Some aspects of non parametric maximum likelihood for normal mixtures

A thesis submitted for the fulfillment of the requirements for the degree of Doctor of Philosophy

Jennifer Wilcox
The University of Sydney

2012

Abstract

In this thesis we look at some aspects of the Non Parametric Maximum Likelihood Estimate (NPMLE) of a mixing distribution in a mixture model. We present a proof that there exists case where the number of components in the NPMLE of a mixture distribution is not consistent for the true number of components in the mixture. This result supports the motivations for using penalised maximum likelihood techniques.

Mixture models are a popular way to approach density estimation problems. We investigate some computational aspects of using NPMLEs of mixture densities in a density estimation setting by looking at an implementation of the Intra Simplex Direction Method (ISDM).

0.1 Acknowledgements

I'd like to thank my supervisor and associate supervisor Michael and John for all their advice and comments. I'd also like to thank my husband Stewart for helping me speed up the slowest parts of my R code by porting them to C code, and for acting like a bonus supervisor to me.

Contents

Chapter 1

Introduction

Mixture models provide a useful generalisation to the simple parametric models used in a wide variety of applications (Lindsay, 1995). Some examples of the various contexts in which they have been proposed and studied are clustering (Fraley and Raftery, 2002), outlier problems (Box and Tiao, 1968), deconvolution problems (Matias, 2002), and density estimation (Jones and Henderson, 2009).

In this thesis we take an interest mainly towards the subject of density estimation, although we provide some results and comments which are not restricted to such a context.

A popular method of fitting mixture models stems from the results about Non Parametric Maximum Likelihood Estimation (NPMLE) by Lindsay (1983). For example, NPMLE methods are one of three common ways to approach density estimation problems, along with Kernel (Silverman, 1986) and Bayesian (Ferguson, 1973) approaches. Moreover, it can be quite easily seen that a Kernel density estimator can be viewed as special case of a mixture density estimator, as we will show in more detail in this thesis.

Several elegant theorems by Lindsay have encouraged the development of various useful algorithms which enable mixture models to be fitted in practice, for example, the Intra Simplex Direction Method (ISDM) by Lesperance and Kalbfleisch (1992). The main theoretical result in this thesis is another direct consequence of Lindsay's work.

The popularity of producing NPMLEs in the mixture model literature has led to the study of various features of these estimates. For example there has been interest in improving the computational speed of such estimates (Wang, 2007), studying the rates of convergence for NPMLEs (Ghosal and van der Vaart, 2001), formulating NPMLEs under various model constraints (Hathaway, 1985), (Jones and Henderson, 2009), and the proposal of penalized maximum likelihood methods to discourage the selection of a maximiser with too many components (Leroux, 1992), (Cathy and Bertrand, 2011).

Without introducing technical notation at the moment, here is an outline of the main result of this thesis. It is a negative result. One aspect of NPMLEs we consider is the nature of the estimated number of components. Suppose the NPMLE of a mixing distribution of a location mixture of unit-variance normals is calculated. Suppose this estimate is calculated based upon a sample of size n from a standard normal distribution. That is, suppose we are estimating a degenerate mixing distribution via NPMLE. Intuition may lead us to believe the number of estimated components should approach 1 as the sample size n increases, since the true density is indeed a 1 component mixture. In the first part of this thesis we show the probability that the number of components of the NPMLE in this example is (strictly) larger than 1 goes to 1 as n goes to ∞. This behaviour of the estimator is not at all desirable!

The main ideas and tools used in the proof in Chapter 2 are as follows. The proof provides a bound on the probability of the supremum of a certain stochastic process exceeding 0 over the range $(-\infty, \infty)$. After defining this stochastic process of interest, we use a three term Taylor expansion to examine the 'main part' of the process and 'remainder parts'. The remainder terms in the actual stochastic process are NOT negligible, however the nature of this particular proof only requires us to consider the event that the supremum is positive. This motivates us to define and examine a modified (scaled) version of the stochastic process. We then use tools from Csörgő et al. (1986) to approximate the main part of the new stochastic process by an empirical process. One of the Taylor expansion remainder terms is easy to deal with, but the part of our proof which deals with the other term requires the use of an idea from Bickel and Chernoff (1993). The lemma which uses the aforementioned idea uses a bound given by Revuz and Yor (1991). Finally, the other important tool we used in this proof was the normal comparison lemma, which was needed to address the introduction of the approximation using the tools from Csörgő et al. (1986).

Our result suggests that it is not sensible (even in the simplest of cases) to consider interpreting the number of components of the NPMLE as a quantity to explore in itself, but rather only as a tool in the process of applying NPMLE in its various contexts. This warning further supports the comments of (for example) Hoff (2003) about maintaining caution about inference about the mixing distribution in a mixture model, beyond the measure it represents.

Moreover, our result provides an example in which the $\geq$ sign of Theorem 4 of Leroux (1992) can be stated with a strict inequality.

Another aspect of non parametric maximum likelihood estimation for normal mixture contexts is how it can be applied in a density estimation setting.

Despite the dangers of extrapolating meaning from estimates of mixture distributions in such models, using NPMLE to produce normal mixture density estimates has been shown to be quite sensible because of the infinite degree of smoothness of the normal density (Ghosal and van der Vaart, 2001). In our thesis we also describe an analogous problem to that of bandwidth selection in kernel density estimation, in the context of mixture density estimation.

The work in this thesis suggests an alternate direction for mixture research away from the focus on estimating a mixture distribution's component number. The negative result in Chapter 2 supports methods which impose constraints on the model parameter spaces, such as the sieve method from Geman and Hwang (1982). This result exemplifies the well known problems concerning standard parametric methods for normal mixtures, since the likelihood surface has many local extrema and is unbounded. Work exists which does not focus intensely on component number estimates in the more generally interesting topic of non parametric maximum likelihood estimation. van de Geer (2003) gives an overview of asymptotic theory for density estimates arising from the non parametric maximum likelihood estimate of the mixing distribution in mixture models.

In this chapter we present the necessary definitions and results to make and refer to throughout our thesis. In Section 1.1 we define mixture models and give a few examples of how they can be useful in practice. In Section 1.2 we describe the major background material related to the estimates produced via non parametric maximum likelihood. In Section 1.3 we discuss some background material related to the estimating the number of mixture components, and why this has motivated the interest in penalised likelihood methods. In Section 1.3 we also state our main theoretical result. In Section 1.4 we show how problems about location-scale mixtures of normals can be reformulated into simpler problems about a location mixture of normals, and we describe why this can be useful towards practical applications such as density estimation.

Following the above sections, we will present some discussion about the way in which the research presented in our thesis fits into the ocean of fascinating mixture model based topics already available. In Section 1.5 we discuss some bonus topics which may be of interest in the wider context in which our theoretical result is a part of. In Section 1.5 we also talk about the relationship between our theoretical results in Chapter 2 with the topics introducted in Chapter 3, and then describe some related areas in the literature that focus on problems or issues of a similar nature. We also comment in Section 1.5 on several issues regarding the relationship between titles and technical results, in order to clarify the areas of focus in this particular body of work in relation to the available literature and broader topics of interest.

In Chapter 2 we formally state our main theoretical result and give our proof.

In Chapter 3 we present the non trivial problem of choosing the optimal component variance to use in location mixture density estimation. We then discuss some computational issues involved in implementing the ISDM in R code, and offer a version which has the heaviest calculations done more quickly via C code. This code is shown to be more than 10 times as fast as the same algorithm written entirely in R.

1.1 Mixture models

In this section we define what we mean by mixture model and describe some motivations for the current interest in mixture model theory to give an idea of how expansive and useful this topic turns out to be.

1.1.1 Mixtures and mixture densities

There are often contexts where measurements have been collected from a population with more than one distinct feature, such that some sort of multimodality is observed in various summaries of the observed dataset.

One example of such a context is bimodality which seems to arise when human heights are measured. Joiner (1975) gave an enthusiastic dramatization as a teaching device, to introduce the concept of bimodalit. Students from two introductory statistics classes and one introductory psychology class were arranged according to their height and photographed to produce a 'living histogram' of their heights.

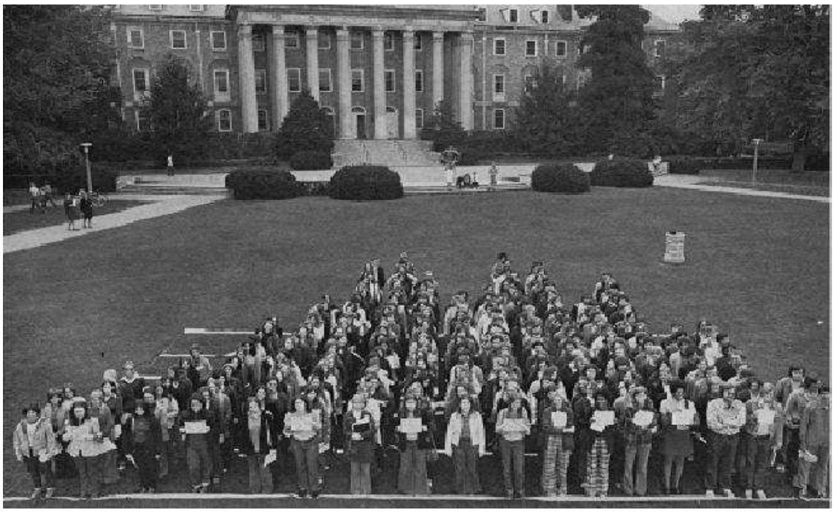


Figure 1.1: Brian Joiner's living histogram.

Joiner claimed that the observed average male height was significantly different from the observed average female height, and implied this was due to the differences in the two subpopulations' (male/female) underlying features.

While the bimodality of human height data has since been disputed (Watkins and Watkins, 2002), Joiner was nevertheless referring to the concept of a nontrivial "mixture" of two densities.

We make the following definitions.

Definition (Family of densities)**.** Let m be a positive integer and let $T \subseteq \mathbb{R}^m$ be some index set of interest. Suppose for each $t \in T$ there is a function f_t, such that f_t is a density. The class of such densities

$$\{f_t : t \in T\}$$

is called a family of densities.

An example of a family of densities is the normal or Gaussian family. If we know the mean μ and variance σ^2 of a normal density $\phi_{\mu,\sigma}$, the function is completely specified by

$$\phi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{1.1}$$

and so the class

$$\{\phi_{\mu,\sigma} : (\mu, \sigma) \in \mathbb{R} \times (0, \infty)\}$$

is a family of densities, parametrised by the set $\mathbb{R} \times (0, \infty)$.

Note: We will use ϕ to mean $\phi_{0,1}$.

Definition (Mass point)**.** Suppose F is a distribution on some $T \subseteq \mathbb{R}^n$. We will say $t \in T$ is a mass point of F if the probability F puts on t is positive. That is, $F(\{t\}) > 0$. We will also say that F puts mass/probability $F(\{t\})$ on $t \in T$.

Definition (Mixture density)**.** Let G be a probability measure on some index set $T \subseteq \mathbb{R}^n$, and let $\{f_t : t \in T\}$ be a family of densities. Then the function given by

$$f = \int_T f_t dG(t)$$

is a mixture density. G is referred to as the mixing distribution of the mixture.

In the case where the mixing distribution G is a discrete distribution with finitely many mass points $t_1, \ldots, t_k$ with (nonzero) probabilities $p_1, \ldots, p_k$ respectively ($\sum_{j=1}^k p_j = 1$), the mixture density f can be written more simply as

$$f = \sum_{j=1}^{k} p_j f_{t_j}, \tag{1.2}$$

and in this case f is called a finite mixture density.

The densities f_{t_j} in (1.2) are called component densities, and the mixture density f is said to have k components.

A further special case of a mixture density would be where the mixing distribution puts the probability 1 on a single point $t \in T$. We will call such a distribution degenerate. We will denote the degenerate distribution which puts probability 1 on the point $t \in T$ by δ_t.

1.1.2 Location and scale parameters

Often, a family of densities may be indexed by a location or scale parameter, or both. Here are some examples of families of densities which can be indexed by location or scale parameters.

The exponential family of densities also has a location parameter μ, and indeed a scale parameter λ:

$$\{f_{\mu,\lambda} : (\mu, \lambda) \in \mathbb{R} \times (0, \infty)\}, \text{ where } f_{\mu,\lambda}(x) = \lambda e^{-\lambda(x-\mu)}, \text{ for } x \geq \mu.$$

Another family of densities with a scale parameter θ (or inverse scale parameter $\beta = \frac{1}{\theta}$) is the gamma family of densities:

$$\{f_{\alpha,\beta} : (\alpha, \beta) \in (0, \infty)^2\}, \text{ where } f_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \text{ for } x \geq 0.$$

The normal family of densities metioned by equation 1.1 can also be parametrized by the location parameter $\mu \in \mathbb{R}$, the scale parameter $\sigma \in (0, \infty)$, or by the coordinates of location-scale parameters $(\mu, \sigma) \in \mathbb{R} \times (0, \infty)$. We will call a mixture density based on such parameterizations as a location-mixture, scale-mixture, or location-scale mixture respectively.

We will show in Section 1.4 that a general location-scale mixture of normal densities with minimum component variance h^2 can be reexpressed as a simpler normal location mixture density. As such, in this thesis we will only focus upon location mixtures of normal densities.

1.1.3 Examples of mixture densities and their applications

Example 1.1.1 (A finite scale-mixture of normal densities)**.** This example is from a popular model for outliers from Box and Tiao (1968), and it also shows one way in which we can interpret mixing distributions.

Suppose we have the iid rvs $X_1, \ldots, X_n$, and suppose they are (usually) standard normals. However occasionally an outlier occurs, which seems to come from a normal with a fatter tail. If we consider the n iid unobservable rv Y_i which take the values

$$Y_i = \begin{cases} 1, & \text{with probability } 0.9 \\ 10, & \text{with probability } 0.1, \end{cases}$$

we can see that the value corresponding to 1 can be interpreted as "not an outlier", while the value corresponding to 10 can be interpreted as "outlier". Thus each X_i can be considered to be the observable rv from the pair (X_i, Y_i), and the distribution of X_1 can be written as:

$$X_1 \sim \begin{cases} N(0, 1), & \text{if } Y_i = 1 \\ N(0, 10), & \text{if } Y_i = 10. \end{cases}$$

Let G be the distribution of Y_1. This scenario could be alternatively described with the notation $X_1 \sim N(0, Y_1)$, where $Y_1 \sim G$. The mixing distribution G can then be interpreted as the distribution of the "(random) variance" Y_1.

An alternate way to interpret this problem is to speak of the conditional distribution of X_1 given the values of Y_1:

$$X_1 | Y_1 = 1 \sim N(0, 1), X_1 | Y_1 = 10 \sim N(0, 10).$$

Another way to write this is to reflect the definition of the density of X_1:

$$X_1 \sim 0.9N(0, 1) + 0.1N(0, 10),$$

or we could simply speak of the density of X_1 in this model, which we could call f (say):

$$f = 0.9\phi + 0.1\phi_{0,\sqrt{10}}.$$

Though there are multiple possible choices for notation in mixture contexts, we will stick to speaking of the densities of mixtures for the sake of consistency.

Example 1.1.2 (A finite location-mixture of normal densities)**.** Like the normal scale-mixture density in Example 1.1.1, a location-mixture of normal densities only varies one type of parameter in the normal family.

Part of the notation in this thesis arises from a context where this sort of location-mixture density shows up. In Kernel density estimation (with the choice of a normal kernel), the n observed values $x_1, \ldots, x_n$ of the random variables $X_1, \ldots, X_n$ are chosen as candidates for n component means, and equal probability is assigned to each "observed component mean". The bandwidth h in kernel density estimation then corresponds to the common component variance of the mixture density

$$\frac{1}{n} \sum_{i=1}^{n} \phi_{x_i,h},$$

and hence the Kernel density estimator is given by

$$\frac{1}{nh} \sum_{i=1}^{n} \phi \left(\frac{x - X_i}{h} \right),$$

where the function ϕ is the standard normal density function $\phi_{0,1}$.

In Chapter 3 we describe how bandwidth selection in the kernel density estimation context shares similarities with the problem of selecting the component variance in a location mixture density estimation setting. In fact, as can be seen in this example, we can view kernel density estimates as an estimate of a mixture density by assuming the probabilities of each mass point are identically $\frac{1}{n}$, and choosing the estimates of the mass points by the observed values $x_1, \ldots, x_n$.

We will adopt the word *bandwidth* for our own purposes in later sections as well, and typically will denote a fixed component variance in a location-mixture of normals by h^2. As in Kernel density estimation (with a Gaussian Kernel), we will use h to refer to the component standard deviation.

Example 1.1.3 (Convolutions of normals)**.** Consider a location-mixture of normals where the mixing distribution is not a discrete distribution. Suppose

$X \sim N(0, \sigma^2)$, and $Y \sim N(0, \tau^2)$. Let F_X, F_Y be the distributions of X and Y respectively. If Z has the mixture density given by

$$f_Z(z) = \int \phi_{0,\sigma}(z - y) dF_Y(y),$$

then since F_Y is a continuous (mixing) distribution, we can write $f_Z(z)$ as

$$f_Z(z) = \int \phi_{0,\sigma}(z - y) dF_Y(y) = \int \phi_{0,\sigma}(z - y)\phi_{0,\tau}(y) dy.$$

The integral $\int \phi_{0,\sigma}(z - y)\phi_{0,\tau}(y) dy$ is simply the convolution of the two original densities at a point z

$$(\phi_{0,\sigma} * \phi_{0,\tau})(z),$$

so with a basic calculation involving completing the square inside the exponential terms of the integral, we can show that $\phi_{0,\sigma} * \phi_{0,\tau} = \phi_{0,\sqrt{\sigma^2+\tau^2}}$. Hence a mixture of normals via a normal mixing distribution yields a normal with a greater variance.

So, normal random variables can be thought of in the mixture context via at least two ways. One way is as a single component mixture with some variance ($\sigma^2 + \tau^2$, say) via some degenerate mixing distribution, and another way is as a location mixture of normals with variance σ^2, where the mixing distribution is itself a normal distribution with variance τ^2.

1.2 Non parametric maximum likelihood estimates (NPMLEs)

In this section we briefly describe a great contribution by Lindsay (1983) which gives us a lot of information about the nature of estimates obtained via maximising a mixture likelihood.

1.2.1 Maximum likelihood

Suppose $\Theta \subseteq \mathbb{R}^m$ and suppose we have an iid sample $X_1, \ldots, X_n$ from a density f_θ, where the parameter $\theta \in \Theta$ is unknown.

The traditional maximum likelihood method for producing an estimate of the parameter θ defines the likelihood function (at least in the case of discrete X_is) as the probability the sample $X_1, \ldots, X_n$ was observed, viewed as a function of the unknown parameter θ

$$\ell(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

Since the domain of ℓ is a subset of $\mathbb{R}^m$, assuming there are sufficient conditions for the derivatives to exist, the maximiser $\widehat{\theta}$ of $\ell(\theta)$ can be found by solving the equations

$$\frac{\partial}{\partial \theta_j}\ell(\theta) = 0, j = 1, \ldots, m$$

using elementary calculus.

In a mixture model context, when we wish to view an iid sample $X_1, \ldots, X_n$ as being from a mixture density

$$f_Q = \int f_\theta dQ(\theta),$$

the unknown 'parameter' is now an unknown distribution. In this thesis we assume the X_i are continuous random variables. The maximum likelihood approach to estimating Q leads to the likelihood function

$$\ell(Q) = \prod_{i=1}^{n} f_Q(X_i) = \prod_{i=1}^{n} \int f_\theta(X_i) dQ(\theta),$$

which cannot be maximised using the traditional techniques from the parametric setting. The mixture model produces a non parametric problem which gives rise to several questions, such as:

1. Does a maximiser $\widehat{Q}$ of $\ell(Q)$ even exist?

2. If it exists, is it unique?

3. If it exists, can we work out what it is (or produce an approximation of it)?

These nontrivial (and once difficult) questions are addressed quite beautifully by Lindsay (1983) with:

1. Yes,

2. Yes under some light conditions,

3. We have a characterisation of it.

We now outline some of Lindsay's results.

1.2.2 Lindsay's theorems

The work of Lindsay (1983) has established that there indeed exists a maximiser $\widehat{Q}$ of $\ell(Q)$. Lindsay also provides conditions under which the uniqueness of $\widehat{Q}$ can be established, and the nature of $\widehat{Q}$ is further revealed to be that of a discrete distribution with finitely many mass points. Moreover he shows the number of mass points K (which is random, depending upon the sample $X_1, \ldots, X_n$) is bounded by the number of distinct observations n in the sample.

The way these results are established is by translating the problem of maximising $\ell(Q)$ into a geometric setting and drawing upon results from convex geometry. Caratheodory's Theorem is applied to establish existence and uniqueness of $\widehat{Q}$ (uniqueness is established under some conditions which are satisfied in the Gaussian mixture context).

A particular function of the observations $X_1, \ldots, X_n$ and n, defined in Lindsay (1983), provides an especially useful characterisation of $\widehat{Q}$.

A definition is now provided here. It is less general than in the original NPMLE setting, because we will focus upon mixtures of normal densities in this thesis.

Definition. Let Q be some probability distribution on $\mathbb{R}$, and suppose $X_1, \ldots, X_n$ is an iid sample from the Gaussian location-mixture density $\int \phi(x-\mu)dQ(\mu)$. Let D_Q be given by

$$D_Q(\theta) \quad = \quad \sum_{i=1}^{n} \left\{ \frac{\phi(X_i - \theta)}{\int \phi(X_i - \mu) dQ(\mu)} - 1 \right\}, \text{for } \theta \in \mathbb{R}. \tag{1.3}$$

Lindsay characterises $\widehat{Q}$ by relating it to $D_Q(\theta)$ with the following equivalence theorem.

Theorem 1.2.1 (by Lindsay (1983))**.** *The following are equivalent:*

1. $\widehat{Q}$ *maximises the log likelihood:* $\sum_{i=1}^{n} \log \left(\int \phi(X_i - \mu) dQ(\mu) \right)$.

2. $\widehat{Q}$ *minimises* $\sup_{\theta \in \mathbb{R}} D_Q(\theta)$.

3. $\sup_{\theta \in \mathbb{R}} D_{\widehat{Q}}(\theta) = 0$.

Moreover, the mass points of $\widehat{Q}$ *are the values* $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ *satisfying:*

$$D_{\widehat{Q}}(\widehat{\theta}_i) = 0, \text{ for } i = 1, 2, \ldots, K.$$

Note that the number of mass points K of the NPMLE $\widehat{Q}$ is random. In cases where the true mixing distribution Q_0 is discrete with k mass points, we may naturally desire K to give us an idea about k, if a sufficiently large number of observations n were available.

1.2.3 Some consequences

As a result of Lindsay's work, the problem of producing NPMLEs of mixing distributions has now been essentially solved, though computational speed is still a current issue. This has led to a renewed interest in applying mixture models to a plethora of statistical problems.

Theorem 2.1.1 by Lindsay (1983) has led to the development of many techniques for computing estimates of $\widehat{Q}$ based upon observations $X_1, \ldots, X_n$. Slow and simplistic computational methods like applying the Expectation Maximisation (EM) algorithm to estimate $\widehat{Q}$ follow immediately since $\widehat{Q}$ is known to have finitely many mass points (bounded by n).

More sophisticated algorithms for computing estimates of $\widehat{Q}$ like the Intra Simplex Direction Method by Lesperance and Kalbfleisch (1992) directly use the definition and properties of D_Q to (more quickly) estimate $\widehat{Q}$. Incidentally, the ISDM is what we have implemented in our simulations in Chapter 3.

The possibility of computing estimates of $\widehat{Q}$ (which itself is an estimate of a mixing distribution Q in a mixture model) has led to interest in various aspects of mixture model applications. For example, questions about the rates of convergence of such estimates arise, along with questions about the properties of mixture model techniques and how to formulate or interpret a mixture model sensibly.

A particular aspect of the Non Parametric Maximum Likelihood Estimation technique relates to the interpretation of the estimated mixing distribution. It is known that the number of components K of the NPMLE typically has very few components (compared to the sample size n), but also that making inferences about K is a nonstandard and difficult problem (Lindsay and Lesperance, 1995). Our own result is an asymptotic inconsistency type result about K. It provides an example where the behaviour of K does not reflect the nature of the true number of components in a mixture. However, functions of $\widehat{Q}$ such as estimated mixture densities $f_{\widehat{Q}}$ can be shown to be sensible and useful to consider in applications such as density estimation.

In the next section, we discuss some issues related to the number of estimated components in a NPMLE of a mixture model's mixing distribution.

1.3 Estimating the number of components via NPMLE

In this section we mention some reasons why it is preferable not to have too many parameters in a mixture model. We then describe a result of

Leroux (1992) about the consistent estimation of the number of components in NPMLE, and then describe our result from Chapter 2. We then show how our result can be used to extend the one in Leroux (1992).

1.3.1 Estimating at least the true number of components

Suppose a finite mixture model assumes m components when the true number of mass points in the mixture is strictly less than m. Let the true mixing distribution in the model be called Q and let $\widehat{Q}$ be an estimator of Q. It was shown by Chen (1995) that the optimal rate of convergence of $\widehat{Q}$ to Q is $n^{-\frac{1}{4}}$, however in the case where the mixture truly does have m components, the optimal rate is $n^{-\frac{1}{2}}$. Using the notation from Chen (1995), suppose the true mixing distribution G_0 is an $m-1$ point mixing distribution and let G be the m point mixing distribution assumed by the model. Let $\widehat{G}$ be a consistent estimator of G in the m point model. Chen (1995) shows that the estimator $\widehat{G}$ cannot converge to G_0 in the $\mathcal{L}_1$ metric any faster than $n^{-\frac{1}{4}}$, where n is the sample size.

Penalised likelihood methods have been motivated by such reasons to address the issue of overparametrisation. For example, Leroux (1992) proposes to choose an estimator of the number of components m as the $\widehat{m}_n$ which maximises

$$\ell_n(\widehat{Q}_m) - a_{m,n}, \tag{1.4}$$

where ℓ_n is the usual log-likelihood function, $\widehat{Q}_m$ is the NPMLE defined by Lindsay (1983) (for an m component mixture), and $a_{m,n}$ is a penalty term which discourages the selection of too many parameters by requiring $a_{m+1,n} \geq a_{m,n}$.

In fact, Leroux (1992) shows that the maximiser $\widehat{m}_n$ of (1.4) is at least as large as the true number of components m.

More precisely, a special case of Theorem 4 of Leroux (1992) (where the penalty term $a_{m,n}$ is removed by setting each $a_{m,n} = 0$) can be stated as (assuming we are talking about normal location-mixtures) follows.

Theorem 1.3.1 (Leroux (1992))**.** *Let (the mixing distribution) F^* have $m*$ components ($m^* = \infty$ if F^* is not a finite distribution). Let $\widehat{m}_n$ be the maximiser of (1.4) with each $a_{m,n} = 0$. We have:*

$$\liminf_{n\to\infty} \widehat{m}_n \geq m^*(\widehat{m}_n \to \infty \textit{ if } m^* = \infty) \textit{ with probability } 1. \tag{1.5}$$

The removal of the penalty term $a_{m,n}$ in the above theorem brings us back into a non penalised maximum likelihood situation. The likelihood in the regular NPMLE setting $\ell(Q)$ is maximised by $\widehat{Q}$ of Theorem 2.1.1, which is discrete and has K components (K could take any of the values $1, \ldots, n$). The maximiser $\widehat{m}_n$ of 1.4 then, is K.

This result tells us that the number of components of the NPMLE for a normal location mixture is at least the true number of components, for large sample sizes.

1.3.2 Can we do any better?

In this thesis we show there exists a case where equality cannot hold in (1.5), and hence no stronger statement of their theorem can be made.

Even if the true number of components is 1, in a normal location-mixture model with mixing distribution Q and NPMLE $\widehat{Q}$, our result shows that the probability that the number of components K is larger than 1, approaches 1, as $n \to \infty$.

At the very least, our result provides a counter example to the idea that the number of components of the estimate $\widehat{Q}$ could be interpreted as an estimate of the number of components of Q.

1.4 Classes of location-scale mixture densities

In this section we show it is possible to reformulate seemingly more general problems about nearly arbitrary normal location-scale mixtures into problems about a location mixture.

Normal location-scale mixture densities and normal location mixture densities can be seen as a generalisation of single normal densities, when the mixing distribution (a distribution on $\mathbb{R} \times (0, \infty)$ or $\mathbb{R}$ respectively) is considered to be not necessarily degenerate. Since mixing distributions on $\mathbb{R}$ are simpler to estimate, compute or visualise than mixing distributions on $\mathbb{R} \times (0, \infty)$, the question about whether simplifying a normal location-scale mixture model to that of a normal location mixture model is sensible arises.

Definition 1.4.1. Suppose $h > 0$ and Q is a probability distribution on $\mathbb{R} \times [\sigma, \infty)$ for all $\sigma \geq h$. Let f_Q denote the density

$$f_Q = \int \phi_{\mu,\sigma} dQ(\mu, \sigma),$$

and let $\mathcal{F}_h^{(a)}$ be the class of densities

$$\mathcal{F}_h^{(a)} = \{f_Q : Q \text{ is prob distribution on } \mathbb{R} \times [\sigma, \infty) \text{ for all } \sigma \geq h\}.$$

Suppose instead that Q is a distribution on $\mathbb{R}$. Let $f_{Q,h}$ denote the density

$$f_{Q,h} = \int \phi_{\mu,h} dQ(\mu),$$

and let $\mathcal{F}_h^{(b)}$ be the class

$$\mathcal{F}_h^{(b)} = \{f_{Q,h} : Q \text{ is a probability distribution on } \mathbb{R}\}.$$

These definitions lead naturally to the following question:

> "Which (out of $\mathcal{F}_h^{(a)}$ and $\mathcal{F}_h^{(b)}$) is a richer class of densities to consider? Normal location-scale mixtures on $\mathbb{R} \times [h, \infty)$ or normal location mixtures on $\mathbb{R}$?".

We now present some examples of what members of $\mathcal{F}_h^{(a)}$ or $\mathcal{F}_h^{(b)}$ might be for any fixed $h > 0$.

Members of $\mathcal{F}_h^{(a)}$ - Location-scale mixtures of normals

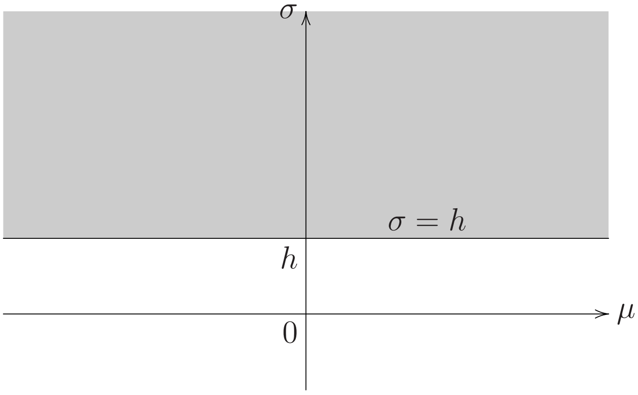
Since for each $f \in \mathcal{F}_h^{(a)}$ there exists a Q on $\mathbb{R} \times [h, \infty)$ such that we can write f as $f = \int \phi_{\mu,\sigma} dQ(\mu, \sigma)$, we will use some specific mixing distributions Q on $\mathbb{R} \times [h, \infty)$ to provide examples of members of $\mathcal{F}_h^{(a)}$.

Example 1.4.2. Suppose $h < 1$ and Q is the degenerate distribution which places probability 1 on the point $(0, 1) \in \mathbb{R} \times [h, \infty)$. Then from the definition of an integral where the measure is a point mass,

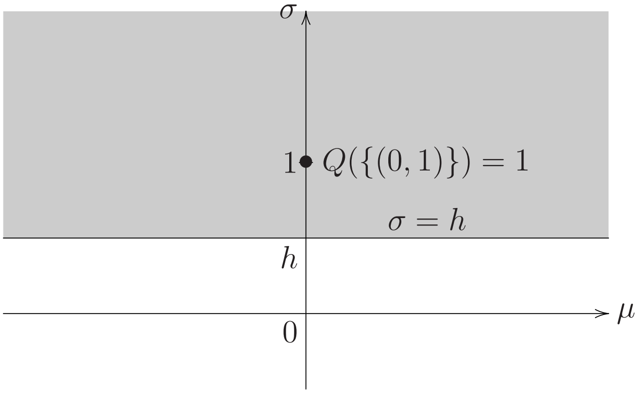
$$\int \phi_{\mu,\sigma} dQ(\mu, \sigma) = \phi.$$

In this sense, we can think of any normal density as a special case of a location-scale mixture density, with a degenerate mixing distribution on $\mathbb{R} \times [h, \infty)$, the possible combinations of locations-scales/means-standard deviations.

Since the mixing distribution Q is defined on $\mathbb{R} \times [h, \infty)$, we can draw a copy of $\mathbb{R}^2$ and label one axis μ to represent possible choices of mean and the other axis σ to represent possible choices of standard deviation. In this example (where $h < 1$), Q would assign probabilities to subsets of $\mathbb{R} \times [h, \infty)$ (the shaded region below).

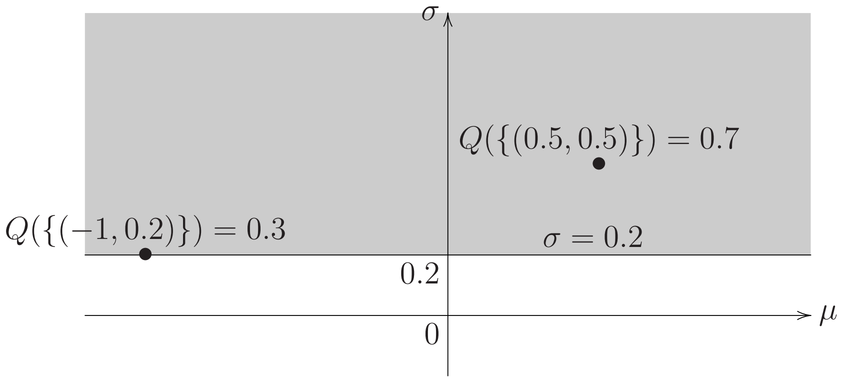


Since Q is a discrete distribution with finitely many points we could depict Q graphically to show it assigns the probability 1 to the set $\{(0, 1)\} \subset \mathbb{R} \times [h, \infty)$.



The point $(\mu, \sigma) = (0, 1)$ with the probability 1 can be seen to specify the mean and standard deviation of a standard normal density with probability 1.

Example 1.4.3. Suppose $h = 0.2$, and let Q be the mixing distribution from Example 3.1.1. We could mark the points where Q puts positive probability graphically on $\mathbb{R} \times [h, \infty)$.



The two points $(\mu, \sigma) = (-1, 0.2)$ and $(\mu, \sigma) = (0.5, 0.5)$ can be seen to correspond to the two component densities $\phi_{-1,0.2}$ and $\phi_{0.5,0.5}$ of the density $\int \phi_{\mu,\sigma} dQ(\mu, \sigma)$.

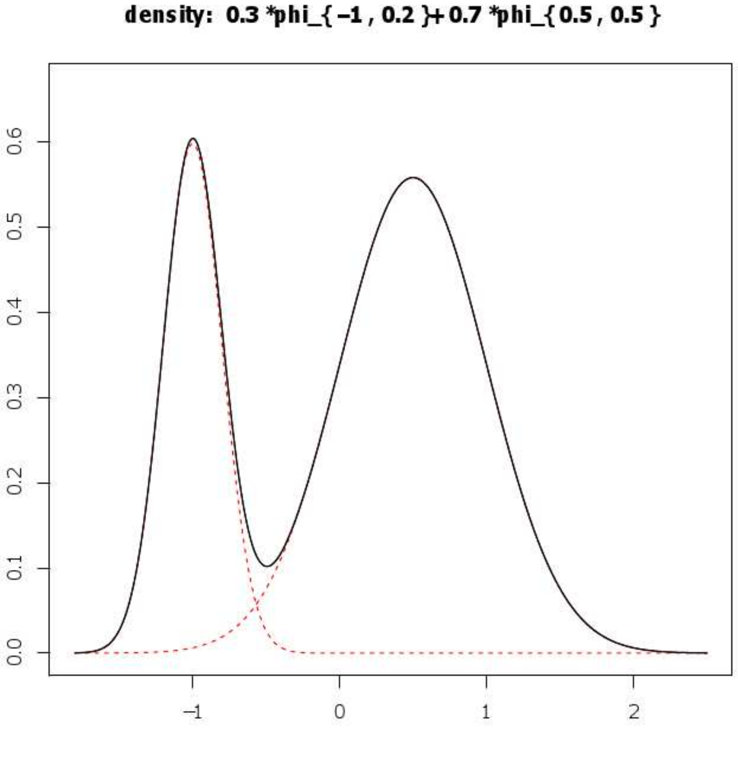
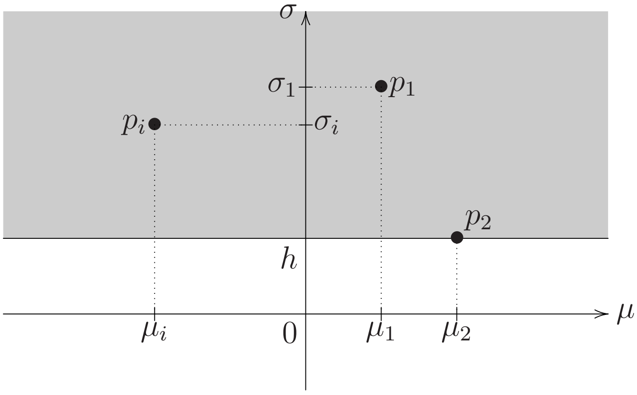


Figure 1.2: Two component densities and the resulting mixture

Example 1.4.4. Suppose Q is a discrete distribution which places probability p_j on the point $(\mu_j, \sigma_j) \in \mathbb{R} \times [h, \infty)$ for $j = 1, 2, \ldots$. Then the mixture density $\int \phi_{\mu,\sigma} dQ(\mu, \sigma)$ is the convex combination of the normal densities ϕ_{μ_j,σ_j}

$$\int \phi_{\mu,\sigma} dQ(\mu, \sigma) = \sum_{i=1}^{\infty} p_j \phi_{\mu_j,\sigma_j}.$$

As with the previous examples, since Q is a distribution defined on the region $\mathbb{R} \times [h, \infty)$ in a simple sort of way, we could depict it graphically.

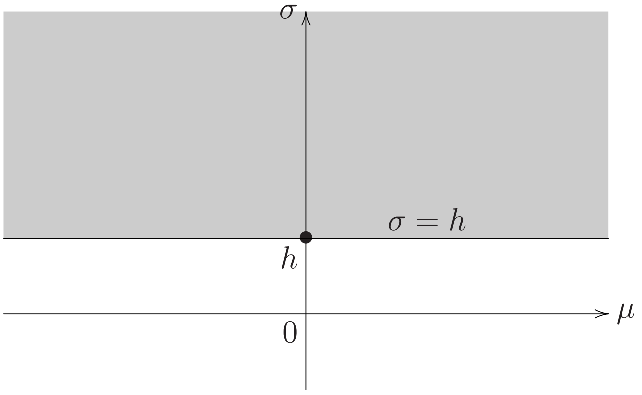


In the usual way in which the integral is defined in measure theory, we can come up with a multitude of measures Q on the possible subsets of $\mathbb{R} \times [h, \infty)$, some of which may not be easily depicted graphically, and we would end up with a plethora of functions $\int \phi_{\mu,\sigma} dQ(\mu, \sigma)$ in $\mathcal{F}_h^{(a)}$.

Members of $\mathcal{F}_h^{(b)}$ - Location mixtures of normals

Example 1.4.5. Suppose Q is the degenerate distribution which puts probability 1 on the point $0 \in \mathbb{R}$. Then the mixture density $\int \phi_{\mu,h} dQ(\mu) = \phi_{0,h}$. Similarly to Example 1.4.2, we can think of any normal density with variance h^2 as a special case of a location mixture density with a degenerate mixing distribution on the possible values of locations/means, $\mathbb{R}$.

Note that the density $\phi_{0,h}$ can be expressed as both $\phi_{0,h} = \int \phi_{\mu,h} dQ(\mu)$ and as $\phi_{0,h} = \int \phi_{\mu,\sigma} d\widetilde{Q}(\mu, \sigma)$, when $\widetilde{Q}$ (defined on $\mathbb{R} \times [h, \infty)$ instead of being defined on $\mathbb{R}$) puts probability 1 on the point $(0, h)$. If we were to draw the graphical representations of this alternate $\widetilde{Q}$ in the same fashion to those in Examples 1.4.2, 1.4.3 and 1.4.4, we would arrive at the following figure.



Note that since $\widetilde{Q}$ puts probability 1 on the subset of $\mathbb{R} \times [h, \infty)$ specified by the line $\sigma = h$, the density in this example is a (single/degenerate) combination of normal densities with a variance of exactly h^2.

From Example 1.4.5 we could express a location mixture of normals $\int \phi_{\mu,h} dQ(\mu)$ in terms of a location-scale mixture of normals as well, by constructing a new mixing distribution $\widetilde{Q}$ (on $\mathbb{R} \times [h, \infty)$) based on the the mixing distribution Q. In fact this is true in general, as shown below.

Lemma 1.4.6. *For any $h > 0$, $\mathcal{F}_h^{(b)} \subseteq \mathcal{F}_h^{(a)}$.*

Proof. Suppose $f \in \mathcal{F}_h^{(b)}$. Then f is expressible in the form

$$f = \int \phi_{\mu,h} dQ(\mu), \text{ for some } Q \text{ on } \mathbb{R}.$$

Let $\widetilde{Q}$ be the distribution on $\mathbb{R} \times [h, \infty)$ given by

$$\widetilde{Q}(\mu, \sigma) = \begin{cases} Q(\mu) & , \sigma = h \\ 0 & , \sigma \neq h \end{cases},$$

then we can rewrite f as

$$f = \int \phi_{\mu,\sigma} d\widetilde{Q}(\mu, \sigma),$$

and thus $f \in \mathcal{F}_h^{(a)}$. □

We next show an example where we reexpress a location mixture from $\mathcal{F}_h^{(b)}$ in terms of an alternate mixing distribution.

Example 1.4.7. Let $h = 0.2$ and Q be the distribution given by

$$Q = 0.3\delta_{-1} + 0.7\Phi_{\frac{1}{2},\sqrt{0.21}},$$

where δ_{-1} is the degenerate distribution putting probability 1 on the point $-1 \in \mathbb{R}$ and $\Phi_{\frac{1}{2},\sqrt{0.21}}$ is the distribution with density $\phi_{\frac{1}{2},\sqrt{0.21}}$. Then the mixture density $f(x) = \int_{\mathbb{R}} \phi_{\mu,0.2}(x)dQ(\mu)$ is given by

$$\begin{aligned}
\int_{\mathbb{R}} \phi_{\mu,0.2}(x)dQ(\mu) &= 0.3\phi_{-1,0.2}(x) + 0.7\int_{\mathbb{R}} \phi_{\mu,0.2}(x)d\Phi_{\frac{1}{2},\sqrt{0.21}}(\mu) \\
&= 0.3\phi_{-1,0.2}(x) + 0.7\int_{\mathbb{R}} \phi_{\mu,0.2}(x)\phi_{\frac{1}{2},\sqrt{0.21}}(\mu)d\mu \\
&= 0.3\phi_{-1,0.2}(x) + 0.7\int_{\mathbb{R}} \phi_{0,0.2}(x-\mu)\phi_{\frac{1}{2},\sqrt{0.21}}(\mu)d\mu \\
&= 0.3\phi_{-1,0.2}(x) + 0.7\phi_{0,0.2} * \phi_{\frac{1}{2},\sqrt{0.21}}(x).
\end{aligned}$$

Since $\sqrt{0.21} = \sqrt{\frac{1}{4} - 0.2^2}$, we have

$$\phi_{0,0.2} * \phi_{\frac{1}{2},\sqrt{0.21}} = \phi_{\frac{1}{2},\sqrt{\frac{1}{4}}} = \phi_{\frac{1}{2},\frac{1}{2}},$$

and so

$$f(x) = \int_{\mathbb{R}} \phi_{\mu,0.2}(x)dQ(\mu) = 0.3\phi_{-1,0.2}(x) + 0.7\phi_{\frac{1}{2},\frac{1}{2}}(x).$$

This is exactly the location-scale mixture density from Example 1.4.3, which is a member of $\mathcal{F}_h^{(a)}$!

As the above example suggests, it turns out that $\mathcal{F}_h^{(a)}$ and $\mathcal{F}_h^{(b)}$ are actually equal. We now show the converse of Lemma 1.4.6 holds.

Lemma 1.4.8. *For any $h > 0$, $\mathcal{F}_h^{(a)} \subseteq \mathcal{F}_h^{(b)}$.*

Proof. Suppose $f \in \mathcal{F}_h^{(a)}$. Then with some Q on $\mathbb{R} \times [h, \infty)$, f may be expressed as

$$f = \int \phi_{\mu,\sigma} dQ(\mu, \sigma),$$

and so for any $x \in \mathbb{R}$, $f(x)$ is given by

$$f(x) = \int \phi_{\mu,\sigma}(x) dQ(\mu, \sigma).$$

For the sake of convenience, we use $\int \phi_{\mu,\sigma} dQ(\mu, \sigma)$ to mean the function given by the integral of $\phi_{\mu,\sigma}$ with respect to the measure Q over the parameters μ

and σ, while we use $Q(A \times B)$ to be the mass Q puts on the set $A \times B \subseteq \mathbb{R} \times (0, \infty)$.

We can rearrange $f(x)$ to be

$$\begin{aligned} f(x) &= \int_{\mathbb{R}\times(0,h)} \phi_{\mu,\sigma}(x) dQ(\mu, \sigma) + \int_{\mathbb{R}\times[h,\infty)} \phi_{\mu,\sigma}(x) dQ(\mu, \sigma) \\ &= 0 + \int_{\mathbb{R}\times[h,\infty)} \phi_{\mu,\sigma}(x) dQ(\mu, \sigma), \end{aligned}$$

since $Q(\mathbb{R} \times [h, \infty)) = 1$. In the above integral, since the relevant σ must satisfy $\sigma \geq h$ we use

$$\phi_{\mu,\sigma} = \phi_{\mu,h} * \phi_{0,\sqrt{\sigma^2-h^2}}$$

to arrive at

$$f(x) = \int_{\mathbb{R}\times[h,\infty)} (\phi_{\mu,h} * \phi_{0,\sqrt{\sigma^2-h^2}})(x) dQ(\mu, \sigma). \tag{1.6}$$

Since the convolution of $\phi_{\mu,h}$ with $\phi_{0,\sqrt{\sigma^2-h^2}}$ is given by

$$\phi_{\mu,h} * \phi_{0,\sqrt{\sigma^2-h^2}}(x) = \int_{\mathbb{R}} \phi_{\mu,h}(x - y)\phi_{0,\sqrt{\sigma^2-h^2}}(y) dy,$$

let us make the substitution $z = y + \mu$ to get

$$\phi_{\mu,h} * \phi_{0,\sqrt{\sigma^2-h^2}}(x) = \int_{\mathbb{R}} \phi_{z,h}(x)\phi_{\mu,\sqrt{\sigma^2-h^2}}(z) dz. \tag{1.7}$$

Using (1.6) together with (1.7) provides

$$f(x) = \int_{\mathbb{R}\times[h,\infty)} \int_{\mathbb{R}} \phi_{z,h}(x)\phi_{\mu,\sqrt{\sigma^2-h^2}}(z) dz dQ(\mu, \sigma). \tag{1.8}$$

The function $\phi_{z,h}(x)\phi_{\mu,\sqrt{\sigma^2-h^2}}(z)$ is nonnegative and bounded, and hence integrable with respect to Q or the Lebesgue measure, and thus satisfies the conditions of Fubini's theorem. We hence swap the order of integration to arrive at

$$\begin{aligned} f(x) &= \int_{\mathbb{R}} \int_{\mathbb{R}\times[h,\infty)} \phi_{z,h}(x)\phi_{\mu,\sqrt{\sigma^2-h^2}}(z) dQ(\mu, \sigma) dz \\ &= \int_{\mathbb{R}} \phi_{z,h}(x) \int_{\mathbb{R}\times[h,\infty)} \phi_{\mu,\sqrt{\sigma^2-h^2}}(z) dQ(\mu, \sigma) dz. \end{aligned}$$

The function g given by $g(z) = \int_{\mathbb{R}\times[h,\infty)} \phi_{\mu,\sqrt{\sigma^2-h^2}}(z) dQ(\mu, \sigma)$ is a normal location-scale mixture density. Let G denote the distribution of a random variable with density g, then G is some distribution on $\mathbb{R}$. Since g is continuous, $\frac{d}{dz}G(z) = g(z)$, so we have

$$\begin{aligned} f(x) &= \int_{\mathbb{R}} \phi_{z,h}(x) g(z) dz \\ &= \int_{\mathbb{R}} \phi_{z,h}(x) dG(z), \end{aligned}$$

and thus $f \in \mathcal{F}_h^{(b)}$. □

Since $\mathcal{F}_h^{(a)} = \mathcal{F}_h^{(b)}$, from now on we drop the superscripts and refer to either as $\mathcal{F}_h$. We will usually prefer to think of $\mathcal{F}_h$ using the definition of $\mathcal{F}_h^{(b)}$, since it is a simpler characterisation. Not only are we able to simplify our notation, we can use this equality (of classes of densities) to perform these tasks:

- Reformulate seemingly more general problems about location-scale mixture densities to a simpler problem concerning a location mixture.
- Use a nesting property of $\mathcal{F}_h^{(a)}$ to deduce a nesting property of $\mathcal{F}_h$. (See Section 3.3)

In Chapter 3 we will describe a bandwidth selection problem which is unsolved to our knowledge. This problem is analogous to one found in Kernel density estimation, and has led us to implement a density estimation procedure via NPMLE using the ISDM. We have implemented the ISDM completely in R, and also written a version which calls C code to do the heaviest computations. The latter version is faster by a factor of 10.

1.5 Our work in a wider context

In this section, we talk about how the work from our thesis fits into the wider scheme of things, in a mixture model context. We first provide further clarification about the nature of our own work, since this thesis has a rather broad title. We describe the relationship between the theoretical result in Chapter 2 with the computational density estimation problem we looked at in Chapter 3.

We then provide a list of several aspects of our work, along with some additional comments about some of the ways in which these ideas fit into the wider literature.

1.5.1 Estimated components and density estimation

Besides both belonging to the huge area of mixture model research, it may appear as though the contents of Chapter 2 and Chapter 3 are quite unrelated. We describe below how the ideas in our thesis developed to show how they sit relative to each other.

Section 1.4 refers to ideas from Magder and Zeger (1996). In fact Lemma 1.4.8 is a less general version of Theorem 1 from Magder and Zeger (1996). The paper by Magder and Zeger (1996) is what motivated the central direction of this thesis. That is, we were interested in the possibilities offered by reformulating problems about general location-scale mixtures of normals into simpler problems about location mixtures of normals. The bandwidth parameter h as discussed in this chapter (and in Chapter 3) corresponds to the one in Magder and Zeger (1996) as well.

Magder and Zeger (1996) mention that their experience suggests the problem of choosing h based upon the data alone would be difficult without extremely large data sets. Our experiences described in Chapter 3 echo this sentiment. In fact, our experiences with this problem have also highlighted how difficult it is to implement a (fast) computational procedure to work with even relatively small data sets in exploring sensible choices of bandwidth h.

The relationship between Chapter 2 and Chapter 3, then, is best described without technicalities or reference to mathematics. We tried to approach this bandwidth selection problem, but failed to produce any satisfying or useful results, in multiple directions of our search. At one point, we noticed in our simulation studies that as we decreased h towards 0, the number of estimated components of the NPMLE given by our code would tend to be large. In the other direction, when we chose ridiculously large h values, the number of estimated components of the NPMLE would tend to decrease.

From this rough and intuitive observation we became interested in studying the distribution of the number of components of the NPMLE. Our motivation was based on the hopes that an understanding about estimated component number would provide insight towards the seemingly related bandwidth parameter h. It was in this direction of focus that we ended up with a proof of a type of inconsistency regarding the number of estimated components of the NPMLE. Note that this notion of 'inconsistency' is described more carefully in Chapter 2.

The result in Chapter 2 has stemmed from interest in the work of Chapter 3, however it also stands alone as a theoretical result independent of our motivations for looking at it. In this sense these two parts of our thesis are quite unrelated. As such, we chose not to express the main result in Chapter 2 under the framework and notation from Chapter 3, in order to avoid overcomplicating the idea with extra context.

Note that our thesis title is 'Some aspects of non parametric maximum likelihood for normal mixtures'. Chapters 2 and 3 both involve NPMLEs, but in different ways. The former chapter concerns a discrete random variable, and the latter is motivated by mostly known properties of density estimation based on NPMLE. To clarify potential confusion between our topics and others' work based upon similar titles; although we provide an asymptotic result about the nature of a discrete random variable K, we do not provide any asymptotic results about density estimates produced via NPMLEs. The article by van de Geer (2003) does an excellent job of describing aspects of asymptotic theory regarding these density estimates.

In fact, this work done by van de Geer (2003), along with others such as Ghosal and van der Vaart (2001) and Geman and Hwang (1982) point to the idea that estimating the density of the data in such models (in the case of Geman and Hwang (1982) there is a sieve model) works quite well. Note again that these are articles about density estimation in a mixture context, not articles about properties of the mixing distribution estimates which lead to the density estimates. For this reason (along with simple fascination with the bandwidth selection comments from Magder and Zeger (1996)), we were motivated to look at density estimation via NPMLE.

1.5.2 Our contribution

There is a large literature on ways of making the NPMLE consistent, for example using sieve methods, penalisation, or other approaches to estimator constraint. Using a sieve (eg Grenander (1981), Geman and Hwang (1982)), penalised likelihood method (Cathy and Bertrand (2011), Leroux (1992)) or some other form of regularisation is known to be sufficient to ensure consistency for estimating the number of mass points.

However the literature does not provide a large amount of discussion about whether it is necessary. Leroux (1992) discusses a result which suggests that the NPMLE might overestimate the number of components of a mixture (the result says the number estimated will be $\geq$ the true number), but it does not confirm whether it definitely will overestimate it. Our result in Chapter 2 extends this idea from Leroux (1992) to say that yes, it might overestimate the number of components as noted by Leroux (1992), and in fact it will overestimate it.

At this time of writing, the literature generally focuses on estimating mixture densities (not mixing distributions in themselves). Otherwise, the focus is generally on providing work to show some kind of regularisation is sufficient, rather than on the question "is this regularisation necessary?". In the context of others' work on such topics, our main theoretical result says "yes, their work is necessary".

Chapter 2

Inconsistency of K for number of components

In this chapter we provide an example in which the Non Parametric Maximum Likelihood Estimate (NPMLE) K of the number of components k of a normal mixture is inconsistent for k. Section 2.1 describes this example which we have called Theorem 2.1.2, and a proof of this theorem is given in Section 2.3. A brief list of the main results we required in our proof is presented in Section 2.2. Some bonus details for this proof are given in the following Section 2.4, since they do not contribute much to the main proof ideas. Section 2.5 contains a recount of a simulation we have done to demonstrate this inconsistency result in practice.

2.1 Introduction

The number of mass points K of the Non Parametric Maximum Likelihood Estimate (NPMLE) $\widehat{Q}$ of an unknown mixing distribution as defined by Lindsay (1983) is known to be random. In this section we describe an inconsistency result regarding K.

Suppose we model $X_1, \ldots, X_n$ as iid with density f given by the normal location-mixture with unit variance

$$f(x) \quad = \quad \int_{\mathbb{R}} \phi(x-\mu) dQ_0(\mu), \tag{2.1}$$

where Q_0 denotes the unknown mixing distribution of the model. As described in Chapter 1 Section 1.2, Lindsay (1983) shows the NPMLE $\widehat{Q}$ of Q_0 exists, is a discrete distribution, and has finitely many mass points. Lindsay defines

$$D_Q(\theta) \quad = \quad \sum_{i=1}^{n} \left\{ \frac{\phi(X_i - \theta)}{\int \phi(X_i - \mu) dQ(\mu)} - 1 \right\}, \tag{2.2}$$

and characterises $\widehat{Q}$ by relating it to $D_Q(\theta)$ with the following equivalence theorem. Note that this is a restatement of an aforementioned result in Chapter 1, for the reader's convenience.

Theorem 2.1.1 (Lindsay (1983))**.**

1. *$\widehat{Q}$ maximises the log likelihood:* $\sum_{i=1}^{n} \log \left(\int \phi(X_i - \mu) dQ(\mu) \right)$.
2. *$\widehat{Q}$ minimises* $\sup_{\theta \in \mathbb{R}} D_Q(\theta)$.
3. $\sup_{\theta \in \mathbb{R}} D_{\widehat{Q}}(\theta) = 0$.

Moreover, the mass points of $\widehat{Q}$ are the values $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ satisfying:

$$D_{\widehat{Q}}(\widehat{\theta}_i) = 0, \ \textit{for } i = 1, 2, \ldots, K.$$

Note that the number of mass points K of the NPMLE $\widehat{Q}$ is random. In cases where the true mixing distribution Q_0 is discrete with k mass points, we may naturally desire K to give us an idea about k, if a sufficiently large number of observations n were available. However in this chapter we present a simple example to show K does not have this desired property.

Theorem 2.1.2. *Suppose $X_1, \ldots, X_n$ are iid with the mixture density given by (2.1). Let K be the number of mass points of the NPMLE $\widehat{Q}$ of the true mixing distribution Q_0, as characterised by Lindsay (1983). Suppose Q_0 has $k = 1$ mass point, and for simplicity suppose this mass point is* 0. *Then*

$$P(K = 1) \to 0, \ \textit{as } n \to \infty.$$

Theorem 2.1.2 provides a simple example which shows the estimated number of mixing distribution components K is not consistent for the true number of components k. This suggests that the classical maximum likelihood approach to estimating an unknown mixing distribution should be handled with caution if the number of components of a population's density are to be estimated using the NPMLE technique. This also justifies the strategy of penalising the likelihood (Leroux, 1992) which can estimate the number of mass points consistently.

2.2 Background

This section contains some definitions and theorems from the literature about stochastic processes, empirical processes and extreme value theory which we will use in the proof of Theorem 2.1.2.

The following theorem is from Bickel and Chernoff (1993, p 88-9) who in turn quote it from Billingsley (1968). It also follows as a consequence of Theorem 2.1 of Chapter 1 of Revuz and Yor (1994, page 25).

Theorem 2.2.1 (Kolmogorov Bound)**.** *Suppose $Z(t)$ is a stochastic process which satisfies*

$$\mathbb{E}|Z(s) - Z(t)|^2 \leq c(s-t)^2,$$

for $0 \leq s \leq t \leq 1$, then

$$P(\sup_{0 \leq t \leq 1} |Z(t) - Z(0)| \geq z) \leq Kc/z^2 \tag{2.3}$$

where K is an absolute constant.

If the interval $[0,1]$ over which $Z(t)$ is defined is replaced by one of length L, then the bound Kc/z^2 in (2.3) is replaced by $Kc(L/z)^2$, and we have

$$P(\sup_{0 \leq t \leq L} |Z(t) - Z(0)| \geq z) \leq Kc(L/z)^2. \tag{2.4}$$

The part of this theorem which talks about intervals of length L is listed as a corollary after the proof, in Bickel and Chernoff (1993).

Csörgő et al. (1986) provides us with theorems about empirical processes which enable us to construct a useful approximation in our proof of Theorem 2.1.2. The following definitions are necessary to state the theorem we wish to use.

Definition (Empirical distribution function)**.** For the rv $X_1, \ldots, X_n$, let $\mathbb{F}_n$ denote their empirical distribution function

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(X_i \leq x)},$$

where $1_{(X_i \leq x)}$ is the indicator function returning 1 if $(X_i \leq x)$ is true and 0 otherwise.

Definition (Uniform empirical process)**.** For $u \in [0,1]$ let $\alpha_n(u)$ be the process

$$\alpha_n(u) = \sqrt{n}\left(\mathbb{F}_n(u) - u\right),$$

where $\mathbb{F}_n(u)$ is the empirical distribution function of n independent uniform $U(0,1)$ random variables.

Definition ($\mathcal{L}^*$)**.** Let $\mathcal{L}^*$ denote the class of functions on $(0,1)$ given by

$$\mathcal{L}^* = \{f : f \text{ is left continuous and non decreasing on } (0,1)\}.$$

Definition ($\mathcal{L}^*$-decomposable)**.** A class of functions $\mathcal{L}$ is said to be $\mathcal{L}^*$-decomposable if each $\ell \in \mathcal{L}$ can be written as $\ell = \ell_1 - \ell_2$, where $\ell_1, \ell_2 \in \mathcal{L}^*$.

Here is an example of a $\mathcal{L}^*$-decomposable class of functions. Let $\mathcal{L}$ be given by

$$\mathcal{L} = \{ax^2 + bx + c, x \in (0,1) | a, b, c \in \mathbb{R}\}.$$

Then the elements of $\mathcal{L}$ are just quadratics defined over $(0,1)$. Each parabola f can be thought of as the difference between two increasing functions g and h. For example, the parabola $f(x) = x(1-x)$ can be written as $f(x) = g(x) - h(x)$, where

$$\begin{aligned} g(x) &= \begin{cases} x(1-x), & x \in (0, \frac{1}{2}] \\ \frac{1}{4}, & x \in (\frac{1}{2}, 1) \end{cases}, \\ h(x) &= \begin{cases} 0, & x \in (0, \frac{1}{2}] \\ x(x-1), & x \in (\frac{1}{2}, 1) \end{cases}. \end{aligned}$$

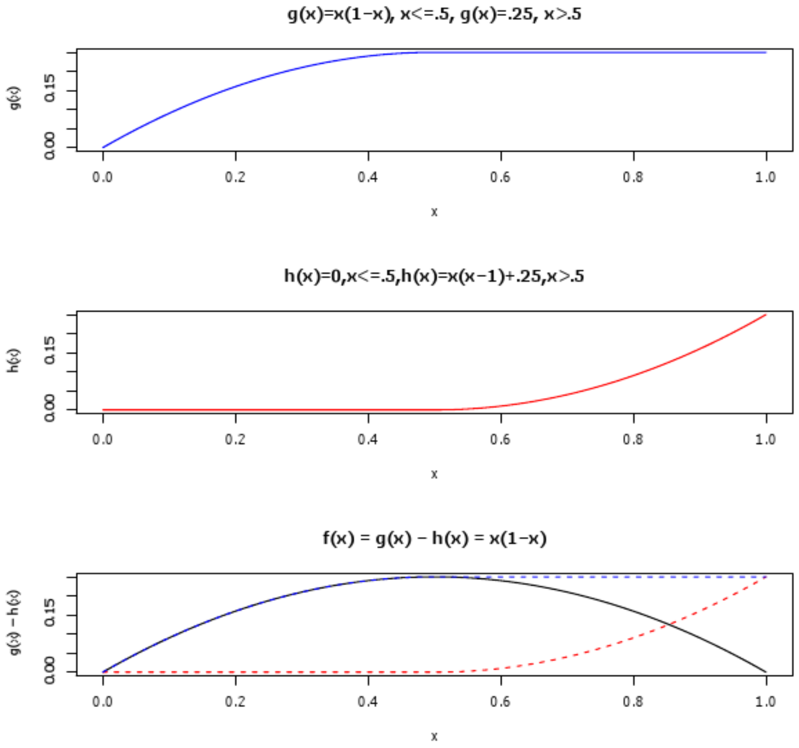


Figure 2.1: $f = g - h$, g (top) is the increasing part of f, h (middle) is the decreasing part

Note that although both g and h are increasing functions, h can be thought of as the function which provides information about the 'decreasing part' of f.

Definition 2.2.2 ($N_n(\delta)$)**.** Let $\mathcal{L}_n$ denote any sequence of $\mathcal{L}^*$-decomposable classes, and let $\delta > 0$.

$$N_n(\delta) = \sup_{\ell \in \mathcal{L}_n} \sup_{0 \le u \le \delta} \left\{ \left(|\ell_1(u)| + |\ell_2(u)| + |\ell_1(1-u)| + |\ell_2(1-u)| \right) u^{\frac{1}{2}} \right\}.$$

We will use the following theorem by Csörgő et al. (1986) in this chapter.

Theorem 2.2.3 (Csörgő et al. (1986))**.** *Let $\mathcal{L}_n$ ($n = 1, 2, \dots$) be any sequence of $\mathcal{L}^*$-decomposable classes, let $\delta_n = (\log n)/n^{\frac{1}{2}}$. If $N_n(\delta_n) = o(1), n \to \infty$, then there exists a probability space $(\Omega, \mathcal{A}, P)$ with independent $U(0,1)$ rv $U_1, U_2, \dots$ and a sequence of Brownian bridges $\{B_i(u); 0 \le u \le 1\}$ ($i =$*

$1, 2, \dots$*) such that*

$$E_n = \sup_{\ell \in \mathcal{L}_n} \left| \int_0^1 \ell(u) d\alpha_n(u) - \int_{1/n}^{1-1/n} \ell(u) dB_n(u) \right| = o_p(1).$$

The proof of Theorem 2.2.3 bounds E_n by three terms $E_n^{(1)}, E_n^{(2)}$, and $E_n^{(3)}$, and proceeds to show that $E_n^{(j)} = N_n(\delta_n) O_p(1)$, for $j = 1, 2, 3$. Thus (if we assume $N_n(\delta_n) \to 0$ as $n \to \infty$) we may state Theorem 2.2.3 as:

Theorem 2.2.4. *Let $\mathcal{L}_n$ $(n = 1, 2, \dots)$ be any sequence of $\mathcal{L}^*$-decomposable classes, let $\delta_n = (\log n)/n^{\frac{1}{2}}$. There exists a probability space $(\Omega, \mathcal{A}, P)$ with independent $U(0,1)$ rv $U_1, U_2, \dots$ and a sequence of Brownian bridges $\{B_i(u); 0 \leq u \leq 1\}$ $(i = 1, 2, \dots)$ such that*

$$E_n = \sup_{\ell \in \mathcal{L}_n} \left| \int_0^1 \ell(u) d\alpha_n(u) - \int_{1/n}^{1-1/n} \ell(u) dB_n(u) \right| = N_n(\delta_n) O_p(1).$$

The following theorem of Leadbetter and Rootzén (1988) gives us a way of comparing the behaviour of the maximum of a sequence of dependent normal random variables with the maximum of a sequence of independent normal random variables.

Theorem 2.2.5 (Normal comparison lemma (Leadbetter and Rootzén, 1988))**.** *Suppose $\zeta_1, \dots, \zeta_n$ are standard normal variables with covariance matrix $\Lambda^1 = \left(\Lambda^1_{ij}\right)$ and $\eta_1, \dots, \eta_n$ are standard normal variables with covariance matrix $\Lambda^0 = \left(\Lambda^0_{ij}\right)$. Let $\rho_{ij} = \max\left(|\Lambda^1_{ij}|, |\Lambda^0_{ij}|\right)$, and let $u_1, \dots, u_n$ be real numbers. Then*

$$\begin{aligned} &P(\zeta_i \leq u_i \textit{ for } i = 1, 2, \dots, n) - P(\eta_i \leq u_i \textit{ for } i = 1, 2, \dots, n) \\ &\quad \leq \frac{1}{2\pi} \sum_{1 \leq i < j \leq n} (\Lambda^1_{ij} - \Lambda^0_{ij})^+ (1 - \rho_{ij}^2)^{-\frac{1}{2}} \exp\left(\frac{u_i^2 + u_j^2}{2(1 + \rho_{ij})} \right), \end{aligned}$$

where $(x)^+ = \max(0, x)$. In particular, if $\max_{i \neq j} |\rho_{ij}| = \delta < 1$, then

$$\begin{aligned} &P(\zeta_i \leq u_i \textit{ for } i = 1, 2, \dots, n) - P(\eta_i \leq u_i \textit{ for } i = 1, 2, \dots, n) \\ &\quad \leq K \sum_{1 \leq i < j \leq n} (\Lambda^1_{ij} - \Lambda^0_{ij})^+ \exp\left(\frac{u_i^2 + u_j^2}{2(1 + \rho_{ij})} \right) \end{aligned}$$

for some constant K depending only on δ. Also,

$$\begin{aligned} &|P(\zeta_i \leq u_i \textit{ for } i = 1, 2, \dots, n) - P(\eta_i \leq u_i \textit{ for } i = 1, 2, \dots, n)| \\ &\quad \leq \frac{1}{2\pi} \sum_{1 \leq i < j \leq n} |\Lambda^1_{ij} - \Lambda^0_{ij}| (1 - \rho_{ij}^2)^{-\frac{1}{2}} \exp\left(\frac{u_i^2 + u_j^2}{2(1 + \rho_{ij})} \right), \end{aligned}$$

and when $\delta < 1$ the factor $\frac{1}{2\pi}(1 - \rho_{ij}^2)^{-\frac{1}{2}}$ can be replaced by K.

2.3 Our proof

The main theorem of this chapter is proven in this section. Some of the calculations and arguments for the proof of this theorem are deferred to Section 2.4 for the sake of clarity.

Proof of Theorem 2.1.2. The general model given by (2.1) says $X_1, \dots, X_n$ are iid with density f given by

$$f(x) = \int_{\mathbb{R}} \phi(x-\mu) dQ_0(\mu).$$

Let δ_μ be the degenerate distribution which places probability 1 on the mass point μ. We will never use the integer n to refer to a mass point of any distribution, so we hope there is no confusion between a distribution δ_μ and a sequence value δ_n in their various contexts.

In our theorem we suppose that Q_0 is δ_0. This largely simplifies the form of the density and tells us that $X_1, \dots, X_n$ are just standard normal random variables.

Since we are interested in a property concerning the NPMLE of $Q_0 = \delta_0$, we make the following definitions. Let G denote the set of degenerate distributions on $\mathbb{R}$

$$G = \{\delta_\mu \text{ on } \mathbb{R} : \mu \in \mathbb{R}\},$$

and

$$\widehat{Q}_1 = \operatorname*{argmax}_{\delta \in G} \sum_{i=1}^{n} \log \int \phi(X_i - t) d\delta(t).$$

For any distribution $\delta_\mu \in G$, or equivalently for any choice of mass point $\mu \in \mathbb{R}$, $\int \phi(X_i - t) d\delta_\mu(t) = \phi(X_i - \mu)$. So $\widehat{Q}_1$ is simply the degenerate distribution $\delta_{\widehat{\mu}}$, where $\widehat{\mu}$ maximises the log likelihood

$$\sum_{i=1}^{n} \log \phi(X_i - \mu).$$

Hence $\widehat{Q}_1 = \delta_{\bar{X}}$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

We remark here that while $\widehat{Q}_1$ is the maximum likelihood estimate of Q_0 over the restrictive class of distributions G, it is not necessarily the NPMLE $\widehat{Q}$ of Q_0, as defined by Lindsay's Theorem 2.1.1. However if $\widehat{Q}_1$ satisfies

$$\sup_{\theta \in \mathbb{R}} D_{\widehat{Q}_1}(\theta) = 0,$$

then by Theorem 2.1.1 $\widehat{Q}_1$ is the NPMLE $\widehat{Q}$. If the NPMLE $\widehat{Q}$ is $\widehat{Q}_1$, then it must have $K = 1$ components since $\widehat{Q}_1$ is degenerate. Thus

$$P(K=1) = P(\widehat{Q}_1 = \widehat{Q}) = P\left(\sup_{\theta\in\mathbb{R}} D_{\widehat{Q}_1}(\theta) = 0\right). \tag{2.5}$$

From (2.2) we have

$$D_{\widehat{Q}_1}(\theta) \;=\; \sum_{i=1}^{n}\left\{\frac{\phi(X_i-\theta)}{\int \phi(X_i-\mu)d\widehat{Q}_1(\mu)} - 1\right\},$$

and since $\widehat{Q}_1$ puts probability 1 on $\bar{X}$, we can simplify $D_{\widehat{Q}_1}(\theta)$ to

$$\begin{aligned} D_{\widehat{Q}_1}(\theta) \;&=\; \sum_{i=1}^{n}\left\{\frac{\phi(X_i-\theta)}{\phi(X_i-\bar{X})} - 1\right\} \\ &=\; \sum_{i=1}^{n}\left\{e^{-(\theta^2-\bar{X}^2)/2+(\theta-\bar{X})X_i} - 1\right\}. \end{aligned} \tag{2.6}$$

Since $D_{\widehat{Q}_1}(\bar{X}) = \sum_{i=1}^{n}\{1-1\} = 0$, $\sup_{\theta\in\mathbb{R}} D_{\widehat{Q}_1}(\theta) \geq 0$, and thus

$$P\left(\sup_{\theta\in\mathbb{R}} D_{\widehat{Q}_1}(\theta) = 0\right) = 1 - P\left(\sup_{\theta\in\mathbb{R}} D_{\widehat{Q}_1}(\theta) > 0\right),$$

so to prove our result it is sufficient to show

$$P\left(\sup_{\theta\in\mathbb{R}} D_{\widehat{Q}_1}(\theta) > 0\right) \to 1 \text{ as } n \to \infty.$$

Let $\{D_n(\theta,\bar{X})\}_{\theta\in\mathbb{R}}$ be the stochastic process given by $D_n(\theta,\bar{X}) = \frac{1}{\sqrt{n}} D_{\widehat{Q}_1}(\theta)$, then

$$P\left(\sup_{\theta\in\mathbb{R}} D_{\widehat{Q}_1}(\theta) > 0\right) = P\left(\sup_{\theta\in\mathbb{R}} D_n(\theta,\bar{X}) > 0\right).$$

The following figures provide some typical realisations of what $D_n(\theta,\bar{X})$ looks like in practice. The R code used to generate these examples is described in more detail in Section 2.5.

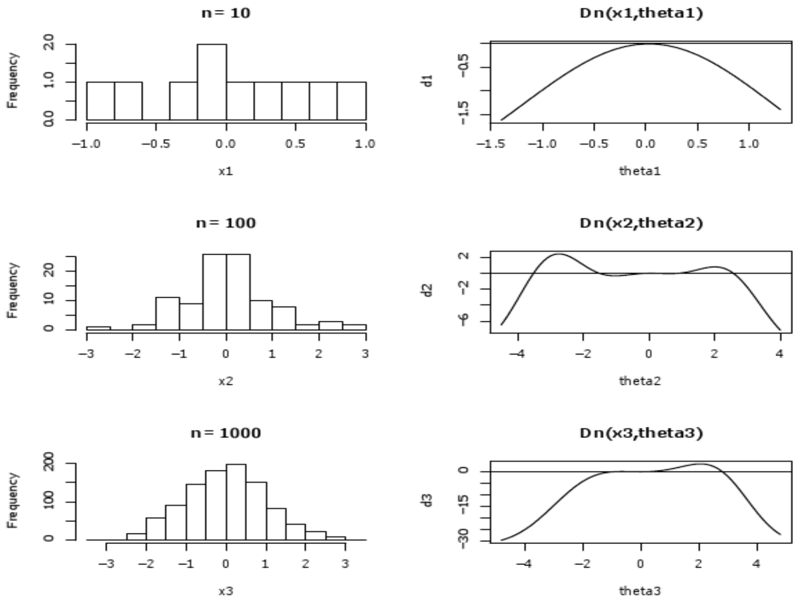


Figure 2.2: Samples to left, $D_n(\theta, \bar{X})$ to right

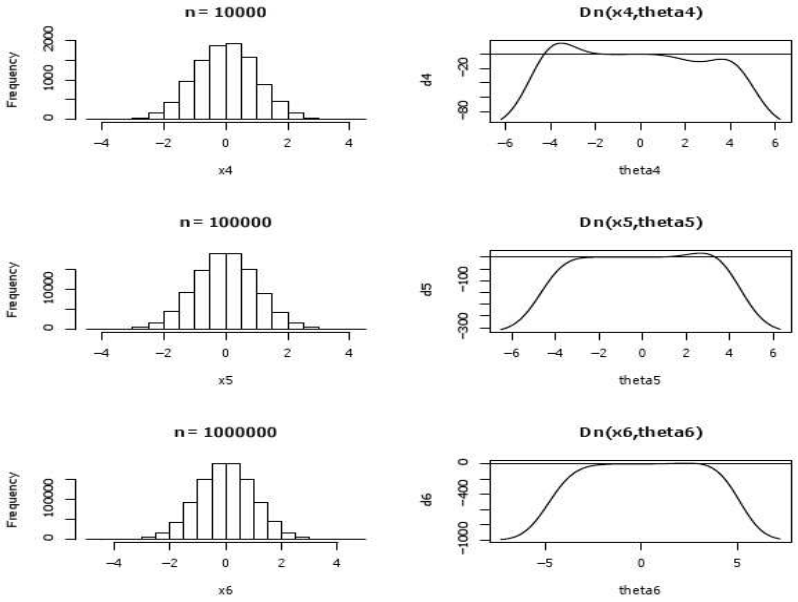


Figure 2.3: Samples to left, $D_n(\theta, \bar{X})$ to right

Let $m = \min_i X_i$ and $M = \max_i X_i$. Note that $D_{\widehat{Q}_1}(m) > D_{\widehat{Q}_1}(\theta)$ for $\theta < m$, because $|X_i - \theta| > |X_i - m|$ for all i in that range. Similarly note that $D_{\widehat{Q}_1}(M) > D_{\widehat{Q}_1}(\theta)$ for $\theta > M$, and so the supremum of $D_n(\theta, \bar{X})$ lies within the range $[m, M]$. This can also be seen graphically in Figures 2.2 and 2.3. Within this proof we will tend to look at $D_n(\theta, \bar{X})$ over intervals for θ where $|\theta| \leq \sqrt{2\log n}$, since $\max X_i$ behaves like $\sqrt{2\log n}$ (see Leadbetter and Rootzén (1988)).

The next part of our proof aims to approximate $P(\sup_{\theta\in\mathbb{R}} D_n(\theta, \bar{X}) > 0)$ by something for which it is easier to obtain the desired type of bounds.

Consider the Taylor expansion of $D_n(\theta, \bar{X})$ about $\bar{X} = 0$

$$D_n(\theta, \bar{X}) \;=\; D_n(\theta, 0) + \bar{X}\,\frac{\partial D_n}{\partial \eta}(\theta, \eta)\bigg|_{\eta=0} + \frac{\bar{X}^2}{2}\,\frac{\partial^2 D_n}{\partial \eta^2}(\theta, \eta)\bigg|_{\eta=\alpha\bar{X}}, \quad (2.7)$$

for some $\alpha \in [0, 1]$. We can write (2.7) as

$$D_n(\theta, \bar{X}) \;=\; D_n(\theta, 0) - \sqrt{n}\bar{X}A_n(\theta) + \frac{n}{2}\bar{X}^2\frac{1}{\sqrt{n}}B_n(\theta, \alpha\bar{X}), \quad (2.8)$$

where $A_n(\theta)$ and $B_n(\theta, \alpha\bar{X})$ are given by the random sums

$$\begin{aligned} A_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} X_i e^{\theta X_i - \frac{\theta^2}{2}}, \\ B_n(\theta, \alpha\bar{X}) &= \frac{1}{n}\sum_{i=1}^{n}\left((\alpha\bar{X} - X_i)^2 + 1\right) e^{(\theta - \alpha\bar{X})X_i - \frac{\theta^2 - (\alpha\bar{X})^2}{2}}. \end{aligned}$$

The details where we show (2.7) can be written as (2.8) are included in Section 2.4. Note that $\mathbb{E}(A_n(\theta)) = \theta$. From (2.8) we can rewrite $D_n(\theta, \bar{X})$ as

$$D_n(\theta, \bar{X}) = D_n(\theta, 0) - \sqrt{n}\bar{X}\theta + \sqrt{n}\bar{X}\left(\theta - A_n(\theta)\right) + \frac{n}{2}\bar{X}^2\frac{1}{\sqrt{n}}B_n(\theta, \alpha\bar{X}). \quad (2.9)$$

We wish to think of $D_n(\theta, 0) - \sqrt{n}\bar{X}\theta$ as the main part of $D_n(\theta, \bar{X})$, and we wish to think of the terms $\sqrt{n}\bar{X}(\theta - A_n(\theta))$ and $\frac{n}{2}\bar{X}^2\frac{1}{\sqrt{n}}B_n(\theta, \alpha\bar{X})$ as negligible remainder terms.

We now state a lemma about the last term in (2.9). The proof is given in Section 2.4. It is insignificant compared to $D_n(\theta, 0) - \sqrt{n}\bar{X}\theta$ in the following sense.

Lemma 2.3.1. *Let* $B_n(\theta, \alpha\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}((\alpha\bar{X} - X_i)^2 + 1)e^{-\frac{1}{2}(\theta^2 - (\alpha\bar{X})^2) + (\theta - \alpha\bar{X})X_i}$. *Let* $C \geq \sqrt{2}, 0 < \epsilon C^2 < \frac{1}{2}$, $R_n \to \infty$ *as* $n \to \infty$ *and* $R_n = o(\sqrt{\frac{n}{\log n}})$. *Let*

$\gamma_n = \frac{R_n^2}{n} + 2\frac{R_n}{\sqrt{n}}C\sqrt{\log n}$. *Suppose* $c_n > 0$ *such that*

$$n^{\epsilon C^2 - \frac{1}{2}} e^{2\gamma_n} = o(c_n).$$

Then for any $|\theta| \leq C\sqrt{\log n}$,

$$P\left(\frac{1}{\sqrt{n}} B_n(\theta, \alpha\bar{X}) > c_n\right) \to 0, \text{ as } n \to \infty.$$

Actually even if $R_n = \sqrt{\frac{n}{\log n}}$, this probability can approach 0 with an appropriate choice of c_n since γ_n approaches a constant in this case.

This lemma gives us quite a bit of freedom when it comes to the choice of c_n. It will turn out that choosing some appropriate $c_n = o(1)$ would do for the purposes of this proof. However choosing certain sequences c_n where $c_n \to \infty$ seems useful for extensions of our result, so we do not wish to think of the c_n in Lemma 2.3.1 as necessarily approaching 0.

It would be nice to provide a similar result about the other remainder-like term in (2.9), however we show a weaker result which suggests to us that a sort of standardisation of our process would be useful for the purposes of this proof.

Lemma 2.3.2. *Let* $A_n(\theta) = \frac{1}{n}\sum_{i=1}^n X_i e^{\theta X_i - \theta^2/2}$. *For* $L > 1$ *and* $z > 0$

$$P(\sup_{-L \leq \theta \leq L} e^{-\theta^2/2} |\theta - A_n(\theta)| \geq z) \leq K\frac{L^4}{nz^2} + 12(1 - \Phi(\frac{\sqrt{n}z}{12L})),$$

where K *is an absolute constant.*

Since $\mathbb{E}(D_n(\theta, 0)) = 0$, $D_n(\theta, 0) - \theta\sqrt{n}\bar{X}$ is a mean 0 process. Section 2.4 contains a calculation showing

$$\mathbb{V}\text{ar}(D_n(\theta, 0) - \sqrt{n}\bar{X}\theta) = e^{\theta^2} - 1 - \theta^2, \tag{2.10}$$

so $\mathbb{V}\text{ar}(D_n(\theta, 0) - \sqrt{n}\bar{X}\theta) \sim e^{\theta^2}$ as $\theta \to \infty$.

We are motivated by (2.10) and Lemma 2.3.2 to consider the stochastic process (2.9) scaled at each point θ by $e^{-\theta^2/2}$. Let

$$\begin{aligned}
\widetilde{D}_n(\theta, \bar{X}) &= e^{-\theta^2/2} D_n(\theta, \bar{X}) \\
&= e^{-\theta^2/2}\left(D_n(\theta, 0) - \sqrt{n}\bar{X}\theta\right) \\
&\quad + \sqrt{n}\bar{X} e^{-\theta^2/2}(\theta - A_n(\theta)) + \frac{n}{2}\bar{X}^2 e^{-\theta^2/2} \frac{1}{\sqrt{n}} B_n(\theta, \alpha\bar{X}),
\end{aligned}$$

and let $R_1(\theta)$ denote the remainder terms

$$R_1(\theta) = \sqrt{n}\bar{X}e^{-\theta^2/2}(\theta - A_n(\theta)) + \frac{n}{2}\bar{X}^2 e^{-\theta^2/2}\frac{1}{\sqrt{n}}B_n(\theta, \alpha\bar{X}). \tag{2.11}$$

If we choose an increasing range to apply Lemma 2.3.2 to, such as $[-L_n, L_n] = [-C\sqrt{\log n}, C\sqrt{\log n}]$, and consider a sequence $c_n > 0$ which works with Lemma 2.3.1 (for example $0 < c_n = \epsilon < 1$ for all n), then Lemma 2.3.2 shows that $P\left(\sup_{|\theta|\le C\sqrt{\log n}} R_1(\theta) \ge c_n\right) \to 0$ as $n \to \infty$. Thus a stochastic process of interest is

$$\widetilde{D}_n(\theta, \bar{X}) = e^{-\theta^2/2}\left(D_n(\theta, 0) - \sqrt{n}\bar{X}\theta\right) + R_1(\theta). \tag{2.12}$$

This scaling of $D_n(\theta, \bar{X})$ by a positive value $e^{-\theta^2/2}$ at each $\theta \in \mathbb{R}$ cannot change whether or not any part of it is positive. Therefore our probability of interest can be reexpressed as

$$P\left(\sup_{\theta\in\mathbb{R}} D_n(\theta, \bar{X}) > 0\right) = P\left(\sup_{\theta\in\mathbb{R}} e^{-\theta^2/2} D_n(\theta, \bar{X}) > 0\right) = P\left(\sup_{\theta\in\mathbb{R}} \widetilde{D}_n(\theta, \bar{X}) > 0\right).$$

The following picture gives an example of what $\widetilde{D}_n(\theta, \bar{X})$ looks like in practice. It was obtained from the same samples described in Figures 2.2 and 2.3.

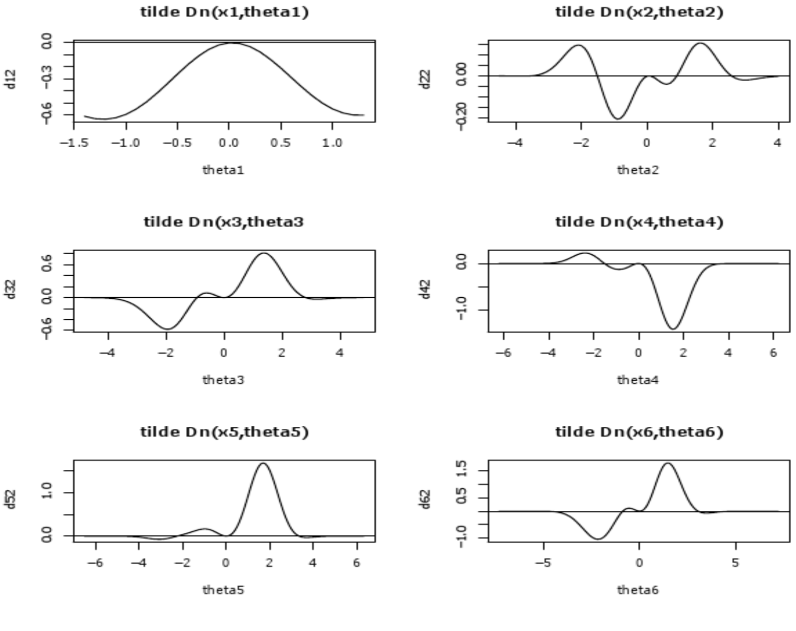


Figure 2.4: $\widetilde{D}_n(\theta, \bar{X})$ for various n

Regarding the first term in (2.12), we now wish to use some results from Csörgő et al. (1986) to approximate $e^{-\theta^2/2}\left(D_n(\theta,0)-\sqrt{n}\bar{X}\theta\right)$ with a stochastic process which is the sum of a Gaussian part and a less significant remainder part. The next section of this proof reexpresses $e^{-\theta^2/2}\left(D_n(\theta,0)-\sqrt{n}\bar{X}\theta\right)$ in terms of an empirical process which has the same distribution.

Let $U_1,\ldots,U_n$ be iid uniform $U(0,1)$, and let $\mathbb{F}_n(u)$ denote their empirical distribution function given by

$$\mathbb{F}_n(u)=\frac{1}{n}\sum_{i=1}^{n}1_{(U_i\le u)},$$

and let $\alpha_n(u)$ be the uniform empirical process defined in Csörgő et al. (1986) by

$$\alpha_n(u)=\sqrt{n}\left(\mathbb{F}_n(u)-u\right).$$

In Section 2.4 we show

$$e^{-\theta^2/2}\left(D_n(\theta,0)-\sqrt{n}\bar{X}\theta\right)$$

has the same distribution as

$$\int_0^1 \ell_\theta(u)d\alpha_n(u), \tag{2.13}$$

where $\ell_\theta(u)$ is the deterministic function on $(0,1)$ given by

$$\ell_\theta(u) \;=\; e^{-\theta^2/2}\left(\frac{\phi(\Phi^{-1}(u)-\theta)}{\phi(\Phi^{-1}(u))}-1-\theta\Phi^{-1}(u)\right). \tag{2.14}$$

We use the above along with (2.12) to arrive at

$$P\left(\sup_{\theta\in\mathbb{R}}\widetilde{D}_n(\theta,\bar{X})>0\right)=P\left(\sup_{\theta\in\mathbb{R}}S_n(\theta)>0\right),$$

where

$$S_n(\theta)=\int_0^1 \ell_\theta(u)d\alpha_n(u)+R_1(\theta). \tag{2.15}$$

Now that our probability of interest is described in terms of the empirical processes $\int_0^1\ell_\theta(u)d\alpha_n(u)$, we wish to use 2.2.4, to approximate $\int_0^1\ell_\theta(u)d\alpha_n(u)$ with a Gaussian process.

We now restrict our attention to θ in the interval $[-C\sqrt{\log n},C\sqrt{\log n}]$. Let $\mathcal{L}_n$ be the sequence of classes of functions on $(0,1)$ given by

$$\mathcal{L}_n=\{\ell_\theta:-C\sqrt{\log n}\le\theta\le C\sqrt{\log n}\}, \tag{2.16}$$

where each ℓ_θ is given by (2.14). We show that each $\mathcal{L}_n$ is $\mathcal{L}^*$-decomposable in Section 2.4. Thus by Theorem 2.2.4 of Csörgő et al. (1986) we have

$$E_n = \sup_{\ell_\theta \in \mathcal{L}_n} \left| \int_0^1 \ell_\theta(u) d\alpha_n(u) - \int_{1/n}^{1-1/n} \ell_\theta(u) dB_n(u) \right| = N_n(\delta_n) O_p(1),$$

where $\delta_n = \frac{\log n}{\sqrt{n}}$, and where $N_n(\delta_n)$ is given by Definition 2.2.2. We have the following lemma about $N_n(\delta_n)$, which is proven is Section 2.4.

Lemma 2.3.3. *Let $\mathcal{L}_n$ be given by (2.16), and $N_n(\delta_n)$ be as in Definition 2.2.2. With $\delta_n = \frac{\log n}{\sqrt{n}}$, we have for $1 - \delta_n \geq \Phi(e^{\frac{1}{2}})$,*

$$N_n(\delta_n) = O(x_n^{-\frac{1}{2}})$$

where $x_n = \Phi^{-1}(1 - \delta_n)$.

We write (2.15) in terms of a nicer Gaussian process and a couple of remainder terms by splitting $\int_0^1 \ell_\theta(u) d\alpha_n(u)$ as follows.

$$\begin{aligned} \int_0^1 \ell_\theta(u) d\alpha_n(u) &= \int_0^1 \ell_\theta(u) d\alpha_n(u) - \int_{1/n}^{1-1/n} \ell_\theta(u) dB_n(u) + \int_{1/n}^{1-1/n} \ell_\theta(u) dB_n(u) \\ &= \int_{1/n}^{1-1/n} \ell_\theta(u) dB_n(u) + R_2(\theta), \end{aligned}$$

where

$$R_2(\theta) = \int_0^1 \ell_\theta(u) d\alpha_n(u) - \int_{1/n}^{1-1/n} \ell_\theta(u) dB_n(u). \tag{2.17}$$

Since $\sup_{\ell_\theta \in \mathcal{L}_n} |R_2(\theta)| = E_n = N_n(\delta_n) O_p(1)$, we know that the approximation by Csörgő et al. (1986) is useful for looking at $S_n(\theta)$ over the interval $[-C\sqrt{\log n}, C\sqrt{\log n}]$. We can also write $\int_{1/n}^{1-1/n} \ell_\theta(u) dB_n(u)$ as

$$\begin{aligned} \int_{1/n}^{1-1/n} \ell_\theta(u) dB_n(u) &= \int_{1/n}^{1-1/n} \ell_\theta(u) dB_n(u) - \int_0^1 \ell_\theta(u) dB_n(u) + \int_0^1 \ell_\theta(u) dB_n(u) \\ &= \int_0^1 \ell_\theta(u) dB_n(u) + R_3(\theta), \end{aligned}$$

where

$$R_3(\theta) = \int_{1/n}^{1-1/n} \ell_\theta(u) dB_n(u) - \int_0^1 \ell_\theta(u) dB_n(u). \tag{2.18}$$

We show that our situation satisfies the conditions of Corollary 3.4 of Csörgő et al. (1986) in Section 2.4. Thus we apply Csörgő et al. (1986)'s corollary to arrive at $\sup_{\ell_\theta \in \mathcal{L}_n} |R_3(\theta)| = o_p(1)$.

Since each $B_n(u)$ is a Brownian bridge, we may consider it as a process $B_n(u) = W_n(u) - uW_n(1)$ where $W_n(u)$ is a Brownian motion. Thus

$$\begin{aligned} \int_0^1 \ell_\theta(u) dB_n(u) &= \int_0^1 \ell_\theta(u) d\left\{W_n(u) - uW_n(1)\right\} \\ &= \int_0^1 \ell_\theta(u) dW_n(u) - W_n(1) \int_0^1 \ell_\theta(u) du \\ &= \int_0^1 \ell_\theta(u) dW_n(u), \end{aligned}$$

since $\int_0^1 \ell_\theta(u)du = \mathbb{E}(\ell_\theta(U_1)) = 0$.

Thus we can write $\int_0^1 \ell_\theta(u)d\alpha_n(u)$ as

$$\int_0^1 \ell_\theta(u) d\alpha_n(u) = \int_0^1 \ell_\theta(u) dW_n(u) + R_2(\theta) + R_3(\theta).$$

From (2.15) we thus have

$$S_n(\theta) = \int_0^1 \ell_\theta(u) dW_n(u) + R_1(\theta) + R_2(\theta) + R_3(\theta). \tag{2.19}$$

Since $W_n(u)$ is a Brownian motion, and each ℓ_θ is a deterministic function, the process $\int_0^1 \ell_\theta(u)dW_n(u)$ is a Gaussian process. Let

$$\begin{aligned} G_n(\theta) &= \int_0^1 \ell_\theta(u) dW_n(u) \text{ and} \\ R_n(\theta) &= R_1(\theta) + R_2(\theta) + R_3(\theta). \end{aligned}$$

From (2.19) we can express $S_n(\theta)$ as the sum of the Gaussian process $G_n(\theta)$ and the remainder process $R_n(\theta)$

$$S_n(\theta) = G_n(\theta) + R_n(\theta).$$

The next idea in this proof makes use of the fact that we only wish to show $P\left(\sup_{\theta \in \mathbb{R}} S_n(\theta) > 0\right)$ tends to 1 as $n \to \infty$. If we choose any grid of values $\Theta_n = \{\theta_1, \ldots, \theta_{a_n}\} \subset \mathbb{R}$, we can obtain the bound

$$P\left(\sup_{\theta \in \mathbb{R}} S_n(\theta) > 0\right) \geq P\left(\max_{\theta \in \Theta_n} S_n(\theta) > 0\right). \tag{2.20}$$

If we can find an appropriate grid of values Θ_n such that the right hand side of (2.20) tends to 1 as $n \to \infty$ then we can prove our result. This allows us to concentrate on the easier problem of examining when the maximum of countably many rv $\{S_n(\theta_i)\}_{\theta_i \in \Theta_n}$ is positive. Before we choose a grid $\Theta_n \subset \mathbb{R}$, here is a lemma which we will use to provide a bound for $P\left(\max_{\theta \in \Theta_n} S_n(\theta) > 0\right)$.

Lemma 2.3.4. *Consider any stochastic process $S_n(\theta)$ over $\Theta_n = \{\theta_1, \ldots, \theta_{a_n}\} \subset \mathbb{R}$ which is the sum of two others*

$$S_n(\theta) = G_n(\theta) + R_n(\theta), \ \text{over } \theta \in \Theta_n.$$

Suppose for some $c_n > 0$

$$P\left(\max_{\theta \in \Theta_n} |R_n(\theta)| \geq c_n\right) \leq \epsilon_n^{(1)}, \tag{2.21}$$

$$P\left(\max_{\theta \in \Theta_n} G_n(\theta) \leq c_n\right) \leq \epsilon_n^{(2)}, \tag{2.22}$$

then

$$P\left(\max_{\theta \in \Theta_n} S_n(\theta) > 0\right) \geq 1 - (\epsilon_n^{(1)} + \epsilon_n^{(2)}).$$

We prove Lemma 2.3.4 in Section 2.4. We wish to choose an appropriate Θ_n such that conditions (2.52) and (2.53) are satisfied, so that Lemma 2.3.4 gives us

$$P\left(\sup_{\theta \in \mathbb{R}} S_n(\theta) > 0\right) \geq 1 - (\epsilon_n^{(1)} + \epsilon_n^{(2)}),$$

for some $\epsilon_n^{(1)}$ and $\epsilon_n^{(2)}$ which will tend to 0 as $n \to \infty$. The remainder of this proof concerns choosing an appropriate Θ_n and describing the terms $\epsilon_n^{(1)}$ and $\epsilon_n^{(2)}$.

We next show the Gaussian part of our process (2.3) satisfies Condition 2.53 of Lemma 2.3.4. To do this we will need the following two lemmas, which are each proven in Section 2.4.

The following lemma provides a bound for $P\left(\max_{\theta \in \Theta_n} G_n(\theta) \leq c_n\right)$ when it is given $G_n(\theta)$, Θ_n and $c_n > 0$. The bound will depend on c_n, the number of elements of Θ_n, the covariance structure of $G_n(\theta)$ and the chosen locations $\theta \in \Theta_n$. The lemma after this one will concern the choice of Θ_n and c_n such that the bound given here is useful.

Lemma 2.3.5. *Let $G_n(\theta) = \int_0^1 \ell_\theta(u) dW_n(u)$, where $\ell_\theta(u)$ is given by (2.14) and $W_n(u)$ is a Brownian motion. Let $\Theta_n = \{\theta_1, \ldots, \theta_{a_n}\} \subset \mathbb{R}$. Let $c_n > 0$*

and for $\theta \in \Theta_n$ let

$$\begin{aligned}
\sigma^2(\theta) &= \mathbb{V}ar(G_n(\theta)), \\
\Lambda^1_{ij} &= \frac{Cov(G_n(\theta_i), G_n(\theta_j))}{\sigma(\theta_i)\sigma(\theta_j)}, \\
u_n(\theta) &= \frac{c_n}{\sigma(\theta)}, \\
\rho_n &= \max_{i \neq j} |\Lambda^1_{ij}|.
\end{aligned}$$

If $\rho_n < 1$ then

$$P\left(\max_{\theta \in \Theta_n} G_n(\theta) \leq c_n\right) \leq \epsilon_n^{(2)},$$

where

$$\epsilon_n^{(2)} = \prod_{i=1}^{a_n} \Phi(u_n(\theta_i)) + \frac{a_n(a_n-1)}{4\pi}\rho_n(1-\rho_n^2)^{-\frac{1}{2}}.$$

This next lemma constructs conditions for choosing Θ_n and $c_n > 0$ such that Lemma 2.3.5 returns a bound $\epsilon_n^{(2)}$ which tends to 0 as $n \to \infty$.

Lemma 2.3.6. *Let $G_n(\theta)$, $\sigma(\theta)^2$, Λ^1_{ij}, $u_n(\theta)$, ρ_n and $\epsilon_n^{(2)}$ be defined as in Lemma 2.3.5, and let $c_n > 0$.*

Let $0 < t_n \leq T_n$, and divide the interval $[t_n, T_n]$ into $a_n/2$ equally spaced points, and divide the interval $[-T_n, -t_n]$ into $a_n/2$ points with the same spacings. Let $\theta_1, \ldots, \theta_{a_n}$ be the names of these a_n points (starting from left to right), and let the spacing between each of these points (in the intervals excluding 0) be $\Delta_n = \theta_j - \theta_{j-1}$.

Suppose we have the following limits as $n \to \infty$:

- *$c_n \to c$, where c is finite*
- *$\Phi(c_n)^{a_n} \to 0$*
- *$t_n \to \infty$*
- *$a_n^2 e^{-\Delta_n^2/2} \to 0$.*

Then for these $\theta_1, \ldots, \theta_{a_n}$, we have $\epsilon_n^{(2)} \to 0$ as $n \to \infty$.

We choose $T_n = C\sqrt{\log n}$, where $C \geq \sqrt{2}$ because we wish to keep in mind Lemma 2.3.1. We choose $t_n \to \infty$ much slower than T_n and a_n roughly equal to $\sqrt{\log\sqrt{\log n}}$ so that $a_n^2 e^{-\Delta_n^2/2}$ looks roughly like

$$e^{\frac{-C^2 \log n + \left(\log\log\sqrt{\log n}\right)\log\sqrt{\log n}}{2\log\sqrt{\log n}}}.$$

Choosing any $c_n \to c < \infty$ will provide us with an $\epsilon_n^{(2)} \to 0$ as $n \to \infty$.

Our remainder terms $R_n(\theta)$ satisfy Condition 2.52 with an $\epsilon_n^{(1)}$ which tends to 0 as $n \to \infty$. We have already shown that $R_1(\theta) \to 0$ as $n \to \infty$, and $R_2(\theta)$ and $R_3(\theta)$ have been uniformly bounded over $\theta \in [-C\sqrt{\log n}, C\sqrt{\log n}]$ using the theorem and corollary from Csörgő et al. (1986) by terms which tend to 0 as $n \to \infty$. Thus we have proven Theorem 2.1.2. □

2.4 Details and calculations

This section contains the details and calculations which were omitted in Section 2.3. These details are listed in the order they were mentioned in the proof of Theorem 2.1.2.

2.4.1 Calculation regarding (2.7)

We first show (2.7) can be rewritten as (2.8).

Proof. Recall from (2.6) that

$$D_n(\theta,\eta) \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(e^{-\frac{1}{2}(\theta^2-\eta^2)}e^{(\theta-\eta)X_i} - 1\right).$$

The first and second derivatives with respect to η, and at $\eta = 0, \alpha\bar{X}$ respectively are thus

$$\begin{aligned}
\frac{\partial}{\partial\eta}D_n(\theta,\eta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\eta - X_i)e^{-\frac{1}{2}(\theta^2-\eta^2)+(\theta-\eta)X_i}, \\
\left.\frac{\partial}{\partial\eta}D_n(\theta,\eta)\right|_{\eta=0} &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i e^{-\frac{1}{2}\theta^2+\theta X_i}. \\
\frac{\partial^2}{\partial\eta^2}D_n(\theta,\eta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}((\eta - X_i)^2+1)e^{-\frac{1}{2}(\theta^2-\eta^2)+(\theta-\eta)X_i}, \\
\left.\frac{\partial^2}{\partial\eta^2}D_n(\theta,\eta)\right|_{\eta=\alpha\bar{X}} &= n\frac{1}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n}((\alpha\bar{X} - X_i)^2+1)e^{-\frac{1}{2}(\theta^2-(\alpha\bar{X})^2)+(\theta-\alpha\bar{X})X_i}.
\end{aligned}$$

□

2.4.2 Proof of Lemma 2.3.1.

After we rewrote (2.7) as (2.8), we stated Lemma 2.3.1, which we used to look at the term $\frac{n}{2}\bar{X}^2\frac{1}{\sqrt{n}}B_n(\theta,\alpha\bar{X})$ more closely by considering the behaviour of $\frac{1}{\sqrt{n}}B_n(\theta,\alpha\bar{X})$ when n was large. Note that this proof provides a uniform bound on the probability in question, independent of θ. This is done to remove the issue of the possibly different rates of convergence at different points of the overall stochastic process, and is achieved since we restrict attention to the range $(-C\sqrt{\log n}, C\sqrt{\log n})$.

Proof of Lemma 2.3.1. Recall that $\alpha \in [0,1]$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The main idea of this proof is that we expect $\alpha\bar{X}$ to be roughly $\mathbb{E}(X_1) = 0$ for large enough n, so we first aim to write $B_n(\theta,\alpha\bar{X})$ as a function of $B_n(\theta,0)$ and a remainder term, so that we may use the behaviour of $\alpha\bar{X}$ to examine how $\frac{1}{\sqrt{n}}B_n(\theta,\alpha\bar{X})$ behaves.

$$\begin{aligned}
B_n(\theta,\alpha\bar{X}) &= \frac{1}{n}\sum_{i=1}^{n}((\alpha\bar{X}-X_i)^2+1)e^{(\theta-\alpha\bar{X})X_i-\frac{1}{2}(\theta^2-(\alpha\bar{X})^2)} \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i^2+1-2\alpha\bar{X}X_i+(\alpha\bar{X})^2)e^{\theta X_i-\frac{1}{2}\theta^2-\alpha\bar{X}X_i+\frac{1}{2}(\alpha\bar{X})^2} \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i^2+1)e^{\theta X_i-\frac{1}{2}\theta^2-\alpha\bar{X}X_i+\frac{1}{2}(\alpha\bar{X})^2} \\
&\quad +\frac{1}{n}\sum_{i=1}^{n}((\alpha\bar{X})^2-2\alpha\bar{X}X_i)e^{\theta X_i-\frac{1}{2}\theta^2-\alpha\bar{X}X_i+\frac{1}{2}(\alpha\bar{X})^2}. \qquad (2.23)
\end{aligned}$$

In (2.23), both summands are functions of the X_i, θ and the common term $(\alpha\bar{X})^2 - 2\alpha\bar{X}X_i$, so we next note that

$$(\alpha\bar{X})^2 - 2\alpha\bar{X}X_i \leq \bar{X}^2 + 2|\bar{X}||X_i|. \qquad (2.24)$$

We next consider two events α_n, β_n where we could expect $P(\alpha_n) \to 1, P(\beta_n) \to 1$ as $n \to \infty$. For any sequence R_n and for any constant $C > 0$, define

$$\begin{aligned}
\alpha_n &= \alpha_n(R_n) = \{\omega \in \Omega : \sqrt{n}|\bar{X}(\omega)| \leq R_n\}, \\
\beta_n &= \beta_n(C) = \{\omega \in \Omega : \left|\max_{1\leq i\leq n} X_i(\omega)\right| \leq C\sqrt{\log n}\},
\end{aligned}$$

and let us choose R_n and C such that $P(\alpha_n) \to 1, P(\beta_n) \to 1$ as $n \to \infty$. For example, any $R_n \to \infty$ will give $P(\alpha_n) \to 1$ as $n \to \infty$. Since X_1 is standard normal, the range of the data $\max_i X_i$ behaves like $\sqrt{2\log n}$ (see (Leadbetter and Rootzén, 1988)), so choosing $C = \sqrt{2}$ would provide us with a β_n such that $P(\beta_n) \to 1$ as $n \to \infty$. For the remainder of this proof we will assume we are refering to α_n, β_n where R_n and C are chosen so that $P(\alpha_n) \to 1, P(\beta_n) \to 1$ as $n \to \infty$.

On such an event $\alpha_n \cap \beta_n$, (2.24) gives us

$$(\alpha\bar{X})^2 - 2\alpha\bar{X}X_i \le \gamma_n, \tag{2.25}$$

where

$$\gamma_n = \frac{R_n^2}{n} + 2\frac{R_n}{\sqrt{n}}C\sqrt{\log n}.$$

From (2.23) we have that on $\alpha_n \cap \beta_n$

$$B_n(\theta, \alpha\bar{X}) \le e^{2\gamma_n}\frac{1}{n}\sum_{i=1}^{n}(X_i^2+1)e^{\theta X_i - \frac{1}{2}\theta^2} + \gamma_n e^{2\gamma_n}\frac{1}{n}\sum_{i=1}^{n}e^{\theta X_i - \frac{1}{2}\theta^2}. \tag{2.26}$$

From (2.26) we can bound $B_n(\theta, \alpha\bar{X})$ on $\alpha_n \cap \beta_n$ with

$$B_n(\theta, \alpha\bar{X}) \le 2\max\left\{e^{2\gamma_n}\frac{1}{n}\sum_{i=1}^{n}(X_i^2+1)e^{\theta X_i - \frac{1}{2}\theta^2}, \gamma_n e^{2\gamma_n}\frac{1}{n}\sum_{i=1}^{n}e^{\theta X_i - \frac{1}{2}\theta^2}\right\}. \tag{2.27}$$

We next deal with the random exponential terms written in the right hand side of (2.27). Pick any $\epsilon > 0$, and rewrite $e^{\theta X_i - \frac{1}{2}\theta^2}$ as

$$e^{\theta X_i - \frac{1}{2}\theta^2} = e^{\epsilon\theta^2}e^{\theta X_i - \frac{1}{2}\theta^2 - \epsilon\theta^2}.$$

Let $\lambda(\theta) = \theta x - \frac{1}{2}\theta^2 - \epsilon\theta^2$. Then λ attains a maximum at $\lambda\left(\frac{x}{1+2\epsilon}\right)$, which can be simplified to

$$\begin{aligned}
\lambda\left(\frac{x}{1+2\epsilon}\right) &= \frac{x^2}{1+2\epsilon} - \frac{1}{2}\frac{x^2}{(1+2\epsilon)^2} - \epsilon\frac{x^2}{(1+2\epsilon)^2} \\
&= \frac{2(1+2\epsilon)x^2 - x^2 - 2\epsilon x^2}{2(1+2\epsilon)^2} \\
&= \frac{x^2\left(2(1+2\epsilon) - 1 - 2\epsilon\right)}{2(1+2\epsilon)^2} \\
&= \frac{x^2}{2(1+2\epsilon)^2}\left(2 + 4\epsilon - 1 - 2\epsilon\right) \\
&= \frac{x^2(1+2\epsilon)}{2(1+2\epsilon)^2} \\
&= \frac{x^2}{2(1+2\epsilon)}.
\end{aligned}$$

So we can conclude that for all θ,

$$e^{\theta X_i - \frac{1}{2}\theta^2} \leq e^{\epsilon\theta^2} e^{\frac{X_i^2}{2(1+2\epsilon)}}, \text{ for } \epsilon > 0.$$

Hence with any choice of $\epsilon > 0$,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(X_i^2+1)e^{\theta X_i - \frac{1}{2}\theta^2} &\leq e^{\epsilon\theta^2}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^2+1)e^{\frac{X_i^2}{2(1+2\epsilon)}}\right) && (2.28) \\ \frac{1}{n}\sum_{i=1}^{n}e^{\theta X_i - \frac{1}{2}\theta^2} &\leq e^{\epsilon\theta^2}\left(\frac{1}{n}\sum_{i=1}^{n}e^{\frac{X_i^2}{2(1+2\epsilon)}}\right). && (2.29) \end{aligned}$$

Note that

$$\sup_{|\theta|\leq C\sqrt{\log n}} e^{\epsilon\theta^2} = e^{\epsilon(C^2\log n)} = n^{\epsilon C^2}. \qquad (2.30)$$

We use (2.28), (2.29) and (2.30), to conclude for all $\epsilon > 0$, for all $|\theta| \leq C\sqrt{\log n}$,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(X_i^2+1)e^{\theta X_i - \frac{1}{2}\theta^2} &\leq n^{\epsilon C^2}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^2+1)e^{\frac{X_i^2}{2(1+2\epsilon)}}\right) && (2.31) \\ \frac{1}{n}\sum_{i=1}^{n}e^{\theta X_i - \frac{1}{2}\theta^2} &\leq n^{\epsilon C^2}\left(\frac{1}{n}\sum_{i=1}^{n}e^{\frac{X_i^2}{2(1+2\epsilon)}}\right). && (2.32) \end{aligned}$$

Let the random variables in (2.31) and (2.32) be called

$$\begin{aligned} Y_{n,\epsilon} &= \frac{1}{n}\sum_{i=1}^{n}(X_i^2+1)e^{\frac{X_i^2}{2(1+2\epsilon)}}, \\ W_{n,\epsilon} &= \frac{1}{n}\sum_{i=1}^{n}e^{\frac{X_i^2}{2(1+2\epsilon)}}. \end{aligned}$$

Since $\mu_\epsilon = \mathbb{E}\left(Y_{n,\epsilon}\right)$ and $\nu_\epsilon = \mathbb{E}\left(W_{n,\epsilon}\right)$ are both finite, the weak law of large numbers tells us that $Y_{n,\epsilon} \xrightarrow{P} \mu_\epsilon$ and $W_{n,\epsilon} \xrightarrow{P} \nu_\epsilon$, and since convergence in probability implies boundedness in probability, $Y_{n,\epsilon}$ and $W_{n,\epsilon}$ are both $O_p(1)$.

We now consider an arbitrary sequence $c_n > 0$. For $|\theta| \leq C\sqrt{\log n}$ we split $P(\frac{1}{\sqrt{n}}B_n(\theta,\alpha\bar{X}) > c_n)$ into

$$\begin{aligned} P(\frac{1}{\sqrt{n}}B_n(\theta,\alpha\bar{X}) > c_n) &= P\left(\left\{\frac{1}{\sqrt{n}}B_n(\theta,\alpha\bar{X}) > c_n\right\} \cap \{\alpha_n \cap \beta_n\}\right) \\ &\quad + P\left(\left\{\frac{1}{\sqrt{n}}B_n(\theta,\alpha\bar{X}) > c_n\right\} \cap \{\alpha_n \cap \beta_n\}^c\right), \end{aligned} \qquad (2.33)$$

and since (2.33) can be bounded by

$$P(\frac{1}{\sqrt{n}}B_n(\theta,\alpha\bar{X}) > c_n) \leq P\left(\left\{\frac{1}{\sqrt{n}}B_n(\theta,\alpha\bar{X}) > c_n\right\} \cap \{\alpha_n \cap \beta_n\}\right) + P(\alpha_n^c \cup \beta_n^c),$$

we use (2.27), (2.31) and (2.32) to obtain

$$\begin{aligned} P(\frac{1}{\sqrt{n}}B_n(\theta,\alpha\bar{X}) > c_n) &\leq P(e^{2\gamma_n}n^{\epsilon C^2-\frac{1}{2}}Y_{n,\epsilon} > \frac{c_n}{2}) + P(\gamma_n e^{2\gamma_n}n^{\epsilon C^2-\frac{1}{2}}W_{n,\epsilon} > \frac{c_n}{2}) \\ &\quad + P(\alpha_n^c) + P(\beta_n^c). \qquad (2.34) \end{aligned}$$

Recall that $\gamma_n = \frac{R_n^2}{n} + 2\frac{R_n}{\sqrt{n}}C\sqrt{\log n}$. When $R_n = o\left(\sqrt{\frac{n}{\log n}}\right)$ and $R_n \to \infty$, we get $\gamma_n = o(1)$ and $P(\alpha_n^c) = o(1)$. For $C \geq \sqrt{2}$, when we choose $0 < \epsilon < \frac{1}{2C^2}$, we have $P(\beta_n^c) = o(1)$ and $n^{\epsilon C^2 - \frac{1}{2}} = o(1)$. So since $Y_{n,\epsilon}$ and $W_{n,\epsilon}$ are $O_p(1)$, if $n^{\epsilon C^2-\frac{1}{2}}e^{2\gamma_n} = o(c_n)$, then (2.34) $= o(1)$, and we are done.

□

2.4.3 Proof of Lemma 2.3.2.

After we stated Lemma 2.3.1, we stated Lemma 2.3.2, which we used to consider the term $|A_n(\theta) - \theta|$ in (2.9).

Proof of Lemma 2.3.2. The idea of this proof uses the fact that $A_n(\theta)$ looks like the derivative of the empirical moment generating function $M_n(\theta)$ of the observations $X_1, \ldots, X_n$.

We use a technique demonstrated in Bickel and Chernoff (1993), where we express $A_n(\theta)$ as a function of a new process $Y_n(\theta)$, which is what looks to be an almost standardised version of $M_n(\theta)$.

We will then apply the Kolmogorov bound from Revuz and Yor (1991) to prove this lemma.

Recall that $X_1, \ldots, X_n$ iid $N(0,1)$ and

$$A_n(\theta) = \frac{1}{n}\sum_{i=1}^n X_i e^{\theta X_i - \frac{\theta^2}{2}}.$$

The empirical moment generating function of $X_1, \ldots, X_n$, and its derivative are

$$\begin{aligned} M_n(\theta) &= \frac{1}{n}\sum_{i=1}^n e^{\theta X_i} \text{ and} \\ M_n'(\theta) &= \frac{1}{n}\sum_{i=1}^n X_i e^{\theta X_i} = e^{\frac{\theta^2}{2}}A_n(\theta). \end{aligned}$$

Let $Z \sim N(0,1)$, and let Z be independent of $X_1, \ldots, X_n$, and let

$$Y_n(\theta) = e^{-\theta^2}\sqrt{n}\left(M_n(\theta) - e^{\frac{\theta^2}{2}}\right) + e^{-\frac{\theta^2}{2}}Z.$$

The derivative of $Y_n(\theta)$ is

$$Y_n'(\theta) = -2\sqrt{n}\theta e^{-\theta^2}(M_n(\theta) - e^{\theta^2/2}) + \sqrt{n}e^{-\theta^2}(M_n'(\theta) - \theta e^{\theta^2/2}) - \theta e^{-\theta^2/2}Z.$$

Since $Y_n(\theta) = e^{-\theta^2}\sqrt{n}\left(M_n(\theta) - e^{\frac{\theta^2}{2}}\right) + e^{-\frac{\theta^2}{2}}Z$, we can rewrite $M_n(\theta)$ and $M_n'(\theta)$ as

$$\begin{aligned}
M_n(\theta) &= \frac{e^{\theta^2}}{\sqrt{n}}\left(Y_n(\theta) - e^{-\theta^2/2}Z\right) + e^{\theta^2/2} \\
&= \frac{e^{\theta^2}}{\sqrt{n}}Y_n(\theta) + e^{\theta^2/2}\left(1 - \frac{Z}{\sqrt{n}}\right), \\
M_n'(\theta) &= \frac{2\theta e^{\theta^2}}{\sqrt{n}}Y_n(\theta) + \frac{e^{\theta^2}}{\sqrt{n}}Y_n'(\theta) + \theta e^{\theta^2/2}\left(1 - \frac{Z}{\sqrt{n}}\right).
\end{aligned}$$

The process $A_n(\theta) = e^{-\theta^2/2}M_n'(\theta)$ can then be written as

$$A_n(\theta) = \frac{2\theta e^{\theta^2/2}}{\sqrt{n}}Y_n(\theta) + \frac{e^{\theta^2/2}}{\sqrt{n}}Y_n'(\theta) + \theta\left(1 - \frac{Z}{\sqrt{n}}\right),$$

and $e^{-\theta^2/2}(\theta - A_n(\theta))$ can be written as

$$e^{-\theta^2/2}(\theta - A_n(\theta)) = \frac{-2\theta}{\sqrt{n}}Y_n(\theta) - \frac{1}{\sqrt{n}}Y_n'(\theta) + \theta e^{-\theta^2/2}\frac{Z}{\sqrt{n}}.$$

Therefore

$$\begin{aligned}
e^{-\theta^2/2}|\theta - A_n(\theta)| &\leq \frac{2}{\sqrt{n}}|\theta Y_n(\theta)| + \frac{1}{\sqrt{n}}|Y_n'(\theta)| + \frac{e^{-\theta^2/2}}{\sqrt{n}}|\theta Z|, \\
\sup_{-L\leq\theta\leq L} e^{-\theta^2/2}|\theta - A_n(\theta)| &\leq \sup_{-L\leq\theta\leq L}\frac{2}{\sqrt{n}}|\theta Y_n(\theta)| + \sup_{-L\leq\theta\leq L}\frac{1}{\sqrt{n}}|Y_n'(\theta)| \\
&\quad + \sup_{-L\leq\theta\leq L}\frac{e^{-\theta^2/2}}{\sqrt{n}}|\theta Z|. \qquad (2.35)
\end{aligned}$$

The right hand side of (2.35) is less than

$$3\max\left(\sup_{-L\leq\theta\leq L}\frac{2}{\sqrt{n}}|\theta Y_n(\theta)|, \sup_{-L\leq\theta\leq L}\frac{1}{\sqrt{n}}|Y_n'(\theta)|, \sup_{-L\leq\theta\leq L}\frac{e^{-\theta^2/2}}{\sqrt{n}}|\theta Z|\right),$$

so over an interval $[-L, L]$ we can bound

$$P(\sup_{-L\leq\theta\leq L} e^{-\theta^2/2}|\theta - A_n(\theta)| \geq z)$$

with

$$\begin{aligned} & P(\sup_{-L\leq\theta\leq L} e^{-\theta^2/2}|\theta - A_n(\theta)| \geq z) \\ \leq \; & P\left(\sup_{-L\leq\theta\leq L} \frac{2}{\sqrt{n}}|\theta Y_n(\theta)| \geq \frac{z}{3}\right) + P\left(\sup_{-L\leq\theta\leq L} \frac{1}{\sqrt{n}}|Y_n'(\theta)| \geq \frac{z}{3}\right) \\ & + P\left(\sup_{-L\leq\theta\leq L} \frac{e^{-\theta^2/2}}{\sqrt{n}}|\theta Z| \geq \frac{z}{3}\right). \end{aligned} \tag{2.36}$$

Let the terms above concerning $[-L, L]$ be called

$$\begin{aligned} p_1 &= P\left(\sup_{-L\leq\theta\leq L} \frac{2}{\sqrt{n}}|\theta Y_n(\theta)| \geq \frac{z}{3}\right) \\ p_2 &= P\left(\sup_{-L\leq\theta\leq L} \frac{1}{\sqrt{n}}|Y_n'(\theta)| \geq \frac{z}{3}\right) \\ p_3 &= P\left(\sup_{-L\leq\theta\leq L} \frac{e^{-\theta^2/2}}{\sqrt{n}}|\theta Z| \geq \frac{z}{3}\right). \end{aligned}$$

From (2.36), we have

$$P(\sup_{-L\leq\theta\leq L} e^{-\theta^2/2}|\theta - A_n(\theta)| \geq z) \leq p_1 + p_2 + p_3.$$

The Kolmogorov bound works with stochastic processes $Z(t)$ defined over an interval $0 \leq t \leq L$, while the situation we have here for each p_i, $i = 1, 2, 3$ involves an interval $-L \leq \theta \leq L$. For p_1 we could define a new stochastic process $\widetilde{Y}_n(\theta) = (\theta + L)Y_n(\theta + L)$ so that $\widetilde{Y}_n(-L) = 0Y_n(0)$ and $\widetilde{Y}_n(L) = 2LY_n(2L)$. Then $\widetilde{Y}_n(\theta)$ is defined over the interval $[0, 2L]$ and we can apply the Kolmogorov bound. Since the effect would be to replace the $2L$ wide interval $[-L, L]$ with the $2L$ wide interval $[0, 2L]$, we will not bother to define $\widetilde{Y}_n(\theta)$ and simply apply the Kolmogorov bound directly by treating $[-L, L]$ as though it were $[0, 2L]$. The same argument can be made for bounding p_2 and p_3.

We now bound p_1.

$$\begin{aligned} p_1 &= P\left(\sup_{0\le\theta\le 2L}\frac{2}{\sqrt{n}}|\theta Y_n(\theta)| \ge \frac{z}{3}\right) \\ &\le P\left(\sup_{0\le\theta\le 2L}\frac{4L}{\sqrt{n}}|Y_n(\theta)| \ge \frac{z}{3}\right) \\ &= P\left(\sup_{0\le\theta\le 2L}\frac{4L}{\sqrt{n}}|Y_n(\theta)-Y_n(0)+Y_n(0)| \ge \frac{z}{3}\right) \\ &\le P\left(\sup_{0\le\theta\le 2L}\frac{4L}{\sqrt{n}}|Y_n(\theta)-Y_n(0)|+\frac{4L}{\sqrt{n}}|Y_n(0)| \ge \frac{z}{3}\right). \end{aligned}$$

Since

$$\sup_{0\le\theta\le 2L}\frac{4L}{\sqrt{n}}|Y_n(\theta)-Y_n(0)|+\frac{4L}{\sqrt{n}}|Y_n(0)| \le 2\max\left(\sup_{0\le\theta\le 2L}\frac{4L}{\sqrt{n}}|Y_n(\theta)-Y_n(0)|, \frac{4L}{\sqrt{n}}|Y_n(0)|\right),$$

we can bound p_1 using

$$\begin{aligned} p_1 &\le P\left(\sup_{0\le\theta\le 2L}\frac{4L}{\sqrt{n}}|Y_n(\theta)-Y_n(0)| \ge \frac{z}{6}\right) \\ &\quad +P\left(\frac{4L}{\sqrt{n}}|Y_n(0)| \ge \frac{z}{6}\right). \end{aligned}$$

We can write this as

$$p_1 \le P\left(\sup_{0\le\theta\le 2L}|Y_n(\theta)-Y_n(0)| \ge \frac{\sqrt{n}z}{24L}\right)+P\left(|Y_n(0)| \ge \frac{\sqrt{n}z}{24L}\right). \tag{2.37}$$

A calculation is provided at the end of this proof which shows for all $s,t\in\mathbb{R}$,

$$\mathbb{E}|Y_n(s)-Y_n(t)|^2 \le (s-t)^2. \tag{2.38}$$

Thus $Y_n(\theta)$ satisfies the condition for using the Kolmogorov bound, so we can arrive at

$$P(\sup_{0\le t\le 2L}|Y_n(t)-Y_n(0)| \ge z) \le K(4L^2/z^2), \tag{2.39}$$

where K is some absolute constant.

Since $Y_n(0) = Z$, we can use (2.39) and (2.37) to arrive at

$$\begin{aligned} p_1 &\leq K\frac{L^2}{\left(\frac{nz^2}{24^2L^2}\right)} + P(|Z| \geq \frac{\sqrt{n}z}{24L}) \\ &= K'\frac{L^4}{nz^2} + 2(1-\Phi(\frac{\sqrt{n}z}{24L})), \end{aligned}$$

where $K' = 24^2K$ is some absolute constant, and Φ is the distribution function of the standard normal random variable.

Similarly,

$$p_2 \leq P\left(\sup_{0\leq\theta\leq 2L} |Y_n'(\theta) - Y_n'(0)| \geq \frac{\sqrt{n}z}{6}\right) + P\left(|Y_n'(0)| \geq \frac{\sqrt{n}z}{6}\right). \tag{2.40}$$

The calculation at the end of this proof also shows for all $s, t \in \mathbb{R}$

$$\mathbb{E}|Y_n'(s) - Y_n'(t)|^2 \leq 3(s-t)^2. \tag{2.41}$$

Thus we can once again use the Kolmogorov bound (listed as Theorem 2.2.1) to arrive at

$$P(\sup_{0\leq t\leq 2L} |Y_n'(t) - Y_n'(0)| \geq z) \leq 3K(4L^2/z^2), \tag{2.42}$$

where K is some absolute constant.

Since $Y_n'(0) = \sqrt{n}\bar{X} \sim N(0,1)$, we can use (2.42) and (2.40) to arrive at (for some absolute constant K)

$$\begin{aligned} p_2 &\leq 12KL^2\left(\frac{36}{nz^2}\right) + P(|\sqrt{n}\bar{X}| \geq \frac{\sqrt{n}z}{6}) \\ &= 432K\frac{L^2}{nz^2} + 2(1-\Phi(\frac{\sqrt{n}z}{6L})). \end{aligned}$$

Finally we have

$$\begin{aligned} p_3 &= P\left(\sup_{0\leq\theta\leq 2L} \frac{e^{-\theta^2/2}}{\sqrt{n}}|\theta Z| \geq \frac{z}{3}\right) \\ &\leq P\left(\left(\sup_{\theta\in\mathbb{R}} e^{-\theta^2/2}|\theta|\right)|Z| \geq \frac{\sqrt{n}z}{3}\right) \\ &= P\left(e^{-\frac{1}{2}}|Z| \geq \frac{\sqrt{n}z}{3}\right) \\ &= 2(1-\Phi(\frac{e^{\frac{1}{2}}}{3}\sqrt{n}z)). \end{aligned}$$

From (2.36), we can thus obtain for any $L > 0$, $z \in \mathbb{R}$

$$\begin{aligned}
P(\sup_{-L\leq\theta\leq L} e^{-\theta^2/2}|\theta - A_n(\theta)| \geq z) &\leq K_1\frac{L^4}{nz^2} + 2(1-\Phi(\frac{\sqrt{n}z}{24L})) \\
&+K_2\frac{L^2}{nz^2} + 2(1-\Phi(\frac{\sqrt{n}z}{6L})) \\
&+2\left(1-\Phi(\frac{e^{\frac{1}{2}}}{3}\sqrt{n}z)\right),
\end{aligned}$$

where K_1, K_2 are absolute constants. Thus for $L > 1$, we have the bound

$$P(\sup_{-L\leq\theta\leq L} e^{-\theta^2/2}|\theta - A_n(\theta)| \geq z) \leq K\frac{L^4}{nz^2} + 12(1-\Phi(\frac{\sqrt{n}z}{24L})),$$

where K is an absolute constant, so the lemma is proven. □

2.4.4 Calculations for the above proof

Here are the calculations used to show (2.38) and (2.41) which were used in the above proof.

Calculation showing (2.38) and (2.41). Note that

$$\begin{aligned}
\mathbb{E}(M_n(t)) &= \int e^{tx}\phi(x)dx \\
&= \int e^{t^2/2}\phi(x-t)dx \\
&= e^{t^2/2}, \text{ and} \\
\mathbb{E}(M_n'(t)) &= \int xe^{tx}\phi(x)dx \\
&= \int x\phi(x-t)e^{t^2/2}dx \\
&= te^{t^2/2}.
\end{aligned}$$

To calculate $\mathbb{E}(Y_n(s)Y_n(t))$, note that (since $Y_n(s) = e^{-s^2}\sqrt{n}\left(M_n(s) - e^{\frac{s^2}{2}}\right) + e^{-\frac{s^2}{2}}Z$):

$$\begin{aligned}
Y_n(s)Y_n(t) &= \left(e^{-s^2}\sqrt{n}\left(M_n(s) - e^{\frac{s^2}{2}}\right) + e^{-\frac{s^2}{2}}Z\right)\left(e^{-t^2}\sqrt{n}\left(M_n(t) - e^{\frac{t^2}{2}}\right) + e^{-\frac{t^2}{2}}Z\right) \\
&= ne^{-s^2-t^2}(M_n(s) - e^{s^2/2})(M_n(t) - e^{t^2/2}) \\
&\quad +\sqrt{n}e^{-s^2-t^2/2}(M_n(s) - e^{s^2/2})Z \\
&\quad +\sqrt{n}e^{-t^2-s^2/2}(M_n(t) - e^{t^2/2})Z \\
&\quad +e^{-s^2/2-t^2/2}Z^2.
\end{aligned}$$

So,

$$\begin{aligned}\mathbb{E}(Y_n(s)Y_n(t)) &= ne^{-s^2-t^2}\mathrm{Cov}\left(M_n(s), M_n(t)\right)\\ &\quad +\sqrt{n}e^{-s^2-t^2/2}\mathrm{Cov}\left(M_n(s), Z\right)\\ &\quad +\sqrt{n}e^{-t^2-s^2/2}\mathrm{Cov}\left(M_n(t), Z\right)\\ &\quad +e^{-s^2/2-t^2/2}.\end{aligned}$$

Since Z is independent of $X_1, \ldots, X_n$, $\mathrm{Cov}\left(M_n(\theta), Z\right) = 0$ for any θ, so

$$\mathbb{E}(Y_n(s)Y_n(t)) = ne^{-s^2-t^2}\mathrm{Cov}\left(M_n(s), M_n(t)\right) + e^{-s^2/2-t^2/2}.$$

Noting that $\mathbb{E}(e^{\theta X_1}) = e^{\theta^2/2}$, and that

$$\begin{aligned}M_n(s)M_n(t) &= \left(\frac{1}{n}\sum_{i=1}^{n} e^{sX_i}\right)\left(\frac{1}{n}\sum_{i=1}^{n} e^{tX_i}\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{sX_i+tX_j}\\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n} e^{sX_i+tX_i} + \sum_{i\neq j} e^{sX_i+tX_j}\right),\end{aligned}$$

we can arrive at

$$\begin{aligned}\mathbb{E}(M_n(s)M_n(t)) &= \frac{1}{n^2}\left(n\mathbb{E}(e^{(s+t)X_1}) + (n^2-n)\mathbb{E}(e^{sX_1+tX_2})\right)\\ &= \frac{1}{n}\left(e^{(s+t)^2/2} + (n-1)e^{s^2/2+t^2/2}\right),\end{aligned}$$

since X_1 and X_2 are independent.

So to summarise,

$$\begin{aligned}\mathrm{Cov}\left(M_n(s), M_n(t)\right) &= \frac{1}{n}\left(e^{(s+t)^2/2} + (n-1)e^{s^2/2+t^2/2}\right) - e^{s^2/2+t^2/2},\\ \mathbb{E}(Y_n(s)Y_n(t)) &= ne^{-s^2-t^2}\mathrm{Cov}\left(M_n(s), M_n(t)\right) + e^{-s^2/2-t^2/2},\\ &= ne^{-s^2-t^2}\left(\frac{1}{n}\left(e^{(s+t)^2/2} + (n-1)e^{s^2/2+t^2/2}\right)\right)\\ &\quad -ne^{-s^2-t^2}\left(e^{s^2/2+t^2/2}\right) + e^{-s^2/2-t^2/2}\\ &= e^{-s^2-t^2+(s+t)^2/2} + (n-1)e^{-s^2-t^2+s^2/2+t^2/2}\\ &\quad -ne^{-s^2-t^2+s^2/2+t^2/2} + e^{-s^2/2-t^2/2}\\ &= e^{-s^2-t^2+s^2/2+t^2/2+st} + (n-1)e^{-s^2/2-t^2/2} - (n-1)e^{-s^2/2-t^2/2}\\ &= e^{-s^2/2-t^2/2+st}\\ &= e^{-(s-t)^2/2}.\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E}|Y_n(s)-Y_n(t)|^2 &= \mathbb{E}(Y_n(s)^2)+\mathbb{E}(Y_n(t)^2)-2\mathbb{E}(Y_n(s)Y_n(t))\\ &= 1+1-2e^{-(s-t)^2/2}\\ &= 2(1-e^{-(s-t)^2/2}),\end{aligned}$$

and since $1-e^{-x}\le x$, $1-e^{-(s-t)^2/2}\le (s-t)^2/2$. So

$$\mathbb{E}|Y_n(s)-Y_n(t)|^2\le (s-t)^2.$$

For the similar result regarding $\{Y_n'(t)\}_{t\in\mathbb{R}}$, recall

$$Y_n'(t)=-2\sqrt{n}te^{-t^2}(M_n(t)-e^{t^2/2})+\sqrt{n}e^{-t^2}(M_n'(t)-te^{t^2/2})-te^{-t^2/2}Z,$$

and denote $Y_n'(t)=a(t)+b(t)+c(t)$, where

$$\begin{aligned}a(t) &= -2\sqrt{n}te^{-t^2}(M_n(t)-e^{t^2/2}),\\ b(t) &= \sqrt{n}e^{-t^2}(M_n'(t)-te^{t^2/2}),\\ c(t) &= -te^{-t^2/2}Z.\end{aligned}$$

Then we can write $Y_n'(s)Y_n'(t)$ as

$$\begin{aligned}Y_n'(s)Y_n'(t)= \;& a(s)a(t) && +a(s)b(t) && +a(s)c(t)\\ & +b(s)a(t) && +b(s)b(t) && +b(s)c(t)\\ & +c(s)a(t) && +c(s)b(t) && +c(s)c(t).\end{aligned}\tag{2.43}$$

We now calculate $\mathbb{E}(Y_n'(s)Y_n'(t))$ one term at a time.

$$\begin{aligned}a(s)a(t) &= 4nste^{-s^2-t^2}(M_n(s)-e^{s^2/2})(M_n(t)-e^{t^2/2}),\\ \mathbb{E}(a(s)a(t)) &= 4nste^{-s^2-t^2}\operatorname{Cov}(M_n(s),M_n(t)),\\ \operatorname{Cov}(M_n(s),M_n(t)) &= \frac{1}{n}\left(e^{(s+t)^2/2}+(n-1)e^{s^2/2+t^2/2}\right)-e^{s^2/2+t^2/2},\\ \mathbb{E}(a(s)a(t)) &= 4ste^{-s^2-t^2}\left(e^{(s+t)^2/2}-e^{s^2/2+t^2/2}\right)\\ &= 4st\left(e^{-(s-t)^2/2}-e^{-s^2/2-t^2/2}\right).\end{aligned}$$

$$\begin{aligned}
b(t) &= \sqrt{n}e^{-t^2}(M_n'(t) - te^{t^2/2}). \\
a(s)b(t) &= -2nse^{-s^2-t^2}(M_n(s) - e^{s^2/2})(M_n'(t) - te^{t^2/2}). \\
\mathbb{E}(a(s)b(t)) &= -2nse^{-s^2-t^2}\operatorname{Cov}\left(M_n(s), M_n'(t)\right). \\
M_n(s)M_n'(t) &= \left(\frac{1}{n}\sum_{i=1}^{n} e^{sX_i}\right)\left(\frac{1}{n}\sum_{i=1}^{n} X_i e^{tX_i}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_j e^{sX_i+tX_j} \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} X_i e^{(s+t)X_i} + \sum_{i\neq j} X_j e^{sX_i+tX_j}\right). \\
\mathbb{E}(M_n(s)M_n'(t)) &= \frac{1}{n}\left(\mathbb{E}(X_1 e^{(s+t)X_1}) + (n-1)\mathbb{E}(X_2 e^{sX_1+tX_2})\right) \\
&= \frac{1}{n}(s+t)e^{(s+t)^2/2} + \frac{n-1}{n}te^{s^2/2+t^2/2}. \\
\operatorname{Cov}\left(M_n(s), M_n'(t)\right) &= \frac{1}{n}(s+t)e^{(s+t)^2/2} + \frac{n-1}{n}te^{s^2/2+t^2/2} - te^{s^2/2+t^2/2} \\
&= \frac{1}{n}(s+t)e^{(s+t)^2/2} - \frac{1}{n}te^{s^2/2+t^2/2}. \\
\mathbb{E}(a(s)b(t)) &= -2se^{-s^2-t^2}\left((s+t)e^{(s+t)^2/2} - te^{s^2/2+t^2/2}\right) \\
&= -2s(s+t)e^{-(s-t)^2/2} + 2ste^{-s^2/2-t^2/2}.
\end{aligned}$$

$$\begin{aligned}
b(t) &= \sqrt{n}e^{-t^2}(M_n'(t) - te^{t^2/2}) \\
b(s)b(t) &= ne^{-s^2-t^2}(M_n'(s) - se^{s^2/2})(M_n'(t) - te^{t^2/2}) \\
\mathbb{E}(b(s)b(t)) &= ne^{-s^2-t^2}\operatorname{Cov}(M_n'(s), M_n'(t)) \\
M_n'(s)M_n'(t) &= \left(\frac{1}{n}\sum_{i=1}^n X_ie^{sX_i}\right)\left(\frac{1}{n}\sum_{i=1}^n X_ie^{tX_i}\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^n X_i^2e^{(s+t)X_i} + \sum_{i\neq j} X_iX_je^{(sX_i+tX_j)}\right) \\
\mathbb{E}(M_n'(s)M_n'(t)) &= \frac{1}{n}\mathbb{E}(X_1^2e^{(s+t)X_1}) + \frac{n-1}{n}\mathbb{E}(X_1e^{sX_1})\mathbb{E}(X_2e^{tX_2}) \\
\mathbb{E}(X_1^2e^{\theta X_1}) &= \int x^2e^{\theta x}\phi(x)dx \\
&= e^{\theta^2/2}\int x^2\phi(x-\theta)dx \\
&= e^{\theta^2/2}(1+\theta^2) \\
\mathbb{E}(M_n'(s)M_n'(t)) &= \frac{1}{n}e^{(s+t)^2/2}(1+(s+t)^2) + \frac{n-1}{n}ste^{s^2/2+t^2/2} \\
\operatorname{Cov}(M_n'(s)M_n'(t)) &= \frac{1}{n}e^{(s+t)^2/2}(1+(s+t)^2) + \frac{n-1}{n}ste^{s^2/2+t^2/2} - ste^{s^2/2+t^2/2} \\
&= \frac{1}{n}e^{(s+t)^2/2}(1+(s+t)^2) - \frac{1}{n}ste^{s^2/2+t^2/2} \\
\mathbb{E}(b(s)b(t)) &= e^{-(s-t)^2/2}(1+(s+t)^2) - ste^{-s^2/2-t^2/2}
\end{aligned}$$

Since Z is independent of $X_1, \ldots, X_n$, each term in (2.43) with only one $c(s)$ or $c(t)$ in it has expectation 0. The remaining term is

$$c(s)c(t) = ste^{-s^2/2-t^2/2}Z^2,$$

and since Z is standard normal, it clearly has expectation

$$\mathbb{E}(c(s)c(t)) = ste^{-s^2/2-t^2/2}.$$

So in summary

$$\begin{array}{llll}
\mathbb{E}(Y_n'(s)Y_n'(t)) = & \mathbb{E}(a(s)a(t)) & +\mathbb{E}(a(s)b(t)) & +0 \\
& +\mathbb{E}(b(s)a(t)) & +\mathbb{E}(b(s)b(t)) & +0 \\
& +0 & +0 & +\mathbb{E}(c(s)c(t)),
\end{array}$$

where

$$\begin{aligned}
\mathbb{E}(a(s)a(t)) &= 4st\left(e^{-(s-t)^2/2} - e^{-s^2/2-t^2/2}\right), \\
\mathbb{E}(a(s)b(t)) &= -2s(s+t)e^{-(s-t)^2/2} + 2ste^{-s^2/2-t^2/2}, \\
\mathbb{E}(b(s)b(t)) &= e^{-(s-t)^2/2}(1+(s+t)^2) - ste^{-s^2/2-t^2/2}, \\
\mathbb{E}(c(s)c(t)) &= ste^{-s^2/2-t^2/2}.
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(a(s)b(t)) + \mathbb{E}(b(s)a(t)) &= -2s(s+t)e^{-(s-t)^2/2} + 2ste^{-s^2/2-t^2/2} \\
&\quad -2t(s+t)e^{-(s-t)^2/2} + 2ste^{-s^2/2-t^2/2} \\
&= -2(s+t)^2e^{-(s-t)^2/2} + 4ste^{-s^2/2-t^2/2}.
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{E}(Y_n'(s)Y_n'(t)) &= 4st\left(e^{-(s-t)^2/2} - e^{-s^2/2-t^2/2}\right) \\
&\quad -2(s+t)^2e^{-(s-t)^2/2} + 4ste^{-s^2/2-t^2/2} \\
&\quad +e^{-(s-t)^2/2}(1+(s+t)^2) \\
&= e^{-(s-t)^2/2}\left(4st - 2(s+t)^2 + 1 + (s+t)^2\right) \\
&\quad +e^{-s^2/2-t^2/2}\left(-4st+4st\right) \\
&= \left(1-(s-t)^2\right)e^{-(s-t)^2/2}
\end{aligned}$$

In summary, we have

$$\begin{aligned}
\mathbb{E}|Y_n'(s) - Y_n'(t)|^2 &= \mathbb{E}(Y_n'(s)^2) + \mathbb{E}(Y_n'(t)^2) - 2\mathbb{E}(Y_n'(s)Y_n'(t)) \\
&= 1 + 1 - 2\left(1-(s-t)^2\right)e^{-(s-t)^2/2} \\
&= 2 - 2\left(1-(s-t)^2\right)e^{-(s-t)^2/2} \\
&= 2 - 2e^{-(s-t)^2/2} + 2(s-t)^2e^{-(s-t)^2/2}.
\end{aligned}$$

As noted in the first part of this calculation, $2 - 2e^{-(s-t)^2/2} \leq (s-t)^2$, and since $e^{-x} \leq 1$, we can bound $\mathbb{E}|Y_n'(s) - Y_n'(t)|^2$ by

$$\mathbb{E}|Y_n'(s) - Y_n'(t)|^2 \leq 3(s-t)^2.$$

□

2.4.5 Calculation for (2.10)

The following is the calculation of $\mathbb{V}\text{ar}(D_n(\theta,0)-\sqrt{n}\bar{X}\theta)$ as mentioned by (2.10) in the proof of Theorem 2.1.2.

Calculation for (2.10). Recall that $D_n(\theta,\eta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\frac{\phi(X_i-\theta)}{\phi(X_i-\eta)}-1)$.

$$\begin{aligned}
\mathbb{V}\text{ar}(D_n(\theta,0)-\sqrt{n}\bar{X}\theta) &= \mathbb{V}\text{ar}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\phi(X_i-\theta)}{\phi(X_i)}-1\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\theta X_i\right)\\
&= \mathbb{V}\text{ar}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\phi(X_i-\theta)}{\phi(X_i)}-1-\theta X_i\right)\right)\\
&= \mathbb{V}\text{ar}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\phi(X_i-\theta)}{\phi(X_i)}-\theta X_i\right)\right)\\
&= \mathbb{V}\text{ar}\left(\frac{\phi(X_1-\theta)}{\phi(X_1)}-\theta X_1\right)\\
&= \mathbb{V}\text{ar}\left(\frac{\phi(X_1-\theta)}{\phi(X_1)}\right)+\theta^2\mathbb{V}\text{ar}(X_1)-2\theta\text{Cov}\left(\frac{\phi(X_1-\theta)}{\phi(X_1)},X_1\right).\\
\text{Cov}\left(\frac{\phi(X_1-\theta)}{\phi(X_1)},X_1\right) &= \int xe^{\theta x-\theta^2/2}\phi(x)dx\\
&= \int(x-\theta+\theta)\phi(x-\theta)dx\\
&= \theta.\\
\mathbb{V}\text{ar}\left(\frac{\phi(X_1-\theta)}{\phi(X_1)}\right) &= \int e^{2\theta x-\theta^2}\phi(x)dx-\left(\int e^{\theta x-\theta^2/2}\phi(x)dx\right)^2\\
&= e^{-\theta^2}\int\phi(x-2\theta)e^{(2\theta)^2/2}dx-\left(\int\phi(x-\theta)dx\right)^2\\
&= e^{\theta^2}-1.\\
\mathbb{V}\text{ar}(D_n(\theta,0)-\sqrt{n}\bar{X}\theta) &= e^{\theta^2}-1-\theta^2.
\end{aligned}$$

□

2.4.6 The distribution of $e^{-\theta^2/2}\left(D_n(\theta,0)-\sqrt{n}\bar{X}\theta\right)$.

The next proof shows that $e^{-\theta^2/2}\left(D_n(\theta,0)-\sqrt{n}\bar{X}\theta\right)$ has the same distribution as (2.14), as mentioned in Section 2.3.

Proof.

$$\begin{aligned} e^{-\theta^2/2}\left(D_n(\theta,0)-\sqrt{n}\bar{X}\theta\right) &= e^{-\theta^2/2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\phi(X_i-\theta)}{\phi(X_i)}-1\right)-\sqrt{n}\bar{X}\theta\right) \\ &= e^{-\theta^2/2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\phi(X_i-\theta)}{\phi(X_i)}-1-\theta X_i\right)\right). \end{aligned}$$

The distribution function of each X_i is Φ, given by

$$\Phi(x)=\int_{-\infty}^{x}\phi(z)dz,$$

so the random variables $\Phi^{-1}(U_i)$ are also $N(0,1)$. The distribution of

$$e^{-\theta^2/2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\phi(X_i-\theta)}{\phi(X_i)}-1-\theta X_i\right)\right)$$

is thus the same as the distribution of

$$e^{-\theta^2/2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\phi(\Phi^{-1}(U_i)-\theta)}{\phi(\Phi^{-1}(U_i))}-1-\theta\Phi^{-1}(U_i)\right)\right). \tag{2.44}$$

For $u\in[0,1]$ let ℓ_θ be given by

$$\ell_\theta(u)=e^{-\theta^2/2}\left(\frac{\phi(\Phi^{-1}(u)-\theta)}{\phi(\Phi^{-1}(u))}-1-\theta\Phi^{-1}(u)\right),$$

then (2.44) is

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell_\theta(U_i).$$

Since for each $\theta\in\mathbb{R}$ we have $\mathbb{E}(D_n(\theta,0)-\sqrt{n}\bar{X}\theta)=0$, we also have $\mathbb{E}(\ell_\theta(U_1))=0$. Note, to show $\mathbb{E}(D_n(\theta,0))-\sqrt{n}\bar{X}\theta)=0$, we only need a direct calculation and to recognise that $\frac{\phi(X_i-\theta)}{\phi(X_i)}$ is a multiple of the moment generating function of X_i, and that each X_i has mean 0 by assumption.

We can thus write $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell_\theta(U_i)$ as

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell_\theta(U_i) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\ell_\theta(U_i) - \mathbb{E}(\ell_\theta(U_i))\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\ell_\theta(U_i) - \int_0^1 \ell_\theta(u)du\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell_\theta(U_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \ell_\theta(u)du \\
&= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\ell_\theta(U_i) - \sqrt{n}\int_0^1 \ell_\theta(u)du \\
&= \sqrt{n}\int_0^1 \ell_\theta(u)d\mathbb{F}_n(u) - \sqrt{n}\int_0^1 \ell_\theta(u)du \\
&= \sqrt{n}\int_0^1 \ell_\theta(u)d\{\mathbb{F}_n(u) - u\} \\
&= \int_0^1 \ell_\theta(u)d\alpha_n(u).
\end{aligned}$$

□

2.4.7 Proof that (2.16) is $\mathcal{L}^*-$decomposable

The next proof shows that (2.16) is $\mathcal{L}^*-$decomposable.

Proof. Let $\mathcal{L}$ be

$$\mathcal{L} = \{\ell_\theta : \theta \in \mathbb{R}\}.$$

To show that (2.16) is $\mathcal{L}^*$-decomposable, we show that $\mathcal{L}$ is $\mathcal{L}^*$-decomposable. The lemma then follows, since (2.16) is a subset of $\mathcal{L}$.

Pick any $\ell_\theta \in \mathcal{L}$. The function ℓ_0 is identically 0 everywhere, so the only nontrivial cases to consider are when $\theta > 0$ or when $\theta < 0$.

Suppose $\theta > 0$. Then $\ell_\theta(u)$ can be written as $\ell_\theta(u) = \ell_1^\theta(u) - \ell_2^\theta(u)$, where

$$\begin{aligned}
\ell_1^\theta(u) &= e^{-\theta^2/2}\left(e^{\theta\Phi^{-1}(u)-\theta^2/2} - 1\right), && (2.45) \\
\ell_2^\theta(u) &= \theta e^{-\theta^2/2}\Phi^{-1}(u). && (2.46)
\end{aligned}$$

Since Φ^{-1} is non decreasing, ℓ_1^θ and ℓ_1^θ are non decreasing.

Suppose $\theta < 0$. Let $\ell_1^{-\theta}(u) = -\ell_2^\theta(u)$, and $\ell_2^{-\theta}(u) = -\ell_1^\theta(u)$. Then $\ell_1^{-\theta}$ and $\ell_2^{-\theta}(u)$ are non decreasing functions of u, and we can write $\ell_\theta(u) = \ell_1^{-\theta}(u) - \ell_2^{-\theta}(u)$.

Hence for all $\theta \in \mathbb{R}$, ℓ_θ can be written as $\ell_\theta = \ell_1 - \ell_2$ for some $\ell_1, \ell_2 \in \mathcal{L}^*$, and hence $\mathcal{L}$ is $\mathcal{L}^*$-decomposable. □

2.4.8 Bound for $N_n(\delta_n)$

In Section 2.3, Lemma 2.3.3 provided a bound for $N_n(\delta_n)$. The proof of the lemma is here.

Proof. Denote the inner sup of $N_n(\delta_n)$ by $I(\theta, \delta_n)$, so that

$$\begin{aligned} N_n(\delta_n) &= \sup_{\ell \in \mathcal{L}_n} \sup_{0 \le u \le \delta_n} \left\{ \left(|\ell_1(u)| + |\ell_2(u)| + |\ell_1(1-u)| + |\ell_2(1-u)|\right) u^{\frac{1}{2}} \right\} \\ &= \sup_{\ell \in \mathcal{L}_n} I(\theta, \delta_n). \end{aligned}$$

Using the functions listed at (2.45) and (2.46) we can split ℓ_θ into

$$\ell_\theta = \begin{cases} \ell_1^\theta - \ell_2^\theta, & \theta > 0 \\ (-\ell_2^\theta) - (-\ell_1^\theta), & \theta < 0. \end{cases}$$

It follows that $I(\theta, \delta_n) = I(-\theta, \delta_n)$, and so $N_n(\delta_n)$ is

$$\begin{aligned} N_n(\delta_n) &= \sup_{\ell \in \mathcal{L}_n} I(\theta, \delta_n) \\ &= \sup_{\ell_\theta, 0 \le \theta \le C\sqrt{\log n}} I(\theta, \delta_n). \end{aligned}$$

Therefore without loss of generality, we wish to consider $I(\theta, \delta_n)$ for $\theta > 0$.

The four terms $|\ell_1(u)|, |\ell_2(u)|, |\ell_1(1-u)|$ and $|\ell_2(1-u)|$ can be bounded as follows:

$$\begin{aligned} |\ell_1(u)| &= |e^{-\theta^2} e^{\theta \Phi^{-1}(u)} - e^{-\theta^2/2}| \\ &\le e^{-\theta^2} e^{\theta \Phi^{-1}(u)} + e^{-\theta^2/2} \\ |\ell_1(1-u)| &\le e^{-\theta^2} e^{\theta \Phi^{-1}(1-u)} + e^{-\theta^2/2} \\ &\le e^{-\theta^2} e^{\theta |\Phi^{-1}(1-u)|} + e^{-\theta^2/2} \\ &= e^{-\theta^2} e^{\theta |\Phi^{-1}(u)|} + e^{-\theta^2/2}, \text{ since } \Phi^{-1}(u) = -\Phi^{-1}(1-u) \\ |\ell_2(u)| &= \theta e^{-\theta^2/2} |\Phi^{-1}(u)| \\ |\ell_2(1-u)| &= \theta e^{-\theta^2/2} |\Phi^{-1}(1-u)| \\ &= \theta e^{-\theta^2/2} |\Phi^{-1}(u)|, \end{aligned}$$

thus $I(\theta, \delta_n)$ may be bounded with

$$I(\theta, \delta_n) \le 2 \sup_{0 \le u \le \delta_n} \left\{ \left(e^{-\theta^2} e^{\theta |\Phi^{-1}(u)|} + e^{-\theta^2/2} + \theta e^{-\theta^2/2} |\Phi^{-1}(u)| \right) u^{\frac{1}{2}} \right\}. \quad (2.47)$$

Let $v = 1 - u$ and let us change variables from $0 \leq u \leq \delta_n$ to $1 - \delta_n \leq v \leq 1$. The sup in (2.47) is

$$\sup_{1-\delta_n \leq v \leq 1} \left\{ \left(e^{-\theta^2} e^{\theta |\Phi^{-1}(1-v)|} + e^{-\theta^2/2} + \theta e^{-\theta^2/2} |\Phi^{-1}(1-v)| \right) (1-v)^{\frac{1}{2}} \right\} . \tag{2.48}$$

Let $v = \Phi(x)$, and let us change variables again from $1 - \delta_n \leq v \leq 1$ to $x \geq \Phi^{-1}(1 - \delta_n)$. Noting that $|\Phi^{-1}(1-v)| = |\Phi^{-1}(v)|$, we can write the sup in (2.48) as

$$\sup_{x \geq \Phi^{-1}(1-\delta_n)} \left\{ \left(e^{-\theta^2} e^{\theta |x|} + e^{-\theta^2/2} + \theta e^{-\theta^2/2} |x| \right) (1 - \Phi(x))^{\frac{1}{2}} \right\} .$$

Let $x_n = \Phi^{-1}(1 - \delta_n)$. If $1 - \delta_n > \frac{1}{2}$ then $x_n > 0$. So if $1 - \delta_n > \frac{1}{2}$, we have

$$I(\theta, \delta_n) \leq 2 \sup_{x \geq x_n} \left\{ \left(e^{-\theta^2} e^{\theta x} + e^{-\theta^2/2} + \theta e^{-\theta^2/2} x \right) (1 - \Phi(x))^{\frac{1}{2}} \right\} .$$

In fact if we assume $1 - \delta_n > \Phi(1)$, we get $x_n > 1$ and may use the inequality

$$1 - \Phi(x) \leq \frac{\phi(x)}{x}$$

to get

$$I(\theta, \delta_n) \leq 2 \sup_{x \geq x_n} \left\{ \left(e^{-\theta^2} e^{\theta x} + e^{-\theta^2/2} + \theta e^{-\theta^2/2} x \right) \left(\frac{\phi(x)}{x} \right)^{\frac{1}{2}} \right\} .$$

Let

$$f_\theta(x) = \left(e^{-\theta^2} e^{\theta x} + e^{-\theta^2/2} + \theta e^{-\theta^2/2} x \right) \left(\frac{\phi(x)}{x} \right)^{\frac{1}{2}} . \tag{2.49}$$

We now show $\sup_{x \geq x_n} f_\theta(x) = O(x_n^{-\frac{1}{2}})$, so $I(\theta, \delta_n) = O(x_n^{-\frac{1}{2}})$ for any $\theta \in \mathbb{R}$. We first note that

$$(2\pi)^{\frac{1}{4}} f_\theta(x) = \left(e^{-\theta^2} e^{\theta x} + e^{-\theta^2/2} + \theta e^{-\theta^2/2} x \right) x^{-\frac{1}{2}} e^{-x^2/4},$$

so since $\theta e^{-\theta^2/2}$ is maximised when $\theta = 1$, we have

$$\begin{aligned} (2\pi)^{\frac{1}{4}} f_\theta(x) &= x^{-\frac{1}{2}} e^{-\theta^2 + \theta x - x^2/4} + x^{-\frac{1}{2}} e^{-\theta^2/2 - x^2/4} + \theta x^{\frac{1}{2}} e^{-\theta^2/2 - x^2/4} \\ &= x^{-\frac{1}{2}} e^{-(\theta - \frac{x}{2})^2} + x^{-\frac{1}{2}} e^{-\theta^2/2 - x^2/4} + \theta x^{\frac{1}{2}} e^{-\theta^2/2 - x^2/4} \\ &\leq x^{-\frac{1}{2}} e^{-(\theta - \frac{x}{2})^2} + x^{-\frac{1}{2}} e^{-x^2/4} + x^{\frac{1}{2}} e^{-\frac{1}{2} - x^2/4} . \end{aligned}$$

So

$$\sup_{x\geq x_n}(2\pi)^{\frac{1}{4}}f_\theta(x) \leq \sup_{x\geq x_n} x^{-\frac{1}{2}}e^{-(\theta-\frac{x}{2})^2} + \sup_{x\geq x_n} x^{-\frac{1}{2}}e^{-x^2/4} + \sup_{x\geq x_n} x^{\frac{1}{2}}e^{-\frac{1}{2}-x^2/4}. \tag{2.50}$$

The functions $x^{-\frac{1}{2}}e^{-x^2/4}$ and $x^{\frac{1}{2}}e^{-\frac{1}{2}-x^2/4}$ are maximised over $[x_n,\infty)$ at their left endpoint x_n, so (2.50) becomes

$$\sup_{x\geq x_n}(2\pi)^{\frac{1}{4}}f_\theta(x) \quad\leq\quad \sup_{x\geq x_n} x^{-\frac{1}{2}}e^{-(\theta-\frac{x}{2})^2} + x_n^{-\frac{1}{2}}e^{-x_n^2/4} + x_n^{\frac{1}{2}}e^{-\frac{1}{2}-x_n^2/4}, \tag{2.51}$$

and since $e^{-(\theta-\frac{x}{2})^2} \leq 1$, we have

$$\sup_{x\geq x_n}(2\pi)^{\frac{1}{4}}f_\theta(x) \leq x_n^{-\frac{1}{2}} + x_n^{-\frac{1}{2}}e^{-x_n^2/4} + x_n^{\frac{1}{2}}e^{-\frac{1}{2}-x_n^2/4},$$

and since $x^{\frac{1}{2}}e^{-x^2/4} = O(x^{-\frac{1}{2}})$, we thus have $\sup_{x\geq x_n} f_\theta(x) = O(x_n^{-\frac{1}{2}})$ for any $\theta \in \mathbb{R}$.

Hence

$$N_n(\delta_n) = \sup_{\ell_\theta, 0\leq\theta\leq C\sqrt{\log n}} I(\theta,\delta_n) = O(x_n^{-\frac{1}{2}}).$$

□

2.4.9 Proof of Lemma 2.3.4.

The next lemma proven here is Lemma 2.3.4. This lemma provided us with a tool in Section 2.3 which motivated us to view our stochastic process of interest as the sum of a Gaussian process and a remainder process. Recall that the statement of Lemma 2.3.4 is as follows:

Consider any stochastic process $S_n(\theta)$ over $\Theta_n = \{\theta_1,\ldots,\theta_{a_n}\} \subset \mathbb{R}$ which is the sum of two others

$$S_n(\theta) = G_n(\theta) + R_n(\theta), \text{ over } \theta \in \Theta_n.$$

Suppose for some $c_n > 0$

$$P\left(\max_{\theta\in\Theta_n}|R_n(\theta)| \geq c_n\right) \quad\leq\quad \epsilon_n^{(1)}, \tag{2.52}$$

$$P\left(\max_{\theta\in\Theta_n}G_n(\theta) \leq c_n\right) \quad\leq\quad \epsilon_n^{(2)}, \tag{2.53}$$

then

$$P\left(\max_{\theta\in\Theta_n}S_n(\theta) > 0\right) \geq 1-(\epsilon_n^{(1)}+\epsilon_n^{(2)}).$$

Proof of Lemma 2.3.4. Consider the $c_n > 0$ for which the conditions in this lemma hold, and define

$$\begin{aligned} A_n &= \left\{ \max_{\theta \in \Theta_n} G_n(\theta) > c_n \right\}, \\ B_n &= \left\{ \min_{\theta \in \Theta_n} R_n(\theta) > -c_n \right\}. \end{aligned}$$

We first show $P(A_n \cap B_n) \leq P\left(\max_{\theta \in \Theta_n} S_n(\theta) > 0\right)$.

Suppose

$$\max_{\theta \in \Theta_n} G_n(\theta) > c_n, \text{ and } \min_{\theta \in \Theta_n} R_n(\theta) > -c_n,$$

then for $\theta \in \Theta_n$, $G_n(\theta) + R_n(\theta) > G_n(\theta) - c_n$, and hence

$$\begin{aligned} \max_{\theta \in \Theta_n} S_n(\theta) &= \max_{\theta \in \Theta_n} (G_n(\theta) + R_n(\theta)) \\ &> \max_{\theta \in \Theta_n} (G_n(\theta) - c_n) \\ &= \max_{\theta \in \Theta_n} G_n(\theta) - c_n \\ &> c_n - c_n \\ &= 0. \end{aligned}$$

Thus $A_n \cap B_n \subseteq \{\max_{\theta \in \Theta_n} S_n(\theta) > 0\}$, and

$$P(A_n \cap B_n) \leq P\left(\max_{\theta \in \Theta_n} S_n(\theta) > 0 \right).$$

We can write $P(A_n \cap B_n)$ as

$$P(A_n \cap B_n) = 1 - P(A_n^c \cup B_n^c),$$

where A_n^c and B_n^c are the complements of A_n and B_n. Since

$$-\min_{\theta \in \Theta_n} R_n(\theta) \leq |\min_{\theta \in \Theta_n} R_n(\theta)| \leq \max_{\theta \in \Theta_n} |R_n(\theta)|,$$

if $\min_{\theta \in \Theta_n} R_n(\theta) \leq -c_n$, then $\max_{\theta \in \Theta_n} |R_n(\theta)| \geq c_n$. Thus

$$P(B_n^c) \leq P(\max_{\theta \in \Theta_n} |R_n(\theta)| \geq c_n),$$

and hence

$$\begin{aligned} P(A_n^c \cup B_n^c) &\leq P(A_n^c) + P(B_n^c) \\ &\leq \epsilon_n^{(2)} + P(\max_{\theta \in \Theta_n} |R_n(\theta)| \geq c_n) \\ &\leq \epsilon_n^{(1)} + \epsilon_n^{(2)}, \end{aligned}$$

so

$$P\left(\max_{\theta\in\Theta_n} S_n(\theta) > 0\right) \geq P(A_n \cap B_n) \geq 1 - \left(\epsilon_n^{(1)} + \epsilon_n^{(2)}\right),$$

and since

$$P\left(\sup_{\theta\in\mathbb{R}} S_n(\theta) > 0\right) \geq P\left(\sup_{\theta\in\Theta_n} S_n(\theta) > 0\right),$$

we are done. □

2.4.10 Proof of Lemma 2.3.5

The next proof uses an application of the Normal Comparison Lemma mentioned in Section 2.2.

Proof of Lemma 2.3.5. Enumerate the a_n elements of Θ_n via $\theta_i, i = 1, 2, \ldots, a_n$.

For each $\theta_i \in \Theta_n$ we have a corresponding normal random variable $G_n(\theta_i) = \int_0^1 \ell_{\theta_i}(u)dW_n(u)$, and a calculation directly after this proof tells us

$$\begin{aligned} \mathbb{E}(G_n(\theta)) &= 0, \text{ and} \\ \operatorname{Cov}(G_n(t), G_n(s)) &= e^{-(s-t)^2/2} - e^{-t^2/2-s^2/2} - ste^{-t^2/2-s^2/2}. \quad (2.54) \end{aligned}$$

To apply the Normal Comparison Lemma, we introduce the notation $\sigma^2(\theta) = \mathbb{V}\text{ar}(G_n(\theta))$ and denote the standardised versions of each $G_n(\theta_i)$ by $Y_n(\theta) = \frac{G_n(\theta)}{\sigma(\theta)}$. Let $Y_1, \ldots, Y_{a_n}$ be given by $Y_i = Y_n(\theta_i)$ so that the covariance structure of the $G_n(\theta_i)$ is directly related to $\Lambda^1 = \left(\Lambda^1_{ij}\right)$, given by

$$\Lambda^1_{ij} = \operatorname{Cov}(Y_i, Y_j) = \operatorname{Cov}\left(\frac{G_n(\theta_i)}{\sigma(\theta_i)}, \frac{G_n(\theta_j)}{\sigma(\theta_j)}\right) = \frac{\operatorname{Cov}(G_n(\theta_i), G_n(\theta_j))}{\sigma(\theta_i)\sigma(\theta_j)}.$$

Let $Z_1, \ldots, Z_{a_n}$ be iid $N(0,1)$ and let Λ^0 be their covariance matrix. Clearly Λ^0 is simply the $a_n \times a_n$ identity matrix I_{a_n}. The probability we are interested in is

$$P\left(\max_{\theta\in\Theta_n} G_n(\theta) \leq c_n\right),$$

for some $c_n > 0$.

Note that

$$\begin{aligned} P\left(\max_{\theta\in\Theta_n} G_n(\theta) \leq c_n\right) &= P\left(\bigcap_{\theta\in\Theta_n} \{G_n(\theta) \leq c_n\}\right) \\ &= P\left(\bigcap_{\theta\in\Theta_n} \left\{\frac{G_n(\theta)}{\sigma(\theta)} \leq \frac{c_n}{\sigma(\theta)}\right\}\right). \end{aligned}$$

If we denote each $\frac{c_n}{\sigma(\theta)}$ by

$$u_n(\theta) = \frac{c_n}{\sigma(\theta)},$$

then

$$P\left(\max_{\theta\in\Theta_n} G_n(\theta) \leq c_n\right) = P\left(\bigcap_{i=1}^{a_n}\{Y_i \leq u_n(\theta_i)\}\right).$$

By applying the Normal Comparison Lemma to $Y_1, \ldots, Y_{a_n}$ and $Z_1, \ldots, Z_{a_n}$, we can arrive at

$$\begin{aligned} &\left|P\left(\bigcap_{i=1}^{a_n} Y_i \leq u_n(\theta_i)\right) - \prod_{i=1}^{a_n}\Phi(u_n(\theta_i))\right| \\ \leq\; & \tfrac{1}{2\pi}\textstyle\sum_{1\leq i<j\leq a_n} |\Lambda_{ij}^1|(1-\rho_{ij}^2)^{-\frac{1}{2}} \exp\left(\frac{-(u_n(\theta_i)^2+u_n(\theta_j)^2)}{2(1+\rho_{ij})}\right). \qquad (2.55)\end{aligned}$$

Since $\rho_{ij} = \max(|\Lambda_{ij}^1|, |\Lambda_{ij}^0|)$, when $i \neq j$ $\rho_{ij} = |\Lambda_{ij}^1|$. Let $\rho_n = \max_{i\neq j}\rho_{ij}$. If $\rho_n < 1$, then

$$\begin{aligned} |\Lambda_{ij}^1|(1-\rho_{ij}^2)^{-\frac{1}{2}} &\leq \rho_n(1-\rho_n^2)^{-\frac{1}{2}}, \\ \text{and } \frac{-1}{1+\rho_{ij}} &\leq \frac{-1}{1+\rho_n}. \end{aligned}$$

Let $u_n = \min_{\theta\in\Theta_n} u_n(\theta)$. Then $u_n(\theta_i)^2 + u_n(\theta_j)^2 \geq 2u_n^2$, and

$$\exp\left(\frac{-(u_n(\theta_i)^2+u_n(\theta_j)^2)}{2(1+\rho_n)}\right) \leq \exp\left(\frac{-u_n^2}{1+\rho_n}\right) \leq 1.$$

So if $\rho_n < 1$ then (2.55) becomes

$$\begin{aligned} &\left|P\left(\bigcap_{i=1}^{a_n} Y_i \leq u_n(\theta_i)\right) - \prod_{i=1}^{a_n}\Phi(u_n(\theta_i))\right| \\ \leq\; & \tfrac{1}{2\pi}\textstyle\sum_{1\leq i<j\leq a_n} \rho_n(1-\rho_n^2)^{-\frac{1}{2}} \\ =\; & \tfrac{a_n(a_n-1)}{4\pi}\rho_n(1-\rho_n^2)^{-\frac{1}{2}}. \qquad (2.56)\end{aligned}$$

Using (2.56) we thus arrive at

$$P\left(\max_{\theta\in\Theta_n} G_n(\theta) \leq c_n\right) \leq \prod_{i=1}^{a_n}\Phi(u_n(\theta_i)) + \frac{a_n(a_n-1)}{4\pi}\rho_n(1-\rho_n^2)^{-\frac{1}{2}}.$$

□

2.4.11 Covariance calculation for above proof

We now provide a calculation of the covariances of the random variables $G_n(\theta)$ mentioned in the above proof.

Calculation to show (2.54). It is established in McKean (1969) that for a stochastic integral of the form $G_n(t) = \int_{[0,1]} f_t(u)dW_n(u)$,

$$\begin{aligned} \mathbb{E}(G_n(t)) &= 0, \\ E(G_n(t)G_n(s)) &= \int_0^1 f_t(u)f_s(u)du. \end{aligned}$$

So when $G_n(\theta) = \int_0^1 \ell_\theta(u)dW_n(u)$,

$$\begin{aligned} \mathbb{E}(G_n(\theta)) &= 0, \\ \mathbb{E}(G_n(t)G_n(s)) &= \int_0^1 \ell_t(u)\ell_s(u)du. \end{aligned}$$

So Cov $(G_n(t)G_n(s)) = \mathbb{E}(G_n(t)G_n(s))$. We now calculate $\mathbb{E}(G_n(t)G_n(s))$, so making the substitution $x = \Phi^{-1}(u)$ we have

$$\mathbb{E}(G_n(t)G_n(s)) = \int_{-\infty}^{\infty} \ell_t(\Phi(x))\ell_s(\Phi(x))\phi(x)dx.$$

Recall that

$$\ell_\theta(u) = e^{-\theta^2/2}\left(\frac{\phi(\Phi^{-1}(u)-\theta)}{\phi(\Phi^{-1}(u))} - 1 - \theta\Phi^{-1}(u)\right).$$

Elementary calculations show

$$\begin{aligned} \ell_t(\Phi(x)) &= e^{-t^2/2}\left(\frac{\phi(x-t)}{\phi(x)} - 1 - tx\right), \\ \ell_t(\Phi(x))\ell_s(\Phi(x)) &= e^{-t^2/2-s^2/2}\left(\frac{\phi(x-t)}{\phi(x)} - 1 - tx\right)\left(\frac{\phi(x-s)}{\phi(x)} - 1 - sx\right), \\ e^{t^2/2+s^2/2}\ell_t(\Phi(x))\ell_s(\Phi(x)) &= \left(e^{-t^2/2+tx} - 1 - tx\right)\left(e^{-s^2/2+sx} - 1 - sx\right) \\ &= e^{-t^2/2-s^2/2+(s+t)x} - e^{-t^2/2+tx} - e^{-s^2/2+sx} + sxe^{-t^2/2+tx} \\ &\quad +1 + (s+t)x + txe^{-s^2/2+sx} + stx^2, \\ e^{t^2/2+s^2/2}\mathbb{E}(G_n(t)G_n(s)) &= e^{-t^2/2-s^2/2+(s+t)^2/2} - 1 - st, \\ \mathbb{E}(G_n(t)G_n(s)) &= e^{-(s-t)^2/2} - e^{-t^2/2-s^2/2} - ste^{-t^2/2-s^2/2}, \end{aligned}$$

and we are now done. □

2.4.12 Proof of Lemma Lemma 2.3.6

We now prove Lemma 2.3.6, which specifies conditions in which the bound provided by Lemma 2.3.5 is useful for our proof of Theorem 2.1.2.

Proof. We first consider conditions for which $\prod_{i=1}^{a_n} \Phi(u_n(\theta_i)) \to 0$ as $n \to \infty$. Recall that $u_n(\theta_i) = \frac{c_n}{\sigma(\theta_i)}$, where $\sigma(\theta)^2 = 1 - e^{-\theta^2}(1+\theta^2)$. For $x > 0$, $\frac{d}{dx}\left(1 - e^{-x}(1+x)\right) > 0$, so when $\theta_i^2 \leq \theta_j^2$ we have $\sigma(\theta_i)^2 \leq \sigma(\theta_j)^2$.

Therefore

$$\max_i u_n(\theta_i) = \frac{c_n}{\min_i \sigma(\theta_i)} = \frac{c_n}{\sigma(t_n)},$$

and so

$$\prod_{i=1}^{a_n} \Phi(u_n(\theta_i)) \leq \Phi(u_n)^{a_n},$$

where $u_n = \frac{c_n}{\sigma(t_n)}$. Note that u_n is defined for the purposes of this proof, and that it is not the same u_n used in the proof of Lemma 2.3.5. Since $t_n \to \infty$ as $n \to \infty$, $\sigma(t_n) \to 1$ as $n \to \infty$, so since $c = \lim_{n\to\infty} c_n$ is finite, $u_n \to c$. Therefore if c_n and a_n satisfy $\Phi(c_n)^{a_n} \to 0$ as $n \to \infty$, then

$$\prod_{i=1}^{a_n} \Phi(u_n(\theta_i)) \leq \Phi(u_n)^{a_n} \to 0, \text{ as } n \to \infty.$$

We next bound the maximum correlation $\rho_n = \max_{i\neq j}\left|\frac{\mathrm{Cov}(G_n(\theta_i),G_n(\theta_j))}{\sigma(\theta_i)\sigma(\theta_j)}\right|$ with

$$\rho_n \leq \left(\max_{i\neq j} \frac{1}{\sigma(\theta_i)\sigma(\theta_j)}\right) \max_{i\neq j} \left|\mathrm{Cov}\left(G_n(\theta_i), G_n(\theta_j)\right)\right|.$$

The two smallest values of θ_i^2 and $\theta_j^2 \neq \theta_i^2$ are t_n^2 and $(t_n + \Delta_n)^2$. Therefore since $\sigma(\theta)$ is increasing as θ^2 increases

$$\max_{i\neq j} \frac{1}{\sigma(\theta_i)\sigma(\theta_j)} = \frac{1}{\sigma(t_n)\sigma(t_n+\Delta_n)}.$$

As $n \to \infty$, $\sigma(t_n) \to 1$ from below, so

$$\rho_n \leq (1+\epsilon_n) \max_{i\neq j} \left|\mathrm{Cov}\left(G_n(\theta_i), G_n(\theta_j)\right)\right|$$

for some $\epsilon_n \to 0$ as $n \to \infty$.

For $s, t \in \mathbb{R}$,

$$st \leq |st| \leq \max\{s^2, t^2\} \leq s^2 + t^2,$$

so we can bound $\mathrm{Cov}\left(G_n(\theta_i), G_n(\theta_j)\right)$ by

$$\begin{aligned}\mathrm{Cov}\left(G_n(s), G_n(t)\right) &= e^{-\frac{1}{2}(s-t)^2} - e^{-t^2/2-s^2/2}(1+st) \\ &\leq e^{-\frac{1}{2}(s-t)^2} + e^{-t^2/2-s^2/2}(1+s^2+t^2),\end{aligned}$$

and since Δ_n is the smallest spacing between any different θ_i, θ_j, we have

$$\rho_n \leq (1+\epsilon_n)\left(e^{-\Delta_n^2/2} + \max_{i\neq j}\left((\theta_i^2+\theta_j^2+1)e^{-\theta_i^2/2-\theta_j^2/2}\right)\right).$$

Let $f(x) = e^{-x}(x+\frac{1}{2})$. Then

$$\max_{i\neq j}\left((\theta_i^2+\theta_j^2+1)e^{-\theta_i^2/2-\theta_j^2/2}\right) = 2f\left(\frac{\theta_i^2+\theta_j^2}{2}\right),$$

and moreover $\frac{d}{dx}f(x) = e^{-x}(\frac{1}{2}-x)$, and since $\frac{\theta_i^2+\theta_j^2}{2} \geq t_n^2$, when $t_n^2 \geq \frac{1}{2}$, $f\left(\frac{\theta_i^2+\theta_j^2}{2}\right) \leq f(t_n^2)$. Therefore

$$\max_{i\neq j}\left((\theta_i^2+\theta_j^2+1)e^{-\theta_i^2/2-\theta_j^2/2}\right) \leq e^{-t_n^2}(2t_n^2+1),$$

and so

$$\rho_n \leq (1+\epsilon_n)\left(e^{-\Delta_n^2/2} + e^{-t_n^2}(2t_n^2+1)\right).$$

If $\frac{t_n^2 e^{-t_n^2}}{e^{-\Delta_n^2/2}} \to 0$ as $n \to \infty$ then ρ_n behaves like $e^{-\Delta_n^2/2}$ for large n.

Recall that

$$\epsilon_n^{(2)} = \prod_{i=1}^{a_n}\Phi(u_n(\theta_i)) + \frac{a_n(a_n-1)}{4\pi}\rho_n(1-\rho_n^2)^{-\frac{1}{2}}.$$

Since $\rho_n(1-\rho_n^2)^{-\frac{1}{2}}$ looks like ρ_n when ρ_n looks like $e^{-\Delta_n^2/2}$, the condition

$$a_n^2\rho_n \to 0, \text{ as } n \to \infty$$

is enough for

$$a_n(a_n-1)\rho_n(1-\rho_n^2)^{-\frac{1}{2}} \to 0, \text{ as } n \to \infty.$$

□

2.4.13 Showing that (2.16) satisfies Corollary 3.4 of Csörgő et al. (1986)

We now show that (2.16) satisfies the conditions of Corollary 3.4 of Csörgő et al. (1986). These satisfied conditions imply that in Section 2.3, $\sup_{\ell_\theta \in \mathcal{L}_n} |R_3(\theta)| = o_p(1)$.

Proof. We first describe the conditions given in Csörgő et al. (1986) that we wish to check. We will need some definitions to do this.

Definition (Positive function)**.** A function q defined on $(0, 1]$ is called positive if

$$\inf_{\delta \leq s \leq \frac{1}{2}} q(s) > 0, \text{ for all } 0 < \delta < \frac{1}{2}.$$

For the purposes of this proof, we will consider the following characterisations as definitions. We refer to Theorems 3.3 and 3.4 of Csörgő et al. (1986) for the actual definitions.

Definition 2.4.1 (EFKP upper-class function of a Brownian bridge)**.** Let q be any positive function defined on $(0, \frac{1}{2}]$, nondecreasing in a neighbourhood of zero. Such a function q will be called an Erdös-Feller-Kolmogorov-Petrovski (EFKP) upper class function of a Brownian bridge $\{B(s); 0 \leq s \leq 1\}$ if and only if the integral

$$I(q, c) = \int_0^{\frac{1}{2}} s^{-1} \exp(-cs^{-1}q^2(s))ds < \infty \tag{2.57}$$

for some $c > 0$. An EFKP upper-class function of a Brownian bridge q is called a Chibisov-O'Reilly function if $I(q, c) < \infty$ for all $c > 0$.

We now are in a position to state Corollary 3.4 of Csörgő et al. (1986).

Corollary 2.4.2 (from Csörgő et al. (1986))**.** *Let $\mathcal{L}$ be any $\mathcal{L}^*-$decomposable class of functions, and let $q_{11}, q_{12}, q_{21}, q_{22}$ be any positive functions defined on $(0, \frac{1}{2}]$, nondecreasing in a neighbourhood of zero and assumed to be right-continuous. For $\delta > 0$ and small enough so that the $q_{ij}(i, j = 1, 2)$ are already nondecreasing on $(0, \delta]$, define*

$$\begin{aligned} N_i^{(1)}(\delta) &= \sup_{\ell \in \mathcal{L}} \int_0^\delta |\ell_i(s)| dq_{1i}(s), \\ N_i^{(2)}(\delta) &= \sup_{\ell \in \mathcal{L}} \int_0^\delta |\ell_i(1-s)| dq_{2i}(s), \end{aligned}$$

$i = 1, 2$, $\ell = \ell_1 - \ell_2 \in \mathcal{L}$.

If these q_{ij} functions are each EFKP upper-class functions of a Brownian bridge and

$$\lim_{\delta\downarrow 0}\max_{1\le i,j\le 2} N_i^{(j)}(\delta)=0, \tag{2.58}$$

then on the probability space of Theorem 1.1 by Csörgő et al. (1986) (listed in this chapter as Theorem 2.2.4 in Section 2.2), we have as $n\to\infty$,

$$\widetilde{E}_n=\sup_{\ell\in\mathcal{L}}\left|\int_0^1 \ell(s)d\alpha_n(s)-\int_0^1 \ell(s)dB_n(s)\right|=o_p(1).$$

The proof of this corollary mentions that if (2.58) holds, then

$$\sup_{\ell\in\mathcal{L}}\left(\left|\int_0^{1/n}\ell(s)dB_n(s)\right|+\left|\int_{1-1/n}^1 \ell(s)dB_n(s)\right|\right)=o_p(1). \tag{2.59}$$

In our chapter here, we wish to apply the above corollary to $\mathcal{L}_n$ (as defined by (2.16)) to arrive at

$$\sup_{\ell\in\mathcal{L}_n}\left(\left|\int_0^{1/n}\ell(s)dB_n(s)\right|+\left|\int_{1-1/n}^1 \ell(s)dB_n(s)\right|\right)=o_p(1).$$

We are now in a position where we can check the conditions of the above corollary. Recall that we wish to apply this corollary to the class

$$\mathcal{L}_n=\{\ell_\theta:\theta\in[-C\sqrt{\log n},C\sqrt{\log n}]\},$$

where for $u\in[0,1]$,

$$\ell_\theta(u)=e^{-\theta^2/2}\left(e^{\theta\Phi^{-1}(u)-\theta^2/2}-1-\theta\Phi^{-1}(u)\right).$$

We have already shown with (2.45) and (2.46) that $\mathcal{L}_n$ is $\mathcal{L}^*-$decomposable. Choose any constant $\lambda>0$ and for $s\in(0,\frac{1}{2}]$ let $q(s)=e^{\lambda\Phi^{-1}(s)}$. The function q is a positive function on $(0,\frac{1}{2}]$, continuous and also non decreasing in a neighbourhood of zero. We show through the remainder of this proof that this choice of q satisfies all the necessary conditions to check, and we use it as our choice of q_{11},q_{12},q_{21} and q_{22}.

We first show that q is a Chibisov-O'Reilly function (and thus an EFKP upper-class function of a Brownian bridge). To reduce clutter in this part of the proof, suppose we have defined q via $q(s)=e^{\frac{\lambda}{2}\Phi^{-1}(s)}$, with some $\lambda>0$.

From Definition 2.4.1, q is Chibisov-O'Reilly if and only if (2.57) holds for all $c > 0$. Choose any $c > 0$. We wish to show

$$I(q,c) = \int_0^{\frac{1}{2}} s^{-1} \exp\left(-cs^{-1}e^{\lambda\Phi^{-1}(s)}\right) ds < \infty.$$

We provide a bound for the integrand $s^{-1}\exp\left(-cs^{-1}e^{\lambda\Phi^{-1}(s)}\right)$ by first comparing the left tail behaviour of ϕ with the tail behaviour of an exponential function.

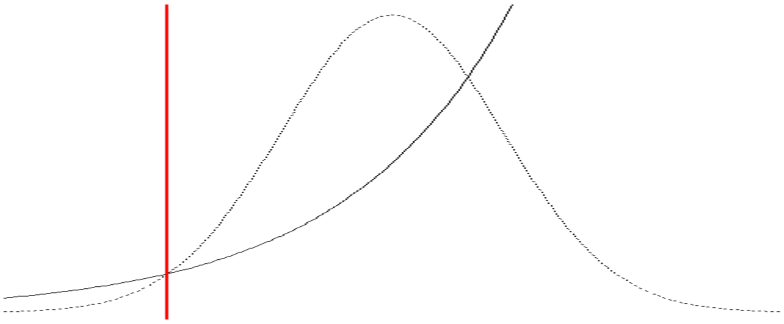


Figure 2.5: An exponential's tail is eventually higher than the left tail of a normal.

Looking at the speeds of e^{-t^2} and e^t as $t \to -\infty$, we see for any $0 < b < \frac{1}{\lambda}$ there exists some $t_0 < 0$ such that we have

$$\phi(t) = e^{-t^2/2-\log(\sqrt{2\pi})} < e^{t/b}, \text{ when } t < t_0.$$

Thus for any $y < t_0 < 0$ we have

$$\Phi(y) = \int_{-\infty}^{y} \phi(t)dt < \int_{-\infty}^{y} e^{t/b}dt = be^{y/b}.$$

Since Φ^{-1} is monotone increasing,

$$y < \Phi^{-1}(be^{y/b}) \text{ for } y < t_0 < 0. \tag{2.60}$$

Let us make the substitution $s = be^{y/b}$, so that $y = b\log\left(\frac{s}{b}\right)$. Using (2.60), for s sufficiently close to 0 we have

$$b\log\left(\frac{s}{b}\right) < \Phi^{-1}(s).$$

Since $\lambda > 0$ and $c > 0$, for $s > 0$ sufficiently close to 0 we have

$$\begin{aligned}
e^{\lambda b \log\left(\frac{s}{b}\right)} &< e^{\lambda \Phi^{-1}(s)}, \\
-e^{\lambda b \log\left(\frac{s}{b}\right)} &> -e^{\lambda \Phi^{-1}(s)}, \\
-ce^{\lambda b \log\left(\frac{s}{b}\right)}/s &> -ce^{\lambda \Phi^{-1}(s)}/s, \\
\frac{1}{s}\exp(-ce^{\lambda b \log\left(\frac{s}{b}\right)}/s) &> \frac{1}{s}\exp(-ce^{\lambda \Phi^{-1}(s)}/s).
\end{aligned}$$

Also, since $b < \frac{1}{\lambda}$ we have

$$\begin{aligned}
e^{\lambda b \log\left(\frac{s}{b}\right)}/s &= s^{\lambda b - 1}/b^{\lambda b} \\
&= \left(\frac{1}{s}\right)^{1-\lambda b}/b^{\lambda b}, \text{ where } 1 - \lambda b > 0.
\end{aligned}$$

Thus for s sufficiently small we have this bound for the integrand of $I(q, c)$

$$\frac{1}{s}\exp\left(-ce^{\lambda \Phi^{-1}(s)}/s\right) < \frac{1}{s}\exp\left(-c\left(\frac{1}{s}\right)^{1-\lambda b}/b^{\lambda b}\right). \tag{2.61}$$

Let a range of s where (2.61) holds be called $0 < s < \epsilon$. We can break $I(q, c)$ into

$$I(q,c) = \int_0^{\epsilon} \frac{1}{s}\exp\left(-ce^{\lambda\Phi^{-1}(s)}/s\right) ds + \int_{\epsilon}^{\frac{1}{2}} \frac{1}{s}\exp\left(-ce^{\lambda\Phi^{-1}(s)}/s\right) ds,$$

and since

$$\int_{\epsilon}^{\frac{1}{2}} \frac{1}{s}\exp\left(-ce^{\lambda\Phi^{-1}(s)}/s\right) ds < \infty,$$

it is sufficient for us to show $\int_0^{\epsilon} \frac{1}{s}\exp\left(-ce^{\lambda\Phi^{-1}(s)}/s\right) ds < \infty$.

$$\begin{aligned}
\int_0^{\epsilon} \frac{1}{s}\exp(-ce^{\lambda\Phi^{-1}(s)}/s)ds &< \int_0^{\epsilon} \frac{1}{s}\exp(-c\left(\frac{1}{s}\right)^{1-\lambda b}/b^{\lambda b})ds, \text{ where } 1 - \lambda b > 0 \\
&= \int_{\frac{1}{\epsilon}}^{\infty} \frac{1}{x} e^{-cx^{1-\lambda b}/b^{\lambda b}} dx < \infty, \text{ using } x = \frac{1}{s}.
\end{aligned}$$

Since this is true for any $c > 0$, we have shown that q given by $q(s) = e^{\frac{\lambda}{2}\Phi^{-1}(s)}$ is Chibisov-O'Reilly. Since $\lambda > 0$ was arbitrary, $q(s) = e^{\lambda\Phi^{-1}(s)}$ is also Chibisov-O'Reilly.

We now show

$$\lim_{\delta \downarrow 0} \max_{1 \le i,j \le 2} N_i^{(j)}(\delta) = 0.$$

We first consider

$$N_1^{(1)}(\delta) = \sup_{\ell_\theta \in \mathcal{L}_n} \int_0^\delta |\ell_1(s)| dq(s).$$

Since the decomposition of any $\ell_\theta \in \mathcal{L}_n$ depends on whether $\theta > 0$ or $\theta \leq 0$, let us split the nontrivial parts of $\mathcal{L}_n$ into the classes

$$\begin{aligned} \mathcal{L}_n^+ &= \{\ell_\theta : 0 < \theta \leq C\sqrt{\log n}\}, \\ \mathcal{L}_n^- &= \{\ell_\theta : -C\sqrt{\log n} \leq \theta < 0\}. \end{aligned}$$

When $\theta = 0$, ℓ_θ is the constant function $\ell_\theta = 0$, so $N_1^{(1)}(\delta)$ is

$$N_1^{(1)}(\delta) = \sup\left\{ \sup_{\ell_\theta \in \mathcal{L}_n^-} \int_0^\delta |\ell_1(s)| dq(s), 0, \sup_{\ell_\theta \in \mathcal{L}_n^+} \int_0^\delta |\ell_1(s)| dq(s) \right\}.$$

We now consider each case $\theta > 0$, $\theta < 0$ seperately. Suppose $\theta > 0$, then $\ell_\theta = \ell_1 - \ell_2$ where

$$\begin{aligned} \ell_1(u) &= e^{-\theta^2/2}\left(e^{\theta\Phi^{-1}(u) - \theta^2/2} - 1\right), \\ \ell_2(u) &= \theta e^{-\theta^2/2}\Phi^{-1}(u). \end{aligned}$$

For convenience we will refer to ℓ_1 as the increasing part of ℓ_θ, and ℓ_2 as the decreasing part. It should be noted that ℓ_2 is also an increasing function but $\ell_\theta = \ell_1 - \ell_2$. We are currently focusing on $N_1^{(1)}(\delta)$ so for now we speak only of the increasing part of ℓ_θ, which is ℓ_1.

We have

$$\int_0^\delta |\ell_1(s)| dq(s) = |\ell_1(s)| q(s)\Big]_0^\delta - \int_0^\delta \left(\frac{d}{ds}|\ell_1(s)|\right) q(s) ds,$$

and the form of ℓ_1 allows us to choose $\delta > 0$ small enough so that $\ell_1(s) < 0$ for $s \in (0, \delta]$, so we may assume $|\ell_1(s)| = -\ell_1(s)$ over $(0, \delta]$ to arrive at

$$\int_0^\delta |\ell_1(s)| dq(s) = -\ell_1(\delta)q(\delta) + \ell_1(0)q(0) + \int_0^\delta \left(\frac{d}{ds}\ell_1(s)\right) e^{\lambda\Phi^{-1}(s)} ds. \tag{2.62}$$

Considering the first group of terms, we note

$$\ell_1(s)q(s) = e^{-\theta^2/2}\left(e^{(\theta+\lambda)\Phi^{-1}(s) - \theta^2/2} - e^{\lambda\Phi^{-1}(s)}\right) \tag{2.63}$$

so $\ell_1(0)q(0) = 0$. Also, $\frac{d}{ds}\ell_1(s)$ is

$$\frac{d}{ds}\ell_1(s) \quad = \quad \theta e^{-\theta^2} e^{\theta\Phi^{-1}(s)} \left(\frac{d}{ds}\Phi^{-1}(s)\right), \tag{2.64}$$

and thus

$$\begin{aligned}\int_0^\delta \left(\frac{d}{ds}\ell_1(s)\right) e^{\lambda\Phi^{-1}(s)} ds \quad &= \quad \theta e^{-\theta^2} \int_0^\delta \left(\frac{d}{ds}\Phi^{-1}(s)\right) e^{(\theta+\lambda)\Phi^{-1}(s)} ds \\ &= \quad \frac{\theta e^{-\theta^2}}{\theta+\lambda} \int_0^\delta \left(\frac{d}{ds} e^{(\theta+\lambda)\Phi^{-1}(s)}\right) ds \qquad (2.65)\end{aligned}$$

Since $\theta + \lambda > 0$, $e^{(\theta+\lambda)\Phi^{-1}(s)} \to 0$ as $s \to 0$, so (2.65) converges, and is

$$\int_0^\delta \left(\frac{d}{ds}\ell_1(s)\right) e^{\lambda\Phi^{-1}(s)} ds \quad = \quad \frac{\theta e^{-\theta^2}}{\theta+\lambda} e^{(\theta+\lambda)\Phi^{-1}(\delta)}. \tag{2.66}$$

Thus

$$\int_0^\delta |\ell_1(s)| dq(s) \quad = \quad -e^{-\theta^2/2}(e^{(\theta+\lambda)\Phi^{-1}(\delta)-\theta^2/2} - e^{\lambda\Phi^{-1}(\delta)}) + \frac{\theta e^{-\theta^2}}{\theta+\lambda} e^{(\theta+\lambda)\Phi^{-1}(\delta)},$$

and so

$$\begin{aligned}\sup_{\ell_\theta \in \mathcal{L}_n^+} \int_0^\delta |\ell_1(s)| dq(s) \quad &\leq \quad \sup_{0<\theta\leq C\sqrt{\log n}} e^{(\theta+\lambda)\Phi^{-1}(\delta)} \\ &\quad\quad + e^{\lambda\Phi^{-1}(\delta)} \\ &\quad\quad + \sup_{0<\theta\leq C\sqrt{\log n}} \frac{\theta e^{-\theta^2}}{\theta+\lambda} e^{(\theta+\lambda)\Phi^{-1}(\delta)} \\ &= \quad 3e^{\lambda\Phi^{-1}(\delta)}.\end{aligned}$$

Suppose $\theta < 0$. Then the increasing and decreasing parts of ℓ_θ are swapped to

$$\begin{aligned}\ell_1(u) \quad &= \quad -\theta e^{-\theta^2/2}\Phi^{-1}(u), \\ \ell_2(u) \quad &= \quad -e^{-\theta^2/2}\left(e^{\theta\Phi^{-1}(u)-\theta^2/2} - 1\right).\end{aligned}$$

In this case we can also choose a $\delta > 0$ small enough so that $\ell_1(s) < 0$ for $0 < s < \delta$, and arrive again at (2.62).

In this case we replace (2.63) with

$$-\theta e^{-\theta^2/2}\Phi^{-1}(s) e^{\lambda\Phi^{-1}(s)},$$

and replace (2.64) with

$$-\theta e^{-\theta^2/2}\left(\frac{d}{ds}\Phi^{-1}(s)\right).$$

The calculation at (2.65) in this case ends with

$$\frac{-\theta e^{-\theta^2/2}}{\lambda}\int_0^\delta e^{\lambda\Phi^{-1}(s)}ds,$$

and since $\lambda > 0$ the integral converges and is

$$\int_0^\delta \left(\frac{d}{ds}\ell_1(s)\right)e^{\lambda\Phi^{-1}(s)}ds = \frac{-\theta e^{-\theta^2}}{\lambda}e^{\lambda\Phi^{-1}(\delta)}.$$

Thus when $\theta < 0$ we arrive at

$$\begin{aligned}\int_0^\delta |\ell_1(s)|dq(s) &= \theta e^{-\theta^2/2}\Phi^{-1}(\delta)e^{\lambda\Phi^{-1}(\delta)} + \frac{\theta e^{-\theta^2}}{\lambda}e^{\lambda\Phi^{-1}(\delta)}\\ &\leq \Phi^{-1}(\delta)e^{\lambda\Phi^{-1}(\delta)} + 0,\end{aligned}$$

so in summary we have (for $\delta > 0$ small enough)

$$N_1^{(1)}(\delta) \leq 3e^{\lambda\Phi^{-1}(\delta)} + \Phi^{-1}(\delta)e^{\lambda\Phi^{-1}(\delta)}.$$

Since

$$N_1^{(1)}(\delta) = \sup\left\{\sup_{\ell_\theta\in\mathcal{L}_n^-}\int_0^\delta |\ell_1(s)|dq(s), 0, \sup_{\ell_\theta\in\mathcal{L}_n^+}\int_0^\delta |\ell_1(s)|dq(s)\right\},$$

we further have

$$0 \leq N_1^{(1)}(\delta) \leq 3e^{\lambda\Phi^{-1}(\delta)} + \Phi^{-1}(\delta)e^{\lambda\Phi^{-1}(\delta)},$$

so in conclusion we must have

$$\lim_{\delta\downarrow 0} N_1^{(1)}(\delta) = 0.$$

A similar argument can be made to show

$$\lim_{\delta\downarrow 0} N_2^{(1)}(\delta) = 0.$$

If $\theta > 0$, the decreasing function ℓ_2 is (up to a $-$ sign) the increasing function ℓ_1 in the $N_1^{(1)}(\delta)$ situation in the $\theta < 0$ case, and the bound $\Phi^{-1}(\delta)e^{\lambda\Phi^{-1}(\delta)}$ is used for $\sup_{\ell_\theta \in \mathcal{L}_n^+} \int_0^\delta |\ell_2(s)| dq(s)$. If $\theta < 0$, the same ideas from the $\theta > 0$ case in the $N_1^{(1)}(\delta)$ part of this proof produce $\sup_{\ell_\theta \in \mathcal{L}_n^-} \int_0^\delta |\ell_2(s)| dq(s) \leq 3e^{\lambda \Phi^{-1}(s)}$.

Similarly again, we can show $\lim_{\delta \downarrow 0} N_i^{(2)}(\delta) = 0$ for $i = 1, 2$. When we replace $\ell_1(s)$ with $\ell_1(1-s)$, we can use the fact that $\Phi^{-1}(s) = -\Phi^{-1}(1-s)$ to arrive back to similar arguments. The signs in some of the earlier equations like (2.62) will be reversed in some cases, but the bounds remain the same (or swapped).

Thus

$$\lim_{\delta \downarrow 0} \max_{1 \leq i,j \leq 2} N_i^{(j)}(\delta) = 0,$$

so all the conditions of Corollary 2.4.2 hold for $\mathcal{L}_n$.

□

2.5 Simulated demonstration

This section details a simulation we ran in R. The names of functions written in R are typeset like `this`, and the full code for each function mentioned can be found in Appendix A.

Demonstrating the inconsistency of K in this chapter only required we provide one example. A practical bonus of our example's simplicity is we do not need to use some of the slower algorithms described in Chapter 3 to illustrate Theorem 2.1.2.

Let $X_1, \ldots, X_n$ be iid with density f given by

$$f(x) = \int \phi(x - \mu) dQ_0(\mu),$$

and let K_n be the number of mass points of the NPMLE $\widehat{Q}$ of Q_0. Recall that in our example in Theorem 2.1.2, Q_0 was the degenerate distribution δ_0 and so the true density of X_1 was ϕ.

Recall from Section 2.3 that

$$P(K_n = 1) = P\left(\sup_{\theta \in \mathbb{R}} D_n(\theta, \bar{X}) = 0 \right),$$

where

$$D_n(\theta, \bar{X}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{\phi(X_i - \theta)}{\phi(X_i - \bar{X})} - 1 \right\}.$$

The condition $\sup_{\theta \in \mathbb{R}} D_n(\theta, \bar{X}) = 0$ is quite easy to check computationally. Given a vector of observations `x=c(`$x_1, \ldots, x_n$`)` and a range `theta=c(`$\theta_1, \ldots, \theta_{a_n}$`)`, the function `Dn` calculates $D_n(\theta, \bar{X})$.

Here is an example of what $D_n(\theta, \bar{X})$ looked like in practice. We simulated $n = 100$ standard normals.

```
> set.seed(98097629)
> n <- 100
> x <- rnorm(n)
> hist(x, n = 50)
```

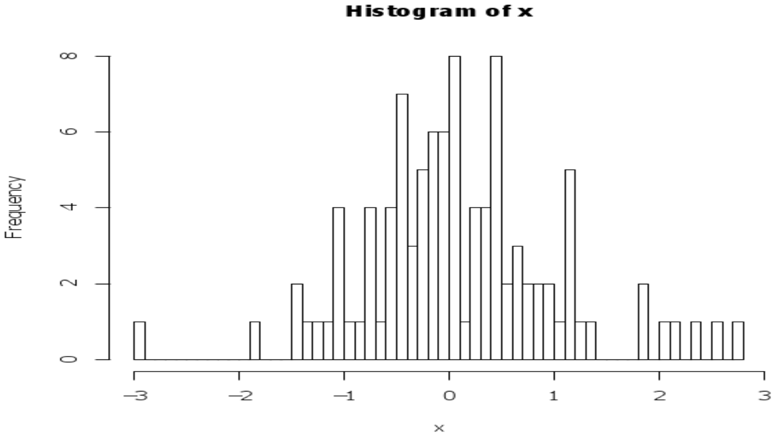


Figure 2.6: $n = 100$ random normals

We then chose a range `theta=`$\Theta \subset \mathbb{R}$ and plotted our observed stochastic process `Dn(x,theta)`.

```
> source("Dn.r")
> theta <- (-50:50)/10
> plot(theta, Dn(x, theta))
> abline(0, 0)
```

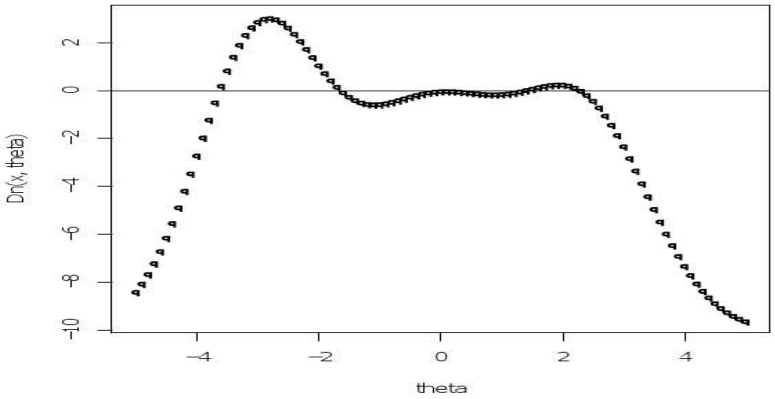


Figure 2.7: The observed stochastic process, sometimes positive.

In this case, $\max_{\theta \in \Theta} D_n(\theta, \bar{X}) > 0$, so $\sup_{\theta \in \mathbb{R}} D_n(\theta, \bar{X}) > 0$ and hence $K_n > 1$.

Here is an example of when $K_n = 1$.

```
> set.seed(98097629)
> n <- 30
> x <- rnorm(n)
> hist(x, n = 10)
```

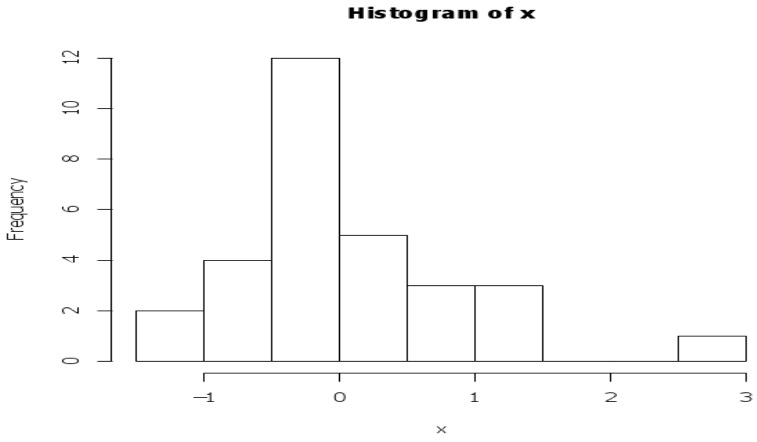


Figure 2.8: $n = 30$ standard normals.

```
> theta <- (-50:50)/10
> plot(theta, Dn(x, theta))
> abline(0, 0)
> abline(v = mean(x))
```

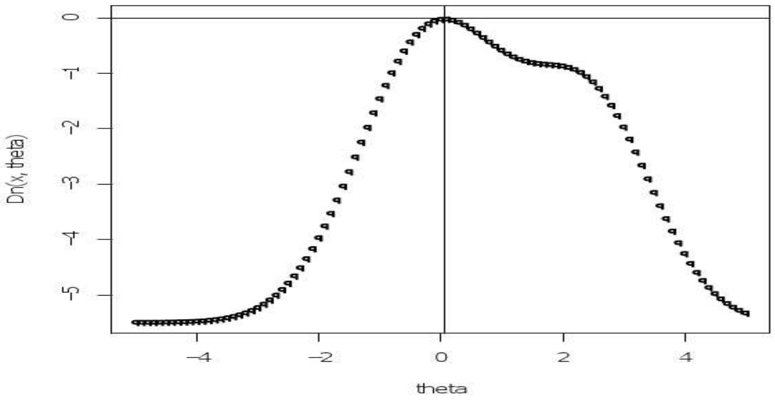


Figure 2.9: The observed stochastic process, which in this case is always ≤ 0.

With $\Theta_n \subset \mathbb{R}$, let

$$p_n = P\left(\max_{\theta \in \Theta_n} D_n(\theta, \bar{X}) = 0\right).$$

Since $\Theta_n \subset \mathbb{R}$, we have the bound $P(K_n = 1) \leq p_n$, and so to demonstrate Theorem 2.1.2 it is sufficient to demonstrate $p_n \to 0$ as $n \to \infty$.

For each $n = 10^i$, $i = 1, \ldots, 8$, we wished to estimate p_n, so we generated n iid standard normals $x_1, \ldots, x_n$ and used the function `Dn_ispositive_C` to check whether any of the function values calculated over a sensible choice of range $\Theta_n \subset \mathbb{R}$ was positive.

The function `Dn_ispositive_C` returned 0 if $\max_{\theta \in \Theta_n} D_n(\theta, \bar{X}) = 0$ and it returned 1 as soon as any positive value was returned (if any).

This was done B times. Let Y_n be the number of times (out of B times) that `Dn_ispositive_C` returned 0. Then $Y_n \sim B(B, p_n)$, so our estimate $\widehat{p}_n = Y_n/B$ had expectation $\mathbb{E}(\widehat{p}_n) = p_n$ and standard error $\left(\frac{\widehat{p}_n(1-\widehat{p}_n)}{B}\right)^{\frac{1}{2}}$.

We noticed that the maximum `max(x)` was like $\sqrt{2\log n}$, so we chose to estimate each p_n using a range Θ_n depending on the range of the data. We used the range $\Theta_n =$`floor(1.5*min(x)):floor(1.5*max(x))` to ensure we checked past the range of the data. We obtained the following estimates of p_n.

```
> load("aug26.RData")
> temp <- p[1:7]
> load("aug28.RData")
> p <- c(temp, p)
> p

[1] 0.546 0.384 0.252 0.204 0.150 0.112 0.102 0.070

> se <- sqrt(p * (1 - p)/500)
> i <- 1:8
> b1 <- p - se
> b2 <- p + se
> plot(1:8, p, main = "B=500, theta=floor(1.5*min(x)):floor(1.5*max(x))",
+     ylab = "pn hat plusminus se(pn hat)", xlab = "i=1:8,n=10^i")
> points(i, b2, pch = 3)
> points(i, b1, pch = 3)
```

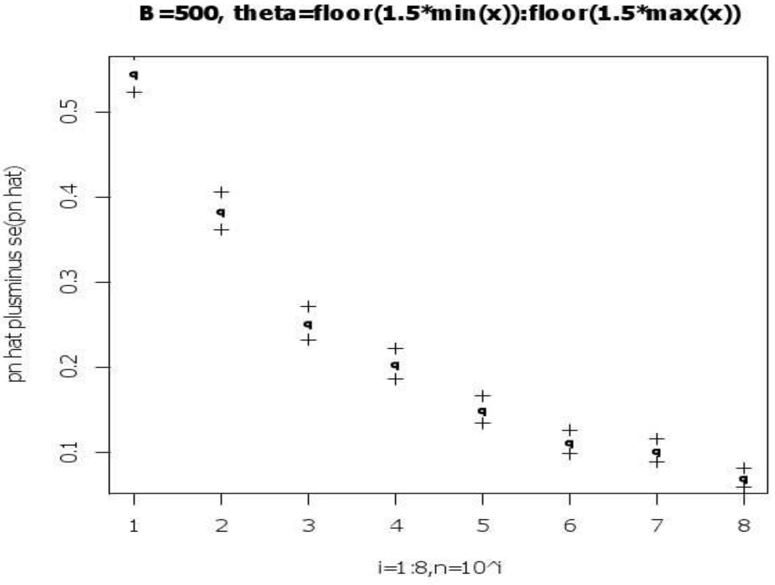


Figure 2.10: Our estimates $\widehat{p}_n$ of $p_n \geq P(K_n = 1)$, obtained by using the ranges $\Theta_n = (-1.5\max_i(X_i), 1.5\max_i(X_i))$.

In the above figure, our estimates $\widehat{p}_n$ of p_n for $n = 10^i$, $i = 1, \ldots, 8$, were made with $B = 500$ repetitions. The values on the horizontal axis are $i = 1, \ldots, 8$. Calculating $\widehat{p}_{10^8}$ took about a week, so we decided not to keep estimating p_n for $i \geq 9$.

Chapter 3

Using NPMLE results for density estimation

In this chapter we investigate a density estimation approach using normal mixtures and Non Parametric Maximum Likelihood Estimation (NPMLE). We first draw an analogy between a bandwidth selection problem in Kernel density estimation, and in density estimation via NPMLE of location mixtures of normals. We have implemented the Intra Simplex Direction Method (ISDM) of Lesperance and Kalbfleisch (1992) in R code. The slowest parts of the ISDM were implemented in C and wrapped into our code, which can be found in Appendix B. This code is faster than the same algorithm implemented in only R by a factor of 10. This chapter concludes with some comments about how the stopping criterion for the code we have written can be calibrated.

3.1 Introduction

The notion of some sort of smoothness parameter in the construction of a non parametric density estimate exists in multiple contexts. Even the construction of a simple histogram depends upon how many breaks are chosen to be drawn or displayed. The following example shows three histograms constructed from the same data set (of size 200). The dataset and the distribution it was sampled from are listed in Appendix D.

Example 3.1.1. The red curve overlaid is the true density from which the data were generated.

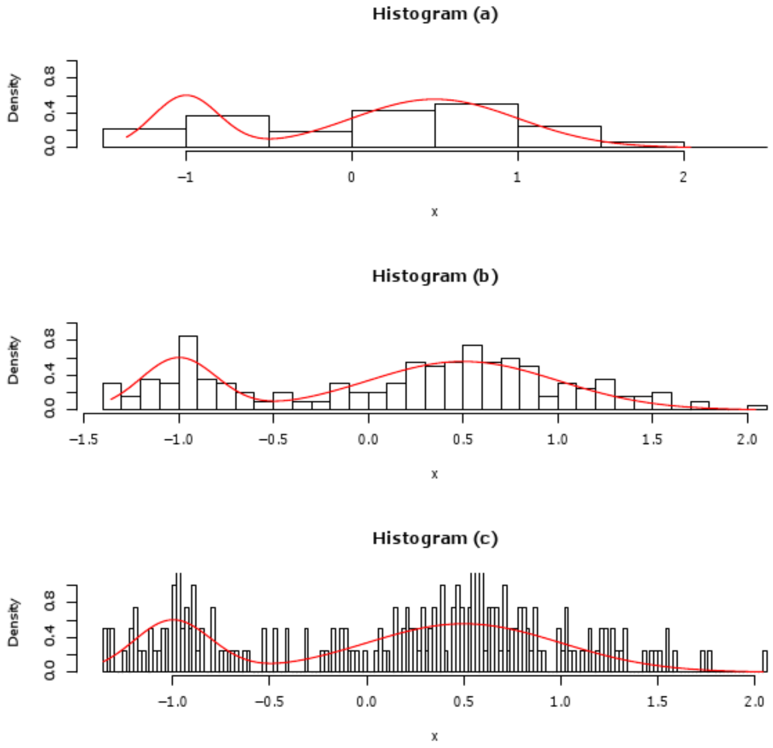


Figure 3.1: Three histograms of the same data set.

The effect of a poor choice of number of breaks may lead to the loss of key features in the data, or too much detail to focus upon. The simplest description of the data which does not lose sight of valuable features is ideal.

This simple example highlights a problem of interest which occurs in the popular kernel density estimation method, and in Section 3.3 we show it also exists in the use of applying normal mixture models to density estimation.

3.2 A multiscale approach in kernel density estimation

In this section we first provide some comments about well known aspects of the kernel density estimation method. We then describe a multiscale approach to bandwidth selection in the kernel density estimation context.

3.2.1 Kernel density estimation

Suppose we model $X_1, \ldots, X_n$ as iid with density f_0 (on $\mathbb{R}$), where f_0 is unknown. We may wish to use the data $X_1, \ldots, X_n$ to construct a function $\widehat{f}$ which in some sense is close to f_0. Let $h > 0$ and let a density k satisfy $k(-x) = -k(x)$ for $x \in \mathbb{R}$. The kernel density estimator $\widehat{f_h}$ of f_0 using the function k and the number $h > 0$ as presented in Silverman (1986) is

$$\widehat{f_h}(x) \quad = \quad \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right), \text{ for } x \in \mathbb{R}. \tag{3.1}$$

In the literature, the function k in (3.1) is called the kernel function and the number h is referred to as the bandwidth of the estimator $\widehat{f_h}$. Figure 3.2 contains some of the popular choices of k in practice.

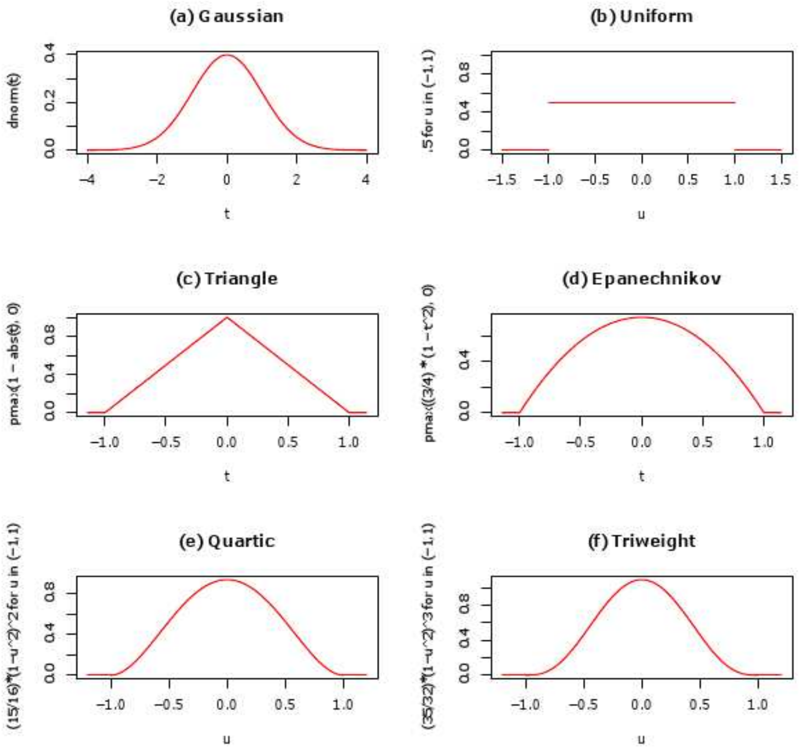


Figure 3.2: Several popular choices of kernel function.

As noted in Wand and Jones (1994), in kernel density estimation the choice of most unimodal kernels does little to alter the performance of the estimator. On the other hand the choice of bandwidth $h > 0$ can have a dramatic effect on the performance of the estimator produced.

Figure 3.3 shows different kernel density estimates constructed from the data set from Example 3.1.1, with either their choice of kernel varied or their choice of bandwidth varied. Both kernels were densities of a random variable with variance 1.

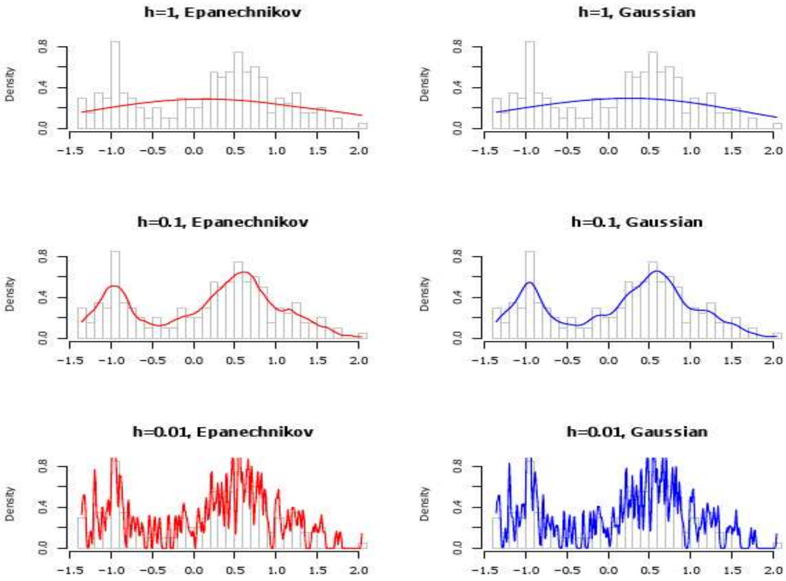


Figure 3.3: Different kernel density estimates with choice of bandwidth or choice of kernel varied.

In Figure 3.3, when the bandwidth is chosen to be $h = 1$, both the Epanechnikov kernel estimate and Gaussian kernel estimate produces a function with only one peak, which misses the multimodal nature of the true density (listed in Appendix D). On the other hand, when the bandwidth is chosen to be $h = 0.01$, the kernel density estimates (using either kernel function) displayed too many peaks which were only reflecting features present from that particular data set. The choice of $h = 0.1$ looks sensible in comparison to either of those extremes, however this choice is neither an obvious one nor a necessarily optimal one, in any sense.

In Section 3.3 we present a more general setting in which kernel density estimates can be viewed as one type of mixture density estimate out of a whole class of such estimates. The following simulation study is one which could be performed in the absence of the known theoretical results about kernel density estimators, and in Section 3.3 we mention a study of a similar flavour.

3.2.2 A multiscale approach to bandwidth selection

Suppose we were interested in estimating the underlying density from Example 3.1.1 via kernel density estimation, and wished to choose a sensible bandwidth h. We could embark upon a naive and computationally intensive procedure where we simply try a range of choices

$$h = 0.01, h = 0.02, \ldots, h = 0.99, h = 1$$

and examine each of the density estimates visually.

We would end up looking at plots (using a Gaussian kernel), the following are 11 examples out of the 100 density estimates constructed:

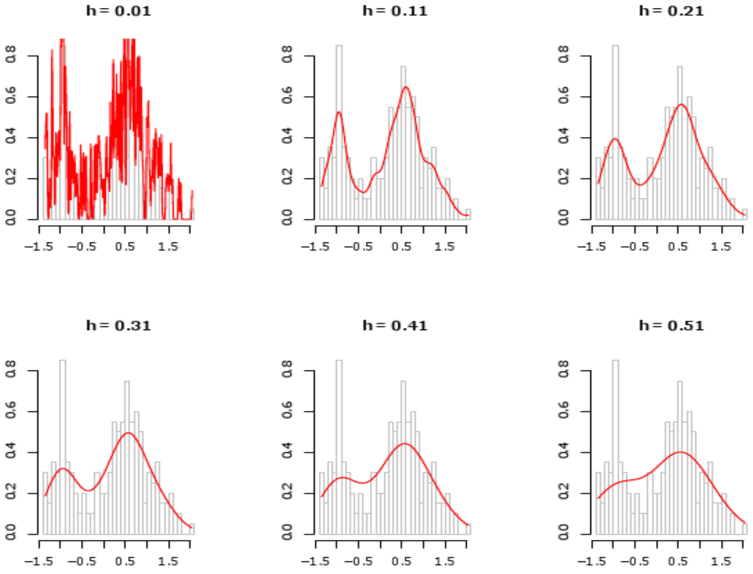


Figure 3.4: Kernel density estimates with varied bandwidth h.

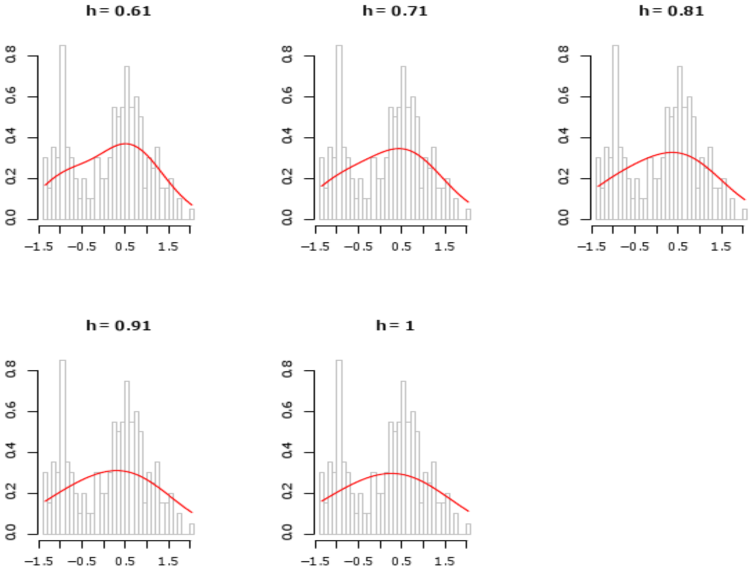


Figure 3.5: Kernel density estimates with varied bandwidth h.

On its own, this theoretically unsatisfying and computationally intense procedure still would not give us any precise idea of how to choose a bandwidth h. However once the introduction of a way to measure distance between the true density and an estimated density is added to this approach, we can plot the distances between the true density and the related estimators against the bandwidth h.

Example 3.2.1. Figure 3.6 shows four plots of an approximate measure of distance against choice of bandwidth. The data set $(X_1, \ldots, X_n)$ is from Example 3.1.1, and some of the density estimates are displayed in Figures 3.4 and 3.5.

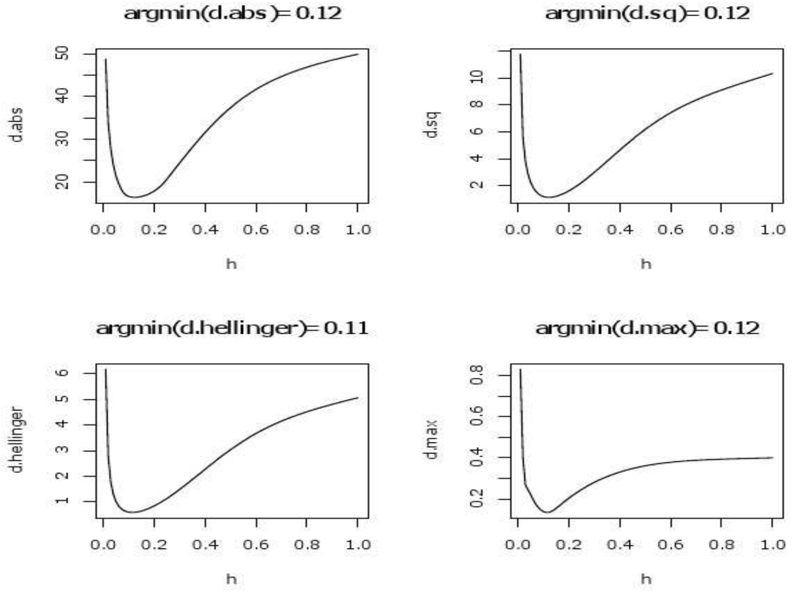


Figure 3.6: Clockwise from the top left: Plot (a), Plot (b), Plot (d), Plot (c).

The true density in this example is

$$f_0 = 0.3\phi_{-1,0.2} + 0.7\phi_{0.5,0.5}.$$

Let the names of the 100 density estimates from Figures 3.4 to 3.5 be $\widehat{f_h}$, with $h = 0.01, 0.11, 0.21, \ldots, 0.91, 1$ corresponding the the bandwidths used. Each of the estimates $\widehat{f_h}$ were calculated over the range of the data at equally spaced values $t_1 = \min_i(x_i), \ldots, t_{100} = \max_i(x_i)$.

The plots in Figure 3.6 were produced as follows:

Plot (a): To examine what $\int_{t_1}^{t_{100}} \left| f_0(x) - \widehat{f_h}(x) \right| dx$ looked like as a function of h, we calculated $\sum_{i=1} |f_0(t_i) - \widehat{f_h}(t_i)|$ (ignoring the constant factor of 100) for each h.

Plot (b): Similarly, to examine what $\int_{t_1}^{t_{100}} (f_0(x) - \widehat{f_h}(x))^2 dx$ looked like we calculated $\sum_{i=1}(f_0(t_i) - \widehat{f_h}(t_i))^2$ for each h.

Plot (c): We used $0.5 \sum_{i=1} \left(\sqrt{f_0(t_i)} - \sqrt{\widehat{f_h}(t_i)} \right)^2$ to approximate Hellinger distance as a function of h.

Plot (d): We used $\max_i\{|f_0(t_i) - \widehat{f}_h(t_i)|\}$ to approximate each $\sup_{t\in[t_1,t_{100}]} |f_0(t) - \widehat{f}_h(t)|$.

In each of the above plots, there existed a clear minimum distance between the true density and an estimate $\widehat{f}_h$ at some h in any of the senses listed, over the range of bandwidth choices $[0.01, 1]$. All four plots suggested that a bandwidth choice of approximately $h = 0.115$ would lead to an optimally close estimate $\widehat{f}_h$ to f_0 in terms of the various notions of distance we used.

We wished to replicate this procedure as a tool for examining bandwidth selection options in the more general setting of mixture density estimation. Unfortunately we did not develop any satisfactory computational technique to aid the choice of bandwidth in the more general case. However, in the next sections we present a basic theoretical bound for the choice of bandwidth.

Given the similarity between kernel density estimation and normal mixture approach it would have been natural to discuss whether standard methods such as cross validation or plug-in methods which are commonly used in kernel density estimation can be used in mixture density context as well. However due to the nature of mixture model work quickly becoming complicated for even two component mixtures, we have not focused on generalising these ideas. Nor have we found much work that has focused on this area. They would provide an interesting topic to further investigate.

3.3 An analogous bandwidth selection problem

In Chapter 1 we defined $\mathcal{F}_h$ to be given by

$$\mathcal{F}_h = \{f_{Q,h} : Q \text{ is on } \mathbb{R}\},$$

where

$$f_{Q,h} = \int \phi_{\mu,h} dQ(\mu).$$

In this section we present a nesting result about the classes $\{\mathcal{F}_h, h > 0\}$. We then show that the bandwidth selection problem described in Section 3.2 appears in the mixture setting described as well, and define a theoretical sense of optimality using the nesting property of $\{\mathcal{F}_h, h > 0\}$. We lastly show how kernel density estimates can be viewed as a special case of mixture density estimates.

3.3.1 A nesting property of $\mathcal{F}_h$

Lemma 3.3.1. *For $0 < s < t$, we have $\mathcal{F}_t \subseteq \mathcal{F}_s$.*

Proof. Recalling the definition of $\mathcal{F}_h^{(a)}$ in Definition 1.4.1, this result follows immediately since the elements of $\{\mathcal{F}_h^{(a)}, h > 0\}$ are easily shown to be nested. Suppose $0 < s < t$ and pick any $f \in \mathcal{F}_t$. From the definition of $\mathcal{F}_h^{(a)}$ there exists a mixing distribution Q on $\mathbb{R} \times [t, \infty)$ such that we can write f as

$$f = \int_{\mathbb{R}\times[t,\infty)} \phi_{\mu,\sigma} dQ(\mu, \sigma).$$

Since $0 < s < t$, $\mathbb{R} \times [t, \infty)$ is a subset of $\mathbb{R} \times [s, \infty)$, so define $\widetilde{Q}(A)$ (for $A \subseteq \mathbb{R} \times [s, \infty)$) via

$$\widetilde{Q}(A) = \begin{cases} Q(A) & , A \subseteq \mathbb{R} \times [t, \infty) \\ Q(A \cap (\mathbb{R} \times [t, \infty))) & , \text{otherwise} \end{cases}.$$

Then f can be written as

$$f = \int_{\mathbb{R}\times[s,\infty)} \phi_{\mu,\sigma} d\widetilde{Q}(\mu, \sigma),$$

and hence $f \in \mathcal{F}_s$. □

3.3.2 Possible values of h

Suppose $\sigma > 0$. Lemma 3.3.1 shows if $f \in \mathcal{F}_\sigma$ then $f \in \mathcal{F}_h$ for all $0 < h < \sigma$. Since it is possible to reexpress normal location-mixture densities in terms of another with a smaller component variance, the following question arises:

> "If $f \in \mathcal{F}_h$, does there exist a $\sigma > h$ such that $f \notin \mathcal{F}_\sigma$?".

The answer is yes. For any $f \in \bigcup_{h>0} \mathcal{F}_h$, we define the following.

Definition. For $f \in \bigcup_{h>0} \mathcal{F}_h$, let h_f be given by

$$h_f = \sup\{\sigma > 0 : f \in \mathcal{F}_\sigma\}.$$

We have the following lemma about h_f.

Lemma 3.3.2. *Suppose $f \in \mathcal{F}_\sigma$ for some $\sigma > 0$. Then*

$$h_f \leq \frac{1}{(\sup_{x\in\mathbb{R}} f(x)) \sqrt{2\pi}}. \tag{3.2}$$

(Note that any $f \in \mathcal{F}_\sigma$ will be bounded, so $h_f < \infty$ for any $f \in \mathcal{F}_\sigma$.)

Proof. We first bound $\sup_{x\in\mathbb{R}} f(x)$ by a quantity depending on σ. Since $f \in \mathcal{F}_\sigma$, there is a Q on $\mathbb{R}$ such that

$$\begin{aligned}
\sup_{x\in\mathbb{R}} f(x) &= \sup_{x\in\mathbb{R}} \int \phi_{\mu,\sigma}(x) dQ(\mu) \\
&\leq \int \left| \sup_{x\in\mathbb{R}} \phi_{\mu,\sigma}(x) \right| dQ(\mu) \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{\mathbb{R}} dQ(\mu) \\
&= \frac{1}{\sigma\sqrt{2\pi}}. \qquad (3.3)
\end{aligned}$$

Let $y = \sup_{x\in\mathbb{R}} f(x)$ and consider the normal density $\phi_{0,s}$ with maximum height $y/2$. We have

$$y/2 = \max_{x\in\mathbb{R}} \phi_{0,s}(x) = \frac{1}{s\sqrt{2\pi}}. \qquad (3.4)$$

We now prove Lemma 3.3.2 by contradiction. Suppose for all $h > 0$, $h_f > h$. Then $f \in \mathcal{F}_s$ where $s = \frac{2}{y\sqrt{2\pi}}$. Using the same argument in (3.3) we arrive at

$$y = \sup_{x\in\mathbb{R}} f(x) \leq \frac{1}{s\sqrt{2\pi}} = y/2.$$

Since $f \geq 0$ almost everywhere, we conclude $f = 0$ almost everywhere, and thus $\int f(x)dx = 0$, contradicting the fact that f is a density. Hence there exists an $s > 0$ such that $f \notin \mathcal{F}_s$, so $h_f \leq s$.

To show the remainder of this lemma, we choose a standard deviation s in (3.4) which gives a density with maximum height $y(1-\epsilon)$ for some $\epsilon \in (0,1)$ (instead of $\frac{1}{2}$ of y) and apply the same argument above to arrive at

$$h_f \leq \frac{1}{(\sup_{x\in\mathbb{R}} f(x))(1-\epsilon)\sqrt{2\pi}}.$$

Since our choice of ϵ was arbitrary, we get

$$h_f \leq \frac{1}{(\sup_{x\in\mathbb{R}} f(x))\sqrt{2\pi}}.$$

□

We can use Lemma 3.3.2 to get an idea of what h_f is for any $f \in \bigcup_{h>0} \mathcal{F}_h \setminus \bigcap_{h>0} \mathcal{F}_h$. Using this chapter's example density

$$f = 0.3\phi_{-1,0.2} + 0.7\phi_{\frac{1}{2},\frac{1}{2}},$$

we know that $f \in \mathcal{F}_{0.2}$, and therefore $0.2 \leq h_f$. Since $\max_{x\in\mathbb{R}} f(x) = f(-1)$, we get

$$0.2 \leq h_f \leq \frac{1}{f(-1)\sqrt{2\pi}} \approx 0.6598253.$$

A special case: single component normals.

When the density f is a single normal, the bound in Lemma 3.3.2 allows us to calculate h_f directly. For example, if $f = \phi$, then $f \in \mathcal{F}_1$ and therefore $h_f \geq 1$. Lemma 3.3.2 gives us $h_f \leq 1$. Note that $f \in \mathcal{F}_h$ for $h \leq 1$, but $f \notin \mathcal{F}_h$ for $h > 1$.

An interval/value of interest

For any $f \in \bigcup_{h>0} \mathcal{F}_h \setminus \bigcap_{h>0} \mathcal{F}_h$, Lemmas 3.3.1 and 3.3.2 together tell us the following two intervals

$$I_1 = (0, h_f],$$

and

$$I_2 = \left(\frac{1}{(\sup_{x\in\mathbb{R}} f(x))\sqrt{2\pi}}, \infty\right)$$

have the following properties:

Lemma	Lemma 3.3.1	Lemma 3.3.2
Interval	$I_1 = (0, h_f]$	$I_2 = \left(\frac{1}{\left(\sup_{x\in\mathbb{R}} f(x)\right)\sqrt{2\pi}}, \infty\right)$
Property	$f \in \mathcal{F}_h$ for any $h \in I_1$	$f \notin \mathcal{F}_h$ for any $h \in I_2$

To our knowledge, little is known about what h_f for a given f could be. Although our Lemma 3.3.2 provides a theoretical upper bound for it, in practical senses it is not very useful or enlightening. The following section provides an example of why this quantity might be useful to estimate or investigate.

3.3.3 Amounts of smoothness in mixture density estimation

Suppose we have data $X_1, \ldots, X_n$ which we wish to model being iid from density f, where $f \in \mathcal{F}_{h_f}$ is unknown. For any $h \in (0, h_f]$, Lemma 3.3.1 implies there exists a mixing distribution Q on $\mathbb{R}$ such that we can write f as

$$f = \int \phi_{\mu,h} dQ(\mu).$$

To estimate f (with such a bandwidth h in mind) it would make sense to produce an estimate $\widehat{Q}$ of Q first and then arrive at

$$\widehat{f_h} = \int \phi_{\mu,h} d\widehat{Q}(\mu). \tag{3.5}$$

Algorithms to calculate such estimates $\widehat{Q}$ (or $\widehat{f_h}$) are widely available; we will discuss several of them in Sections 3.5 and 3.6. We will call estimates such as (3.5) mixture density estimates with bandwidth h, or more specifically, normal location-mixture density estimates with bandwidth h.

However, for $h > h_f$, there is no Q on $\mathbb{R}$ such that $\int \phi_{\mu,h} dQ(\mu)$ is the same as f. It is not sensible to find an estimate of f based on the approach described in (3.5).

In either case, the following questions are equivalent:

> "For which values of $h > 0$ is there a Q on $\mathbb{R}$ such that f can be written as $\int \phi_{\mu,h} dQ(\mu)$?",

and

> "What is h_f?".

For unknown f, the quantity h_f is also unknown, so a given estimate $\widehat{f_h}$ of f may not even have been calculated over a large enough space $\mathcal{F}_h$ of densities. Choice of bandwidth h in (3.5) is thus an important problem to consider in mixture density estimation.

We now describe the behaviour of mixture density estimates in two scenarios. The first scenario is when the choice of bandwidth $h > 0$ is decreased to a small enough number so that $h < h_f$. The second is when the choice of bandwidth $h > 0$ is chosen to be large, such that $h > h_f$.

h is chosen to be too small

When the bandwidth h of a mixture density estimate is very small relative to h_f, the estimate tends to look more 'wiggly' than the original density f. While choosing h small enough so that $\mathcal{F}_h$ contains f is a sensible approach, there is a tradeoff between how small h can be chosen. The time taken to compute estimates tends to increase as h comes closer to 0, and unnecessary features in the estimated density appear more often as the bandwidth is reduced towards 0.

h is chosen to be too large

On the other hand, if the bandwidth h is chosen to be too large, the estimate loses too much detail and looks 'oversmooth'. A larger choice of h tends to result in a faster computation of the density estimate $\widehat{f}_h$, however the risk of estimating over a set $\mathcal{F}_h$ which does not contain the true density f is higher as h increases.

As indicated above, we do not wish to choose a bandwidth h to be too small or large, as in either case problems arise. This analogous behaviour of the mixture density bandwidth parameter h to that of the bandwidth parameter in the kernel density estimation context (from Section 3.2) is the motivator for our choice of notation h and terminology. In fact, we next show that kernel density estimation can be interpreted as a special case of mixture density estimation.

3.3.4 A special case of mixture density estimation

Recall the kernel density estimation setting, and suppose we are using the Gaussian kernal ϕ. The random variables $X_1, \ldots, X_n$ with unknown density f are observed to be $x_1, \ldots, x_n$. The kernel density estimate of f with bandwidth h is given by

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \phi\left(\frac{x - X_i}{h}\right).$$

Suppose now we were to produce a mixture density estimate $\widehat{f}$ of f with (mixture) bandwidth h. The mixture density (if $h < h_f$) would be expressible as

$$f = \int \phi_{\mu,h} dQ(\mu)$$

for some Q on $\mathbb{R}$. Instead of estimating Q, suppose we assumed Q placed equal probability on each of the n observed x_i as a choice of component mean; $Q(\{x_i\}) = \frac{1}{n}$ for $i = 1, \ldots, n$, and since it is our guess for Q, let the name of this assumed mixing distribution be $\widehat{Q}$. Since $\widehat{Q}$ is discrete with the n mass points $x_1, \ldots, x_n$ and probabilities $\frac{1}{n}, \ldots, \frac{1}{n}$, the mixture density estimate is given by

$$\begin{aligned}\widehat{f}(x) &= \int \phi_{\mu,h}(x) d\widehat{Q}(\mu) \\ &= \sum_{i=1}^{n} \frac{1}{n} \phi_{x_i,h}(x) \\ &= \frac{1}{nh} \sum_{i=1}^{n} \phi\left(\frac{x - x_i}{h}\right),\end{aligned}$$

which is just the observed value of the kernel density estimator $\widehat{f}_h$. Thus kernel density estimation can be viewed as a special case of mixture density estimation where the mixing distribution Q is estimated by a specific distribution $\widehat{Q}$.

In general we may wish to relax this estimator $\widehat{Q}$ of Q to be the maximiser of the log likelihood over a larger class of distributions on $\mathbb{R}$. The very useful and elegant results from Lindsay (1983) allow us to do so via algorithms such as the ones mentioned in Section 3.5.

3.4 A multiscale approach in normal mixture density estimation

Lemma 3.3.2 suggests a way to estimate an upper bound for h_f, however we do not know of any way to estimate a lower bound, which would be more useful in practice for reducing computational times in simulation studies such as the one we have performed.

Based upon the similarities between Kernel density estimation bandwidth selection and the mixture density estimation bandwidth selection problem, we ran a simulation study using the multiscale approach to examine whether we could produce a data driven technique for getting some intuition about h_f.

Multiscale approach

The procedure we used in our simulation study was as follows.

1. We generated data $x_1, \ldots, x_n$ from a known mixture density.

2. For a sensible set of $\{h_1, h_2, \cdots > 0\}$, we calculated the fitted density $\widehat{f}_{h_j}, j = 1, \ldots$ according to (3.5), via NPMLE.

3. We examined the goodness of fit between $\widehat{f}_h$ and f as a function of h to see if we could mimic a technique similar to the one illustrated by Figure 3.6.

Unfortunately our simulation study provided us with little intuition towards how a sensible interval or estimate for h_f could be produced in practice, so we do not describe the details here. It would be interesting to see future research shed light towards this problem.

In the following section, we describe the algorithms we used in the context of this multiscale approach. The bandwidth parameter h will be fixed in the next section.

3.5 Outline of algorithms used

In this section we describe three algorithms from the literature which are useful for the multiscale approach mentioned in Section 3.3.

Recall from Chapter 1 that Lindsay (1983) showed the log likelihood

$$\sum_{i=1}^{n} \int \phi_{\mu,h}(X_i) dQ(\mu)$$

had a maximiser $\widehat{Q}$ which exists, is unique, is a discrete distribution, and has up to n mass points. Also recall that the following statements were equivalent:

1. $\widehat{Q}$ maximises $\sum_{i=1}^{n} \int \phi_{\mu,h}(X_i) dQ(\mu)$
2. $\widehat{Q}$ minimizes $\sup_{\theta \in \mathbb{R}} D_Q(\theta)$, where $D_Q(\theta) = \sum_{i=1}^{n} \left\{ \frac{\phi_{\theta,h}(X_i)}{\int \phi_{\mu,h}(X_i) dQ(\mu)} - 1 \right\}$
3. $\sup_{\theta \in \mathbb{R}} D_{\widehat{Q}}(\theta) = 0$,

and moreover, the mass points of $\widehat{Q}$ are the values $\widehat{\theta}_1$, $\widehat{\theta}_2$, …, $\widehat{\theta}_K$ satisfying

$$D_{\widehat{Q}}(\widehat{\theta}_i) = 0, \text{ for } i = 1, 2, \ldots, K.$$

These results allow the problem of finding

$$\widehat{Q} = \arg\max_{Q} \sum_{i=1}^{n} \int \phi_{\mu,h}(X_i) dQ(\mu)$$

to be transformed from a theoretical problem to an implementable computational problem. A particular consequence of these results is that algorithms such as the Expectation Maximisation (EM) algorithm (Laird, 1978) may be applied to find the K mass points of $\widehat{Q}$ along with their associated probabilities, and thus be used to calculate density estimates of the form given by

$$\widehat{f}_h(x) = \int \phi_{\mu,h}(x) d\widehat{Q}(\mu).$$

The discreteness of the NPMLE $\widehat{Q}$ can be used to suggest a way to implement a method of computing density estimates $\widehat{f}_h$ in practice. Algorithms such as the Intra Simplex Direction Method (ISDM) by Lesperance and Kalbfleisch (1992) make use of Lindsay's characterisation of $\widehat{Q}$ in terms of the minimizer of $\sup_{\theta \in \mathbb{R}} D_Q(\theta)$ to provide a hill-search type method of computing such estimates.

The next subsection discusses how density estimates can be calculated via NPMLE using the EM algorithm.

3.5.1 The EM algorithm

In this subsection we outline the Expectation Maximisation (EM) algorithm as it can be used in the context of applying Non Parametric Maximum Likelihood Estimation (NPMLE) to the problem of mixture density estimation for a fixed choice of (mixture) bandwidth h.

In our code we iteratively applied the EM algorithm to estimate $\widehat{Q}$ for a non random number of mass points k, where k was a value chosen by the steps within the Intra Simplex Direction Method (ISDM), presented by Lesperance and Kalbfleisch (1992).

Alternatively we could regard $\widehat{Q}$ to have exactly n mass points (some points with 0 probability) and apply the EM algorithm directly to the problem of maximising the log likelihood, however in practice the k chosen by the ISDM is smaller than n and leads to a faster computational EM algorithm run time. Assuming n mass points also garuntees our model to be from the overspecified scenario mentioned in Chen (1995).

Suppose we are iteratively computing estimates $\widehat{Q}_1, \widehat{Q}_2, \ldots$ of the NPMLE $\widehat{Q}$ of the model's true mixing distribution Q. When the estimated density associated with any estimate $\widehat{Q}_j$ of $\widehat{Q}$ has k mass points $(\mu_1^{(j)}, \ldots, \mu_k^{(j)}$ with probabilities $p_1^{(j)}, \ldots, p_{k-1}^{(j)})$, the estimated mixture density would be

$$\int \phi_{\mu,h} d\widehat{Q_j}(\mu) = \sum_{\ell=1}^{k} p_\ell^{(j)} \phi_{\mu_\ell^{(j)},h}.$$

If we regarded the n observations $X_1, \ldots, X_n$ as being drawn from a mixture density with a discrete mixing distribution (like $\widehat{Q}_j$ above), we could consider them to be the observed parts of the bivariate random variables (X_i,M_i), where the random component means M_i were unobservable.

In such a context, since the conditional random variables $X_i|M_i = \mu$ are $N(\mu, h^2)$, the joint density of any X_i and M_i is hence

$$f_{X_1M_1}(x, \mu_1) = P(M_1 = \mu_j)f_{X_1|M_1=\mu_j}(x|\mu_j) = P(M_1 = \mu_j)\phi_{\mu_j,h}(x).$$

The log likelihood function of a model with random variables $(X, M) = X_1, \ldots, X_n, M_1, \ldots, M_n)$, parameters $\theta = (\mu_1, \ldots, \mu_k, p_1, \ldots, p_{k-1})$ is

$$\begin{aligned} \ell(\theta) \quad &:= \quad \log\left\{\prod_{i=1}^{n} f_{X_1M_1}(X_i, M_i)\right\} \\ &= \quad \sum_{i=1}^{n}\log\left\{\prod_{j=1}^{k}(p_j\phi_{M_i,h}(X_i))^{\mathbb{I}(M_i=\mu_j)}\right\}, \end{aligned}$$

where

$$\mathbb{I}(M_i = \mu_j) = \begin{cases} 1 & , \text{ if } M_i = \mu_j \\ 0 & , \text{ if } M_i \neq \mu_j \end{cases}$$

and $p_k = 1 - \sum_{j=1}^{k-1} p_j$. Note that we want each $p_j \geq 0$. We can rearrange $\ell(\theta)$ to

$$\ell(\theta; X, M) = \sum_{i=1}^{n}\sum_{j=1}^{k}\mathbb{I}(M_i = \mu_j)\left\{\log(p_j) + \log(\phi_{\mu_j,h}(X_i))\right\}.$$

Suppose we currently estimate θ to be $\theta_0 = (\nu_1, \ldots, \nu_k, q_1, \ldots, q_{k-1})$ and the observed values of $X = X_1, \ldots, X_n$ are $x = x_1, \ldots, x_n$. The EM algorithm's expectation step computes the EM log likelihood $\ell_{EM}(\theta|\theta_0)$, which is the expected value of $\ell(\theta)$ with respect to the conditional distribution of M given $X = x$ under the current estimate θ_0 of θ, to be

$$\begin{aligned} \ell_{EM}(\theta|\theta_0) \quad &= \quad \mathbb{E}_{M|X=x,\theta_0}(\ell(\theta; X, M)) \\ &= \quad \mathbb{E}_{M|X=x,\theta_0}(\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbb{I}(M_i = \mu_j)\left\{\log(p_j) + \log(\phi_{\mu_j,h}(X_i))\right\}) \\ &= \quad \sum_{i=1}^{n}\sum_{j=1}^{k}\mathbb{E}_{M_i|X_i=x_i,\theta_0}(\mathbb{I}(M_i = \mu_j))\left\{\log(p_j) + \log(\phi_{\mu_j,h}(x_i))\right\}. \end{aligned} \tag{3.6}$$

The conditional expectation $\mathbb{E}_{M_i|X_i=x_i}(I(M_i = \mu_j))$ is

$$\mathbb{E}_{M_i|X_i=x_i}(I(M_i = \mu_j)) = P(M_i = \mu_j | X_i = x_i) = \frac{p_j \phi_{\mu_j,h}(x_i)}{\sum_{j=1}^{k} p_j \phi_{\mu_j,h}(x_i)},$$

so under the current estimate θ_0 of θ, that is

$$\mathbb{E}_{M_i|X_i=x_i,\theta_0}(\mathbb{I}(M_i = \mu_j)) = \frac{q_j \phi_{\nu_j,h}(x_i)}{\sum_{j=1}^{k} q_j \phi_{\nu_j,h}(x_i)}.$$

Let $\pi_{j|i}$ be given by

$$\pi_{j|i} := \frac{q_j \phi_{\nu_j,h}(x_i)}{\sum_{j=1}^{k} q_j \phi_{\nu_j,h}(x_i)}.$$

From (3.6) we can write the EM log likelihood as

$$\ell_{EM}(\theta|\theta_0) = \sum_{i=1}^{n} \sum_{j=1}^{k} \pi_{j|i} \log(p_j) + \sum_{i=1}^{n} \sum_{j=1}^{k} \pi_{j|i} \log(\phi_{\mu_j,h}(x_i)). \qquad (3.7)$$

Since the possible values of p_j are constrained by $\sum_{j=1}^{k} p_j = 1$, we use the method of Lagrange multipliers to maximise (3.7) with respect to $p_1, \ldots, p_{k-1}$. The possible values of $\mu_1, \ldots, \mu_k$ are unconstrained so we can simply maximize (3.7) with respect to $\mu_1, \ldots, \mu_k$. The EM estimates for θ based on the initial estimate θ_0 are thus

$$\widehat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \pi_{j|i}, \qquad (3.8)$$

$$\widehat{\mu}_j = \frac{\sum_{i=1}^{n} x_i \pi_{j|i}}{n \widehat{p}_j}. \qquad (3.9)$$

The following property of the EM algorithm provides a way to implement it in practice.

Remark 3.5.1. Suppose θ_0 and θ_1 are two estimates of θ. If

$$\ell_{EM}(\theta_1|\theta_0) \geq \ell_{EM}(\theta_0|\theta_0),$$

then

$$\sum_{i=1}^{n} \log(f_{X_1}(x_i|\theta_1)) \geq \sum_{i=1}^{n} \log(f_{X_1}(x_i|\theta_0)), \qquad (3.10)$$

where $f_{X_1}(\cdot|\theta)$ is the marginal density of X_1 in this bivariate interpretation of our observations.

The proof of Remark 3.5.1 is in Appendix C.

Note our original aim in the multiscale method from Section 3.3 was to estimate the parameters θ of the NPMLE $\widehat{Q}$. The log likelihood

$$\ell(Q) = \sum_{i=1}^{n} \log \left(\int \phi_{\mu,h}(X_i) dQ(\mu) \right)$$

when restricted to being defined over a class of discrete distributions Q could be regarded as

$$\ell(\theta) = \sum_{i=1}^{n} \log(f_{X_1}(X_i|\theta)), \tag{3.11}$$

where $f_{X_1}(\cdot|\theta)$ is given in Remark 3.5.1. Thus if we start with an initial estimate θ_0 of θ, and then apply the EM algorithm to produce

$$\theta_1 := \arg\max_{\theta} \ell_{EM}(\theta|\theta_0)$$

(which is by definition $\geq \ell_{EM}(\theta_0|\theta_0)$), then the log likelihood given by (3.11) at θ_1 is at least as big as the log likelihood at θ_0.

Thus iteratively applying the EM algorithm to our mixture distribution estimation problem within our multiscale bandwidth selection approach for our mixture density estimation application to produce estimates $\theta_1, \theta_2, \ldots$ will always produce a non decreasing sequence of log likelihood values $\ell(\theta_0) \leq \ell(\theta_1) \leq \ell(\theta_2) \leq \ldots$. In the code for our simulation study, we implemented the EM estimates (3.8) in the main ISDM algorithm to estimate the probabilities associated with the means chosen by the ISDM, rather than implementing a full blown EM to estimate $\widehat{Q}$ directly. One issue with the ISDM as mentioned by Lesperance and Kalbfleisch (1992) is that the estimated distribution $\widehat{Q}$ tends to be hairy, though the estimated density was good.

3.5.2 The ISDM

In this subsection we outline how we implemented the Intra Simplex Direction Method (ISDM) from Lesperance and Kalbfleisch (1992) to address our mixture density estimation problem.

Recall the density in our mixture model is of the form

$$\int \phi_{\mu,h} dQ(\mu)$$

and we wish to estimate Q by the NPMLE $\widehat{Q}$ as defined by Lindsay (1983). The ISDM is based upon the theorem by Lindsay (1983) which states that the following are equivalent

1. $\widehat{Q}$ maximises $\ell(Q) = \sum_{i=1}^{n} \int \phi_{\mu,h}(X_i) dQ(\mu)$
2. $\widehat{Q}$ minimizes $\sup_{\theta\in\mathbb{R}} D_Q(\theta)$, where $D_Q(\theta) = \sum_{i=1}^{n} \left\{ \frac{\phi_{\theta,h}(X_i)}{\int \phi_{\mu,h}(X_i) dQ(\mu)} - 1 \right\}$
3. $\sup_{\theta\in\mathbb{R}} D_{\widehat{Q}}(\theta) = 0$,

and moreover, the mass points of $\widehat{Q}$ are the values $\widehat{\theta}_1$, $\widehat{\theta}_2$, $\ldots$, $\widehat{\theta}_K$ satisfying

$$D_{\widehat{Q}}(\widehat{\theta}_i) = 0, \text{ for } i = 1, 2, \ldots, K.$$

The ISDM is described as follows.

0. Start with an initial estimate Q_j of Q, with iteration counter $j := 1$.
1. Compute all the local maxima of $D_{Q_j}(\theta)$. Suppose the k_j values calculated are $\theta_1, \ldots, \theta_{k_j}$.
   - If $\max_{s=1,\ldots,k_j} \theta_s = 0$, stop.
2. Compute the proportions $p_0, p_1, \ldots, p_{k_j}$ which maximise

   $$\ell(p_0 Q_j + \sum_{s=1}^{k_j} p_s \delta_{\theta_s}),$$

   subject to the constraints $\sum_{s=0}^{k_j} p_s = 1$ and $p_s \geq 0$ for all $s = 0, \ldots, k_j$.
3. The new estimate Q_{j+1} becomes

   $$Q_{j+1} = p_0 Q_j + \sum_{s=1}^{k_j} p_s \delta_{\theta_s},$$

   and $j := j + 1$. Return to Step 1.

3.6 Discussion of algorithms

In this section we discuss how we calibrated the settings used in the ISDM with its internal EM step. The code we used is listed in Appendix B.

Since the stopping criterion for Step 1 of the ISDM is:

$$\text{“If } \max_{s=1,\ldots,k_j} \theta_s = 0, \text{ stop”},$$

we implemented the following criterion for our code to check whether:

$$\max_{s=1,\ldots,k_j} \theta_s < \texttt{epsilon1},$$

where `epsilon1` was specified as some small positive number, for example 0.1. In order to prevent the code from potentially never stopping, we added a counter which stopped the algorithm if more than `giveUp` iterations of the ISDM were performed.

Figure 3.7 displays a graphical comparison of $\max_s \theta_s$ over (up to) `giveUp` iterations of the ISDM, with `giveUp`$= 50$ or `giveUp`$= 100$. The ISDM in this comparison was applied to the same data set with the same vector of initial density estimate values. The only difference is the variation of the parameter `giveUp`.

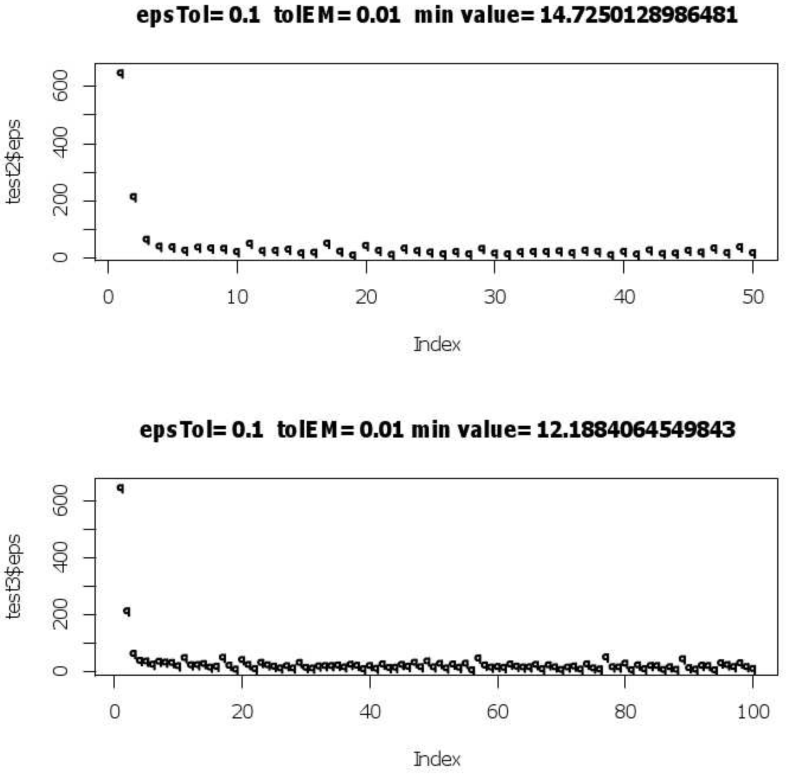


Figure 3.7: Above: $\max_s \theta_s$ over 50 iterations of the ISDM. Below: $\max_s \theta_s$ over 100 iterations of the ISDM

As can be seen in the above figure, iterating the steps in the ISDM 50 or 100 times did not tend to dramatically change the distance between $\max_s \theta_s$

and epsilon1$= 0.1$. The smallest value of $\max_s \theta_s$ in both scenarios were only about 14.7250 and 12.1884, which seemed rather larger than epsilon1$= 0.1$, let alone 0.

However, by decreasing the parameter associated with the stopping condition of the EM algorithm in step 2 of the ISDM, tolEM, we found the code produced values of $\max_s \theta_s$ which were closer to 0. This is illustrated in Figure 3.8.

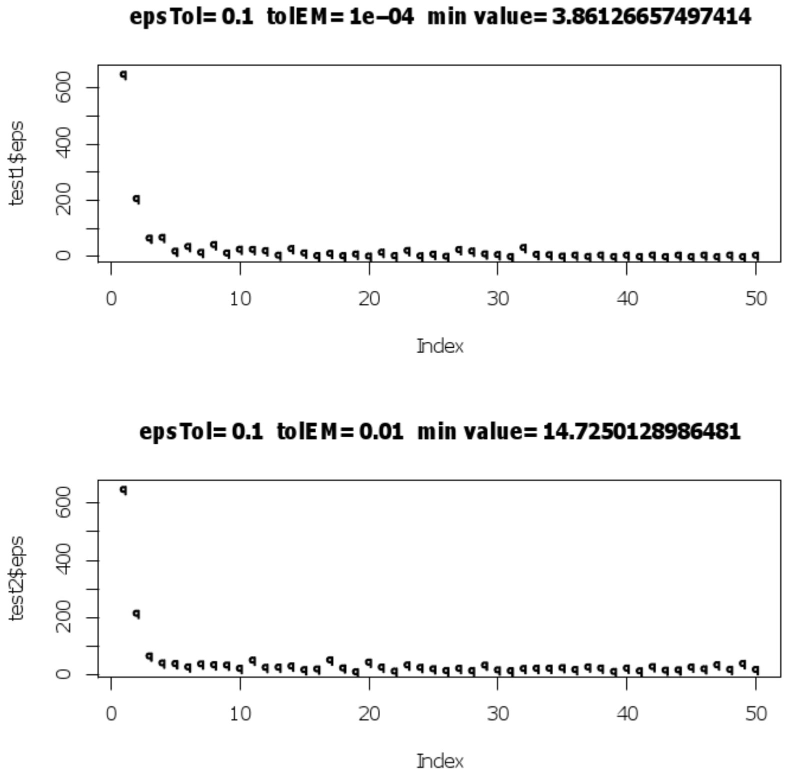


Figure 3.8: Varying the stopping condition of the EM algorithm affects how quickly the stopping condition of the overall ISDM is approached.

In all cases illustrated by Figures 3.7 and 3.8, the dataset used to compare the ISDM was of size $n = 1000$. The next figure contains the same plot from the top of Figure 3.8, compared to the plot produced via the same choice of stopping condition parameters, by applying the ISDM to a dataset of size $n = 10000$. Both datasets were generated from the same mixture density.

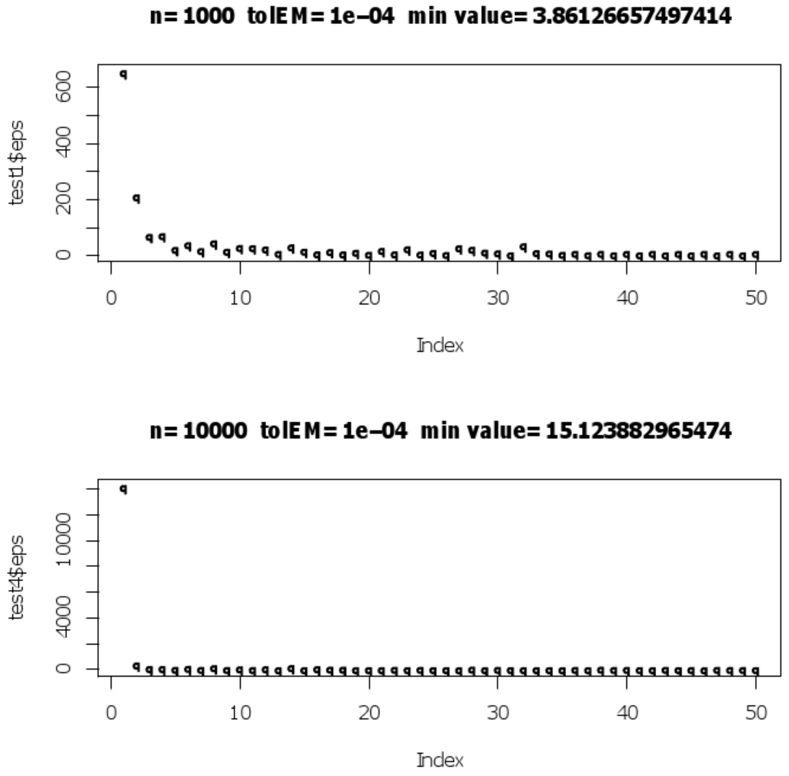


Figure 3.9: Calibration of stopping condition parameters seems dependent upon the dataset size.

As Figures 3.7, 3.8 and 3.9 suggest, the global stopping criterion

$$\sup_s \theta_s = 0$$

is approached at speeds depending upon the size of the dataset, as well as the EM stopping tolerance value `tolEM`. A choice of a maximum of `giveUp`$= 50$ iterations of the ISDM seemed more than enough in all cases, and in all the above cases the closest any $\sup_s \theta_s$ value came to 0 was unfortunately larger than our choice of `epsilon1`$= 0.1$.

The main parameter of interest regarding the calibration of the ISDM settings in our code was thus the tolerance `tolEM`. This parameter appeared in the following condition in our code within our EM step in step 2 of the ISDM:

While $\ell(Q_j) - \ell(Q_{j-1}) >$ `tolEM`,
iterate the EM algorithm.

In the above condition, the index j refers to the EM iterations within step 2, and the quantity $\ell(Q_j)$ refers to the value of the log likelihood function at the j^{th} EM algorithm estimate of the weights. Note that $\ell(Q_j) - \ell(Q_{j-1}) > 0$, by (C.1).

Unfortunately, while the distance between the NPMLE $\widehat{Q}$ and the estimates $\widehat{Q}_{ISDM}$ produced by the ISDM seems to be sensitive to the choice of `tolEM`, the speed of the ISDM is also sensitive to `tolEM`.

Figure 3.10 shows the value of the log likelihood over (what turned out to be) 79 applications of the EM algorithm in one iteration of Step 2 of the ISDM, with the EM stopping tolerance set to `tolEM`= 0.0001. Figure 3.11 is a closer look at Figure 3.10 with the first 10 values ommitted.

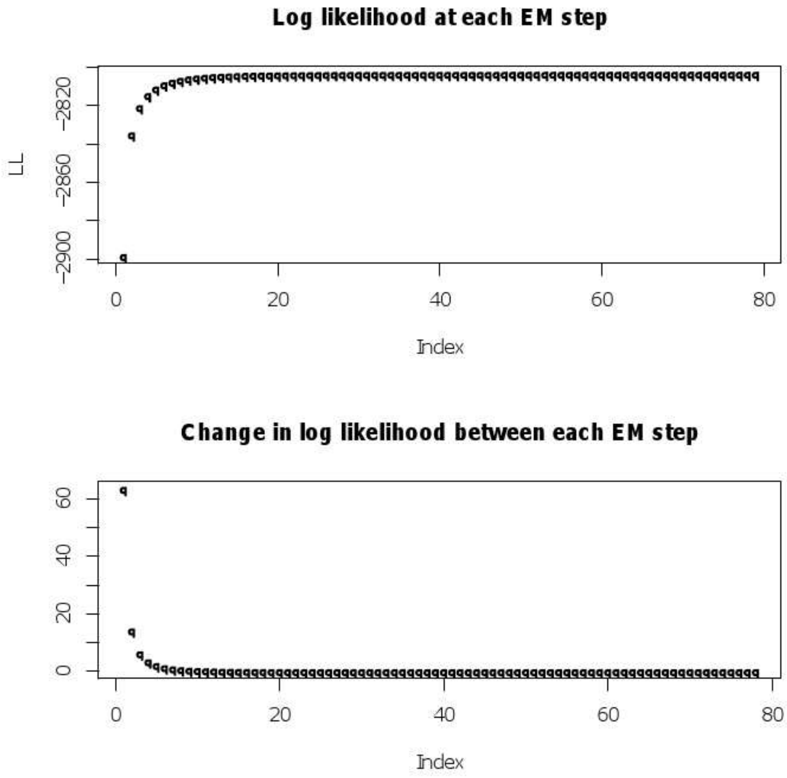


Figure 3.10: Change in log likelihood between each EM step quickly becomes slow.

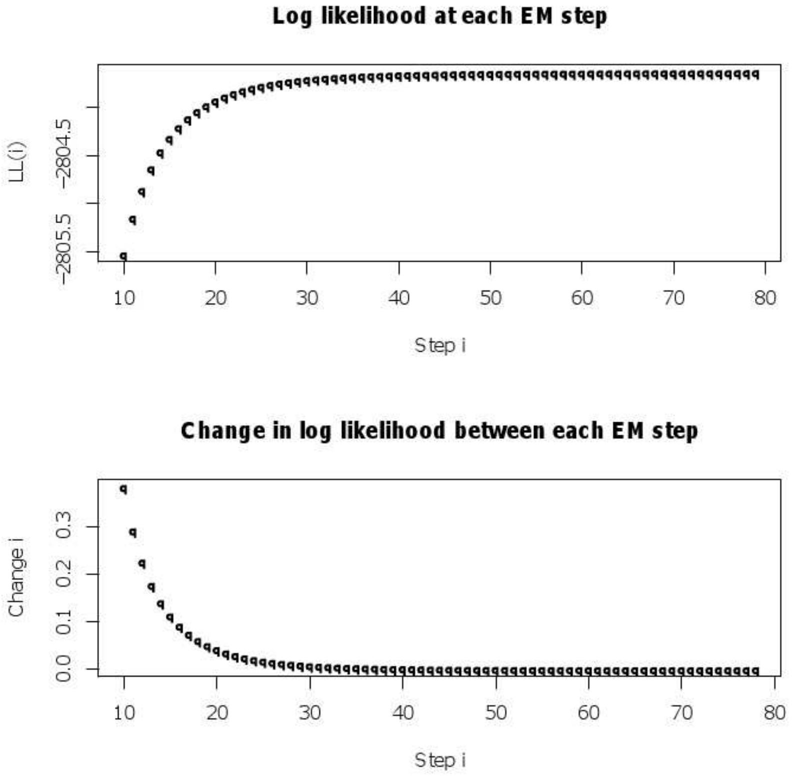


Figure 3.11: Same figure as 3.10, with the first 10 values ommitted.

As visualised by the above figures, convergence of the EM algorithm can become quite slow when change in the log likelihood is required to be less than `tolEM`$= 0.0001$. When this stopping parameter was changed to `tolEM`$= 0.001$ (in this same example), the EM algorithm was only applied 49 times, yet the final calculated values of the log likelihood were -2803.631 (`tolEM`$= 0.0001$) and -2803.64 (`tolEM`$= 0.001$).

This slowness of convergence of the EM algorithm suggested that the difference between choosing `tolEM`$= 0.0001$ versus `tolEM`$= 0.001$ did not seem to significantly impact how close the EM algorithm came to converging.

We wished to calibrate the choice of `tolEM` to be large enough to produce a satisfactory computational speed, yet small enough to produce density estimates $\int \phi_{\mu,h} d\widehat{Q}_{ISDM}(\mu)$ close enough to

$$\int \phi_{\mu,h} d\widehat{Q}(\mu).$$

Fortunately, as mentioned in Lesperance and Kalbfleisch (1992), the distribution estimates $\widehat{Q}_{ISDM}$ produced by the ISDM tend to be rougher than the density estimates.

In the following example, we applied the ISDM to the same dataset (of size $n = 1000$) to produce two density estimates, f_1 and f_2. The first estimate f_1 was produced with the parameter `tolEM`$= 0.0001$ and the second estimate f_2 was produced with `tolEM`$= 0.001$. The top plot in the above figure actually displays f_1 and f_2, but the estimates are so similar that it looks like only one curve. We have plotted the difference $f_1 - f_2$ underneath. Figure 3.12 shows the two estimated densities (at a reasonable choice of bandwidth $h = 1$) look similar when `tolEM`$= 0.0001$ and `tolEM`$= 0.001$. The value h_f in this example was $\sqrt{2}$.

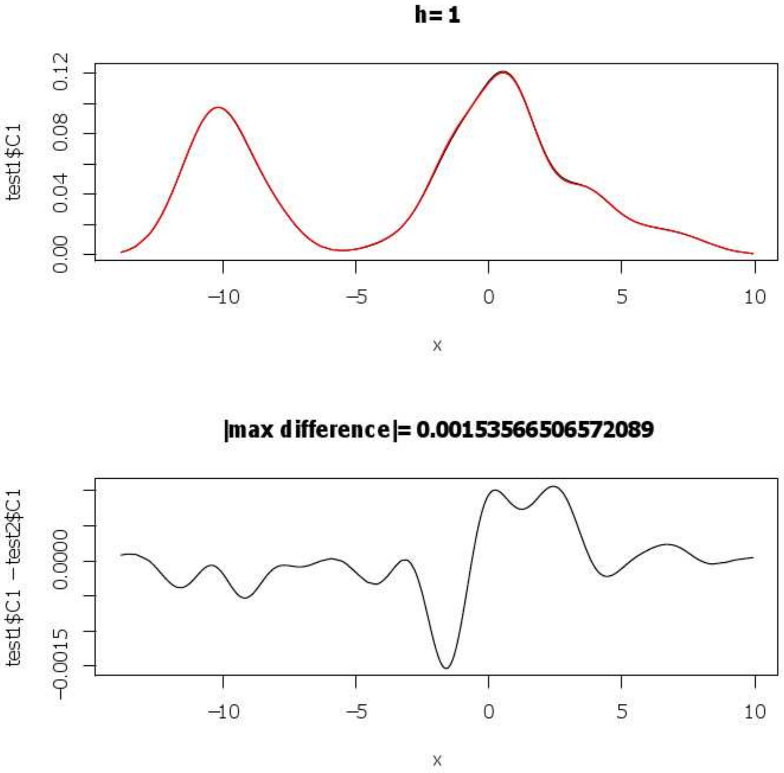
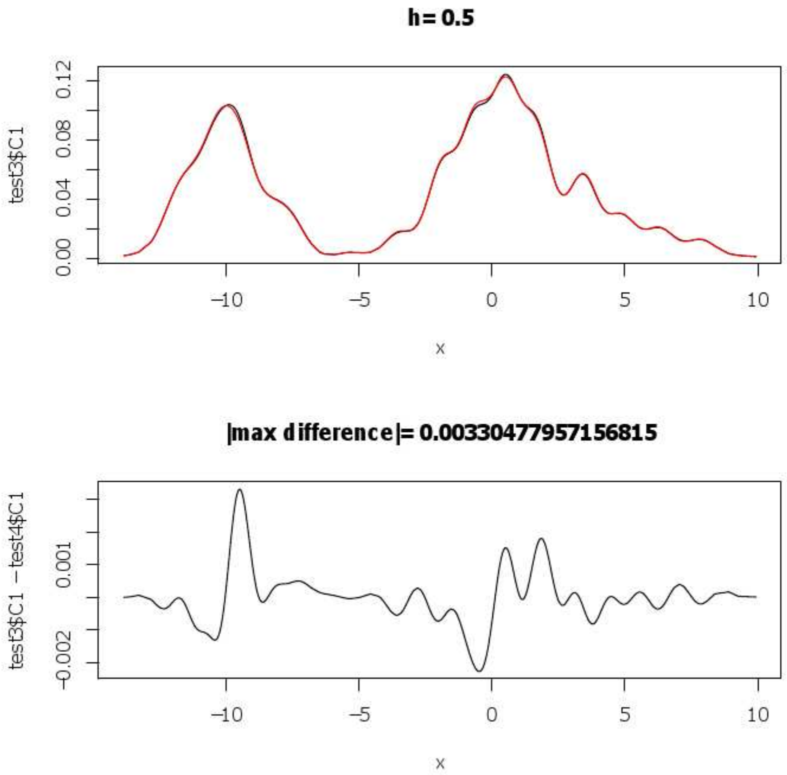


Figure 3.12: When a sensible bandwidth of $h = 1$ was chosen, f_1 and f_2 look similar, even when the parameter ‘tolEM’ is varied from 0.0001 to 0.001.

The following Figures 3.13 and 3.14 were produced in the same way as Figure 3.12, except the bandwidth h was chosen to be $h = 0.5$ and $h = 3$

respectively. Note that the true density f is a member of $\mathcal{F}_{0.5}$ but not a member of $\mathcal{F}_3$ (so $h = 0.5$ could have been chosen to be larger, and $h = 3$ was too large), yet in all cases the variation of the stopping parameter `tolEM` did not seem to produce problematic consequences for this density estimation application.



h= 0.5

|max difference|= 0.00330477957156815

Figure 3.13: When an unnecessarily small bandwidth of $h = 0.5$ was chosen, f_1 and f_2 still look similar when 'tolEM' is varied.

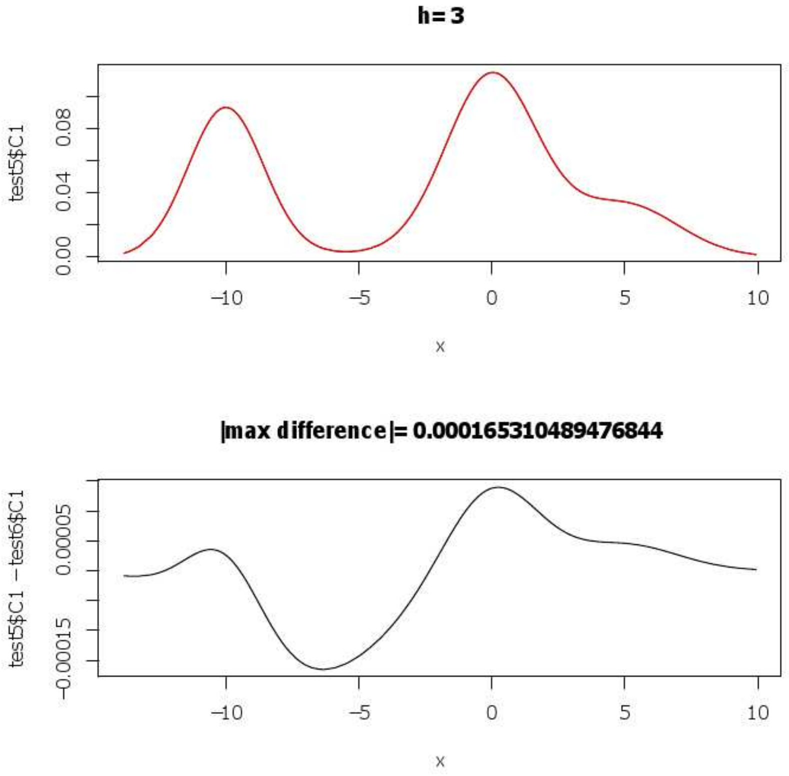


Figure 3.14: When the bandwidth was chosen to be too large ($h = 3$), f_1 and f_2 still look similar when 'tolEM' is varied.

Once these details were calibrated, we addressed the main speed problem of this algorithm by implementing the slowest parts in C (as specified in Appendix B). We found the new code ran more than 10 times faster than the old. Below is the output for the two versions of code run on the same data set `x` with the same settings. The dataset had $n = 1000$ observations.

```
> source("isdm")
> source("isdmC")
> h <- 1.4
> settings <- c(100, 0.1, 50, 0.01, 0.001)
> C0 <- make.C0(x, h)
> print(system.time(isdm(x, h, C0, settings)))

   user  system elapsed
  40.27    0.02   40.47
```

```
> print(system.time(isdmC(x, h, C0, settings)))

   user  system elapsed
   2.99    0.02    3.01
```

The full R and C code for this implementation of the ISDM can be found in Appendix B. There are several packages with tools for analyzing (or fitting) finite models (using the EM or newton type algorithms but not the ISDM). To the author's knowledge there does not exist a package which provides tools as an aid towards the multiscale mixture bandwidth selection problem outlined in this thesis.

Appendix A

Code for demonstration in Chapter 2

Given vectors θ and X, the following code calculates (viewed as a function of θ)

$$D_n(\theta, X) = \sum_{i=1}^{n} \left\{ \frac{\phi(X_i - \theta)}{\phi(X_i - \bar{X})} - 1 \right\}.$$

```
Dn <- function(x,theta){
  n <- length(x)
  m <- length(theta)
  xmean <- mean(x)
  temp <- matrix(0, nrow = 1, ncol = m)
  for(i in 1:n){
    denom <- dnorm(x[i]-xmean)
    for(j in 1:m){
      temp[j] <- temp[j] + dnorm(x[i]-theta[j])/denom-1
    }
  }
  temp
}
```

The code above was then translated into C for a speed increase, as follows.

```
#include<R.h>
#include<Rmath.h>

// if(*mode), cast output to double* and write all
//  *pm function values to it
// otherwise cast to int* and return 1 or 0 if the function
//   does or doesn't exceed 0

void Dn(double* x,int* pn,double* theta,int* pm,
  int* mode,void* output){

  int n=*pn, m=*pm;
  double xbar=0;
  for(int i=0;i<n;++i) xbar+=x[i];
  xbar/=n;
  for(int j=0;j<m;++j){
    double tmp=-n;
    double a=theta[j]-xbar;
    double b=(xbar*xbar-theta[j]*theta[j])/2;
    for(int i=0;i<n;++i) tmp+=exp(x[i]*a+b);
    if(*mode) ((double*)output)[j]=tmp;
    else if(tmp>0){
      *((int*)output)=1;
      return;
    }
  }
  if(*mode==0) *((int*)output)=0;
  return;
}

// phi(x_i-theta_j)/phi(x_i-xbar)-1 =
// exp(x_i(theta_j-xbar)+(xbar^2-theta_j^2)/2)-1
```

The above was compiled into dn.dll and the following functions are wrappers for that C code.

```
dyn.load("dn.dll")
Dn_ispositive_C <- function(x,theta){
  .C("Dn",as.double(x),as.integer(length(x)),
  as.double(theta),as.integer(length(theta)),
  as.integer(0),out=integer(1))$out
}
Dn_C <- function(x,theta){
  .C("Dn",as.double(x),as.integer(length(x)),
  as.double(theta),as.integer(length(theta)),
  as.integer(1),out=double(length(theta)))$out
}
```

The simulated demonstration in Chapter 2 was then implemented in R as follows.

```
# first demo with plot of Dn(theta)
#set.seed(98097629)

#cat("seed is 98097629\n")
#n<-100;
#x<-sort(rnorm(n))
#cat("The simple dataset for a picture is:\n",x,"\n")
#hist(x,n=10)

# this function is used only to quickly plot a Dn(theta),
# not for estimating probs.
Dn <- function(x,theta){
  n <- length(x)
  m <- length(theta)
  xmean <- mean(x)
  temp <- matrix(0, nrow = 1, ncol = m)
  for(i in 1:n){
    denom <- dnorm(x[i]-xmean)
    for(j in 1:m){
      temp[j] <- temp[j] + dnorm(x[i]-theta[j])/denom-1
    }
  }
  (temp/sqrt(n))
}

# plot of Dn(theta) using function Dn
#theta<-(-500:500)/100
#Dntheta<-Dn(x,theta)
#plot(theta,Dntheta,type="l")
#abline(0,0)

# testing prob estimate without going too crazy first..
# choosing n=100 and B=10
#b<-10;n<-100;
```

```
#set.seed(98097629)
#theta<-(-500:500)/100
#D<-0
#for(i in 1:b){
#  x<-rnorm(n)
#  D[i]<-Dn_C(x,theta)
#}
#pn100b10<-sum(D==0)/b



 #this took from about 11am to 9:44pm to run
 # B was 500, and theta was (-50:50)/10. Dn_C was used

 # Update: now this uses Dn_ispositive_C and the range
 # depends on the range of the data
 # started about 7:58pm aug 24, finished sometime before
 # the next morning
#B<-100
#N<-c(10,100,1000,10000,100000,1000000,10000000,100000000)
#p<-matrix(0, nrow=length(B),ncol=length(N))
#set.seed(98097629)
#  for(j in 1:length(N)){
#    D<-0
#    for(k in 1:B){
#      x<-rnorm(N[j])
#      theta<-floor(1.5*min(x)):floor(1.5*max(x))
#      D[k]<-Dn_ispositive_C(x,theta)
#    }
#    p[j]<-sum(D==0)/B

#  }

#save(p,file="aug24.RData")



# new edit: making theta finer to get a tighter bound. B=500
# started:2:03pm aug26
# finished:didn't finish by aug 28

# new edit: same as just above, but theta is kept the
# same as the aug24 run
# started: about lunchtime aug 28
# finished: accidentally closed before the last point was calculated
#B<-500
#N<-c(10,100,1000,10000,100000,1000000,10000000)
#p<-matrix(0, nrow=length(B),ncol=length(N))
#set.seed(98097629)
#  for(j in 1:length(N)){
#    D<-0
#    for(k in 1:B){
#      x<-rnorm(N[j])
#      theta<-floor(1.5*min(x)):floor(1.5*max(x))
#      D[k]<-Dn_ispositive_C(x,theta)
#    }
#    p[j]<-sum(D==0)/B
#    save(p,file="aug26.RData")
#  }

# doing the last point on aug 28
#set.seed(98097629)
#B<-500
#N<-100000000
#p<-matrix(0, nrow=length(B),ncol=length(N))
#set.seed(98097629)
#  for(j in 1:length(N)){
#    D<-0
#    for(k in 1:B){
#      x<-rnorm(N[j])
#      theta<-floor(1.5*min(x)):floor(1.5*max(x))
#      D[k]<-Dn_ispositive_C(x,theta)
#      save(D,file="aug28_log.RData")
#    }
#    p[j]<-sum(D==0)/B
#    save(p,file="aug28.RData")
#  }


# started 9:55pm aug17
#set.seed(98097629)
#B<-500
#N<-100000000
```

```
#theta<-(-50:50)/10
#p<-0
#D<-0
#for(j in 1:B){
#  x<-rnorm(N)
#  D[j]<-Dn_C(x,theta)
#}
#pn8<-sum(D==0)/B
# prob with n=10^8
#print(pn8)
#save(pn8,file="aug17b.RData")
```

Appendix B

Code for Chapter 3

The following R code was used to implement the ISDM.

```
# ISDM code

# Contents

# (1) Mixture density functions
# rmixnorm - generates random vector of observations
# dmixnorm - calculates density values
# pmixnorm - cdf of mixture


# (2) Setting up initial estimate of density for ISDM
# form.f - returns a matrix where each element is
# \frac1h\phi(\frac{X_i-\theta_j}h)
# make.C0 - makes Kernel density estimate with normal kernel
# and bandwidth h, based on data x, evaluated at each x


# (3) Setting up a grid to plot a function over
# form.gr - creats a non decreasing vector of points of
# length m and endpoints min(x) and max(x).


# (4) Functions for step1 to call
# get.d - given C1(x), observations x and a grid gr, calculates
# the function \Sum1ni (\frac{\phi_h(X_i-gr_j)}{C1_i} -1)
# examine.d - for a given set of function values of a
# function d=d(theta) and corresponding values theta, calculates
# the locations of the local maxima of d



# ~~~
# (1) Mixture density functions

# rmixnorm - generates random vector of observations
# returns a random sample from a mixture density (assumes a
# default mixture)

rmixnorm<-function(n,mu=c(-10,0,5),sig=sqrt(c(2,3,4)),
  Prob=c(2,3,1)/6){

  K<-length(mu)
  which<-sample(size=n,x=1:K,prob=Prob,repl=T)
  as.vector(rnorm(n,mu[which],sig[which]))
}

# dmixnorm - calculates density values
# returns the density of a normal mixture, evaluated at
# each grid point gr

dmixnorm<-function(gr,mu=c(-10,0,5),sig=sqrt(c(2,3,4)),
  prob=c(2,3,1)/6){

  K<-length(mu)
  n<-length(gr)
  dmat<-matrix(0,n,K)
  for(j in 1:K){
    dmat[,j]<-dnorm(gr,mu[j],sig[j])
  }
  as.vector(dmat%*%prob)
}

# pmixnorm - cdf of mixture
# returns the cdf of a normal mixture, evaluated at each gr
# gr should be an increasing vector for this to be sensical

pmixnorm<-function(gr,mu=c(-10,0,5),sig=sqrt(c(2,3,4)),
```

```
  prob=c(2,3,1)/6){

  gr<-sort(gr)
        K<-length(mu)
        n<-length(gr)
        pmat<-matrix(0,n,K)
        for(j in 1:K){
                pmat[,j]<-pnorm(gr,mu[j],sig[j])
        }
        as.vector(pmat%*%prob)
}



# ~~~
# (2) Setting up initial estimate of density for ISDM

# form.f - returns a matrix where each element is
# \frac1h\phi(\frac{X_i-\theta_j}h)

form.f<-function(x,th,h){
  z<-outer(x,th,FUN="-")  # removing means
  sdevs<-matrix(h,length(x),length(th))
  sdevs<-1/sdevs
  z<-z*sdevs  # dividing by the scale sd
  f<-sdevs*dnorm(z)
  (f)
}

# form.f is in the innermost loop of the code so it
# might be good to optimize this

# make.C0 - makes Kernel density estimate with
# normal kernel and bandwidth h, based on data x,
# evaluated at each x

# Only for creating initial estimates for feeding
# into the ISDM, not for general kernel density
# estimation since the grid is only x each time

make.C0<-function(x,h){
```

```
  x<-sort(x)
  f<-form.f(x,x,h)
  apply(f,2,mean)
}


# ~~~
# (3) Setting up a grid to plot a function over
# form.gr - creats a non decreasing vector of points of
# length m and endpoints min(x) and max(x).

# x is the data, m is the number of points in the grid
# alternatively, x is a vector c(min,max) which defines
# where the boundary of the grid is
form.gr<-function(x,m){
  as.vector(seq(from=min(x),to=max(x),len=m))
}


# ~~~
# (4) Functions for step1 to call
# get.d - given C1(x), observations x and a grid gr, calculates
# the function \Sum1ni (\frac{\phi_h(X_i-gr_j)}{C1_i} -1)

get.d<-function(x,gr,h,C1){
  z<-outer(x,gr,FUN="-")
  z<-z/h
  f<-dnorm(z)/h
  d<-(1/C1)%*%f-length(C1)
  as.vector(d)
}

# examine.d - for a given set of function values of a function
# d=d(theta) and corresponding values theta, calculates the
# locations of the local maxima of d

examine.d<-function(theta,d){

  # constructing vectors to store info in:
  maxima<-where.max<-rep(-1,length(d))
```

```
  thresh=-1
  # checking the end point
  if (d[1]>=d[2] & d[1]>=thresh){
    maxima[1]<-d[1]
    where.max[1]<-1
  }

  # looking through the function values
  for (i in 2:(length(d)-1)){
    if (d[i]>=d[i+1] & d[i]>=d[i-1] & d[i]>=thresh){
      maxima[i]<-d[i]
      where.max[i]<-i
    }
  }

  # checking the other end point
  if (d[length(d)]>=d[length(d)-1] & d[length(d)]>=thresh){
    maxima[length(d)]<-d[length(d)]
    where.max[length(d)]<-length(d)
  }

  maxima<-maxima[maxima>thresh]
  where.max<-where.max[where.max>0]
  where.max<-theta[where.max]

  list(global=max(maxima),where.max=where.max)
}


# ~~~
# (5) Functions for step2 to call
# applyEM - one iteration of EM for probabilities

applyEM<-function(fmat,e0){
  # e0 is the initial vector of probabilities
  # fmat is a matrix with the first column being C1, and the rest
  # of the matrix being form.f(x,thetas,h), where thetas is
  # the output of step1's estimates of the means
  n<-dim(fmat)[1]
  m<-dim(fmat)[2]
  E<-matrix(e0,n,m,byrow=TRUE)
```

```
  numer<-E*fmat
  denom<-apply(numer,1,sum)
  denom<-matrix(denom,n,m,byrow=FALSE)
  piji<-numer/denom
  e<-apply(piji,2,sum)/n
  as.vector(e)
}


iter.EM<-function(fmat,e0,tolEM){
  LL<-0
  di<-tolEM+1
  count<-1
  while(di>tolEM){
    temp<-log(fmat%*%e0)
    LL[count]<-sum(temp[temp>-Inf])
    e0<-applyEM(fmat,e0)

    if(count>1){
      di<-LL[count]-LL[count-1]
    }
    count<-count+1
  }
  (e0)
}

# Main code section:

# x is the data
# h is the bandwidth (component sd, not component variance)
# settings is a vector of scalar values which determine
# stopping conditions, etc within the ISDM

isdm<-function(x,h,C0,settings){

  # Pulling out the scalars from settings
  gridNum<-settings[1]  # integer, maybe about 100 or so
  epsilon1<-settings[2]  # should be positive and close to zero.
  # ^ Criteria to halt ISDM.

  giveUp<-settings[3]  # integer, number of times to
```

```
# iterate ISDM before giving up

smallbit<-settings[4]  # if there are an expected maximum of
# k=k(h) mass points then smallbit should be around 1/(2k),
# not so important, this is to give an initial EM
# estimate of the weights.

tolEM<-settings[5]  # tolerance for the EM algorithm's
    # stopping criteria.
# The ISDM is quite sensitive to this.
# It should be positive, and ideally if computers
# were super fast it should be very small.
# About 0.01 for now (Edit: investigation finished).

# Making initial density estimate
x<-sort(x)
C1<-C0

# Making grid for step 1
gr<-form.gr(x,gridNum)

# Starting ISDM
eps<-epsilon1+1  # Stopping criteria 1 (real one)
count<-1  # Stopping criteria 2 (to ensure ISDM finishes)
LL<-0    # To keep track of the log likelihood value
epsTrack<-0  # To keep track of the epsilon values to
# see if the code gave up

while((eps > epsilon1)&(count<=giveUp)){

  # Step 1
  s1<-step1(x,gr,h,C1)
  thetas<-s1$thetas
  eps<-s1$global
  epsTrack[count]<-eps

  # Step 2
  s2<-step2(x,thetas,h,C1,smallbit,tolEM)
  e<-s2$e
  fmat<-s2$fmat
```

```
    # Step 3
    C1<-as.vector(fmat%*%e)

    # The LL
    LL[count]<-sum(log(C1))
    count<-count+1
  }

  list(C1=C1, e=e, thetas=thetas, LL=LL, epsTrack=epsTrack)
}

iter.ISDM<-function(x,H,settings){
  # H is a vector of bandwidths h_j now
  # x are the observations x_i
  # settings is the same thing to feed into the ISDM

  x<-sort(x)  # arranging in ascending order
  H<-rev(sort(H))  # arranging in descending
  #order (since it is faster to compute estimates on larger h)

  # For storage
  Out<-list()
  LL<-0

  C0<-make.C0(x,H[1])
  for(i in 1:length(H)){
    cat("Starting iteration number ",i," out of ",
    length(H)," iterations.\n")
    Contents<-isdm(x,H[i],C0,settings)
    Out[[i]]<-Contents
    names(Out)[i]<-paste("When bandwidth h =",H[i])
    C0<-Contents$C1
    LL[i]<-Contents$LL[length(Contents$LL)]
    cat("Log likelihood for bandwidth h =",H[i]
    ," was ",LL[i],"\n")
  }
  cat("Finished running code.\n")
  list(Out=Out,x=x,H=H,LLh=LL)
}

step1<-function(x,gr,h,C1){
  d<-get.d(x,gr,h,C1)

  #plot(gr,d,type="l")
  #abline(0,0)

  temp<-examine.d(gr,d)
  thetas<-temp$where.max
  global<-temp$global
  list(thetas=thetas,global=global)
}
step2<-function(x,thetas,h,C1,smallbit,tolEM){
  # thetas is the vector of mean estimates output from step 1
  fmat<-cbind(C1,form.f(x,thetas,h))

  # making an initial weight vector
  temp<-1/(2*length(thetas))
  small<-min(smallbit,temp)
  e0<-rep(c(1-length(thetas)*small,small),c(1,length(thetas)))
  e<-iter.EM(fmat,e0,tolEM)
  list(fmat=fmat,e=e)
}
```

The slowest step within the ISDM was rewritten in C for speed purposes. The file `code.dll` was compiled from the following C code.

```
#include<R.h>
#include<Rmath.h>

/*
finds all indices i for which x[i] is a local maximum >= thresh
positions should initially be a list the same length as x
on return n will be the number of such indices,
and the first n entries of positions will be the indices
 themselves (R style, ie starting at 1)
*/
void examine_d(double* x,int* xn,double* thresh,
double* positions,int* n){
  *n=0;  // the number of valid indices found so far
  for(int i=0;i<*xn;++i)
```

```
        // check it's a local max
    if((i==0||x[i-1]<=x[i])&&(i==*xn-1||x[i+1]<=x[i])

        //check it's at least thresh
        &&x[i]>=*thresh){
      positions[(*n)++]=i+1;
  }
}

//computes the value of the normal density with
//standard deviation h at x

double phi(double x,double h){
  static double one_over_sqrt_2_pi=1.0/sqrt(2.0*M_PI);
  x/=h; x*=x;
  return one_over_sqrt_2_pi/h*exp(-x/2);
}

/*
Uses the EM method to approximate the weighted average

p_0 C_i + \sum_j p_j(phi_{h,\theta_j}(x_i))

which maximises

\sum_i log[p_0 C_i + \sum_j p_j(phi_{h,\theta_j}(x_i))]

Input:

x = the observed measurements (vector of length xn)
C = the current density, evaluated at the x_i (vector of
  length xn)
theta = means of the normals being added to the
  density (vector of length thetan)
h = the standard deviation of the normals being added
  to the density (double)
smallbit = initial weights of new components

Output:
```

```
C will be changed to the above weighted average
p = the weights (vector of length thetan+1)

*/
void step2(double* x,int* xn,double* theta,int* thetan,
  double* h,double* C,double* tol,double* p,double* smallbit){

  // extra space for weights, used in calculations
  double* p2=(double*)Calloc(*thetan+1,double);

  // space to store the phi_{h,\theta_j}(x_i)
  double* phi_matrix=(double*)Calloc(*thetan*(*xn),double);

  // space to store phi_{h,\theta_j}(x_i) p_j
  double* phi_p=(double*)Calloc(*thetan+1,double);

  // compute all the phi's
  for(int i=0;i<*xn;++i)
    for(int j=0;j<*thetan;++j)
      phi_matrix[i+j*(*xn)]=phi(x[i]-theta[j],*h);

  // initialise the weights arbitrarily
  double small=0.5/(*thetan);
  if(small>*smallbit) small=*smallbit;
  p[*thetan]=1-small*(*thetan);
  for(int j=0;j<*thetan;++j) p[j]=small;

  double old_ll=-INFINITY;
  while(1){
    double new_ll=0;
    for(int j=0;j<=*thetan;++j) p2[j]=0;
    for(int i=0;i<*xn;++i){
      double tot=0;
      for(int j=0;j<=*thetan;++j)
        tot+=(phi_p[j]=((j<*thetan)?
          phi_matrix[i+j*(*xn)]:C[i])*p[j]);
      for(int j=0;j<=*thetan;++j) p2[j]+=phi_p[j]/tot;
      new_ll+=log(tot);
    }
    for(int j=0;j<=*thetan;++j) p[j]=p2[j]/(*xn);
    // ll didn't change much, so stop
    if(old_ll+*tol>new_ll) break;
    old_ll=new_ll;
  }
  // replace C by the weighted average
  for(int i=0;i<*xn;++i){
    C[i]*=p[*thetan];
    for(int j=0;j<*thetan;++j)
      C[i]+=phi_matrix[i+j*(*xn)]*p[j];
  }

  // free the memory we allocated
  Free(p2);
  Free(phi_matrix);
  Free(phi_p);
}
```

After the above code was compiled into `code.dll`, the R function `isdm` was then replaced with the following function (`isdmC`).

```
dyn.load("code.dll")

# Wrapper for the C code
examine_d<-function(theta,d,thresh=-1){
  positions<-rep(0,length(d))
  result_of_code=.C("examine_d",as.double(d),
    as.integer(length(d)),as.double(thresh),
    positions=as.double(positions),n=integer(1))

# truncate positions to the number of valid indices
  positions=result_of_code$positions[1:result_of_code$n]

    maxima<-d[positions]
#  list(maxima=maxima,global=max(maxima),
# where.max=theta[positions],theta=theta)
  list(global=max(maxima),where.max=theta[positions])
}

# returns the new C1 vector
step2and3<-function(x,theta,h,C1,tol,smallbit){
  p<-rep(0,length(theta)+1)
  output<-.C("step2",as.double(x),as.integer(length(x)),
  as.double(theta),as.integer(length(theta)),as.double(h),
  C1=as.double(C1),as.double(tol),p=as.double(p),
  as.double(smallbit))
  list(C1=output$C1,p=output$p)
}

# End of wrappers


# ISDM code

# x is the data
# h is the bandwidth (component sd, not component variance)
# settings is a vector of scalar values which determine
# stopping conditions, etc within the ISDM

isdmC<-function(x,h,C0,settings){
```

⋮ (Same as original code in the function `isdm`)

```
  # Starting ISDM
  eps<-epsilon1+1  # Stopping criteria 1 (real one)
  count<-1  # Stopping criteria 2 (to ensure ISDM finishes)
  LL<-0    # To keep track of the log likelihood value
  epsTrack<-0  # To keep track of the epsilon values to
  # see if the code gave up

  while((eps > epsilon1)&(count<=giveUp)){

    # Step 1
    s1<-step1(x,gr,h,C1)
    thetas<-s1$thetas
    eps<-s1$global
    epsTrack[count]<-eps

    # Step 2 and 3

    coutput<-step2and3(x,thetas,h,C1,tolEM,smallbit)
    e<-coutput$p
    C1<-coutput$C1

    # The LL
    LL[count]<-sum(log(C1))
    count<-count+1
  }

  list(C1=C1, e=e, thetas=thetas, LL=LL, epsTrack=epsTrack)
}
```

Appendix C

Miscellaneous Appendices

Remark 3.5.1 was as follows.
Suppose θ_0 and θ_1 are two estimates of θ. If

$$\ell_{EM}(\theta_1|\theta_0) \geq \ell_{EM}(\theta_0|\theta_0),$$

then

$$\sum_{i=1}^{n} \log(f_{X_1}(x_i|\theta_1)) \geq \sum_{i=1}^{n} \log(f_{X_1}(x_i|\theta_0)), \tag{C.1}$$

where $f_{X_1}(\cdot|\theta)$ is the marginal density of X_1 in this bivariate interpretation of our observations.

Here is a proof.

Proof. We prove this in the case where we have $n = 1$ observation, since the general case follows a similar proof, with the letters x and X representing vectors and the integrals over $\mathbb{R}$ being over $\mathbb{R}^n$ instead.

Suppose (X, Y) is a bivariate random vector with joint density $f_{XY}(x, y|\theta)$ and X marginal given by $f_X(x|\theta) = \int_{\mathbb{R}} f_{XY(x,y|\theta)}dy$. We assume each $f_X(\cdot|\theta) \neq 0$ a.e, and so we define the conditional density of Y given $X = x$ to be given by $f_{Y|X=x}(y|x, \theta) = 1_{(f_X(x|\theta)>0)}\frac{f_{XY(x,y|\theta)}}{f_X(x|\theta)}$.

Fix $\theta_0 = (\theta_{01}, \ldots, \theta_{0k})$ and let θ be arbitrary. Given the observed value of X is x, the conditional expectation of the full data log likelihood $\log(f_{XY}(X, Y|\theta))$, evaluated under the probability corresponding to θ_0 is:

$$\begin{aligned}
\ell_{EM}(\theta|\theta_0) &= \mathbb{E}_{Y|X=x,\theta_0}\left(\log(f_{XY}(X,Y|\theta))\right) \\
&= \int_{\mathbb{R}} \log\left\{f_{XY}(x,y|\theta)\right\} f_{Y|X=x}(y|x,\theta_0)dy \\
&= \int_{\mathbb{R}} \left(\log\{f_{Y|X=x}(y|x,\theta)\} + \log\{f_X(x|\theta)\}\right) f_{Y|X=x}(y|x,\theta_0)dy \\
&= \int_{\mathbb{R}} \log\{f_{Y|X=x}(y|x,\theta)\} f_{Y|X=x}(y|x,\theta_0)dy \\
&\quad + \int_{\mathbb{R}} \log\{f_X(x|\theta)\} f_{Y|X=x}(y|x,\theta_0)dy \\
&= \int_{\mathbb{R}} \log\{f_{Y|X=x}(y|x,\theta)\} f_{Y|X=x}(y|x,\theta_0)dy + \log\{f_X(x|\theta)\},
\end{aligned} \tag{C.2}$$

since $f_{Y|X=x}(y|x,\theta_0)$ is a density and since $\log\{f_X(x|\theta)\}$ does not depend on y.

Using (C.2), we have

$$\begin{aligned}
\log\left\{f_X(x|\theta_1)\right\} - \log\left\{f_X(x|\theta_0)\right\} &= \ell_{EM}(\theta_1|\theta_0) - \ell_{EM}(\theta_0|\theta_0) \\
&\quad + \int_{\mathbb{R}} \log\left\{f_{Y|X=x(y|x,\theta_0)}\right\} f_{Y|X=x}(y|x,\theta_0)dy \\
&\quad - \int_{\mathbb{R}} \log\left\{f_{Y|X=x(y|x,\theta_1)}\right\} f_{Y|X=x}(y|x,\theta_0)dy.
\end{aligned}$$

Suppose now that $\ell_{EM}(\theta_1|x,\theta_0) - \ell_{EM(\theta_0|x,\theta_0)} \geq 0$. Then

$$\begin{aligned}
\log\left\{f_X(x|\theta_1)\right\} - \log\left\{f_X(x|\theta_0)\right\} &\geq \int_{\mathbb{R}} \log\left\{f_{Y|X=x}(y|x,\theta_0)\right\} f_{Y|X=x}(y|x,\theta_0)dy \\
&\quad - \int_{\mathbb{R}} \log\left\{f_{Y|X=x}(y|x,\theta_1)\right\} f_{Y|X=x}(y|x,\theta_0)dy \\
&= -\int_{\mathbb{R}} \log\left\{\frac{f_{Y|X=x}(y|x,\theta_1)}{f_{Y|X=x}(y|x,\theta_0)}\right\} f_{Y|X=x}(y|x,\theta_0)dy \\
&= \mathbb{E}_{Y|X=x}\left(-\log\left\{\frac{Y|X=x,\theta_1}{Y|X=x,\theta_0}\right\}\right), \\
&\geq -\log\left\{\mathbb{E}_{Y|X=x}\left(\frac{Y|X=x,\theta_1}{Y|X=x,\theta_0}\right)\right\},
\end{aligned}$$

by Jensen's Inequality, and so

$$\begin{aligned}
\log\left\{f_X(x|\theta_1)\right\} - \log\left\{f_X(x|\theta_0)\right\} &\geq -\log\left\{\int_{\mathbb{R}} \frac{f_{Y|X=x}(y|x,\theta_1)}{f_{Y|X=x}(y|x,\theta_0)} f_{Y|X=x}(y|x,\theta_0)dy\right\} \\
&= -\log\left\{\int_{\mathbb{R}} f_{Y|X=x}(y|x,\theta_1)dy\right\} \\
&= -\log\{1\}, \text{ since } f_{Y|X=x}(y|x,\theta_1) \text{ is a density.}
\end{aligned}$$

Thus if $\ell_{EM}(\theta_1|x,\theta_0) - \ell_{EM(\theta_0|x,\theta_0)} \geq 0$ then

$$\log\left\{f_X(x|\theta_1)\right\} - \log\left\{f_X(x|\theta_0)\right\} \geq 0,$$

and we are done. □

Appendix D

Example data set

The data set in Example 3.1.1 of Chapter 3 is an iid sample of size $n = 200$, generated from the distribution with mixture density

$$0.3\phi_{-1,0.2} + 0.7\phi_{0.5,0.5}$$

via the following `R` code:

```
> rmixnorm <- function(n, weights, means, sds) {
+     J <- sample(size = n, x = 1:length(weights), prob = weights,
+         repl = TRUE)
+     as.vector(rnorm(n, mean = means[J], sd = sds[J]))
+ }
> dmixnorm <- function(x, prob, means, sds) {
+     z <- matrix(0, length(x), length(prob))
+     for (j in 1:length(prob)) {
+         z[, j] <- dnorm(x, means[j], sds[j])
+     }
+     as.vector(z %*% prob)
+ }
> set.seed(98097629)
> x <- rmixnorm(200, c(0.3, 0.7), c(-1, 0.5), c(0.2, 0.5))
> x <- sort(x)
```

Bibliography

P. Bickel and H. Chernoff. Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In J. K. Ghosh, S. K. Mitra, K. R. Parthasararthy, and B. L. S. Prakasa Rao, editors, *Statistics and Probability: A Raghu Raj Bahadur Festschrift*, pages 83–96. Wiley Eastern Limited, 1993.

P. Billingsley. Convergence of probability measures. *New York*, 1968.

G.E.P. Box and G.C. Tiao. A Bayesian approach to some outlier problems. *Biometrika*, 55(1):119–129, 1968.

M. Cathy and M. Bertrand. Adaptive density estimation for clustering with Gaussian mixtures. *Arxiv preprint arXiv:1103.4253*, 2011.

J. Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23(1):221–233, 1995.

M. Csörgő, S. Csörgő, L. Horváth, and D.M. Mason. Weighted empirical and quantile processes. *Ann. Probab.*, 14(1):31–85, 1986. ISSN 0091-1798.

T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

S. Geman and C.R. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 10(2):401–414, 1982.

S. Ghosal and A.W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263, 2001. ISSN 0090-5364.

U. Grenander. Abstract inference. *Wiley, New York*, 1981.

R.J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2):795–800, 1985.

P.D. Hoff. Nonparametric estimation of convex models via mixtures. *Annals of Statistics*, pages 174–200, 2003.

B.L. Joiner. Living histograms. *International Statistical Institute (ISI)*, 43 (3):339–340, 1975.

M.C. Jones and D.A. Henderson. Maximum likelihood kernel density estimation: On the potential of convolution sieves. *Computational Statistics & Data Analysis*, 53(10):3726–3733, 2009.

N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, pages 805–811, 1978.

M.R. Leadbetter and Holger Rootzén. Extremal theory for stochastic processes. *Ann. Probab.*, 16(2):431–478, 1988. ISSN 0091-1798.

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

M.L. Lesperance and J.D. Kalbfleisch. An algorithm for computing the nonparametric mle of a mixing distribution. *J. Amer. Stat. Assoc.*, 87(417): 120–126, 1992.

B.G. Lindsay. The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, 11(1):86–94, 1983. ISSN 0090-5364.

B.G. Lindsay. *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and American Statistical Association, 1995.

B.G. Lindsay and M.L. Lesperance. A review of semiparametric mixture models. *Journal of Statistical Planning and Inference*, 47(1-2):29–39, 1995.

L.S. Magder and S.L. Zeger. A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *J. Amer. Statist. Assoc.*, 91(435): 1141–1151, 1996. ISSN 0162-1459.

C. Matias. Semiparametric deconvolution with unknown noise variance. *ESAIM: Probability and Statistics*, 6(1):271–292, 2002.

H.P. McKean, Jr. *Stochastic integrals.* Probability and Mathematical Statistics, No. 5. Academic Press, New York, 1969.

D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences].* Springer-Verlag, Berlin, 1991. ISBN 3-540-52167-4.

D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences].* Springer-Verlag, Berlin, second edition, 1994. ISBN 3-540-57622-3.

B.W. Silverman. *Density estimation for statistics and data analysis.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986. ISBN 0-412-24620-1.

S. van de Geer. Asymptotic theory for maximum likelihood in nonparametric mixture models. *Computational statistics & data analysis*, 41(3):453–464, 2003.

M.P. Wand and M.C. Jones. *Kernel Smoothing (Monographs on Statistics and Applied Probability).* Chapman & Hall, London, 1994.

Y. Wang. On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):185–198, 2007.

A.E. Watkins, M.F. Schilling and W. Watkins. Is human height bimodal? *The American Statistician*, 56(3):223–229, 2002.