\exp(-\exp(x))$ for large positive x. This tail behavior remains unaffected when h_0 is convolved with the

density m_j. In fact a direct calculation shows that for $j \geq 2$

<5.22> $$[h_0 * m_j](x) \leq C h_0(x),$$

for a finite constant, C, not depending upon j. In particular, it is possible to show

<5.23> $$\int \frac{[h_0 * (m_1(\cdot,\tau) - m_1)]^2}{h_0} = O\left(\tau^{\delta/2\epsilon}\right),$$

which is of order $O\left(\tau^2\right)$ by our choice for δ and ϵ. Conclude that the first term on the right side of <5.21> has squared $L^2(P_0)$-norm of order $O\left(\tau^4\right)$.

To show that the squared $L^2(P_0)$-norm of the third term on the right side of <5.21> is of order $O\left(\tau^4\right)$, use Lemma <3.38> with $F_1 = \sqrt{h_0}$, $F_2 = \sqrt{q_0}$, $G_1 = h_2(\cdot,\tau)/\sqrt{h_0}$ and $G_2 = q_0/\sqrt{q_0}$, to bound the squared $L^2(P_0)$-norm of this term by:

$$\tau^4 \int \frac{[h_2(\cdot,\tau) * q_0]^2}{f_0} = \tau^4 \int \frac{\left[\left(\sqrt{h_0}\ \ h_2(\cdot,\tau)/\sqrt{h_0}\right) * q_0\right]^2}{\left(\sqrt{h_0}\right)^2 * \left(\sqrt{q_0}\right)^2}$$
$$\leq \tau^4 \int \frac{h_2(\cdot,\tau)^2}{h_0}.$$

The $L^2(h_0)$-differentiability of $h(\cdot,\tau)/h_0$ shows that the right-hand side is finite.

The remaining two terms on the right side of <5.21> are dealt with by a similar argument. To illustrate the method, consider the squared $L^2(P_0)$-norm of the second term

<5.24> $$\tau^4 \int \frac{[h_1 * q_1(\cdot,\tau)]^2}{h_0 * q_0}.$$

From the expression <5.20> for $q_1(\cdot,\tau)$,

<5.25> $$h_1 * q_1(\cdot,\tau) = \gamma\left(h_1 * q_0^{(1)}\right) + h_1 * m_1(\cdot,\tau) * q_0^{(2)} + \sum_{j\geq 2} \frac{1}{j^2}\left(h_1 * m_j * q_0^{(2)}\right).$$

To show that <5.24> is of order $O\left(\tau^4\right)$ divide each of the terms on the right side of <5.25> by f_0 and bound their squared $L^2(P_0)$-norms by Lemma <3.38>. For example

$$\int \frac{[h_1 * m_j * q_0^{(2)}]^2}{h_0 * q_0} \leq \int \frac{[h_1 * m_j]^2}{h_0} \int \frac{(q_0^{(2)})^2}{q_0}.$$

We have already observed that the second integral on the right side is finite. The first integral can be shown to be finite by an argument similar to the one which led to inequalities <5.22> and <5.23>. The other terms in <5.24> are dealt with by using Lemma <3.38> and the fact that h_1^2/h_0 and $(q_0^{(1)})^2/q_0$ are integrable (see <5.17>). Deduce that <5.24> is of order $O\left(\tau^4\right)$. A similar argument takes care of the remaining term in <5.21>

All the terms on the right side of <5.21> have squared $L^2(P_0)$-norms of order $O\left(\tau^4\right)$. We find that

$$P_0\left(\frac{f_\tau}{f_0} - 1\right)^2 = O\left(\tau^4\right).$$

Invoke Lemma <2.10> to deduce that a regular estimator for the shape parameter in an identifiable Weibull mixture model cannot have rate of convergence faster than $O_p\left(n^{-1/4}\right)$.

<5.26> **Remarks:** To verify that <5.9> is a valid identity, first establish the integrability of $\widehat{h}_0$ by using the representation <5.7> to bound the modulus:

$$|\widehat{h}_0(t)| = \prod_{j\geq 1}\left(1+\left(\frac{t}{j}\right)^2\right)^{-1/2}$$

<5.27>
$$\leq \left(1+\left(\frac{t}{k}\right)^2\right)^{-k/2},$$

for any positive integer, k. The terms in the expansion <5.4> and the expansion <5.8> are both obtained by differentiation with respect to τ. In particular, $h_1(x) = \partial h(x,0)/\partial\tau$ and

$$-\widehat{h}_0(t)tl'(t) = \frac{\partial\widehat{h}(t,0)}{\partial\tau}.$$

The bound <5.27> implies that $\widehat{h}(\cdot,\tau)$ is integrable and enables us to express $h(\cdot,\tau)$ as an integral representing the Fourier inverse of its transform. Differentiation applied to the outside of the integral, with respect to τ, can be taken inside the integral by appealing to the dominated convergence theorem and the fact that it is possible to find an integrable function which dominates $\partial\widehat{h}(\cdot,\tau)/\partial\tau$ for small τ (use the bound <5.27> and the expression for l). Therefore, h_1 can be evaluated by differentiating the Fourier transform $\widehat{h}(\cdot,\tau)$ and computing the Fourier inverse of the resulting function as τ goes to zero. By the uniqueness of the Fourier transform, deduce that <5.9> is a valid identity.

References

Ahlfors, L. V. (1979). *Complex analysis* (Third edition). McGraw-Hill, Inc., New York.

Begun, J. M., Hall, W. J., Huang, W-M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432-452.

Beran, R. (1977). Estimating a distribution function. *Ann. Statist.* **5** 400-404.

Bickel, P. J. and Ritov Y. (1987). Efficient estimation in the errors in variables model. *Ann. Statist.* **15** 513-540.

Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381-393.

Birgé, L. and Massart, P. (1992). Estimation of integral functionals of a density. Preprint 024-92, Mathematical Sciences Research Institute, Berkeley, California.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. American Statistical Association* **83** 1184-1186.

Cartan, H. (1973). *Elementary theory of analytic functions of one or several complex variables*. (Translated from "théorie élémentaire des fonctions analytiques d'une ou plusieurs variables complexes".) Addison-Wesley, Inc., Massachusetts.

Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *J. Econometrics* **32** 189-218.

Donoho, D. L. and Liu, R. C. (1987). Geometrizing rates of convergence, I. (University of California, Berkeley, Statistics Department Technical Report.)

Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence, II. *Ann. Statist.* 633-667.

Edelman, D. (1988). Estimation of the mixing distribution for a normal mean with applications to the compound decision problem. *Ann. Statist.* **16** 1609-1622.

Feller, W. (1971). *An introduction to probability theory and its applications, Vol. II* (Second edition). John Wiley & Sons, New York.

Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 175-194. University of California Press, Berkeley.

Hall, P. (1989). On convergence rates in nonparametric problems. *International Statist. Review* **57** 45-58.

Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52** 271-320.

Heckman, J. and Singer, B. (1984). Econometric duration analysis. *J. Econometrics* **24** 63-132.

Jewell, N. P. (1982). Mixtures of exponential distributions. *Ann. Statist.* **10** 479-484.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist.* **27** 887-906.

Koshevnik, Y. A. and Levit, B. Ya. (1976). On a non-parametric analogue of the information matrix. *Theory Probab. Appl.* **21** 738-753.

Lambert, D. and Tierney, L. (1984). Asymptotic properties of maximum likelihood estimates in the mixed poisson model. *Ann. Statist.* **12** 1388-1399.

Le Cam, L. (1972). Limits of experiments. In *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 245-261. University of California Press, Berkeley.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38-53.

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer-Verlag, New York.

Le Cam, L. and Yang, G. (1990). *Asymptotics in statistics: some basic concepts*. Springer-Verlag, New York.

Levit, B. Ya. (1975). On the efficiency of a class of non-parametric estimates. *Theory Probab. Appl.* **20** 723-740.

Lindsay, B. G. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Philos. Trans. Roy. Soc. London* **296** 639-665.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11** 86-94.

Lindsay, B. G. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11** 486-497.

Millar, P. W. (1981). The minimax principle in asymptotic statistical theory. *Ecole d'eté de Probabilités de Saint-Flour XI* 76-265.

Pfanzagl, J. and Wefelmeyer, W. (1982). *Contributions to a general asymptotic statistical theory*. Lecture Notes in Statistics, Vol.13. Springer-Verlag, New York.

Pfanzagl, J. (1988). Consistency of maximum likelihood estimators for certain non-parametric families, in particular: mixtures. *J. Statist. Plann. Inference* **19** 137-158.

Pfanzagl, J. (1990). *Estimation in semiparametric models: some recent developments*. Lecture Notes in Statistics, Vol. 63. Springer-Verlag, New York.

Pollard, D. (1993). Hypercubes and minimax rates of convergence. (Preprint, Department of Statistics, Yale University.)

Ritov, Y. and Bickel, P. J. (1990). Achieving information bounds in non and semiparametric models. *Ann. Statist.* **18** 925-938.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. American Statistical Association* **85** 617-624.

Roeder, K. (1992). Semiparametric estimation of normal mixture densities. *Ann. Statist.* **20** 929-943.

Rudin, W. (1987). *Real and complex analysis* (Third edition). McGraw-Hill, Inc., New York.

Simar, L. (1976). Maximum likelihood estimation of a compound poisson process. *Ann. Statist.* **4** 1200-1209.

Stein, C. (1956). Efficient nonparametric testing and estimation. In *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 187-195. University of California Press, Berkeley.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348-1360.

Strasser, H. (1985). *Mathematical theory of statistics: statistical experiments and asymptotic decision theory*. De Gruyter, Berlin.

van der Vaart, A. W. (1988). Estimating a real parameter in a class of semiparametric models. *Ann. Statist* **16** 1450-1474.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595-602.

Zhang, C-H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann. Statist.* **18** 806-831.