Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

On the Mathematical Foundations of Likelihood Theory

por

Pedro Franklin Cardoso Silva

Belo Horizonte

December 16, 2017

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

On the Mathematical Foundations of Likelihood Theory

por

Pedro Franklin Cardoso Silva *
Orientador: Flávio Bambirra Gonçalves

Tese apresentada ao Departamento de Estatística da Universidade Federal de Minas Gerais como parte dos requisitos para obtenção do grau de Doutor em

ESTATÍSTICA

Belo Horizonte - MG
2017

*O autor foi bolsista da CAPES durante a elaboração deste trabalho.

Para minha mãe, como se dissesse vida

“ Se esse mundo não fosse uma região de espíritos desatentos, era escusado lembrar ao leitor que eu só afirmo certas leis, quando as possuo deveras; em relação a outras restrinjo-me à admissão da probabilidade.”

Abertura do Capítulo 100 de Memórias Póstumas de Brás Cubas, Machado de Assis.

Abstract

We discuss a general definition of likelihood function in terms of Radon-Nikodým derivatives. The definition is validated by the Likelihood Principle once we establish a result regarding the proportionality of likelihood functions under different dominating measures. This general framework is particularly useful when there exists no or more than one obvious choice for a dominating measure as in some infinite-dimensional models. We also discuss some versions of densities which are specially important when obtaining the likelihood function. In particular, we argue in favor of continuous versions of densities and highlight how these are related to the basic concept of likelihood. Finally, we present a method, based on the concept of differentiation of measures, to obtain a valid likelihood function, i.e., which is in accordance with the Likelihood Principle. Some examples are presented to illustrate the general definition of likelihood function and the importance of choosing particular dominating measures in some cases.

Keywords: Statistical model, Likelihood Principle, dominating measure, Radon-Nikodým derivative, proportional likelihood, continuous densities, differentiation of measures.

Agradecimentos

Esta tese é o resultado de dois anos de muito trabalho. Os teoremas obtidos neste período estão nos Capítulos 2, 3 e 4. Por trás destes teoremas, estão algumas pessoas. E aqui, com o meu genuíno agradecimento, estão elas:

Obrigado, mãe, você é a pessoa mais bonita do mundo;

Mari, irmã, como você é importante;

Em minha trajetória na UFMG eu pude conhecer muita gente especial. Ana, Caio, Gabi, Henrique, Lilian, Marina, Rodrigo, Tamires: obrigado.;

Denise, pelo incentivo e pela poesia, obrigado de coração;

Flávio, obrigado por todo o conhecimento que você compartilhou comigo e por todos os conselhos; aprendi muito contigo e sou muito grato por isso.

E me cerro, aqui, mire e veja. Isto não é o de um relatar passagens de sua vida, em toda admiração. Conto o que fui e vi, no levantar do dia. Auroras.

Cerro. O senhor vê. Contei tudo. Agora estou aqui, quase barranqueiro. Para a velhice vou, com ordem e trabalho. Sei de mim? Cumpro. O Rio de São Francisco - que de tão grande se comparece - parece é um pau grosso, em pé, enorme... Amável o senhor me ouviu, minha ideia confirmou: que o Diabo não existe. Pois não? O senhor é um homem soberano, circunspecto. Amigos somos. Nonada. O diabo não há! É o que eu digo, se for... Existe é homem humano. Travessia.

Grande Sertão: Veredas, Guimarães Rosa.

Contents

Chapter 1

Introduction

In this thesis, we shall discuss some mathematical foundations of Likelihood Theory, more specifically, the definition of likelihood function. Likelihood-based methodologies are undoubtedly the most common and efficient ones to perform statistical inference - in particular, maximum-likelihood estimation and Bayesian inference. This is due to general strong properties of the likelihood function that stem from a solid mathematical foundation, based on Measure/Probability Theory.

The concept of likelihood goes back to Fisher, with the actual term first appearance in Fisher (1921), and therefore before Kolmogorov's probability axioms (Kolmogorov, 1933) and the Radon-Nikodým Theorem (Nikodým, 1930) (Radon, in 1913 proved the theorem for $\mathbb{R}^n$ Radon (1913), but Fisher did not mention him in his work). Nevertheless, the intuition given by Fisher to construct the concept of likelihood made it straightforward to extend the definition of likelihood function (LF) in terms of Radon-Nikodým derivatives. The earliest explicit version of such definition we could find is from Lindley (1953) [Definition 2.4], however, it is implicitly assumed for example in Halmos and Savage (1949). It consists of defining the likelihood function as any Radon-Nikodým (RN) derivative (see Definition 2.2 in Section 2.2), i.e. using any σ-finite dominating measure.

Since any model that has a dominating measure admits an uncountable number of dominating measures, the aforementioned definition of likelihood function could only be admissible if the choice of the dominating measure has no influence in the inference process. Under the Likelihood Principle (LP), it means that any two distinct dominating measures should lead to proportional likelihood functions. Although such a result is accepted by the statistical community, it has not yet been properly stated, proven or explored. This is one of the specific aims of this thesis. In fact, this issue has never been properly raised in the literature. The general definition of likelihood is always approached by assuming the existence of a common dominating measure and there is no mention of other measures or what

would be the implications of making a different choice. Reid (2013) mentions that "Some books describe the likelihood function as the Radon-Nikodým derivative of the probability measure with respect to a dominating measure. Sometimes the dominating measure is taken to be P_{θ_0} for a fixed value $\theta_0 \in \Theta$. When we consider probability spaces and/or parameter spaces that are infinite dimensional, it is not obvious what to use as a dominating measure."

We state and prove what we call the Likelihood Proportionality Theorem, which validates (in terms of the LP) the general definition of likelihood function in terms of Radon-Nikodým derivatives. Moreover, we discuss how continuous RN derivatives are relevant when obtaining the likelihood function. More specifically, we show that the continuity property guarantees the proportionality result and leads to likelihood functions that carry the intuitive concept of likelihood.

Finally, we discuss and provide several examples where the choice of the dominating measure requires special attention. Namely, situations: i) that require some effort to find a valid dominating measure that can be used to obtain a valid likelihood function; ii) in which more than one obvious dominating measure is available but a particular choice may significantly easy the inference process. We also emphasise that we work with Likelihood Theory in a general context and not just for parametric models. This context is considered in several relevant inference problems nowadays, specially infinite-dimensional problems under the Bayesian approach, as we illustrate in some of the examples provided.

We discuss five general classes of widely used models. The first example considers general finite-dimensional models and describes how to obtain a valid likelihood function when dealing with point-mass mixtures. In the second example we discuss exponential family models. The following two examples consider classes of infinite-dimensional models, non-homogeneous Poisson processes and diffusions/jump-diffusions. Finally, the fifth example explores possibly important implications of the choice of the dominating measure in general missing data problems.

Other works in the context of mathematical aspects of the likelihood function but that pursue different directions can be found in Barndorff-Nielsen et al. (1976), Fraser and Naderi (1996), Fraser et al. (1997) and Fraser and Naderi (2007).

This thesis is organised as follows: Section 1.1 presents some background on Measure Theory and Topology, which are essential to develop the work in the following chapters. Chapter 2 presents the Likelihood Proportionality Theorem in Section 2.2 and discusses the importance of continuous densities and likelihoods in Section 2.3. Section 2.4 discusses the use of the prior predictive measure as a dominating measure, in a Bayesian context. Chapter 3 describes some examples regarding

the choice of dominating measure and shows how the Likelihood Proportionality Theorem is to be used in practical inference problems. Chapter 4 discusses how likelihood functions can be defined as a derivative of measures. Chapter 5 presents some conclusions and future work.

1.1 Fundamentals

In this section we present the mathematical results and the notation needed to understand the work in this thesis.

1.1.1 Fundamentals of Measure Theory

Let $(\Omega, \mathcal{F}, \mu)$ denote a measure space and $M(\Omega, \mathcal{F})$ the collection of all measurable functions $f : \Omega \longrightarrow \mathbb{R}$.

Definition 1.1 (Absolute continuity)**.** A measure λ on $(\Omega, \mathcal{F})$ is said to be absolutely continuous with respect to a measure μ on $(\Omega, \mathcal{F})$ if $E \in \mathcal{F}$ and $\mu(E) = 0$ imply that $\lambda(E) = 0$. In this case we write $\lambda << \mu$ and say that λ is dominated by μ. A family $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$ of probability measures on $(\Omega, \mathcal{F})$ is said to be absolutely continuous with respect to a measure μ on $(\Omega, \mathcal{F})$ if $P_\theta << \mu,\ \forall \theta \in \Theta$. In this case we write $\mathcal{P} << \mu$ and say that $\mathcal{P}$ is dominated by μ.

If μ and ν are two measures on the same measurable space $(\Omega, \mathcal{F})$ such that $\mu << \nu$ and $\nu << \mu$, then μ and ν are said to be equivalent measures.

Theorem 1.1 (Radon-Nikodým Theorem)**.** *Let λ and μ be σ-finite measures on $(\Omega, \mathcal{F})$ such that $\lambda << \mu$. Then there exists a nonnegative function f in $M(\Omega, \mathcal{F})$ such that*

$$\lambda(A) = \int_A f d\mu,\ A \in \mathcal{F}.$$

Furthermore, f is uniquely determined μ-almost everywhere.

Function f of the Radon-Nikodým Theorem is often called the Radon-Nikodým derivative of λ with respect to μ and is denoted by $d\lambda/d\mu$.

Proposition 1.1 (Radon-Nikodým chain rule)**.** *Let λ, μ, ν be σ-finite measures on $(\Omega, \mathcal{F})$ such that $\lambda << \mu$ and $\mu << \nu$. Then*

$$\frac{d\lambda}{d\nu} = \frac{d\lambda}{d\mu}\frac{d\mu}{d\nu},\ \nu - a.e.$$

Proposition 1.2. *Let $A \in \mathcal{F}$ be a nonempty set. If we denote*

$$\mathcal{F}(A) = \{B \cap A;\ B \in \mathcal{F}\},$$

then $\mathcal{F}(A)$ is a σ-algebra of subsets of A and $\mathcal{F}(A) \subset \mathcal{F}$.

Proof. Since $\emptyset,\ A \in \mathcal{F}$, it follows that $\emptyset,\ A \in \mathcal{F}(A)$. Let $D \in \mathcal{F}(A)$. Then, there exists $B \in \mathcal{F}$ such that $D = A \cap B$. Since $B^c \in \mathcal{F}$, it follows that $A \cap B^c$, the complement of $A \cap B$ in A, belongs to $\mathcal{F}(A)$. Now, let $\{D_n\}_{n=1}^{\infty}$ be a sequence of sets in $\mathcal{F}(A)$. Then, there exists a sequence $\{B\}_{n=1}^{\infty}$ of sets in $\mathcal{F}$ such that $D_n = A \cap B_n$ for every $n \in \mathbb{N}$. Since $\cup_{n=1}^{\infty} B_n \in \mathcal{F}$, it follows that $\cup_{n=1}^{\infty} D_n = \cup_{n=1}^{\infty}(A \cap B_n) = A \cap (\cup_{n=1}^{\infty} B_n) \in \mathcal{F}(A)$. Hence, $\mathcal{F}(A)$ is a σ-algebra of subsets of A. To see that $\mathcal{F}(A) \subset \mathcal{F}$, simply note that a σ-algebra is closed under countable intersections. □

Definition 1.2. Let $A \in \mathcal{F}$ be a nonempty set. We denote $\mu\big|_A$ as the restriction of the measure μ on $(A, \mathcal{F}(A))$, i.e., $\mu\big|_A$ is the measure defined on $(A, \mathcal{F}(A))$ such that $\mu\big|_A(B) = \mu(B),\ \forall B \in \mathcal{F}(A)$.

Definition 1.3. Let $f \in M(\Omega, \mathcal{F})$ and $A \in \mathcal{F}$ a nonempty set. We denote $f\big|_A$ as the restriction of the measurable function f on A, i.e.,

$$\begin{array}{rccc} f\big|_A : A & \longrightarrow & \mathbb{R} \\ \omega & \longmapsto & f(\omega). \end{array}$$

Proposition 1.3. *Let $f \in M(\Omega, \mathcal{F})$ and $A \in \mathcal{F}$ be a nonempty set. Then, $f\big|_A \in M(A, \mathcal{F}(A))$.*

Proof. Let $B \in \mathcal{B}(\mathbb{R})$. Since $f\big|_A^{-1}(B) = A \cap f^{-1}(B)$, the proof is complete. □

Definition 1.4. Let f and g be two functions in $M(\Omega, \mathcal{F})$ and let μ be a measure on $(\Omega, \mathcal{F})$. We say that f and g are equivalent functions with respect to μ if $f = g$ μ-a.e, in which case we write $f \equiv_\mu g$.

The relation $\equiv_\mu$ is an equivalence relation, that is, if f, g and h are functions in $M(\Omega, \mathcal{F})$, then

a. $f \equiv_\mu f$, the reflexive property;

b. if $f \equiv_\mu g$ then $g \equiv_\mu f$, the symmetric property;

c. if $f \equiv_\mu g$ and $g \equiv_\mu h$ then $f \equiv_\mu h$, the transitive property.

Given a function f in $M(\Omega, \mathcal{F})$, the equivalence class of f with respect to μ, $[f]_\mu$, is the collection of all functions g in $M(\Omega, \mathcal{F})$ such that $f \equiv_\mu g$, i.e.,

$$[f]_\mu = \{g \in M(\Omega, \mathcal{F});\ f \equiv_\mu g\}.$$

In words, μ-equivalent functions are indistinguishable from the point of view of a measure μ.

Proposition 1.4. *Let λ and μ be σ-finite measures on $(\Omega, \mathcal{F})$ such that $\lambda << \mu$ and let $f \in [\frac{d\lambda}{d\mu}]_\mu$. Then, $f\big|_A \in [\frac{d\lambda|_A}{d\mu|_A}]_{\mu|_A}$.*

Proof. This follows from noting that

$$\int_B f d\mu = \int_B f|_A d\mu|_A,\ B \in \mathcal{F}(A). \tag{1.1}$$

□

Proposition 1.5. *Let λ and μ be σ-finite measures on $(\Omega, \mathcal{F})$ such that $\lambda << \mu$ and let $f \in [\frac{d\lambda|_A}{d\mu|_A}]_{\mu|_A}$. If $\mu(A^c) = 0$, then $g = fI_A \in [\frac{d\lambda}{d\mu}]_\mu$.*

Proof. Let $B \in \mathcal{F}$. Since λ is dominated by μ and $\mu(A^c) = 0$, it follows that $\lambda(A^c) = 0$. Hence, $\lambda(B) = \lambda(A \cap B)$. Moreover,

$$\lambda(A \cap B) = \lambda\big|_A(A \cap B) = \int_{A\cap B} f|_A d\mu|_A = \int_{A\cap B} f d\mu,$$

where the last equality follows from (1.1). Then, since $\mu(A^c) = 0$, we have that

$$\lambda(B) = \int_{A\cap B} f d\mu = \int_B f d\mu.$$

Since B was taken arbitrary, the proof is complete. □

1.1.2 Fundamentals of Topology

In this section, we present all the definitions and results from Topology needed to state and prove our results. The first part is dedicated to some general concepts, which will be useful when the σ-algebra of the problem is induced by a topology. The second part presents the formal definition of continuous functions. Finally, in the last part of this section, we consider measures defined on abstract spaces.

General Topology

Definition 1.5. A family $\mathcal{T}$ of subsets of a set Ω is said to be a topology on Ω in case:

(i) $\emptyset$ and Ω are in $\mathcal{T}$.

(ii) The union of the elements of any subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

(iii) The intersection of the elements of any finite subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

An ordered pair $(\Omega, \mathcal{T})$ consisting of a set Ω and a topology $\mathcal{T}$ on Ω is called a topological space. Any set in $\mathcal{T}$ is called an open set of Ω. Sometimes, we prefer say "U is a neighborhood of ω" rather than "U is an open set containing ω".
If U is an open set of Ω, then the set $U^c = \Omega - U$ is called an closed set of Ω. Therefore, $\emptyset$ and Ω are open and closed sets of Ω.

Theorem 1.2. *(Munkres (2014), Theorem 17.1) Let $(\Omega, \mathcal{T})$ be a topological space. Then we have that*

(1) Arbitrary intersections of closed sets are closed.

(2) Finite unions of closed sets are closed.

Definition 1.6. A basis for a topology on a set Ω is a collection $\mathcal{B}$ of subsets of Ω such that

(i) if $\omega \in \Omega$, then there exists $B \in \mathcal{B}$ such that $\omega \in B$.

(ii) If $\omega \in B_1 \cap B_2$, $B_1, B_2 \in \mathcal{B}$, then there exists $B_3 \in \mathcal{B}$, $B_3 \subset B_1 \cap B_2$, such that $\omega \in B_3$.

Proposition 1.6. *(Munkres (2014), Section 13) Let $\mathcal{B}$ a collection of subsets of a set Ω such that $\mathcal{B}$ satisfies the two conditions in Definition 1.6. Let $\mathcal{T}$ a family of subsets of Ω and suppose that a set U is in $\mathcal{T}$ if and only if for each $\omega \in U$, there exists $B \in \mathcal{B}$ such that $\omega \in B$ and $B \subset U$. Then, $\mathcal{T}$ is a topology on Ω.*

A topology $\mathcal{T}$ given by the proposition above is called the topology generated by $\mathcal{B}$. The elements of the collection $\mathcal{B}$ are called basis elements. Hence, note that if a topology $\mathcal{T}$ on a set Ω is generated by $\mathcal{B}$, then each basis element is itself an open set of Ω.

Definition 1.7. If d is a metric on Ω, then the collection of all ϵ-balls $B_d(x, \epsilon) = \{\upsilon \in \Omega;\ d(\omega, \upsilon) < \epsilon\}$, for each $\omega \in \Omega$ and for each $\epsilon > 0$, is a basis for a topology on Ω. This topology generated by the ϵ-balls is called the metric topology induced by d.

Given $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$, we define the norm of the vector x by

$$\|x\| = \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}}$$

and we define the euclidean distance in $\mathbb{R}^n$ by

$$d(x, y) = \|x - y\| = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{\frac{1}{2}}.$$

This is indeed a metric. The topology generated by the euclidean distance induces the standard topology on $\mathbb{R}^n$. The space $\mathbb{R}^n$ with its usual topology will be considered in Chapter 4.
From now, we will simplify the notation. If $(\Omega, \mathcal{T})$ is a topological space, we will omit $\mathcal{T}$ and just say that Ω is a topological space.

Definition 1.8. If Ω is a topological space, Ω is said to be metrizable if there exists a metric d on Ω that induces the topology of Ω.

Proposition 1.7. *(Munkres (2014), Section 16) Let $(\Omega, \mathcal{T})$ be a topological space. If Γ is a subset of Ω, the collection*

$$\mathcal{T}_\Gamma = \{\Gamma \cap U;\ U \in \mathcal{T}\}$$

is a topology on Γ.

The topology $\mathcal{T}_\Gamma$ is called the subspace topology and $(\mathcal{T}_\Gamma)$ is called a subspace of $(\Omega, \mathcal{T})$.

Proposition 1.8. *Any subspace of a metrizable space is metrizable.*

Proof. Let $(\Omega, \mathcal{T})$ be a metrizable space by some metric d and let $(\Gamma, \mathcal{T}_\Gamma)$ a subspace of $(\Omega, \mathcal{T})$. Then, the metric $d|_{\Gamma\times\Gamma} = d_\Gamma$ defined by $d_\Gamma(\omega, \upsilon) = d(\omega, \upsilon)$ for all $\omega, \upsilon \in \Gamma$ is a metric on Γ which induces the topology $\mathcal{T}_\Gamma$. □

Definition 1.9. Given a subset A of a topological space Ω, the interior of A is defined as the union of all open sets contained in A, and the closure of A is defined as the intersection of all closed sets containing A.

The interior of A is denoted by A^o and the closure of A is denoted by $\bar{A}$. By the definition of a topology, A^o is an open set and $\bar{A}$ is a closed set.

Definition 1.10. A subset A of a space Ω is said to be dense in Ω if $\bar{A} = \Omega$.

Definition 1.11. A collection $\mathcal{A}$ of subsets of a space Ω is said to cover Ω, or to be a covering of Ω, if the union of the elements of $\mathcal{A}$ is equal to Ω. The collection $\mathcal{A}$ is called an open covering of Ω if each of its elements is an open set of Ω.

Definition 1.12. A space for which every open covering contains a countable subcovering is called a Lindelöf space. A space having a countable dense subset is often said to be separable.

Proposition 1.9. *An arbitrary intersection of closed sets in a Lindelöf space can be writen as a countable intersection of closed sets.*

Proof. Let Ω be a Lindelöf space and $\{B_\theta\}_{\theta\in\Theta}$ be a collection of closed sets of Ω, where Θ is a nonempty index set. Then, $\Omega - B_\theta$ is open for each $\theta \in \Theta$ and, consequently, $\bigcup_{\theta\in\Theta}(\Omega - B_\theta)$ is open. Hence, since Ω is Lindelöf, there exists a countable sequence $\{\theta_j\}_{j\in\mathbb{N}} \subset \Theta$ such that $\bigcup_{\theta\in\Theta}(\Omega - B_\theta) = \bigcup_{j\in\mathbb{N}}(\Omega - B_{\theta_j})$. Therefore,

$$\bigcap_{\theta\in\Theta} B_\theta = \Omega - \left(\bigcup_{\theta\in\Theta}(\Omega - B_\theta)\right) = \Omega - \left(\bigcup_{j\in\mathbb{N}}(\Omega - B_{\theta_j})\right) = \bigcap_{j\in\mathbb{N}} B_{\theta_j}$$

and the proof is complete. □

Theorem 1.3. *(Heinonen et al. (2015), Section 3.3) Every subspace of a separable metric space is separable.*

Theorem 1.4. *(Heinonen et al. (2015), Section 3.3) A metric separable space is Lindelöf.*

Continuous functions

In Section 2.3.1 we investigate how continuous RN derivatives can lead to proportional likelihood functions. Next, we present the definition and properties that will be necessary to understand our results in that section.

Definition 1.13. Let Ω and Υ be two topological spaces. A function $f : \Omega \longrightarrow \Upsilon$ is said to be continuous if for each open set V of Υ, the set $f^{-1}(V)$ is an open set of Ω.

Theorem 1.5. *(Munkres (2014), Theorem 18.1) Let Ω and Υ be two topological spaces and let $f : \Omega \longrightarrow \Upsilon$. Then, f is continuous if and only if for each $\omega \in \Omega$ and each neighborhood V of $f(\omega)$, there is a neighborhood U of ω such that $f(U) \subset V$.*

Proposition 1.10. *(Munkres (2014), Theorem 18.2) If $f : \Omega \longrightarrow \Upsilon$ is continuous and Γ is a subspace of Ω, then the restricted function $f\big|_\Gamma : \Gamma \longrightarrow \Upsilon$ is continuous.*

Definition 1.14. Let $f : \Omega \longrightarrow \Upsilon$ and let $\omega \in \Omega$. If for each neighborhood V of $f(\omega)$, there exists a neighborhood U of ω such that $f(U) \subset V$, then we say that f is continuous at ω.

Theorem 1.6. *(Munkres (2014), Theorem 21.1) Let $f : \Omega \longrightarrow \Upsilon$ and let Ω and Υ be metrizable with metrics d_Ω and d_Υ, respectively. Then, f is continuous at $\omega \in \Omega$ if, and only if, for each $\epsilon > 0$ there exists $\delta > 0$ such that*

$$d_\Omega(\omega, v) < \delta \Longrightarrow d_\Upsilon(f(\omega), f(v)) < \epsilon.$$

Theorem 1.7. *(Munkres (2014), Theorem 21.3) Let $f : \Omega \longrightarrow \Upsilon$. If f is continuous at ω, then for every convergent sequence $\omega_n \longrightarrow \omega$, the sequence $f(\omega_n) \longrightarrow f(\omega)$. The converse holds if Ω is a metrizable space.*

Measures and topological spaces

In Chapter 2, our sample space is a separable metric space. This space is very special since we can extract a countable subcovering from every open covering. In particular, since the σ-algebra will be generated by the open balls of this metric separable space it will be possible to calculate the probability of an arbitrary union of events. This is the key of the proof of Theorem 2.2.

Definition 1.15. Let Ω be a topological space. The Borel σ-algebra $\mathcal{B}(\Omega)$ is the smallest σ-algebra in Ω that contains all open subsets of Ω. The elements of $\mathcal{B}(\Omega)$ are called Borel sets of Ω.

Since a topology is closed for arbitrary union of open sets and all open sets of Ω are in $\mathcal{B}(\Omega)$, the union of any collection of open sets is a Borel set. On the other hand, since a σ-algebra is closed under complement, all closed sets of Ω are Borel sets and, consequently, the intersection of any collection of closed sets is a Borel set.

Definition 1.16. A topological space Ω is called a Hausdorff space if for each pair ω_1, ω_2 of distinct points of Ω, there exist neighborhoods U_1 and U_2 of ω_1 and ω_2, respectively, such that $U_1 \cap U_2 = \emptyset$.

Definition 1.17. Let Ω be a Hausdorff space. A measure μ on the σ-algebra $\mathcal{B}(\Omega)$ is called:

1. a Borel measure on Ω if

$$\mu(K) < +\infty$$

for every compact $K \subset \Omega$;

2. locally finite if every point ω of Ω has a neighborhood U_ω such that $\mu(U_\omega) < +\infty$;

3. inner regular if for every $B \in \mathcal{B}(\Omega)$

$$\mu(B) = \sup\{\mu(K);\ K \subset B,\ K \text{ is compact}\};$$

4. outer regular if for every $B \in \mathcal{B}(\Omega)$

$$\mu(B) = \inf\{\mu(U);\ B \subset U,\ U \text{ is open}\};$$

5. regular if it is both inner regular and outer regular.

Definition 1.18. A measure defined on the Borel σ-algebra $\mathcal{B}(\Omega)$ of a Hausdorff space Ω is called a Radon measure on Ω if its is both locally finite and inner regular.

Definition 1.19. A topological space Ω is called Polish when its topology has a countable base and can be defined by a complete metric.

A metric is called complete when it induces a complete space. In turn, a space Ω is said to be complete if every Cauchy sequence in Ω converges.

Theorem 1.8. *(Bauer (2001), Theorem 26.3) On a Polish space Ω every locally finite Borel measure μ is a σ-finite Radon measure.*

Every finite measure is locally finite. Hence, on a Polish space, every probability measure is a Radon Measure. We end this chapter with some results from Piccioni (1982). These results will be usefull in Section 2.3.

Let Ω be a metric separable space and let $\mathcal{B}(\Omega)$ the Borel σ-algebra of Ω

Theorem 1.9. *(Piccioni (1982), Theorem I) Any locally finite measure on $(\Omega, \mathcal{F})$ is σ-finite.*

Theorem 1.10. *(Piccioni (1982), Theorem II) If μ is a locally finite measure on $(\Omega, \mathcal{B}(\Omega))$, then the support of μ, say S_μ, has total measure, i.e., $\mu(S_\mu) = \mu(\Omega)$. In particular, any probability measure defined on a metric separable space has a support with total measure.*

Proposition 1.11. *Let ν and μ be measures on $(\Omega, \mathcal{F})$ and let S_ν and S_μ the supports of ν and μ, respectively. If $\nu << \mu$, then $S_\nu \subset S_\mu$.*

Proof. For any $\omega \notin S_\mu$, there exists an open set U_ω such that $\mu(U_\omega) = 0$. Because $\nu << \mu$, it follows that $\nu(U_\omega) = 0$. Then, $\omega \notin S_\nu$ and $S_\mu^c \subset S_\nu^c$. □

The following result from Piccioni (1982) guarantees the uniqueness of continuous versions of densities under some mild conditions.

Theorem 1.11. *Let μ and ν be LF measures on $(\Omega, \mathcal{F})$ such that $\mu << \nu$ and $S_\mu = S_\nu = \Omega$. If there exists a continuous version of $d\mu/d\nu$ on Ω, it is unique.*

The following variate of the previous theorem will be of particular interest in the results presented further ahead in this thesis.

Theorem 1.12. *Let μ and ν be LF measures on $(\Omega, \mathcal{F})$ such that $\mu << \nu$. If there exists a continuous version of $d\mu/d\nu$ on S_μ, it is unique.*

Proof. Simply use Proposition 1.11, consider the measures $\mu\big|_{S_\mu}$ and $\nu\big|_{S_\mu}$ and apply the previous theorem. □

Chapter 2

Likelihood Proportionaly Theorem

In this Chapter we provide a rigorous definition of the likelihood function. If a population $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$, Θ a nonempty set, is dominated by a σ-finite measure ν, the likelihood function is a function of θ and will be defined in terms of the Radon-Nikodým derivatives of P_θ with respect to ν. Since this definition depends on the choice of ν, we ask ourselves whether the choice of the dominanting measure has any influence in the inference process.

2.1 Motivation

Consider the following definition.

Definition 2.1 (Statistical model)**.** A statistical model is a family of probability measures $\mathcal{P}$ on $(\Omega, \mathcal{F})$, i.e $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$, where the P_θ's are probability measures and Θ is an arbitrary index set. In the particular case where $\Theta \subset \mathbb{R}^d$ for $d \in \mathbb{N}$, $\mathcal{P}$ is called a parametric model, θ a parameter and Θ the parameter space. In any other case $\mathcal{P}$ is called a non-parametric model.

A statistical inference problem can be generally described as follows. Given a model $\mathcal{P}$, one wants to estimate a population $P_{\theta^*} \in \mathcal{P}$ based on a sample (realization(s) from P_{θ^*} - a random experiment). The likelihood function is one way to quantify the likelihood of each P_θ having generated the data. We formally define the likelihood function as follows.

Definition 2.2 (Likelihood function)**.** Let $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$ be a statistical model and ν any σ-finite measure such that $\mathcal{P} << \nu$. For a given observed sample point ω, the likelihood function $l(\theta; \omega)$ for $P_\theta \in \mathcal{P}$ is given by the Radon-Nikodým derivative $\dfrac{dP_\theta}{d\nu}(\omega)$, for all $\theta \in \Theta$.

Note that the only condition imposed to the measure ν in the definition above is that it dominates the model $\mathcal{P}$. In many cases, the choice for ν is natural and no other possibility is even considered. For example, the Lebesgue measure for continuous random variables and the counting measure for discrete ones. This is in fact how the likelihood function is defined in many books. Nevertheless, it may be the case that the choice of the dominating measure is not obvious and it is then natural to question how one should proceed. Consider, for example, a homogeneous Poisson processes (PP) in a region $S \in \mathbb{R}^d$, i.e. $\mathcal{P} = \{PP(\lambda),\ \lambda \in \mathbb{R}^+\}$. There are two well-known choices for the dominating measure in this case. The first one is to use the measure of a homogeneous Poisson process with rate λ_0, for any fixed $\lambda_0 > 0$. The second one, based on the factorisation of the process in terms of the number of points N and their locations, is the product measure between the counting and the N-dimensional Lebesgue measures. The two likelihood functions induced by those measures are, respectively,

$$\exp\left\{-\int_S(\lambda-\lambda_0)ds\right\}\prod_{j=1}^{N}(\lambda/\lambda_0) \quad \text{and} \quad \frac{e^{-\lambda\mu(S)}(\lambda\mu(S))^N}{N!}(\mu(S))^{-N}, \tag{2.1}$$

where $\mu(S)$ is the volume of S.
Note that both functions are proportional w.r.t. λ and, therefore, under the Likelihood Principle, lead to the same inference. It is then natural to ask ourselves if different choices for the dominating measure always lead to proportional likelihoods. A positive answer for this question validates Definition 2.2 in terms of the Likelihood Principle.
The Likelihood Principle specifies how the likelihood function ought to be used for data reduction - a detailed addressing of the LP can be found in Berger and Wolpert (1988).

The Likelihood Principle. *All the information about P_θ obtainable from an experiment is contained in the likelihood function for P_θ given the sample. Two likelihood functions contain the same information about P_θ if they are proportional to one another.*

2.2 Likelihood Proportionality Theorem

The proportionality mentioned in the LP stated above means that $l_1(\theta;\omega) = h(\omega)l_2(\theta;\omega)$, with l_1 and l_2 being the two likelihood functions. This is a general version of the LP, which is a variant from the version presented in Berger and Wolpert (1988) [page 19]. It may be contextualised in different cases and stated in particular ways. For example, considering different data points ω or even different experiments which, in our construction, could be characterised as the sample consisting of observing different functions $X \in M(\Omega, \mathcal{F})$. However, as motivated by the example in the previous section, we consider the LP under the perspective of different dominating measures. This way, Definition 2.2 is validated by the LP if different dominating measures lead to proportional likelihood functions. Such a result is stated in its details in the Likelihood Proportionality Theorem further ahead in this section.

Before stating and proving the theorem, we need some auxiliary results. The first one is a neat result from Halmos and Savage (1949) (Lemma 7) considering dominated families of measures.

Lemma 2.1. *(Halmos and Savage (1949)) Let $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$ be a family of probability measures and ν a σ-finite measure on $(\Omega, \mathcal{F})$. If $\mathcal{P} << \nu$ then there exists a probability measure Q, such that $\mathcal{P} << Q$ and $Q = \sum_{i=1}^{\infty} c_i P_{\theta_i}$, where the c_i's are nonnegative constants with $\sum_{i=1}^{\infty} c_i = 1$ and $P_{\theta_i} \in \mathcal{P}$, $i \in \mathbb{N}$.*

Proof. First, consider the case where ν is a finite measure. Let

$$\mathcal{P}_0 = \left\{ \sum_{i=1}^{\infty} c_i P_i;\ P_i \in \mathcal{P},\ c_i \geq 0 \text{ and } \sum_{i=1}^{\infty} c_i = 1 \right\}$$

so $\mathcal{P} \subset \mathcal{P}_0$ and if $Q \in \mathcal{P}_0$, then $Q << \nu$. Now, let $\mathcal{C}$ be the class of all measurable sets C for which there exists $Q \in \mathcal{P}_0$ such that $Q(C) > 0$ and $dQ/d\nu > 0$ ν-a.e. on C. To see that $\mathcal{C}$ is not empty, take any $P_0 \in \mathcal{P}$ and note that $\{\omega \in \Omega;\ dP_0/d\nu(\omega) > 0\} \in \mathcal{C}$. Since ν is a finite measure, it follows that $\sup_{C\in\mathcal{C}} \nu(C) < \infty$. Moreover, there exists a sequence $\{C_i\}_{i=1}^{\infty} \subset \mathcal{C}$ such that $\nu(C_i) \longrightarrow \sup_{C\in\mathcal{C}} \nu(C)$. For each C_i, let $Q_i \in \mathcal{P}_0$ such that $Q_i(C_i) > 0$ and $dQ_i/d\nu > 0$ ν-a.e. on C_i. Let $Q_0 = \sum_{i=1}^{\infty} 2^{-i} dQ_i/d\nu \in \mathcal{P}_0$. It follows that $dQ_0/d\nu = \sum_{i=1}^{\infty} 2^{-i} dQ_i/d\nu$. Let $C_0 = \cup_{i=1}^{\infty} C_i$. Since

$$\bigcup_{i=1}^{\infty} \left\{ \omega \in C_i;\ \frac{dQ_i}{d\nu}(\omega) > 0 \right\} \subset \left\{ \omega \in C_0;\ \frac{dQ_0}{d\nu}(\omega) > 0 \right\},$$

it follows that $C_0 \in \mathcal{C}$ and, consequently, $\sup_{C\in\mathcal{C}} \nu(C) = \nu(C_0)$.

We now prove that $P << Q_0$ for all $P \in \mathcal{P}$. Suppose that $Q_0(A) = 0$. Let $P \in \mathcal{P}$ and $B = \{\omega \in \Omega;\ dP/d\nu(\omega) > 0\}$. Since $Q_0(A \cap C_0) = 0$ and $dQ_0/d\nu > 0$ ν-a.e. on C_0, it follows that $\nu(A \cap C_0) = 0$ and, consequently, $P(A \cap C_0) = 0$. Then, $P(A) = P(A \cap B) = P(A \cap B \cap C_0^c)$. If $P(A \cap B \cap C_0^c) > 0$, then $\nu(A \cap B \cap C_0^c) > 0$. But $C_0 \cup (A \cap B \cap C_0^c) \in \mathcal{C}$ and $\nu(C_0 \cup (A \cap B \cap C_0^c)) = \nu(C_0) + \nu(A \cap B \cap C_0^c) > \nu(C_0)$, which contradicts $\nu(C_0) = \sup_{C \in \mathcal{C}} \nu(C)$. Hence, $P(A) = 0$.
For the case where ν is a σ-finite measure, it suffices, in view of the preceding case, to show that there exists a finite measure μ that dominates the family $\mathcal{P}$. Since ν is a σ-finite measure, there exists a partition $\{A_n\}_{n=1}^{\infty}$ of Ω such that $\nu(A_n) < \infty$ for all $n \in \mathbb{N}$. For each $B \in \mathcal{F}$, let $\mu(B) = \sum_{n=1}^{\infty} \nu(B \cap B_n)/(2^n \nu(B_n))$. It follows that μ is a finite measure on $(\Omega, \mathcal{F})$ and $P << \mu$ for every $P \in \mathcal{P}$. □

We now present a definition regarding sets of dominating measures and a proposition which will play an important role in the proof of the Likelihood Proportionality Theorem.

Definition 2.3. For a family of probability measures $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$, suppose that the family $\Upsilon = \{\nu;\ \mathcal{P} << \nu \text{ and } \nu \text{ is } \sigma\text{-finite}\}$ is nonempty. If there exists $\lambda \in \Upsilon$ such that $\lambda << \nu$ for all $\nu \in \Upsilon$, then we say that λ is a minimal dominating measure for the family $\mathcal{P}$.

Note that a minimal dominating measure is not necessarily unique. However, by definition, two minimal dominating measures are always equivalent.

Proposition 2.1. *Let $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$ be a family of probability measures defined on the measurable space $(\Omega, \mathcal{F})$. Suppose that the family $\Upsilon = \{\nu;\ \mathcal{P} << \nu\}$ is nonempty. Then, there exists a minimal dominating measure λ for $\mathcal{P}$.*

Proof. Since $\Upsilon \neq \emptyset$, there exists $\nu \in \Upsilon$ such that $\mathcal{P} << \nu$ and ν is a σ-finite measure. Then, it follows from Lemma 2.1 that there exists a measure λ such that $\mathcal{P} << \lambda$ and where $\lambda = \sum_{i=1}^{\infty} c_i P_{\theta_i}$, where the c_i's are nonnegative constants with $\sum_{i=1}^{\infty} c_i = 1$ and $P_{\theta_i} \in \mathcal{P}$. We now show that measure λ is a minimal dominating measure w.r.t. $\mathcal{P}$, i.e. if $\nu \in \Upsilon$, then $\lambda << \nu$. Take any $\nu \in \Upsilon$ and let $A \in \mathcal{F}$ such that $\nu(A) = 0$. Then, $P_\theta(A) = 0$ for all $\theta \in \Theta$ and, particularly, $P_{\theta_i}(A) = 0$, for all $i \in \mathbb{N}$. Thus, $\lambda(A) = \sum_{i=1}^{\infty} c_i P_{\theta_i}(A) = 0$. □

For a function f in $M(\Omega, \mathcal{F})$, define $[f]_\mu$ as the equivalence class of f with respect to μ, i.e. the collection of all functions g in $M(\Omega, \mathcal{F})$ such that $g = f$ μ-a.s. We now state and prove the Likelihood Proportionality Theorem.

Theorem 2.1 (The Likelihood Proportionality Theorem)**.** *Let $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$ be a family of probability measures and ν_1, ν_2 σ-finite measures on $(\Omega, \mathcal{F})$, where*

Θ is a nonempty set. Suppose that $\mathcal{P} << \nu_1$ and $\mathcal{P} << \nu_2$. Then, there exists a measurable set A such that $P_\theta(A) = 1$, for all $\theta \in \Theta$, and there exist $f_{1,\theta} \in [\frac{dP_\theta}{d\nu_1}]_{\nu_1}$, $f_{2,\theta} \in [\frac{dP_\theta}{d\nu_2}]_{\nu_1}$, for all $\theta \in \Theta$, and a measurable function h such that

$$f_{1,\theta}(\omega) = h(\omega) f_{2,\theta}(\omega), \ \forall \theta \in \Theta, \ \forall \omega \in A. \tag{2.2}$$

Proof. Let ν be a minimal dominating measure for $\mathcal{P}$ (its existence is guaranteed by Proposition 2.1). Now, take $h_1 \in [\frac{d\nu}{d\nu_1}]_{\nu_1}$, $h_2 \in [\frac{d\nu}{d\nu_2}]_{\nu_1}$ and, for each $\theta \in \Theta$, take $g_\theta \in [\frac{dP_\theta}{d\nu}]_{\nu_1}$. Define, for each $\theta \in \Theta$, $f_{1,\theta}(\omega) = g_\theta(\omega)h_1(\omega)$ and $f_{2,\theta}(\omega) = g_\theta(\omega)h_2(\omega)$. It follows that $f_{1,\theta} \in [\frac{dP_\theta}{d\nu_1}]_{\nu_1}$ and $f_{2,\theta} \in [\frac{dP_\theta}{d\nu_2}]_{\nu_1}$. Let

$$A = \{\omega \in \Omega; \ h_2(\omega) > 0\}$$

so that $\nu(A^c) = 0$ and consequently $P_\theta(A) = 1$ for all $\theta \in \Theta$. Let h be defined to be

$$h(\omega) = \begin{cases} \frac{h_1(\omega)}{h_2(\omega)}, & \text{if} \quad \omega \in A, \\ 0, & \text{if} \quad \omega \in A^c. \end{cases}$$

Then, $h \in M(\Omega, \mathcal{F})$ and

$$f_{1,\theta}(\omega) = h(\omega) f_{2,\theta}(\omega), \ \forall \theta \in \Theta, \ \forall \omega \in A.$$

□

Discussion of Theorem 2.1. Note that equation (2.2) implies that $f_{1,\theta}(\omega) \propto_\theta f_{2,\theta}(\omega)$, $\forall \theta \in \Theta$, $\forall \omega \in A$, which validates Definition 2.2 in terms of the Likelihood Principle i.e., independent of the choice of the dominating measure the inference will (a.s.) be the same. Furthermore, the proportionality result is valid a.s. P_θ, for all $\theta \in \Theta$, in particular, for the true θ.

Note, however, that Theorem 2.1 states the existence of versions of RN derivatives that satisfies (2.2), which means that not all versions necessarily do. In this sense, it would be useful to define a class of versions that always satisfy (2.2) and, possibly, lead to a well-behaved likelihood function, for example, that satisfy the classical regularity conditions. We further explore this issue in the first part of Section 2.3 and in Chapter 4, considering versions of RN derivatives that satisfy some continuity properties.

In some cases, $[\frac{dP_\theta}{d\nu_1}]_{\nu_1}$ and $[\frac{dP_\theta}{d\nu_2}]_{\nu_2}$ are unitary sets. For example, in a family $\mathcal{P} = \{P_\theta; \ \theta \in \Theta\}$ of discrete distributions, i.e. $P_\theta(\omega) > 0$, for all $\theta \in \Theta$ and for all $\omega \in \Omega$. Another interesting particular example is the case where the family of

probability measures is a countable set. In this case, any pair of versions of the RN derivative satisfies (2.2).
Proposition 2.2 below relates Theorem 2.1 to the Factorisation Theorem.

Proposition 2.2. *Consider $\mathcal{P}$, ν_1 and ν_2 from Theorem 2.1, Q from Lemma 2.1 and let T be a sufficient statistic for $\mathcal{P}$ with range space $(\mathcal{T}, \mathcal{B})$. Then:*

i) For each version $g_\theta^* \in [\frac{dP_\theta}{dQ}]_{\nu_1}$ in $(\Omega, \sigma(T))$ and $h_1 \in [\frac{dQ}{d\nu_1}]_{\nu_1}$ in $(\Omega, \mathcal{F})$, there exists a $\mathcal{B}$-measurable function g_θ such that $g_\theta^* = g_\theta \circ T$ and the function $f_{1,\theta} = (g_\theta \circ T)h_1$ is a version in $[\frac{dP_\theta}{d\nu_1}]_{\nu_1}$, for all θ.

ii) If we obtain $f_{1,\theta}$ and $f_{2,\theta}$ as in i), for ν_1 and ν_2, respectively, from the same g_θ^*, then $f_{1,\theta} \propto f_{2,\theta}$ in a measurable set A, for all θ, such that $\nu_1(A^c) = 0$.

Proof. To prove i), for each $\theta \in \Theta$, take $g_\theta^* \in [\frac{dP_\theta}{dQ}]_{\nu_1}$ in $(\Omega, \sigma(T))$ and $h_1 \in [\frac{dQ}{d\nu_1}]_{\nu_1}$ in $(\Omega, \mathcal{F})$. Then, there exists a $\mathcal{B}$-measurable function g_θ such that $g_\theta^* = g_\theta \circ T$ (see Shao, 2003, Section 1.4, Lemma 1.2). Now, since T is a sufficient statistic for $\mathcal{P}$, it follows that $g_\theta \circ T \in [\frac{dP_\theta}{dQ}]_{\nu_1}$ in $(\Omega, \mathcal{F})$ (see Lehmann, 1986, Section 2.6, Theorem 8). Define the function $f_{1,\theta}$ as

$$f_{1,\theta}(\omega) = g_\theta(T(\omega))h_1(\omega), \ \forall \omega \in \Omega.$$

Thus, it follows from the RN chain rule, that $f_{1,\theta} \in [\frac{dP_\theta}{d\nu_1}]_{\nu_1}$ for all $\theta \in \Theta$.
To prove ii), let $f_{1,\theta}(\omega) = g_\theta(T(\omega))h_1(\omega)$ and $f_{2,\theta}(\omega) = g_\theta(T(\omega))h_2(\omega)$ for all $\omega \in \Omega$ and $\theta \in \Theta$, where $h_2 \in [\frac{dQ}{d\nu_2}]_{\nu_2}$. Let $A = \{\omega \in \Omega;\ h_1(\omega) > 0\}$. Then, $\nu_1(A^c) = 0$ and $f_{1,\theta} \propto f_{2,\theta}$ in A, for all $\theta \in \Theta$. □

Part *i*) from Proposition 2.2 can be seen as a stronger version of the Factorisation Theorem as it states that the density representation is valid for all θ in the whole Ω, i.e. it holds P_θ a.s., for all θ. The classical version of the Factorisation Theorem is a consequence since all versions in $[\frac{dP_\theta}{d\nu_1}]_{\nu_1}$ are P_θ equivalent.
Finally, note that the result in Theorem 2.1 is valid for any topological structure induced in the sample space Ω, in particular, if Ω is non-separable and/or non-metric. In the next section, a specific topological space will be needed.

2.3 Properties under continuity assumptions

We now discuss two particular versions of RN derivatives that always satisfy the Likelihood Proportionality Theorem.

2.3.1 Continuous versions of Radon-Nikodým derivatives

As we have mentioned before, we would like to define a subclass of RN versions that would always satisfy the proportionality relation (2.2) and, therefore, provide a practical way to obtain a likelihood function. That is achieved by considering continuous versions of densities. We state two results (Theorem 2.2 and Proposition 2.3) that, under different assumptions, guarantee that continuous versions of the RN derivatives, when these exist, do satisfy (2.2). In fact, in Piccioni (1982) and Piccioni (1983), the likelihood function is defined as a continuous version of the RN derivative. The author proves, under some additional assumptions, that, if such a version exists, it is unique and this particular definition is justified by the fact that such a version is related to a limit that builds on the intuition of likelihood. Finally, regarding well-behaved versions of the likelihood function, continuity is a particular property of interest. In particular, most of the important results regarding properties of the MLE rely on assumptions that include continuity. In some cases, specially for parametric models, continuity of the likelihood is implied by continuity of the RN density.

For the whole of this section, let Ω be a metric separable space with a distance that induces the topology $\mathcal{T}$. As usual, $\mathcal{F}$ is the smallest σ-algebra containing $\mathcal{T}$ - the Borel σ-algebra of Ω. Since the topology of Ω is induced by a metric, Ω is metrizable.

We now discuss why continuous versions of densities lead to likelihood functions that carry the true intuition of likelihood. In the simplest case where Ω is discrete, the likelihood is proportional to the probability of the observed sample which gives a clear interpretation to the concept of likelihood. This concept is extended to the continuous case by defining the likelihood ratio in a point ω_0 as the limit

$$\lim_{A \longrightarrow \omega_0} \frac{P_\theta(A)}{\nu(A)}, \tag{2.3}$$

where A is a neighborhood of ω_0 such that the diameter of A tends to zero. Piccioni (1982) shows that there exists a continuous version f_θ^c of $dP_\theta/d\nu$ if and only if there exists the limit in (2.3), in which case $f_\theta^c(\omega_0)$ is exactly this limit.

It is natural to expect that continuous versions will satisfy the proportionality relation (2.2). This is established in Theorem 2.2 and Proposition 2.3 below. In order to prove these two results, we require the following Lemma and Definition, which are valid for general sample spaces Ω.

Lemma 2.2. *Let $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$ be a family of probability measures and ν_1 and ν_2 σ-finite measures on $(\Omega, \mathcal{F})$, where Θ is a nonempty set. Suppose that $\mathcal{P} << \nu_1$ and $\mathcal{P} << \nu_2$. Then, there exists a measurable set A such that*

(i) $P_\theta(A) = 1$, *for all* $\theta \in \Theta$ *and*

(ii) $\nu_1\big|_A$ *and* $\nu_2\big|_A$ *are equivalent measures, that is,* $\nu_1\big|_A << \nu_2\big|_A$ *and* $\nu_2\big|_A << \nu_1\big|_A$.

Proof. Since $\mathcal{P} << \nu_1$, it follows from Lemma 2.1 that there exists a sequence $\{c_i\}_{i=1}^\infty$ of nonnegative constants such that $\sum_{i=1}^\infty c_i = 1$ and there exists a sequence $\{\theta_i\}_{i=1}^\infty \subset \Theta$ such that $\mathcal{P} << Q$, where $Q = \sum_{i=1}^\infty c_i P_{\theta_i}$. Now define for each $i \in \mathbb{N}$ the following sets

$$A_{1,i} = \left\{\omega \in \Omega;\ \frac{dP_{\theta_i}}{d\nu_1}(\omega) > 0\right\} \text{ and } A_{2,i} = \left\{\omega \in \Omega;\ \frac{dP_{\theta_i}}{d\nu_2}(\omega) > 0\right\}.$$

Thus, $P_{\theta_i}(A_{1,i}) = 1$ and $P_{\theta_i}(A_{2,i}) = 1, \forall i \in \mathbb{N}$. Let $A_i = A_{1,i} \cap A_{2,i}$ (A_i is measurable since every $A_{1,i}$ and $A_{2,i}$ are measurable), then $P_{\theta_i}(A_i) = 1,\ \forall i \in \mathbb{N}$. Let $A = \bigcup_{i\in\mathbb{N}} A_i$, then clearly $P_{\theta_i}(A) = 1,\ \forall i \in \mathbb{N}$ (simply note that $A_i \subset A$) and, since $Q = \sum_{i=1}^\infty c_i P_{\theta_i}$, it follows that $Q(A) = 1$. Now, since $Q(A^c) = 0$ and $\mathcal{P} << Q$, $P_\theta(A^c) = 0,\ \forall \theta \in \Theta$. Therefore, $P_\theta(A) = 1,\ \forall \theta \in \Theta$.
To prove (ii), suppose that there exists a measurable set $B \in \mathcal{F}$ such that $\nu_2(A \cap B) = 0$ but $\nu_1(A \cap B) > 0$. Since $A \cap B = \bigcup_{i\in\mathbb{N}}(A_i \cap B)$, there exists $i_0 \in \mathbb{N}$ such that $\nu_1(A_{i_0} \cap B) > 0$. Hence,

$$0 < \int_{A_{i_0}\cap B} \frac{dP_{\theta_{i_0}}}{d\nu_1} d\nu_1 = P_{\theta_{i_0}}(A_{i_0} \cap B). \tag{2.4}$$

On the other hand, since $\nu_2(A \cap B) = 0$, $\nu_2(A_i \cap B) = 0,\ \forall i \in \mathbb{N}$, and then, by hypothesis, $P_\theta(A_i \cap B) = 0,\ \forall(\theta, i) \in \Theta \times \mathbb{N}$. In particular, $P_{\theta_{i_0}}(A_{i_0} \cap B) = 0$, which contradicts (2.4). This implies that, for all $B \in \mathcal{F}$, if $\nu_2(A \cap B) = 0$, then $\nu_1(A \cap B) = 0$, which means that $\nu_1\big|_A << \nu_2\big|_A$. The proof that $\nu_2\big|_A << \nu_1\big|_A$ is symmetrically analogous.

□

Definition 2.4 (Dominating pair)**.** Consider $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$, where Θ is a nonempty set, to be a family of probability measures and let ν_1 and ν_2 be σ-finite measures on $(\Omega, \mathcal{F})$ such that $\mathcal{P} << \nu_1$ and $\mathcal{P} << \nu_2$. A pair (A, ν_3) is called a dominating pair for the triple $(\mathcal{P}, \nu_1, \nu_2)$, where $A \in \mathcal{F}$ and $\nu_3 = \sum_{i=1}^\infty c_i P_{\theta_i}$ for some sequences $\{\theta_i\}_{i=1}^\infty$ and $\{c_i\}_{i=1}^\infty$ such that $\sum_{i=1}^\infty c_i = 1$, if $\nu_i\big|_A << \nu_j\big|_A,\ i, j = 1, 2, 3$ and $\nu_3(A) = 1$.

Note that a dominating pair for $(\mathcal{P}, \nu_1, \nu_2)$ always exists. That is guaranteed by Lemma 2.1 and Lemma 2.2.

Theorem 2.2. *Let (A, ν_3) be a dominating pair for $(\mathcal{P}, \nu_1, \nu_2)$. If there exist continuous versions of Radon-Nikodým derivatives $f_{1,\theta} \in [\frac{dP_\theta|_A}{d\nu_1|_A}]_{\nu_1|_A}$, $f_{2,\theta} \in [\frac{dP_\theta|_A}{d\nu_2|_A}]_{\nu_1|_A}$, $\forall \theta \in \Theta$, then, for all $h \in [\frac{d\nu_2|_A}{d\nu_1|_A}]_{\nu_1|_A}$, there exists a measurable set $B_h \in \mathcal{F}(A)$ such that $P_\theta(B_h) = 1$, for all $\theta \in \Theta$, h is continuous on B_h and*

$$f_{1,\theta}(\omega) = h(\omega) f_{2,\theta}(\omega), \ \forall \theta \in \Theta, \ \forall \omega \in B_h.$$

Proof. Let $\{P_{\theta_i}\}$ be a family of probability measures used in the construction of the measure ν_3. Now, define measures $\dot{P}_\theta$, $\dot{\nu_1}$, $\dot{\nu_2}$ and $\dot{\nu_3}$ to be the restriction of the respective measures on $(A, \mathcal{F}(A))$, for all $\theta \in \Theta$. For each $i \in \mathbb{N}$, consider the continuous derivatives $f_{1,\theta_i} \in [\frac{d\dot{P}_{\theta_i}}{d\dot{\nu_1}}]_{\dot{\nu_1}}$, $f_{2,\theta_i} \in [\frac{d\dot{P}_{\theta_i}}{d\dot{\nu_2}}]_{\dot{\nu_1}}$ and take any $h \in [\frac{d\dot{\nu_2}}{d\dot{\nu_1}}]_{\dot{\nu_1}}$. For each $i \in \mathbb{N}$, define

$$A_i = \{\omega \in A \ f_{1,\theta_i}(\omega) = h(\omega) f_{2,\theta_i}(\omega)\}.$$

Note that the RN chain rule implies that $\nu_1(A_i^c) = 0$ for all $i \in \mathbb{N}$. Now, let

$$B_i = \{\omega \in A; \ f_{2,\theta_i}(\omega) > 0\},$$

and let $B = \bigcup_{i=1}^\infty B_i$, $D_h = \bigcap_{i=1}^\infty A_i$ and $S_h = D_h \cap B$. Hence, it follows that $\dot{\nu_3}(B) = 1 = \dot{\nu_3}(S_h) = 1$ and, consequently, $\dot{P}_\theta(S_h) = 1$, for all $\theta \in \Theta$. We claim that h is continuous on the subspace S_h. To see that, note first that since Ω is a metrizable and separable space we know from Theorem 1.8 that any subspace of a metrizable space is metrizable and from Theorem 1.3 that every subspace of a metric separable space is separable. Hence, the subspace S_h of Ω is metrizable and separable. Therefore, we can use Theorem 1.7 to prove that

$$h : S_h \longrightarrow \mathbb{R}$$

is continuous.Fix $\omega_0 \in S_h$ and let $\{\omega_n\}_{n=1}^\infty \subset S_h$ such that $\lim_n \omega_n = \omega_0$. Thus, by the definition of S_h, it follows that $\omega_0 \in D_h$ and there exists $i_0 \in \mathbb{N}$ such that $\omega_0 \in B_{i_0}$. This implies that

$$h(\omega_0) = \frac{f_{1,\theta_{i_0}}(\omega_0)}{f_{2,\theta_{i_0}}(\omega_0)}. \tag{2.5}$$

Now, since $f_{2,\theta_{i_0}}$ is continuous in A and $(0, +\infty)$ is an open set in $\mathbb{R}$, we have by Definition 1.14 that $B_{i_0} = f_{2,\theta_{i_0}}^{-1}((0, +\infty))$ is an open set in A. Hence, by the definition of a subspace, we have that $S_h \cap B_{i_0}$ is an open set in S_h. Thus, by the convergence of the sequence $\{\omega_n\}_{n=1}^\infty$, there exists $n_0 \in \mathbb{N}$ such that, for $n \geq n_0$,

$\omega_n \in S_h \cap B_{i_0}$ and

$$h(\omega_n) = \frac{f_{1,\theta_{i_0}}(\omega_n)}{f_{2,\theta i_0}(\omega_n)}. \tag{2.6}$$

Finally, from (2.5), (2.6) and the continuity of $f_{1,\theta_{i_0}}$ and $f_{2,\theta_{i_0}}$, it follows that

$$\lim_n h(\omega_n) = h(\omega_0),$$

which establishes the continuity of h in S_h.
Now, for each $\theta \in \Theta$, define the following set

$$B_\theta = \{\omega \in S_h;\ f_{1,\theta}(\omega) = h(\omega) f_{2,\theta}(\omega)\}.$$

It follows, by the RN chain rule, that $\dot{\nu_1}(B_\theta^c \bigcap S_h) = 0$, for all $\theta \in \Theta$. Furthermore, since the function $(f_{1,\theta} - h f_{2,\theta})$ is continuous on S_h, we have that $B_\theta = (f_{1,\theta} - h f_{2,\theta})^{-1(\{0\})}$ is a closed set in S_h for each $\theta \in \Theta$ and, consequently, $B_h = \bigcap_{\theta\in\Theta} B_\theta$ is also a closed set in S_h. Since S_h is metric separable, it follows from Theorem 1.4 that S_h is Lindelöf. This implies, by Theorem 1.9, that there exists a sequence $\{\theta_j\} \subset \Theta$ such that $B_h = \bigcap_{j=1}^{\infty} B_{\theta_j}$. Moreover, since $\dot{\nu_1}(B_\theta^c \bigcap S_h) = 0$, for all $\theta \in \Theta$, it follows that $\dot{\nu_1}(B_h^c \bigcap S_h) = 0$ which, in turn, implies that $\dot{P_\theta}(B_h) = 1$ for all $\theta \in \Theta$, and

$$f_{1,\theta}(\omega) = h(\omega) f_{2,\theta}(\omega),\ \forall\theta \in \Theta,\ \forall\omega \in B_h.$$

Moreover, $\dot{P_\theta}(B_h) = P_\theta\big|_A(B_h) = P_\theta(A \cap B_h) = 1$ for all $\theta \in \Theta$. Since $P_\theta(A) = 1$, we have that $P_\theta(B_h) = P_\theta(A \cap B_h) = 1$ for all $\theta \in \Theta$. The proof is complete. □

Theorem 2.2 defines a specific subclass of RN versions that always satisfies the proportionality relation (2.2). Moreover, if the dominating measures under consideration are locally finite (LF), Theorem 1.12 guarantees that the continuous version (w.r.t. each of the measures) is unique. In many statistical models, there exist, and it is straightforward to obtain, continuous versions of $f_{1,\theta}$ and $f_{2,\theta}$ in Ω, for all $\theta \in \Theta$.
Let S_ν be the support of a measure ν on $(\Omega, \mathcal{F})$. The following corollary applies to several examples of statistical models.

Corollary 2.1. *Suppose that ν_1 and ν_2 are LF measures with $S_{\nu_1} = \Omega$. Suppose also that there exist continuous versions of Radon-Nikodým derivatives $f_{1,\theta} \in [\frac{dP_\theta}{d\nu_1}]_{\nu_1}$, $f_{2,\theta} \in [\frac{dP_\theta}{d\nu_2}]_{\nu_1}$, for all $\theta \in \Theta$, and that $f_{1,\theta}(\omega) > 0$ and $f_{2,\theta}(\omega) > 0$, for all $\omega \in \Omega$ and $\theta \in \Theta$. Then*

$$f_{1,\theta}(\omega) \propto_\theta f_{2,\theta}(\omega),\quad \forall\omega \in \Omega,\ \forall\theta \in \Theta.$$

Proof. Since $f_{1,\theta}$ and $f_{2,\theta}$ are strictly positive in Ω, for all $\theta \in \Theta$, it follows that all the P_θ's, ν_1 and ν_2 are equivalent and, by Proposition 1.11, $S_\theta = S_{\nu_2} = S_{\nu_1} = \Omega$, for all $\theta \in \Theta$. For each $\theta \in \Theta$, define $h_\theta(\omega) = \frac{f_{1,\theta}(\omega)}{f_{2,\theta}(\omega)}$, for all $\omega \in \Omega$, and note that, for all $\theta \in \Theta$, $h_\theta \in [\frac{d\nu_2}{d\nu_1}]_{\nu_1}$ and h_θ is continuous in Ω. Since, ν_1 and ν_2 are LF measures, Theorem 1.12 guarantees that all the h_θ's coincide in Ω, i.e. $h_\theta = h$, for all $\theta \in \Theta$. The result follows from the fact that $f_{1,\theta}(\omega) = h(\omega) f_{2,\theta}(\omega)$, for all $\omega \in \Omega$ and for all $\theta \in \Theta$. □

Proposition 2.3. *Let $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$ be a family of probability measures and ν_1 and ν_2 LF measures on $(\Omega, \mathcal{F})$, where Θ is a nonempty set, $\mathcal{P} << \nu_1$, $\mathcal{P} << \nu_2$. Let (ν_3, A) be a dominating pair for $(\mathcal{P}, \nu_1, \nu_2)$ and S_θ, S_1, S_2 and S_3 be the supports of P_θ (for each $\theta \in \Theta$), ν_1, ν_2 and ν_3, respectively. If there exists a continuous version on S_θ of the Radon-Nikodým derivative $f_{2,\theta} \in [\frac{dP_\theta|_{S_\theta}}{d\nu_2|_{S_\theta}}]_{\nu_1|_{S_\theta}}$, $\forall \theta \in \Theta$, and there exists a continuous version on S_3 of the Radon-Nikodým derivative $h \in [\frac{d\nu_2|_{S_3}}{d\nu_1|_{S_3}}]_{\nu_1|_{S_3}}$, then $f_{2,\theta}$ and h are unique in S_θ and S_3, respectively, and there exists an unique continuous version of $f_{1,\theta} \in [\frac{dP_\theta|_{S_\theta}}{d\nu_1|_{S_\theta}}]_{\nu_1|_{S_\theta}}$ on S_θ, for all $\theta \in \Theta$. Moreover, we have that $f_{1,\theta}(\omega)$ and $f_{2,\theta}(\omega)$ are proportional for every $\theta \in \Phi_\omega = \{\theta \in \Theta;\ \omega \in S_\theta\}$.*

Proof. Simply note that $S_\theta \subset S_3$ (Proposition 1.11) and define $f_{1,\theta}(\omega) = h(\omega) f_{2,\theta}(\omega)$, $\forall \omega \in S_\theta$, $\forall \theta \in \Theta$. The uniqueness of $f_{1,\theta}$, $f_{2,\theta}$ and h is guaranteed by Theorem 1.12. □

2.3.2 Continuous likelihood functions

For each $\theta \in \Theta$, take a version $f_\theta \in [\frac{dP_\theta}{d\nu}]$ and let $\mathcal{F} = \{f_\theta;\ \theta \in \Theta\}$.

Definition 2.5. (Fraser and Naderi (2007)) $\mathcal{F}$ is said to be continuous on Θ if Θ is a separable metric space and if, for each ω in Ω, $f_\theta(\omega)$ is continuous in θ.

Lemma 2.3. *Let ν_1 and ν_2 be two σ-finite measures on $(\Omega, \mathcal{B})$ such that $\mathcal{P} << \nu_i$, $i = 1, 2$. Suppose that Θ is a separable metric space and for each $\theta \in \Theta$ there exists versions $f_{1,\theta} \in [\frac{dP_\theta}{d\nu_1}]$ and $f_{2,\theta} \in [\frac{dP_\theta}{d\nu_2}]$ such that $\mathcal{F}_1 = \{f_{1,\theta};\ \theta \in \Theta\}$ and $\mathcal{F}_2 = \{f_{2,\theta};\ \theta \in \Theta\}$ are continuous on Θ. Then, there exist a measurable set A and a measurable function h such that $P_\theta(A) = 1$ for all $\theta \in \Theta$ and*

$$f_{1,\theta}(\omega) = f_{2,\theta}(\omega) h(\omega), \forall \omega \in A,\ \forall \theta \in \Theta. \tag{2.7}$$

Proof. Since $\mathcal{P} << \nu_1$, it follows from Lemma 2.1 that there exists a sequence $\{c_i\}_{i=1}^\infty$ of nonnegative constants such that $\sum_{i=1}^\infty c_i = 1$ and there exists a sequence $\{\theta_i\}_{i=1}^\infty \subset \Theta$ such that $\mathcal{P} << Q$, where $Q = \sum_{i=1}^\infty c_i P_{\theta_i}$. Since Θ is separable there exists a countable subset $\Theta_0 = \{\theta_n\}_{n=1}^\infty \subset \Theta$ such that Θ_0 is dense in Θ. Then,

since $\mathcal{P} << Q$ and Θ_0 is a countable set, it follows from the RN chain rule that there exist a measurable set F and measurable functions $h_1 \in [\frac{dQ}{d\nu_1}]$ and $h_2 \in [\frac{dQ}{d\nu_2}]$ such that $Q(F) = 1$ and

$$f_{1,\theta}(\omega) = g_\theta(\omega)h_1(\omega), \forall\omega \in F, \forall\theta \in \Theta_0, \tag{2.8}$$

$$f_{2,\theta}(\omega) = g_\theta(\omega)h_2(\omega), \forall\omega \in F, \forall\theta \in \Theta_0, \tag{2.9}$$

where g_θ is the Radon-Nikodým derivative of P_θ with respect to Q. Let $G = \{\omega \in \Omega;\ h_2(\omega) > 0\}$. Define $A = F \cap G$ and let $l(\omega) = h_2(\omega)I_A(\omega) + I_{A^c}(\omega)$. Thus, $Q(A) = 1$ and

$$f_{1,\theta}(\omega) = f_{2,\theta}(\omega)\frac{h_1(\omega)}{l(\omega)}, \forall\omega \in A, \forall\theta \in \Theta_0. \tag{2.10}$$

Now, fix $\theta \in \Theta$. Then, there exists a sequence $\{\theta_n\}_{n=1}^{\infty} \subset \Theta_0$ such that $\lim_n \theta_n = \theta$. Since for each $\omega \in A$ the derivatives $f_{1,\theta}(\omega)$ and $f_{2,\theta}(\omega)$ are continuous on θ, taking the limit in (2.10), it follows that

$$f_{1,\theta}(\omega) = f_{2,\theta}(\omega)h(\omega), \forall\omega \in A,$$

where $h = h_1/l$. Since θ is arbitrary,

$$f_{1,\theta}(\omega) = f_{2,\theta}(\omega)h(\omega), \forall\omega \in A, \forall\theta \in \Theta,$$

where $P_\theta(A) = 1$ for all $\theta \in \Theta$ and h is a measurable function. □

Lemma 2.3 says that if $l_1(\theta|\omega)$ and $l_2(\theta|\omega)$ are continuous likelihood functions with respect to the dominating measures ν_1 and ν_2, respectively, then, they are necessarily proportional in θ. In summary, continuous likelihood functions satisfy the LPT.

Theorem 2.3. *Let ν_1 and ν_2 be two σ-finite measures on $(\Omega, \mathcal{B})$ such that $\mathcal{P} << \nu_i$, $i = 1, 2$. Suppose that Θ is a separable metric space and for each $\theta \in \Theta$ there exists a version $f_{1,\theta} \in [\frac{dP_\theta}{d\nu_1}]$ such that $\mathcal{F}_1 = \{f_{1,\theta};\ \theta \in \Theta\}$ is continuous on Θ. Then, for each $\theta \in \Theta$, there exists a version $f_{2,\theta} \in [\frac{dP_\theta}{d\nu_2}]$such that*

1. *$\mathcal{F}_2 = \{f_{2,\theta};\ \theta \in \Theta\}$ is continuous on Θ;*
2. *there exists a measurable set A and a measurable function h such that*

$$f_{1,\theta}(\omega) = f_{2,\theta}(\omega)h(\omega), \forall\omega \in A, \forall\theta \in \Theta,$$

where $P_\theta(A) = 1$ for all $\theta \in \Theta$.

Proof. We will first prove the assertion 1. Let $Q = \sum_{i=1}^{\infty} c_i P_{\theta_i}$ and $\Theta_0 \subset \Theta$ as in Lemma 2.3. Then, there exist a measurable set F and a measurable function h_1 sucht that $Q(F) = 1$, h_1 is strictly positive on F and

$$f_{1,\theta}(\omega) = g_\theta(\omega)h_1(\omega), \forall \omega \in F, \forall \theta \in \Theta_0, \tag{2.11}$$

where g_θ is the Radon-Nikodým derivative of P_θ with respect to Q. Now, take $h_2 \in [\frac{dQ}{d\nu_2}]$ and, for each $\theta \in \Theta$, define

$$f_{2,\theta}(\omega) = I_F(\omega)g_\theta(\omega)h_2(\omega). \tag{2.12}$$

Hence, for each $\theta \in \Theta$, $f_{2,\theta} \in [\frac{dP_\theta}{d\nu_2}]$. We claim that, for each $\omega \in F$, the likelihood function with respect to ν_2 is continuous on Θ. To see that, fix $\theta_0 \in \Theta$ and $\omega \in F$. Since Θ_0 is a dense subset of Θ, there exists a sequence $\{\theta_n\}_{n=1}^{\infty}$ such that $\lim_n \theta_n = \theta_0$. By the continuity of the likelihood function with respect to ν_1, we have that $f_{1,\theta_0}(\omega) = \lim_n f_{1,\theta_n}(\omega)$ and, consequently, it follows from (2.11) that

$$g_{\theta_0}(\omega) = \lim_n g_{\theta_N}(\omega). \tag{2.13}$$

Finally, it follows from (2.12) and (2.13) that

$$\lim_n f_{2,\theta_n}(\omega) = \lim_n [I_F(\omega)g_{\theta_n}(\omega)h_2(\omega)] = I_F(\omega)g_{\theta_0}(\omega)h_2(\omega) = f_{2,\theta_0}(\omega).$$

Thus, we conclude that $\mathcal{F}_2 = \{f_{2,\theta};\ \theta \in \Theta\}$ is continuous on Θ and the assertion 1 is proved. Since we know that $\mathcal{F}_1$ and $\mathcal{F}_2$ are continuous on Θ, the assertion 2 follows from Lemma 2.3. □

Suppose that the population $\mathcal{P}$ is dominated by a σ-finite measure ν_1 such that the likelihood function obtained from ν_1 is a continuous function from Θ to $\mathbb{R}$. Theorem 2.3 states that if we choose another σ-finite measure ν_2 to dominate the population then the new likelihood function obtained from ν_2 will also be a continuous function and consequently, by Lemma 2.3, these two different functions will be proportional in θ. In other words, continuity is preserved when we change the dominating measure.

2.4 The predictive measure as a dominating measure

Izbicki et al. (2014) propose a novel methodology for nonparametric density ratio

estimation and show how this general framework can be extended to address the problem of estimating the likelihood function when this is intractable. In particular, the authors use the density of the prior predictive measure in the denominator of the ratio and, therefore, obtain an approximation for the likelihood function induced by the use of this particular dominating measure. We now investigate when the prior predictive measure can be used as a dominating measure for the model.

Let X be a sample from a population in a parametric family $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$ for a fixed $k \in \mathbb{N}$ and $\mathcal{X}$ be the range of X. Let R be a non zero prior distribution on Θ and denote by $\mathcal{B}_\mathcal{X}$ and $\mathcal{B}_\theta$ the σ-fields on $\mathcal{X}$ and Θ, respectively. Suppose that the function

$$\begin{aligned} \Theta &\longmapsto [0,1] \\ \theta &\longmapsto P_\theta(B) \end{aligned}$$

is Borel for any fixed $B \in \mathcal{B}_\mathcal{X}$. Then, there is a unique probability measure P on $(\mathcal{X} \times \Theta, \mathcal{B}_\mathcal{X} \times \mathcal{B}_\Theta)$ (Shao (2003), Chapter 4) such that, for $B \in \mathcal{B}_\mathcal{X}$ and $C \in \mathcal{B}_\Theta$,

$$P(B \times C) = \int_C P_\theta(B) dR.$$

The posterior distribution of θ, given $X = x$, will be denoted by $P_{\theta|x}$. The next theorem provides a formula for the p.d.f of the posterior distribution $P_{\theta|x}$.

Theorem 2.4. *(Bayes Formula) Assume that $\mathcal{P}$ is dominated by a σ-finite measure ν and $f_\theta(x) = \frac{dP_\theta}{d\nu}(x)$ is a Borel funtion on $(\mathcal{X} \times \Theta, \mathcal{B}_\mathcal{X} \times \mathcal{B}_\Theta)$. Suppose that $m(x) = \int_\Theta f_\theta(x) dR > 0$. Then, the posterior distribution $P_{\theta|x}$ is dominated by R and*

$$\frac{dP_{\theta|x}}{dR}(x) = \frac{f_\theta(x)}{m(x)}.$$

The function m in Theorem 2.4 is called the marginal p.d.f. of X with respect to ν. Observe that the Bayes Formula is well defined only for the points $X = x$ such that $m(x) > 0$. If $m(x) = 0$ we may have problems. To see that, suppose there exists $x \in \mathcal{X}$ such that $m(x) = 0$. Then, by the definition of m, it follows that

$$\int_\Theta f_\theta(x) dR = 0,$$

and since $R(\Theta) > 0$, $R(Z_x^c) = 0$, where $Z_x = \{\theta \in \Theta;\ f_\theta(x) = 0\}$. In words, given $X = x$ such that $m(x) = 0$, the likelihood function vanishes R-almost everywhere. We will see later that the zero set of the function m plays an important role for the predictive measure. Before that, we formalise the concept of this measure.

Definition 2.6. The measure λ defined on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ by

$$\lambda(A) = \int_A m d\nu, \ \forall A \in \mathcal{B}_{\mathcal{X}}$$

is called (prior) predictive measure.

Proposition 2.4. *The predictive measure is independent of the choice of the measure that dominates the population $\mathcal{P}$.*

Proof. Let μ be a σ-finite measure such that $\mathcal{P} << \mu$. Let $g_\theta(x) = \frac{dP_\theta}{d\mu}(x)$ and define

$$m^*(x) = \int_\Theta g_\theta(x) dR.$$

Now consider the predictive measure ξ obtained from m^*, i.e.,

$$\xi(A) = \int_A m^* d\mu, \ \forall A \in \mathcal{B}_{\mathcal{X}}.$$

We claim that $\lambda = \xi$. For any $A \in \mathcal{B}_{\mathcal{X}}$,

$$\begin{aligned} \lambda(A) &= \int_A m d\nu = \int_A \int_\Theta f_\theta(x) dR d\nu \overset{(i)}{=} \int_\Theta \int_A f_\theta(x) d\nu dR = \int_\Theta P_\theta(A) dR \\ &= \int_\Theta \int_A g_\theta(x) d\mu dR \overset{(ii)}{=} \int_A \int_\Theta g_\theta(x) dR d\mu = \int_A m^* d\mu = \xi(A), \end{aligned}$$

where the equalities (i) and (ii) follow from Fubini's theorem. □

From now, let N denote the zero set of the function m. As we discussed previously, we have a problem when $m(x) = 0$, since the likelihood function, given $X = x$, is zero almost everywhere in this case. Therefore, the ideal marginal p.d.f. of X with respect to ν is a function m such that $m(x) > 0$ for all $x \in \mathcal{X}$.

Proposition 2.5. *If $m(x) > 0$ for all $x \in \mathcal{X}$, then the predictive measure λ dominates $\mathcal{P}$.*

Note that, for the previous proposition to be valid, it is enough $\nu(N) = 0$. Nevertheless, the result is not guaranteed if we only have that $P_\theta(N) = 0$ R-almost everywhere.

Theorem 2.5. *The predictive measure λ dominates P_θ if and only if $P_\theta(N) = 0$. In particular, λ dominates $\mathcal{P}$ if and only if $P_\theta(N) = 0$ for all $\theta \in \Theta$.*

Proof. If λ dominates P_θ, the result follows immediately since $\lambda(N) = 0$. Suppose now that $P_\theta(N) = 0$ and take $A \in \mathcal{B}_{\mathcal{X}}$ such that $\lambda(A) = 0$. We have to show that

$P_\theta(A) = 0$. Note that

$$0 = \lambda(A) = \lambda(A \cap N^c) = \int_{A \cap N^c} m d\nu. \tag{2.14}$$

Then, since m is strictly positive in $A \cap N^c$, equation (2.14) is true only if $\nu(A \cap N^c) = 0$. Hence, $P_\theta(A \cap N^c) = 0$. But, by hypothesis, $P_\theta(A) = P_\theta(A \cap N^c)$ and the result follows. □

Theorem 2.6. *If* $M_\theta = \{x \in \mathcal{X};\ f_\theta(x) > 0\}$ *does not depend on* θ*, then* $\mathcal{P} << \lambda$.

Proof. Let $M = M_\theta$ for all $\theta \in \Theta$ and let $A \in \mathcal{B}_\mathcal{X}$ such that $\lambda(A) = 0$. To show that $P_\theta(A)$ for all $\theta \in \Theta$ is sufficient to show that $\nu(A \cap M) = 0$, since $\mathcal{P} << \nu$ and $P_\theta(A) = P_\theta(A \cap M)$ for all $\theta \in \Theta$. Suppose that $\nu(A \cap M) > 0$. Hence, since f_θ is strictly positive on $A \cap M$,

$$P_\theta(A) = P_\theta(A \cap S) = \int_{A \cap S} f_\theta d\nu > 0,\ \forall \theta \in \Theta. \tag{2.15}$$

On the other hand,

$$\lambda(A) = \int_A m d\nu = \int_A \int_\Theta f_\theta dR d\nu = \int_\Theta P_\theta(A) dR = \int_\Theta P_\theta(A \cap S) dR, \tag{2.16}$$

where the penultimate equation follows from Fubini's theorem. Then, since $R(\Theta) > 0$, it follows from (2.15) and (2.16) that $\lambda(A) > 0$, contradicting the assumption that $\lambda(A) = 0$. So, $\nu(A \cap M) = 0$ and the proof is complete. □

Chapter 3

Exploring some model classes

We now explore the results from Chapter 2, specially the Likelihood Proportionality Theorem, considering some classes of statistical models. We highlight special aspects of that theorem and illustrate its importance in different contexts.

3.1 Finite-dimensional random variables

It is often the case in which the statistical model under consideration is a family of probability measures consisting of a finite dimensional random variable with discrete and/or continuous coordinates. This covers a wide range of models from iid univariate random variables to highly structured hierarchical Bayesian models with mixture components. In this case, the most common choice for dominating measure is the appropriate product of the counting and Lebesgue measures. Nevertheless any probability measure with common support is a valid dominating measure and, therefore, admits versions that lead to proportional likelihoods. A particularly interesting example, that goes beyond a purely discrete or continuous random variable, are point-mass mixtures.

Consider the probability measure of a r.v. Y such that $P(Y = a_i) = p_i > 0$, for $i = 1, \ldots, m$ and $\sum_{i=1}^{m} p_i = p < 1$, and $Y = Z_j$ w.p. q_j, such that Z_j is a continuous r.v. on $B \subset \mathbb{R}$ with Lebesgue density f_j, for $j = 1, \ldots, n$ and $\sum_{j=1}^{n} q_j = 1 - p$. In this case, Gottardo and Raftery (2009) show that the probability measure P of Y is dominated by the measure $\nu_1 + \nu_2$, where ν_1 is the counting measure and ν_2 is the Lebesgue measure and

$$\frac{dP}{d(\nu_1 + \nu_2)}(y) = \sum_{i=1}^{m} p_i \mathbb{I}_{\{a_i\}}(y) + \sum_{j=1}^{n} q_j f_j(y) \mathbb{I}_{B \setminus A}(y), \tag{3.1}$$

where $A = \{a_1, \ldots, a_n\}$. The use of a non-valid RN derivative, in particular by ignoring the indicator functions in (3.1), leads to misspecified likelihood functions with possibly serious consequences in the inference process. The density in (3.1) is uniquely defined on A and one should always consider continuous versions of the f_j's (in B) when these exist. These versions not only guarantee the proportionality of likelihoods obtained for different dominating measures (see Theorem 2.2) as it also guarantees that the likelihood obtained is the limit in (2.3).
The result from Gottardo and Raftery (2009) is actually more general and provides a valid dominating measure with the respective RN derivative for probability measures consisting of a countable mixture of mutually singular probability measures.

3.2 Exponential families

In this section, we discuss how the choice of the measure that will dominate the model can affect an exponential family. Moreover, we will see how Exponential families are related to the LPT.

Definition 3.1. (Shao (2003)) A parametric family $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$ dominated by a σ-finite measure ν on $(\Omega, \mathcal{F})$ is called an exponential family if and only if

$$\frac{dP_\theta}{d\nu}(\omega) = \exp\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\}h(\omega),\ \omega \in \Omega, \tag{3.2}$$

where T is a random p-vector with $p \in \mathbb{N}$, η is a function from Θ to $\mathbb{R}^p$, h is a nonnegative Borel function and

$$\xi(\theta) = \log\{\int_\Omega \exp\{[\eta(\theta)]^\tau T(\omega)\}h(\omega)d\nu(\omega).$$

Note that the Definition 3.1 depends on the measure ν. Then, if we change the measure that will dominate the family $\mathcal{P}$, the representation given in (3.2) will be different. Thus, it is natural to ask if the exponential representation is independent of the choice of the dominating measure, i.e., if $\mathcal{P}$ is dominated by a σ-finite measure μ, then $\frac{dP_\theta}{d\mu}$ has the form given in (3.2). The aim of this section is to answer this question. Theorem 3.1 states that the definition of a exponential family is independent of the choice of a dominating measure, i.e., if a family $\mathcal{P}$ has densities with respect to a σ-finite measure ν given by (3.2), then for every σ-finite measure μ that dominates $\mathcal{P}$ there will be exist functions g_θ and h_μ (that do not depend on θ) such that

$$\frac{dP_\theta}{d\mu}(\omega) = g(\omega, \theta)h_\mu(\omega).$$

Furthermore, the function $g(\omega, \theta)$, named here kernel function, will be the same for all the σ-finite measures that dominate the model. This kernel is:

$$g(\omega, \theta) = exp\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\}.$$

The unique difference between the densities of a model with respect to two different measures, say ν and μ, will be the measurable functions h_ν and h_μ. Since h_ν and h_μ do not depend on θ, the class of Exponential families whose the densities are given by (3.2) always satisfies the LPT. Before states the main theorem of this section, we draw attention to a particular dominating measure. This dominanting measure is discussed briefly in Shao (2003).
For any $A \in \mathcal{F}$, define

$$\lambda(A) = \int_A h d\nu,$$

where the function h is the same measurable function in (3.2). Hence, λ is a σ-finite measure on $(\Omega, \mathcal{F})$. We claim that $\mathcal{P} << \lambda$. To see this, let $B = \{\omega;\ h(\omega) > 0\}$ and let $\lambda(A) = 0$ for some $A \in \mathcal{F}$. Then, $\lambda(A) = \lambda(A \cap B)$ and consequently

$$\int_{A \cap B} h d\nu = 0.$$

Since the function h is strictly positive on $A \cap B$, it follows that $\nu(A \cap B) = 0$ and $P_\theta(A) = P_\theta(A \cap B) = 0$ for all $\theta \in \Theta$. Therefore, λ dominates $\mathcal{P}$ and, since $\lambda << \nu$, we have by RN chain rule

$$\frac{dP_\theta}{d\lambda}(\omega) = exp\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\},\ \omega \in \Omega. \tag{3.3}$$

We can conclude two importants things from the result above: (i) the change of the dominating measure has preserved the exponential representation and (ii) there exists a measure λ such that λ dominates the model and the RN derivatives of P_θ with respect to λ are strictly positive for all $\omega \in \Omega$ and for all $\theta \in \Theta$. Observation (ii) can be useful to demonstrate when a family $\mathcal{F}$ is not an exponential family and it may also be usefull to prove when the predictive measure can be used as a dominating measure for the model. Particularly, if m is the marginal p.d.f. of a sample X with respect to ν, then m is strictly positive. Hence, the hypothesis from Proposition 2.5 is satisfied.

Example 3.1. *Let $\mathcal{P} = \{P_\theta\ \ \theta = (a, \beta) \in \mathbb{R} \times (0, \infty)\}$ be a family of distributions where each P_θ has an exponential distribution with two unknown parameters a and*

β, i.e.,

$$P_\theta(A) = \int_A \frac{1}{\beta} exp\{\frac{-(x-a)}{\beta}\} I_{(a,\infty)}(x) d\mu(x),$$

where $A \in \mathcal{B}(\mathbb{R})$ and μ is the Lebesgue measure on the real line. Then $\mathcal{P}$ is not an exponential family.

Proof. We have from (3.3) that there exists a non zero measure λ such that P_θ has a positive density with respect to λ given by

$$\frac{dP_\theta}{d\lambda}(x) = exp\{[\eta(a,\beta)]^\tau T(x) - \xi(a,\beta)\},\ x \in \Omega.$$

Let $r \in \mathbb{R}$ and take any $a > r$. If we define $A = (-\infty, r)$, then for all $\beta \in (0, \infty)$, it follows that

$$P_{(a,\beta)}(A) = \int_{(-\infty,r)\cap(a,\infty)} \frac{1}{\beta} exp\{\frac{-(x-a)}{\beta}\}(x) d\mu(x) = 0.$$

.
On the other hand, since

$$0 = P_{(a,\beta)}(A) = \int_A exp\{[\eta(a,\beta)]^\tau T(x) - \xi(a,\beta)\} d\lambda(x),$$

it must be $\lambda((-\infty, r)) = 0$. Since r is arbitrary, we conclude that λ must be a zero measure, which is a contradiction. □

Theorem 3.1. *Being an Exponential family is a property of the model $\mathcal{P}$, i.e., it is independent of the dominating measure ν from Definition 3.1. Moreover, if $\mathcal{P}$ is an Exponential family, then, for all σ-finite mesure ν such that $\mathcal{P} << \nu$, there exist functions η, T and ξ and there exists a measurable function h_ν such that*

$$\frac{dP_\theta}{d\nu}(\omega) = exp\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\} h_\nu(\omega),\ \omega \in \Omega, \forall \theta \in \Theta.$$

Proof. Suppose that $\frac{dP_\theta}{d\nu}$ is given by (3.2). Consider the measure Q given by Lemma 2.1 and let $q \in [\frac{dQ}{d\nu}]$. Remember that Q is minimal and so $Q << \nu$. Without loss of generality we may assume that $q > 0$. Define, for each $\theta \in \Theta$, the following function:

$$b_\theta(\omega) = exp\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\} m(\omega),\ \omega \in \Omega, \tag{3.4}$$

where $m = h_\nu / q$. By RN chain rule, it follows that

$$exp\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\} h(\omega) = \frac{dP_\theta}{dQ}(\omega) q(\omega),\ \nu - a.e. \tag{3.5}$$

Consequently, from (3.4) and (3.5), $b_\theta = dP_\theta/dQ$ ν-almost-everywhere. Hence, $b_\theta \in [\frac{dP_\theta}{dQ}]$. Now, let μ be a σ-finite measure such that $\mathcal{P} << \mu$ and let $\mu \neq \nu$. Again, by the minimality of Q, $Q << \mu$. Let $s \in [\frac{dQ}{d\mu}]$ and define, for each $\theta \in \Theta$,

$$p_\theta(\omega) = exp\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\}h_\mu(\omega),\ \omega \in \Omega, \tag{3.6}$$

where $h_\mu = ms$. Hence, by RN chain rule, $p_\theta \in [\frac{dP_\theta}{d\mu}]$ and the proof is complete. □

Let $\mathcal{C} = \{\nu;\ \nu \text{ is } \sigma - \text{finite and } \mathcal{P} << \nu\}$. It follows from Theorem 3.1 that for all $\nu \in \mathcal{C}$, there exists a version $dP_\theta/d\nu$ such that

$$\frac{dP_\theta}{d\nu}(\omega) = g(\omega, \theta)h_\nu(\omega), \forall\omega \in \Omega, \forall\theta \in \Theta,$$

where the function g will be the same for any $\nu \in \mathcal{C}$. Hence, in particular, if $\mathcal{P}$ is an Exponential family, the versions $\frac{dP_\theta}{d\nu}$ and $\frac{dP_\theta}{d\mu}$ given by (3.2) and (3.6) satisfies the LPT.

3.3 Missing data problems

Consider a statistical model $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$ on $(\Omega, \mathcal{F})$, such that $\Omega = \Omega_1 \times \Omega_2$ and $\mathcal{F} = \sigma(\mathcal{F}_1, \mathcal{F}_2)$. Suppose, however, that only $\omega_1 \in \Omega_1$ is observed. This is the general formulation of a statistical missing data problem and may be motivated by modelling reasons and/or because the marginal density of P_θ (w.r.t. some dominating measure) on $(\Omega_1, \mathcal{F}_1)$ is not available but the joint density on $(\Omega, \mathcal{F})$ is. A likelihood-based inference approach considers the (pseudo-)likelihood, which is obtained from the density of P_θ w.r.t. some dominating measure, and integrates out the missing data somehow. This is typically done via EM (or Monte Carlo EM) in the frequentist approach or via MCMC in the Bayesian approach. Both methodologies involve dealing with the conditional measure of the missing data given the data ω_1 and the parameters θ.

Suppose that two dominating measures ν_1 and ν_2 for $\mathcal{P}$ are available. Each of them may be used to obtain a RN derivative for measures P_θ and, consequently, a (pseudo-)likelihood. If ω_1 is observed, we have

$$\pi_i(\omega_2|\omega_1, \theta) \propto \pi_i(\omega_1, \omega_2|\theta), \quad i = 1, 2, \tag{3.7}$$

where the right hand side is the RN derivative of P_θ w.r.t. ν_i. This way, the left hand side is the density of the conditional measure of the missing data given data

and θ w.r.t. some dominating measure which is induced by ν_i and, therefore, may be different for ν_1 and ν_2.

Theorem 2.1 guarantees that the (pseudo-)likelihood is proportional w.r.t. θ only and not w.r.t. ω_2, which also needs to be estimated. As a consequence, although both measures can be used, this choice may have great influence when devising the inference methodology. The EM algorithm requires computing an expectation w.r.t. the conditional measure of the missing data whilst the Monte Carlo EM and the MCMC require sampling from this measure. If the conditional densities $\pi_i(\omega_2|\omega_1, \theta)$ are different for $i = 1$ and $i = 2$, it may be the case that the required tasks are harder or even not feasible for one of them - although both densities are valid ones.

An interesting example can be found in Gonçalves and Gamerman (2017), where (ω_1, ω_2) is the realisation of a homogeneous Poisson process in a compact region $S \subset \mathbb{R}^d$. Furthermore, ω_1 are the Poisson events remaining after performing a Poisson thinning and ω_2 are the thinned events. Therefore, the missing data consists of a discrete random variable which represents the number of thinned events and a vector of continuous r.v.'s representing their locations. Gonçalves and Gamerman (2017) devise an MCMC algorithm to perform inference in their model (which also involves a Gaussian process) which is a Gibbs sampling that samples the missing data, from its full conditional distribution, in one of its steps. As we have mentioned in Section 3.4, there are two obvious dominating measures to obtain a density for the homogeneous Poisson process. Gonçalves and Gamerman (2017) argue that their algorithm is feasible if and only if the measure consisting of the product of the counting and multidimensional Lebesgue measure is used. This leads to a conditional density where the marginal p.m.f. of the number of points and the conditional Lebesgue density of their locations can be devised.

3.4 Poisson processes

Poisson process (PP) is the most common statistical model to fit point pattern data. Consider some region $S \subset \mathbb{R}^d$, for $d \in \mathbb{N}$. Poisson processes can actually be defined in more general measurable spaces (see Kingman, 1993, Chp. 2). Let us first consider a homogeneous PP on S with intensity $\lambda(s) = \lambda,\ \forall s \in S,\ \lambda \in \mathbb{R}^+$, which defines a probability measure P_λ. In this case, we have two obvious dominating measures for P_λ. The first one represents a realization ω as $(N, s_1, \ldots, s_N)$, where N is the number of points and the s_j's are their respective locations. We can factor their joint density as $\pi(N)\pi(s_1, \ldots, s_N|N)$ and use the product measure $\nu_1 \otimes \nu_2$ as a dominating measure, where ν_1 is the counting measure and ν_2 is the

N-dimensional Lebesgue measure. We get that

$$\frac{dP_\lambda}{d(\nu_1 \otimes \nu_2)}(\omega) = \frac{e^{-\lambda\mu(S)}(\lambda\mu(S))^N}{N!}(\mu(S))^{-N}, \tag{3.8}$$

where $\mu(S)$ is the volume of S.

Another valid dominating measure is the probability measure ν of any PP for which the intensity function is positive everywhere in S, in particular constant and equals to 1. In that case, the RN derivative is given by Jacod's formula (see Andersen et al., 1993, Chp. II):

$$\frac{dP_\lambda}{d\nu}(\omega) = \exp\left\{-\int_S (\lambda - 1)ds\right\}\prod_{j=1}^{N}(\lambda/1). \tag{3.9}$$

Note that the densities in (3.8) and (3.9) are proportional in λ. In a standard inference problem where ω is observed and λ is to be estimated, there is no practical difference in considering one or the other. In a more complex context, however, it may be a crucial choice. Gonçalves and Gamerman (2017) propose a methodology to make exact (discretisation-free) inference for spatio-temporal Cox processes which is based on an augmented model consisting of a homogeneous PP. In their case, choosing the dominating in (3.8) to obtain the (pseudo-)likelihood of the augmented model is crucial to devise a valid MCMC algorithm to perform (Bayesian) inference - this example is described in more details in Section 3.3.

In the case of a non-homogeneous PP, the measure of another PP is the only obvious choice for a dominating measure. The density of a PP with intensity function $\lambda := \{\lambda(s),\ s \in S\}$ - measure P_λ, w.r.t. the measure ν of a unit intensity PP is given by

$$\frac{dP_\lambda}{d\nu}(\omega) = \exp\left\{-\int_S \lambda(s) - 1ds\right\}\prod_{j=1}^{N}(\lambda(s_j)/1). \tag{3.10}$$

If we consider the Skorokhod space D of càdlàg functions with the respective Skorokhod topology, we get that D is a separable space and the likelihood function in (3.10) is continuous on D.

Note that, for a fixed ω, the expression in (3.10) is proportional (in λ) to the following function - known as the Poisson process likelihood:

$$l(\lambda) = \exp\left\{-\int_S \lambda(s)ds\right\}\prod_{j=1}^{N}\lambda(s_j). \tag{3.11}$$

Finally, note that using the dominating measure of any other PP (with positive intensity over S) would lead to a function proportional to the one in (3.11).

3.5 Diffusions and jump-diffusions

Brownian motion driven stochastic differential equations (SDE), known as diffusion processes, are quite popular in the statistical literature to model a variety of continuous time phenomena. Formally, a diffusion is defined as the continuous time stochastic process which is the unique solution of a well-defined SDE. Making statistical inference for diffusions is a challenging problem due to the complex nature of such processes. The continuous time feature implies that they lie on infinite-dimensional space and typically have unknown intractable transition densities. As a consequence, an exact likelihood in a discretely observed context is unavailable. The most promising solutions available stand out for treating the inference problem without resorting to discretisation schemes (see Beskos et al., 2006). These methodologies, called exact, rely on the (pseudo-)likelihood function of a continuous-time trajectory and give rise to interesting issues related to the context of this thesis. We discuss the case where the processes are univariate and the diffusion process $Y := \{Y_s,\ s \in [0,t]\}$ is defined as the solution of an SDE of the type:

$$dY_s = a(Y_s,\theta)ds + \sigma(Y_s,\theta)dW_s,\ s \in [0,t],\quad Y_0 = y_0, \tag{3.12}$$

where W_s is a Brownian motion and functions a and σ are suppose to satisfy some regularity conditions to guarantee the existence of an unique solution (see Kloeden and Platen, 1995). Diffusion processes trajectories are a.s. continuous and non-differentiable everywhere. An interesting generalisation considers the possibility of discontinuity points stochastically defined by a marked Poisson process, possibly non-homogeneous and depending on the state and time of the original diffusion process. Such processes are called jump-diffusions and are also quite appealing in a variety of applications. Formally, a jump-diffusion is the solution of the SDE:

$$dY_s = a(Y_s,\theta)ds + \sigma(Y_s,\theta)dW_s + dJ_s,\ s \in [0,t],\quad Y_0 = y_0, \tag{3.13}$$

where J_s is a marked Poisson process with intensity function $\lambda(Y_s,s,\theta)$ and jump size density $f(\cdot;Y_s,s,\theta)$ - in its most general form.
In a typical statistical problem, one is interested in estimating the functions $a(Y_s,\theta)$, $\sigma(Y_s,\theta)$ and, in a jump-diffusion context, also $\lambda(Y_s,s,\theta)$ and $f(\cdot;Y_s,s,\theta)$. These are typically defined parametrically, as it is done here, but non-parametric approaches may be considered. In the parametric case, the aim is to estimate the

parameter set θ. As it was mentioned above, exact methodologies rely on the likelihood of a complete trajectory which can only be obtained if a valid dominating measure is available. It turns out, however, that processes with distinct diffusion coefficient σ define mutually singular probability measures. As a consequence, there exists no σ-finite measure that simultaneously dominates the family of probability measures if this is uncountable, which is often the case (if it is countable, a countable sum of measures would dominate - see Gottardo and Raftery (2009)). Therefore, different values of θ define mutually singular measures and no likelihood function can be obtained. The clever solution for this problem, firstly proposed in Roberts and Stramer (2001), considers two transformations of the diffusion path. This is decomposed as $(Y_{obs}, \dot{X})$, where Y_{obs} are the discrete observations of Y and $\dot{X}$ are transformed bridges between the observations. More specifically, for (time-ordered) observations $y_0, \ldots, y_n$ at times $t_0, t_1, \ldots, t_n$, consider the Lamperti transform $X_s = \eta(Y_s, \theta) = \int_y^{X_s} \frac{1}{\sigma(u,\theta)} du$, for some element y of the state space of Y. This implies that X is the solution of a SDE with unit diffusion coefficient and drift $\alpha(X_s, \theta)$, which depends on function σ. Now, defining $x_i(\theta) = \eta(y_i, \theta)$, $i = 0, \ldots, n$, consider the following transformation of the bridges of X between the $x_i(\theta)$ points, $\dot{X}_s = \varphi^{-1}(X_s) = X_s - \left(1 - \frac{s-t_{i-1}}{t_i - t_{i-1}}\right) x_{i-1}(\theta) - \left(\frac{s-t_{i-1}}{t_i - t_{i-1}}\right) x_i(\theta)$, for $s \in (t_{i-1}, t_i)$. This implies that the transformed bridges start and end in zero and are, therefore, dominated by the measure of standard Brownian bridges. The density of $(Y_{obs}, \dot{X})$ is decomposed as $\pi(Y_{obs}, \dot{X}) = \pi(Y_{obs})\pi(\dot{X}|Y_{obs})$ and obtained w.r.t. to the parameter-free dominating measure $\nu^n \otimes \mathbb{W}^n$ - the product measure of the n-dimensional Lebesgue measure and the measure of a standard Brownian bridges of respective time lengths. Lemma 2 from Beskos et al. (2006) gives that:

$$\begin{aligned} \pi(Y_{obs}, \dot{X}) \;=\; & \prod_{i=1}^{n} \eta'(y_i;\theta)\phi\left((x_i(\theta) - x_{i-1}(\theta))/\sqrt{t_i - t_{i-1}}\right) \\ & \exp\left\{A(x_n(\theta);\theta) - A(x_0(\theta);\theta)\right\} \\ & \exp\left\{-\int_0^T \left(\frac{\alpha^2 + \alpha'}{2}\right)(\varphi_\theta(\dot{X}_s);\theta)\right\}, \end{aligned} \tag{3.14}$$

where $A(u) = \int_0^u \alpha(z, \theta) dz$ and ϕ is the standard Gaussian density.
Assuming that σ is continuously differentiable, one can show that, under the supremum norm, the density in (3.14) is continuous in C - the space of continuous functions on $[0, t]$. The sup norm on C also defines a separable space.
In the case of jump-diffusions, the dominating measure combines the dominating measure from (3.14) with the measure of a marked Poisson process with unit jump intensity and some known jump size measure that dominates the jump size measure of all measures in the family. Lemma 1 from (see Gonçalves et al., 2017) gives the

desired likelihood function. As in the Poisson process example, if we consider the Skorokhod space D of càdlàg functions with the respective Skorokhod topology, the density obtained is continuous.

Chapter 4

Differentiation

Let $(\Omega, \mathcal{B})$ be a measurable space and let ν and μ be σ-measures defined on $(\Omega, \mathcal{B})$. Suppose that $\nu << \mu$. The Radon-Nikodým theorem guarantees that there exists an integrable function f, called Radon-Nikodým derivative, such that

$$\nu(E) = \int_E f d\mu, \ E \in \mathcal{F}.$$

Note that the Radon-Nikodým theorem only guarantees the existence of f. It does not suggest any method to obtain this derivative. Suppose that Ω is a metrizable space. Let $x \in \Omega$ and $I \in \mathcal{F}$. We write $I \Longrightarrow x$ (I contracts to x) to say that $x \in I$ and the diameter of I tends to zero. An interesting question is: can the Radon-Nikodým derivative f behave like a genuine derivative?, i.e., can we find a differentiation basis $\mathcal{I} = \{I; \ I \in \mathcal{F}\}$ such that

$$\lim_{I \Longrightarrow x} \frac{\nu(I)}{\mu(I)} = f(x). \tag{4.1}$$

μ-almost everywhere? Section 1 presents an answer to this question. The purpose of this chapter is to provide a method to obtain likelihood functions from a genuine derivative as in (4.1). This is discussed in Section 2. Section 1 presents the results on the theory of differentiation of measures needed to develop the theory in Section 2. All the theorems and definitions in Section 1 are from Evans and Gariepy (1991).

4.1 Differentiation of Radon measures

Let μ and ν be Radon measures on $\mathbb{R}^n$. Since R^n is a Polish space, it follows from Theorem 1.8 that μ and ν are σ-finite measures.

Definition 4.1. For each point $x \in \mathbb{R}^n$, define

$$\overline{D}_\mu\nu(x) \equiv \begin{cases} \limsup_{r\to 0} \frac{\nu(B(x,r))}{\mu(B(x,r))} & \text{if} \quad \mu(B(x,r)) > 0 \text{ for all } r > 0, \\ +\infty & \text{if} \quad \mu(B(x,r)) = 0 \text{ for some } r = 0. \end{cases}$$

$$\underline{D}_\mu\nu(x) \equiv \begin{cases} \liminf_{r\to 0} \frac{\nu(B(x,r))}{\mu(B(x,r))} & \text{if} \quad \mu(B(x,r)) > 0 \text{ for all } r > 0, \\ +\infty & \text{if} \quad \mu(B(x,r)) = 0 \text{ for some } r = 0. \end{cases}$$

Definition 4.2. If $\overline{D}_\mu\nu(x) = \underline{D}_\mu\nu(x) < +\infty$, we say ν is differentiable with respect to μ at x and write

$$D_\mu\nu(x) \equiv \overline{D}_\mu\nu(x) = \underline{D}_\mu\nu(x).$$

$D_\mu\nu(x)$ is the derivative of ν with respect to μ. We also call $D_\mu\nu$ the density of ν with respect to μ.

Definition 4.3. If $\overline{D}_\mu\nu(x) = \underline{D}_\mu\nu(x) < +\infty$, we call x a density point of ν with respect to μ. The set of the density points of ν with respect to μ will be denoted by $D(\nu, \mu)$.

Theorem 4.1. *Let μ and ν be Radon measures on $\mathbb{R}^n$. Then $D_\mu\nu$ exists and is finite μ a.e. Futhermore, $D_\mu\nu$ is measurable.*

Theorem 4.1 says that $\mu(D(\nu,\mu)^c) = 0$.

Theorem 4.2. *Let ν and μ be Radon measures on $\mathbb{R}^n$, with $\nu << \mu$. Then*

$$\nu(A) = \int_A D_\mu\nu \, d\mu$$

for all Borel sets $A \subset \mathbb{R}^n$. In other words

$$D_\mu\nu(x) = \frac{d\nu}{d\mu}(x), \ \mu \text{ a.e.}$$

4.2 Defining the likelihood function as a derivative of measures

Let $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$ be a family of probability measures on $\mathbb{R}^n$ such that $\mathcal{P} << \mu$. Since R^n is a Polish space and any probability measure is a locally finite measure, it follows from Theorem 1.8 that each P_θ is a Radon measure. Therefore, in view of Section 4.1, we can define $D_\mu P_\theta$, the derivative of P_θ with respect to μ. The aim of this section is to define the likelihood as a derivative $D_\mu P_\theta$. To do that, we should be able to exhibit a measurable set A such that $P_\theta(A) = 1$ and $D_\mu P_\theta$

exists for all $\theta \in \Theta$ and for all $x \in A$. In view of Theorem 4.2, $D_\mu P_\theta$ exists μ almost everywhere, i.e., there exists, for each $\theta \in \Theta$, a measurable set A_θ such that $\mu(A_\theta^c) = 0$ and $D_\mu P_\theta$ exists in A_θ. Therefore, if Θ is uncountable (as usual), Theorem 4.2 does not imply that $D_\mu P_\theta$ is a likelihood function. This result is established by Theorem 4.3. More specifically, it states, under certain conditions, that $D_\mu P_\theta$ is a valid likelihood for θ, i.e., for any two dominating measures μ and ν there exists a measurable set A such that

(a) $P_\theta(A) = 1$ for all $\theta \in \theta$;

(b) $D_\mu P_\theta$ and $D_\nu P_\theta$ exist for all $\theta \in \Theta$ and for all $x \in A$ and

(c) $D_\mu P_\theta(x) \propto_\theta D_\nu P_\theta(x)$ for each $x \in A$ and for all $\theta \in \Theta$.

In order to prove Theorem 4.3, we need the following four lemmas.

Lemma 4.1. *Let x be a point in the support of ν. Then, $D_\mu \nu(x) = s \in \mathbb{R}^+$ if and only if for every $\epsilon > 0$ there exists $N_0 \in \mathbb{N}$ such that*

$$\left| \frac{\nu(B(x, \frac{1}{n}))}{\mu(B(x, \frac{1}{n}))} - s \right| < \epsilon \tag{4.2}$$

for all $n \geq N_0$.

Proof. Suppose that $D_\mu \nu(x) = s \in \mathbb{R}^+$ and take $\epsilon > 0$. For each $N \in \mathbb{N}$, define

$$A_N = \sup_{n \geq N} \left\{ \frac{\nu(B(x, \frac{1}{n}))}{\mu(B(x, \frac{1}{n}))} \right\} \quad \text{and} \quad B_N = \inf_{n \geq N} \left\{ \frac{\nu(B(x, \frac{1}{n}))}{\mu(B(x, \frac{1}{n}))} \right\}. \tag{4.3}$$

Hence, $\inf_N\{A_N\} = s = \sup_N\{B_N\}$. Thus, there exists $N_1 \in N$ such that $A_{N_1} < s + \epsilon$. Therefore, by the definition of the sequence A_N,

$$\frac{\nu(B(x, \frac{1}{n}))}{\mu(B(x, \frac{1}{n}))} < s + \epsilon, \ \forall n \geq N_1. \tag{4.4}$$

A similar argument for the sequence B_N shows that there exists $N_2 \in N$ such that

$$s - \epsilon < \frac{\nu(B(x, \frac{1}{n}))}{\mu(B(x, \frac{1}{n}))}, \ \forall n \geq N_2. \tag{4.5}$$

Taking $N_0 = \max\{N_1, N_2\}$, (4.2) follows from (4.4) and (4.5). To go the other way, we will show firstly that $\overline{D}_\mu \nu(x) < +\infty$. Consider the sequence $\{A_N\}_N$ given by (4.3). By hypothesis, x is a point in the support of ν. Hence, since the support of ν is a subset of the support of μ, it follows that $\mu(U) > 0$ for every neighborhood of

x and, consequently, $A_N \in \mathbb{R}$ for every $N \in \mathbb{N}$. Again, by hypothesis, there exists some integer K_0 such that the sequence $0 \leq A_N \leq s+1$. Then, since $\{A_N\}_N$ is a nonincreasing sequence, we have that the sequence $\{A_N\}_{N \geq K_0}$ is bounded. Hence, $\inf_N A_N = \inf_{N \geq K_0} A_N$ exists and is finite. But

$$\inf_N A_N = \inf_{N \geq K_0} A_N = \limsup_{r \to 0} \frac{\nu(B(x,r))}{\mu(B(x,r))}.$$

Thus, $\overline{D}_\mu \nu(x) < +\infty$. We claim that $\overline{D}_\mu \nu(x) \leq s$. Suppose $s < \overline{D}_\mu \nu(x)$. Let $\epsilon = (\overline{D}_\mu \nu(x) - s)/2$. Hence, $\epsilon > 0$. Then, it follows from (4.2) that there exists $N_0 \in \mathbb{N}$ such that

$$\frac{\nu(B(x, \frac{1}{n}))}{\mu(B(x, \frac{1}{n}))} < s + \epsilon < \overline{D}_\mu \nu(x), \ \forall n \geq N_0.$$

Therefore, we have that $A_{N_0} \leq s + \epsilon < \overline{D}_\mu \nu(x)$, contradicting the fact that $\overline{D}_\mu \nu(x) \leq A_N$ for all $N \in \mathbb{N}$. So, it must be $\overline{D}_\mu \nu(x) \leq s$. Similarly, $\underline{D}_\mu \nu(x) \geq s$. Since $\underline{D}_\mu \nu(x) \leq \overline{D}_\mu \nu(x)$, we have $\underline{D}_\mu \nu(x) = s = \overline{D}_\mu \nu(x)$ and the proof is complete. □

From now Let $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$ be a family of probability measures on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ such that $\mathcal{P} << \nu$ and $\mathcal{P} << \mu$. We write $Q = \sum_{k=1}^{\infty} c_k P_{\theta_k}$ for the minimal dominating measure given by Lemma 2.1. Let S_θ, S_μ, S_ν and S_Q denote the support of P_θ, μ, ν and Q, respectively, for each $\theta \in \Theta$. Remember that $S_\theta \subset S_Q \subset S_\mu$ and $S_\theta \subset S_Q \subset S_\nu$ for all $\theta \in \Theta$.

Lemma 4.2. *Fix $\theta \in \Theta$ and suppose that $D_\mu P_\theta(x)$ and $D_\mu Q(x)$ exist and are finite for some $x \in S_\theta$. If $D_\mu Q(x) > 0$, then $D_Q P_\theta(x)$ exists and is finite. Moreover, $D_Q P_\theta(x) = D_\mu P_\theta(x)/D_\mu Q(x)$.*

Proof. Suppose that $D_\mu P_\theta(x) = p$ and $D_\mu Q(x) = q > 0$. From Lemma 4.1 there exists $N_1 \in \mathbb{N}$ such that

$$\left| \frac{Q(B(x, \frac{1}{k}))}{\mu(B(x, \frac{1}{k}))} - q \right| < \frac{q}{2},$$

for every $k \geq N_1$. Set $M = 2/q^2$. Hence, $M > 0$ and

$$\frac{\mu(B(x, \frac{1}{k}))}{qQ(B(x, \frac{1}{k}))} < M, \tag{4.6}$$

for all $k \geq N_1$. Now take $\epsilon > 0$. We will find some $N_0 \in \mathbb{N}$ such that

$$\left| \frac{P_\theta(B(x, \frac{1}{n}))}{Q(B(x, \frac{1}{n}))} - \frac{p}{q} \right| < \epsilon,$$

for all $n \geq N_0$. There are two cases to consider: (i) the case $p = 0$ and (ii) the case $p > 0$. We consider only the case when $p > 0$. The proof when p is zero is quite similar. By Lemma 4.1, there exist $N_1, N_2 \in \mathbb{N}$ such that

$$\left|\frac{P_\theta(B(x,\frac{1}{n}))}{\mu(B(x,\frac{1}{n}))} - p\right| < \frac{\epsilon}{2qM} \quad \text{and} \quad \left|\frac{Q(B(x,\frac{1}{k}))}{\mu(B(x,\frac{1}{k}))} - q\right| < \frac{\epsilon}{2pM}. \tag{4.7}$$

for all $n \geq N_1$ and for all $k \geq N_2$. Let $N_0 = \max\{N_1, N_2\}$ and let $n \geq N_0$. Since

$$\left|\frac{P_\theta(B(x,\frac{1}{n}))}{\mu(B(x,\frac{1}{n}))}\frac{\mu(B(x,\frac{1}{n}))}{Q(B(x,\frac{1}{n}))} - \frac{p}{q}\right| = \left|\left(\frac{P_\theta(B(x,\frac{1}{n}))}{\mu(B(x,\frac{1}{n}))}q - \frac{Q(B(x,\frac{1}{n}))}{\mu(B(x,\frac{1}{n}))}p\right)\frac{\mu(B(x,\frac{1}{n}))}{qQ(B(x,\frac{1}{n}))}\right|$$

and $N_0 \geq N_1$, we have from (4.6) that

$$\left|\frac{P_\theta(B(x,\frac{1}{n}))}{\mu(B(x,\frac{1}{n}))}\frac{\mu(B(x,\frac{1}{n}))}{Q(B(x,\frac{1}{n}))} - \frac{p}{q}\right| \leq M\left|\frac{P_\theta(B(x,\frac{1}{n}))}{\mu(B(x,\frac{1}{n}))}q - \frac{Q(B(x,\frac{1}{n}))}{\mu(B(x,\frac{1}{n}))}p + pq - pq\right|, \tag{4.8}$$

Finally, it follows from (4.7) and (4.8) that

$$\left|\frac{P_\theta(B(x,\frac{1}{n}))}{Q(B(x,\frac{1}{n}))} - \frac{p}{q}\right| \leq Mq\left|\frac{P_\theta(B(x,\frac{1}{n}))}{\mu(B(x,\frac{1}{n}))} - p\right| + Mp\left|\frac{Q(B(x,\frac{1}{n}))}{\mu(B(x,\frac{1}{n}))} - q\right| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

for all $n \geq N_0$. □

Since $D_\mu Q(x) > 0$ Q almost surely (and then P_θ almost surely for every $\theta \in \Theta$), we can consider that the density of Q with respect to μ is strictly postive for all $x \in A$.

Lemma 4.3. *If $D_Q P_\theta(x)$ and $D_\mu Q(x)$ exist and are finite for some $x \in S_\theta$, then $D_\mu P_\theta(x)$ exists and is finite. Furthermore, $D_\mu P_\theta(x) = D_Q P_\theta(x) D_\mu Q(x)$.*

Proof. Similar to the proof of Lemma 4.2. □

Lemma 4.4. *Let $x_0 \in S_\theta$. Suppose that there exists a version $f_\theta \in [\frac{dP_\theta}{d\mu}]$ such that f_θ is continuous at x_0. Then, $D_\mu P_\theta(x_0) = f_\theta(x_0)$.*

Proof. Given $\epsilon > 0$, there exists, by the continuity of f_θ at x_0, a positive integer N_0 such that

$$x \in B\left(x_0, \frac{1}{n}\right) \quad \Longrightarrow \quad f_\theta(x) \in B(f_\theta(x_0), \epsilon),$$

for all $n \geq N_0$. Hence,

$$P_\theta(B(x_0, 1/n)) = \int_{B(x_0,1/n)} f_\theta(x) d\nu(x) \leq (f_\theta(x_0) + \epsilon)\nu(B(x_0, 1/n)).$$

The last formula shows that, for all $n \geq N_0$,

$$\frac{P_\theta(B(x_0, 1/n))}{\nu(B(x_0, 1/n))} - f_\theta(x_0) \leq \epsilon,$$

and using the fact that $f_\theta(x_0) - \epsilon < f_\theta(x)$ for every $x \in B(x_0, 1/n)$, $n \geq N_0$, we get that

$$\left| \frac{P_\theta(B(x_0, 1/n))}{\nu(B(x_0, 1/n))} - f_\theta(x_0) \right| \leq \epsilon,$$

for all $n \geq N_0$. The result follows from Lemma 4.1. □

Lemma 4.4 also appears in Piccioni (1982) (Theorem V). The two proofs are quite similar. Whilst our proof uses Lemma 4.1, Piccioni's proof uses the definition of lim sup and lim inf. Moreover, Piccioni's result is established for any metric separable space and locally finite measures.
We are now ready to state Theorem 4.3.

Theorem 4.3. *Let $\mathcal{P}$ be a family of probability measures on $\mathbb{R}^n$ and suppose that:*

(i) $S = S_\theta$ for all $\theta \in \Theta$;

(ii) there exists a Radon measure μ on $\mathbb{R}^n$ and a version $f_{\mu,\theta} \in [\frac{dP_\theta}{d\mu}]$, for each $\theta \in \Theta$, that is continuous on S.

Then, for any other Radon measure ν that dominates $\mathcal{P}$, there exists a measurable set A, with $P_\theta(A) = 1$ for all $\theta \in \Theta$, such that

(I) the derivatives $D_\mu P_\theta$ and $D_\nu P_\theta$ exist for all $x \in A$ and for all $\theta \in \Theta$;

(II) $D_\mu P_\theta$ and $D_\nu P_\theta$ are versions that satisfy the Likelihood Proportionality Theorem, i.e.,

$$D_\mu P_\theta(x) \propto_\theta D_\nu P_\theta(x), \ \forall \theta \in \Theta, \ \forall x \in A.$$

Proof. Theorem 1.10 implies that $Q(S) = 1$, where Q is the minimal dominanting measure for $\mathcal{P}$ given by Lemma 2.1. Also, condition (ii) and Lemma 4.4 imply that $D_\mu P_\theta$ exists for all $\theta \in \Theta$ and for all $x \in S$. By Theorem 4.1, there exists a measurable set B_μ such that $Q(B_\mu) = 1$ and $D_\mu Q$ exists for all $x \in B_\mu$. Hence, $Q(B_\mu \cap S) = 1$ and $D_\mu Q$ exists for all $x \in (B_\mu \cap S)$. Remember that (see comment just before Lemma 4.2) we can consider that $D_\mu Q$ is strictly positive on $(B_\mu \cap S)$. Thus, by Lemma 4.3, it follows that $D_Q P_\theta(x)$ exists for all $x \in (B_\mu \cap S)$ and for all $\theta \in \Theta$. Furthermore

$$D_\mu P_\theta(x) = D_Q P_\theta(x) D_\mu Q(x), \ \forall x \in (B_\mu \cap S), \ \forall \theta \in \Theta. \tag{4.9}$$

Now note that, by Theorem 4.1, there exists a measurable set B_ν such that $Q(B_\nu) = 1$ and $D_\nu Q$ exists and is strictly positive for all $x \in B_\nu$. Define $A = (B_\mu \cap B_\nu \cap S)$ and note that $Q(A) = 1$. Since $D_Q P_\theta(x)$ and $D_\nu Q$ exist for all $\in A$, it follows from Lemma 4.3 that $D_\nu P_\theta$ exists for all $x \in A$ and

$$D_\nu P_\theta(x) = D_Q P_\theta(x) D_\nu Q(x), \ \forall \theta \in \Theta. \tag{4.10}$$

Finally, it follows from (4.9) and (4.10) that

$$D_\mu P_\theta(x) = D_\nu P_\theta(x) \frac{D_\mu Q(x)}{D_\nu Q(x)}$$

for all $x \in A$ and for all $\theta \in \Theta$. Since $D_\mu Q(x)/D_\nu Q(x)$ does not depend on θ, the proof is complete. □

4.3 Extension for general spaces

In this section, we will extend the result in Theorem 4.3 for any Vitali metric measure space.

Definition 4.4. A metric measure space id defined to be a triple (Ω, d, μ), where (Ω, d) is a separable metric space and μ is a nontrivial locally finite Borel regular mesure on Ω.

Definition 4.5 (Fine covering)**.** A covering $\mathcal{B}$ of a set $A \subset \Omega$ by closed balls[1] is called fine if

$$\inf\{r;\ r > 0 \text{ and } \overline{B}(x, r) \in \mathcal{B}\} = 0, \tag{4.11}$$

for each $x \in A$.

Definition 4.6. A metric measure space (Ω, d, μ) is called a Vitali metric measure space, and the measure μ a Vitali measure, if, and only if, for every subset A of Ω and for every covering $\mathcal{B}$ of A by closed balls satisfying (4.11) for each $x \in A$ there exists a pairwise disjoint subcollection $\mathcal{C} \subset \mathcal{B}$ such that

$$\mu\left(A \setminus \bigcup_{B \in \mathcal{C}} B\right) = 0.$$

Theorem 4.4 (Lebesgue differentiation theorem)**.** *(Heinonen et al. (2015), Section 3.4) Let (Ω, d, μ) be a Vitali metric space and let f be a locally integrable*

[1] $\overline{B}(x, r) = \{y \in \Omega;\ d(x, y) \leq r\}$

function. Then

$$\lim_{r\to 0} \frac{1}{\mu(\overline{B}(x,r))} \int_{\overline{B}(x,r)} f(y)d\mu(y) = f(x) \tag{4.12}$$

for almost every $x \in \Omega$.

Let P be a probability measure dominated by μ. Then, if $f \in [\frac{dP}{d\mu}]$, the Lebesgue differentiation theorem implies that

$$\lim_{r\to 0} \frac{P(\overline{B}(x,r))}{\mu(\overline{B}(x,r))} = f(x) \ a.e.$$

As in Section 4.1, we call x a density point of P with respect to μ if the limit in (4.12) exists. The set of density points of P with respect to μ will be denoted by $D(P,\mu)$.
NOte that if μ is any Radon measure in $\mathbb{R}^n$, then $(\mathbb{R}^n, \mu)$ is a Vitali metric space (Heinonen (2001), Remark 1.13). Therefore, the next theorem is an extension of Theorem 4.3.

Theorem 4.5. *Let* $\mathcal{P}$ *be a family of probability measures on* Ω *and suppose that:*

(i) $S = S_\theta$ *for all* $\theta \in \Theta$*;*

(ii) there exists a Vitali measure μ *on* (Ω, d) *and a version* $f_{\mu,\theta} \in [\frac{dP_\theta}{d\mu}]$*, for each* $\theta \in \Theta$*, that is continuous on* S*.*

Then, for any other Vitali measure ν *that dominates* $\mathcal{P}$*, there exists a measurable set* A*, with* $P_\theta(A) = 1$ *for all* $\theta \in \Theta$*, such that*

(I) the derivatives $D_\mu P_\theta$ *and* $D_\nu P_\theta$ *exist for all* $x \in A$ *and for all* $\theta \in \Theta$*;*

(II) $D_\mu P_\theta$ *and* $D_\nu P_\theta$ *are versions that satisfy the Likelihood Proportionality Theorem, i.e.,*

$$D_\mu P_\theta(x) \propto_\theta D_\nu P_\theta(x), \ \forall \theta \in \Theta, \ \forall x \in A.$$

Proof. The proof is analogous to that from Theorem 4.3. The Lebesgue differentiation theorem guarantees the validity of Lemmas 4.1, 4.2, 4.3 and 4.4 for Vitali metric measure spaces. □

Chapter 5

Final remarks

5.1 Conclusion

In this thesis, we discussed some mathematical foundations of Likelihood Theory, more specifically, the definition of likelihood function, in both parametric and non-parametric contexts. We consider the general definition of likelihood function in terms of the Radon-Nikodým derivative of each probability measure in the model w.r.t. any dominating measure, evaluated at the observed sample and, therefore, seen as a function of θ. The Likelihood Proportionality Theorem validates this definition in terms of the Likelihood Principle by guaranteeing the existence of versions of the densities that are, almost surely, proportional for any two dominating measures.

Whilst the Likelihood Proportionality Theorem only guarantees the existence of versions that are proportional, a practical strategy to find such versions is provided by considering densities which satisfy at least one of two continuity properties. First, the density is continuous in ω and second, the density defines a continuous likelihood function. Namely, those versions are always in accordance with the Likelihood Principle.

The decision of which dominating measure to use is particularly interesting in cases where there exists no or more than one obvious choice. Both cases are illustrated and discussed in Chapter 3 for some general classes of models. In particular, we presented appealing versions of RN derivatives and discussed how different choices, although leading to the same result, may have influence in the complexity of the inference process. We also discussed, in Section 2.4, the choice of the prior predictive measure as a dominating measure.

Finally, a method to obtain valid likelihood functions was proposed in Chapter

4 by discussing how Radon-Nikodým derivatives behave, under some mild conditions, like a genuine derivative that can be determined from the differentiation of measures. Moreover, those genuine derivatives are shown to satisfy the Likelihood Proportionality Theorem.

5.2 Future work

This thesis studies some aspects of likelihood theory from a point of view never actually considered in the literature before. It is then natural that the work developed here instigates further investigation in the area.

A first problem would be to extend the results about differentiation of measure from Chapter 4 for more general spaces. For example, the space of càdlàg functions. In this context, it would be interesting to investigate whether likelihood functions for continuous time/space models could be devised based on discrete approximations. In particular, Gaussian process driven models.

Given the generality of the definition of likelihood function, it would be interesting to study its properties in non-parametric contexts, i.e., when the parametric space is infinite. In particular, what properties can be established for maximum likelihood estimators of finite-dimensional functions of the parameter? Or, in a Bayesian context, what are the properties of the posterior distribution of those functions. What are the implications of adopting improper prior measures?

Bibliography

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

Barndorff-Nielsen, O., Hoffmann-Jørgensen, and Pedersen, K. (1976). On the minimal sufficiency of the likelihood function. *Scandinavian Journal of Statistics*, 3:115–127.

Bauer, H. (2001). *Measure and Integration Theory*. De Gruyter studies in mathematics. W. de Gruyter.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*. Lecture Notes-Monograph Series. Institute of Mathematical Statistics, Hayward, California, 2nd edition.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based inference for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society, Series B*, 68(3):333–382.

Evans, L. C. and Gariepy, R. F. (1991). *Measure Theory and Fine Propoerties of Functions*. Studies in Advanced Mathematics. Taylor & Francis.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron. I*, part 4:3–32.

Fraser, D., McDunnough, P., Naderi, A., and Plante, A. (1997). From the likelihood map to euclidean minimal sufficiency. *Journal of Probability and Mathematical Statistics*, 17:223–230.

Fraser, D. and Naderi, A. (1996). On the definition of conditional probability. *Research Developments in Probability and Statistics*, pages 23–26.

Fraser, D. and Naderi, A. (2007). Minimal sufficient statistics emerge from the observed likelihood functions. *International Journal of Statistical Sciences*, 6:55–61.

Gonçalves, F. B. and Gamerman, D. (2017). Exact Bayesian inference in spatio-temporal Cox processes driven by multivariate Gaussian processes. *To appear in Journal of the Royal Statistical Society - Series B*.

Gonçalves, F. B., Roberts, G. O., and Łatuszyński, K. G. (2017). Exact Monte Carlo likelihood-based inference for jump-diffusion processes. *arXiv preprint arXiv:1707.00332*.

Gottardo, R. and Raftery, A. E. (2009). Markov chain Monte Carlo with mixtures of mutually singular distributions. *Journal of Computational and Graphical Statistics*, 17:949–975.

Halmos, P. R. and Savage, L. J. (1949). Application of the Radon-Nikodym Theorem to the theory of sufficient statistics. *The Annals of Mathematical Statistics*, 20:225–241.

Heinonen, J. (2001). *Lectures on Analysis on Metric Spaces*. Hochschultext / Universitext. Springer New York.

Heinonen, J., Koskela, P., Shanmugalingam, N., and Tyson, J. (2015). *Sobolev Spaces on Metric Measure Spaces*. New Mathematical Monographs. Cambridge University Press.

Izbicki, R., Lee, A. B., and Schafer, C. M. (2014). High-dimensional density ratio estimation with extensions to approximate likelihood computation. *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 33.

Kingman, J. F. C. (1993). *Poisson Processes*. Oxford University Press, New York.

Kloeden, P. and Platen, E. (1995). *Numerical Solution of Stochastic Differential Equations*. Springer, New York.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung (in German)*. Julius Springer, Berlin.

Lehmann, E. (1986). *Testing Statistical Hypotheses*. Springer, New York.

Lindley, D. V. (1953). Statistical Inference. *Journal of the Royal Statistical Society. Series B*, 15:131–179.

Munkres, J. (2014). *Topology*. Featured Titles for Topology Series. Prentice Hall, Incorporated, 2nd edition.

Nikodým, O. (1930). Sur une gÃ©nÃ©ralisation des intÃ©grales de M. J. Radon. *Fundamenta Mathematicae (in French)*, 15:131–179.

Piccioni, M. (1982). On the definition of likelihood in abstract spaces. *Journal of the Franklin Institute*, 313:1–15.

Piccioni, M. (1983). Continuous versions of Radon-Nikodym derivatives as likelihood ratios. *Systems & Control Letters*, 2:369–374.

Radon, J. (1913). Theoric undanwendungen der absolut additiven mengenfunktionen. *Sitz Akad Wiss*, 122:1295–1438.

Reid, N. (2013). Likelihood formalities. STA3000 Lecture notes - http://www.utstat.utoronto.ca/reid/sta3000y/likelihood-formal.pdf.

Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, 88:603–621.

Shao, J. (2003). *Mathematical Statistics*. Springer, New York.