Doctoral Thesis

Statistical decisions based directly on the likelihood function

Author(s):
Cattaneo, Marco E.G.V.

Publication Date:
2007

Permanent Link:
https://doi.org/10.3929/ethz-a-005463829

Rights / License:
In Copyright - Non-Commercial Use Permitted
Statistical Decisions Based Directly on the Likelihood Function

A dissertation submitted to
ETH Zurich
for the degree of
Doctor of Mathematics

presented by
Marco E. G. V. Cattaneo
Dipl. Math. ETH
born September 19, 1976
citizen of Capriasca, Ticino

accepted on the recommendation of
Prof. em. Dr. Frank Hampel, examiner
Prof. Dr. Sara van de Geer, co-examiner
Prof. Dr. Thomas Augustin, co-examiner

2007
## Contents

Abstract ................................................................. iii  
Sunto ................................................................. v  
Notation ............................................................... vii  

1 Introduction .......................................................... 1  
  1.1 Statistical Decisions ............................................. 1  
      1.1.1 Classical Decision Theory .......................... 3  
      1.1.2 Bayesian Decision Theory ......................... 5  
  1.2 Likelihood Function .......................................... 8  
      1.2.1 Likelihood-Based Inference ......................... 10  
  1.3 Likelihood-Based Statistical Decisions .................. 13  
      1.3.1 MPL Criterion ........................................ 13  
      1.3.2 Other Likelihood-Based Decision Criteria .......... 20  

2 Nonadditive Measures and Integrals .......................... 25  
  2.1 Nonadditive Measures ........................................ 25  
      2.1.1 Types of Measures .................................. 26  
      2.1.2 Derived Measures .................................. 28  
  2.2 Nonadditive Integrals ........................................ 30  
      2.2.1 Shilkret Integral ................................... 31  
      2.2.2 Invariance Properties ............................... 35  
      2.2.3 Regular Integrals .................................... 38  
      2.2.4 Subadditivity ......................................... 51  
      2.2.5 Continuity ............................................. 55  
      2.2.6 Choquet Integral ....................................... 58
3 Relative Plausibility .................................................. 63
  3.1 Description of Uncertain Knowledge .......................... 63
    3.1.1 Representation Theorem .................................. 66
    3.1.2 Updating .................................................. 78
  3.2 Hierarchical Model .............................................. 81
    3.2.1 Classical and Bayesian Models ............................ 83
    3.2.2 Imprecise Probability Models ............................ 86

4 Likelihood-Based Decision Criteria .............................. 95
  4.1 Decision Theoretic Properties ................................ 95
    4.1.1 Attitudes Toward Ignorance ................................ 102
    4.1.2 Sure-Thing Principle .................................... 116
  4.2 Statistical Properties ......................................... 127
    4.2.1 Invariance Properties .................................... 128
    4.2.2 Asymptotic Optimality .................................... 138

5 Point Estimation .................................................. 149
  5.1 Likelihood-Based Estimates .................................... 149
    5.1.1 Asymptotic Efficiency ..................................... 152
  5.2 Shilkret Integral and Quasiconvexity .......................... 157
    5.2.1 Impossible Estimates ....................................... 160

References ............................................................ 171

Index ................................................................. 179

Curriculum Vitae .................................................... 183
Abstract

This thesis studies the possibility of basing statistical decisions directly on the likelihood function. By describing the problems of statistical inference as particular decision problems, the usual likelihood-based inference methods can be obtained as special cases. These are among the most appreciated methods of statistical inference, although in general they are not optimal from the repeated sampling point of view. In fact, their principal advantages are that, being based on the likelihood function, they are intuitive, generally applicable, and asymptotically optimal. These properties are shared by all the methods that base decisions on the likelihood function in some reasonable way.

The decision criteria based on the likelihood function are analyzed, and some of their decision theoretic properties are related to their representations by means of nonadditive integrals; in this connection, the theory of integration with respect to nonadditive measures is reviewed and extended. Some statistical properties of the likelihood-based decision criteria are studied as well, and a paradigmatic example of statistical decision problem is examined in more detail: the problem of point estimation in Euclidean spaces. A simple, intuitive decision criterion emerges as particularly interesting: the MPL criterion, which consists in using the likelihood of the statistical models to weight the respective losses, before applying the minimax criterion.

The likelihood function can be interpreted as a measure of the relative plausibility of the statistical models considered, and thus as the second level of a hierarchical description of uncertain knowledge, of which the statistical models form the first level. There is a similarity with the Bayesian model, in which the second level consists of a probability measure, but the measure of relative plausibility considered in this thesis can also describe
iv Abstract

ignorance; in particular, in the case of complete ignorance we obtain the classical model, which in fact consists only of the first level. The measure of relative plausibility and the MPL criterion are linked by a simple representation theorem for preferences between randomized decisions.
Sunto

Questa tesi studia la possibilità di basare le decisioni statistiche direttamente sulla funzione di verosimiglianza. Descrivendo i problemi di inferenza statistica come particolari problemi di decisione, i consueti metodi di inferenza basati sulla verosimiglianza possono essere ottenuti come casi particolari. Essi sono tra i più apprezzati metodi di inferenza statistica, benché in generale non siano ottimali secondo il principio del campionamento ripetuto. Infatti, i loro principali vantaggi, implicati dall’essere basati sulla funzione di verosimiglianza, sono l’intuitività, l’applicabilità generale, e l’ottimalità asintotica. Queste proprietà sono condivise da ogni metodo che basi in modo ragionevole le decisioni sulla funzione di verosimiglianza.

I criteri di decisione basati sulla funzione di verosimiglianza sono analizzati, e alcune delle loro proprietà inerenti alla teoria della decisione sono poste in relazione con le loro rappresentazioni per mezzo di integrali non additivi; a questo proposito, la teoria dell’integrazione rispetto a misure non additive è passata in rassegna e ampliata. Alcune proprietà statistiche dei criteri di decisione basati sulla funzione di verosimiglianza sono pure studiate, e un esempio paradigmatico di problema di decisione è esaminato più dettagliatamente: il problema della stima puntuale negli spazi euclidei. Un criterio di decisione semplice e intuitivo risulta essere particolarmente interessante: il criterio MPL, che consiste nell’utilizzare la verosimiglianza dei modelli statistici per pesare le relative perdite prima di applicare il criterio minimax.

La funzione di verosimiglianza può essere interpretata come una misura della relativa plausibilità dei modelli statistici considerati, e quindi come il secondo livello di una rappresentazione gerarchica della conoscenza incerta, di cui i modelli statistici costituiscono il primo livello. C’è una chiara
analoga con il modello bayesiano, nel quale il secondo livello consiste in una misura di probabilità, ma la misura di plausibilità relativa considerata in questa tesi può anche descrivere l’ignoranza; in particolare, in caso di ignoranza totale otteniamo il modello classico, che di fatto consiste unicamente nel primo livello. La misura di plausibilità relativa è associata al criterio MPL da un semplice teorema di rappresentazione per le preferenze tra decisioni randomizzate.
Notation

Let $A$ and $B$ be two sets. The inclusion of $A$ in $B$ is denoted by $A \subseteq B$, while $A \subset B$ denotes the strict inclusion; the set $B \setminus A$ is the difference of $B$ and $A$. The power set of $A$ is denoted by $2^A$, and $|A|$ denotes the cardinality of $A$. If $A$ is a set of sets, then $\bigcup A$ is the union of the elements of $A$, with the convention that $\bigcup \emptyset = \emptyset$.

The set of all functions $f : A \rightarrow B$ is denoted by $B^A$. If $f \in B^A$, then $f^{-1} : 2^B \rightarrow 2^A$ is the mapping assigning to each subset of $B$ its inverse image under $f$; but when $f$ is bijective, $f^{-1}$ can also denote its inverse function (the meaning of $f^{-1}$ should be clear from the context). The image of $a \in A$ under $f \in B^A$ is denoted by $f(a)$ or $f[a]$, but the brackets can be omitted when there is no ambiguity; for instance, the inverse image of $\{b\} \subseteq B$ can be expressed as $f^{-1}\{b\}$. If $f \in B^A$, and $C \subseteq A$, then $f|_C$ is the restriction of $f$ to $C$; and if $D$ is a set, and $g \in D^B$, then $g \circ f$ is the composition of $f$ and $g$. The identity function on $A$ is denoted by $\text{id}_A$, while $I_A$ denotes the indicator function of $A$ (the domain of $I_A$ should be clear from the context).

The set of real numbers is denoted by $\mathbb{R}$, while $\overline{\mathbb{R}}$ denotes the set of (affinely) extended real numbers: $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty, -\infty\}$. In the interval notation, a round bracket indicates the exclusion of the corresponding endpoint, while a square bracket indicates its inclusion; for instance, if $x, y \in \overline{\mathbb{R}}$, then $(x, y] = \{z \in \overline{\mathbb{R}} : x < z \leq y\}$. The set of positive real numbers is denoted by $\mathbb{P}$, while $\overline{\mathbb{P}}$ denotes the set of nonnegative extended real numbers:

$$\mathbb{P} = (0, \infty) \quad \text{and} \quad \overline{\mathbb{P}} = [0, \infty].$$

The supremum and infimum of $S \subseteq \overline{\mathbb{P}}$ are denoted respectively by $\sup S$ and $\inf S$, with the convention that $\sup \emptyset = 0$ and $\inf \emptyset = \infty$. If $f \in \overline{\mathbb{P}}^A$, 

and $C \subseteq A$, then $\sup f$ is the supremum of $f$, and $\sup_C f$ is the supremum of $f|_C$:

$$\sup_C f = \sup_{a \in C} f(a) = \sup \{ f(a) : a \in C \} = \sup f|_C;$$

and analogously for the infimum. When the supremum is attained, $\max$ can be substituted for $\sup$; and when the infimum is attained, $\min$ can be substituted for $\inf$.

Expressions involving functions are to be interpreted pointwise, when they do not make sense otherwise; for instance, if $f, g \in \mathbb{P}^A$, then $f \leq g$ means that $f(a) \leq g(a)$ for all $a \in A$. Accordingly, if $b \in B$, then $b$ can also denote the function in $B^A$ with constant value $b$ (the domain $A$ should be clear from the context); for instance, if $C \subseteq A$, and $I_C$ has domain $A$, then $1 - I_C$ is the indicator function of $A \setminus C$. A pair of curly brackets enclosing an expression involving functions denote the subset of the domain of the functions consisting of all arguments for which the expression is satisfied; for instance, if $f \in \mathbb{P}^A$, and $x \in \mathbb{P}$, then

$$\{ f \geq x \} = \{ a \in A : f(a) \geq x \}.$$

As usual in measure theory, we define $\infty - \infty = \infty \cdot 0 = 0$; but otherwise we adopt the standard definitions about extended real numbers. In particular, $\infty - \infty$ is undefined, and thus in general the difference of two functions $f, g \in \mathbb{P}^A$ is undefined: with a slight abuse in notation we define

$$\| f - g \| = \sup_{a \in \{ f \neq g \}} | f(a) - g(a) |.$$

The notation is reasonable, because when $f$ and $g$ are bounded, $\| f - g \|$ is the supremum norm of their difference; hence, the function $(f, g) \mapsto \| f - g \|$ on $\mathbb{P}^A \times \mathbb{P}^A$ extends the metric induced by the supremum norm. The notion of uniform convergence can be extended accordingly: the sequence $f_1, f_2, \ldots \in \mathbb{P}^A$ converges uniformly to $f \in \mathbb{P}^A$ when $\lim_{n \to \infty} \| f_n - f \| = 0$.

If $f \in \mathbb{P}^\mathbb{R}$, and $y \in \mathbb{R}$, then $\lim_{x \downarrow y} f(x)$ and $\lim_{x \uparrow y} f(x)$ denote the limits of $f$ at $y$ from the left and from the right, respectively (when the limits exist). If $x \in \mathbb{P}$, then $\log x$ is the (extended) natural logarithm of $x$. If $k$ is a positive integer, and $\delta \in \mathbb{P}$, then $B_\delta = \{ x \in \mathbb{R}^k : |x| \leq \delta \}$ is the closed ball in $\mathbb{R}^k$ with center $0$ and radius $\delta$, and $B_\delta^c = \mathbb{R}^k \setminus B_\delta$ is its complement (the value of $k$ should be clear from the context). If $k$ is a positive integer, and $y \in \mathbb{R}^k$, then the function $t_y : x \mapsto x - y$ on $\mathbb{R}^k$ is the
translation by \(-y\) (the value of \(k\) should be clear from the context); for instance, if \(\delta \in \mathbb{P}\), then \(t_y^{-1}(B_\delta)\) is the closed ball with center \(y\) and radius \(\delta\).

Let \((\Omega, \mathcal{C}, P)\) be a probability space. A random object is a measurable function \(X: \Omega \rightarrow \mathcal{X}\), where \(\mathcal{X}\) is a set equipped with a \(\sigma\)-algebra containing all singletons of \(\mathcal{X}\); the random object \(X\) is continuous if \(P\{X = x\} = 0\) for all \(x \in \mathcal{X}\), and it is discrete if there is a finite or countable \(\mathcal{Y} \subseteq \mathcal{X}\) such that \(P\{X \in \mathcal{Y}\} = 1\). A random variable is a random object \(X: \Omega \rightarrow \overline{\mathbb{R}}\), where \(\overline{\mathbb{R}}\) is equipped with the Borel \(\sigma\)-algebra.

The symbol \(\Box\) marks the end of a proof, while the symbol \(\Diamond\) marks the end of an example.
Introduction

In the present chapter the basic concepts of statistical decision problem and of likelihood function are introduced. The usual approaches to statistical decision problems and the principal likelihood-based inference methods are briefly described (avoiding foundational issues). Finally, the likelihood-based decision criteria appearing in the literature are presented.

1.1 Statistical Decisions

A statistical decision problem is described by a loss function

\[ L : \mathcal{P} \times \mathcal{D} \rightarrow \mathbb{P}, \]

where \( \mathcal{P} \) is a set of probability measures on a measurable space \((\Omega, \mathcal{A})\), and \( \mathcal{D} \) is a set.

\( \mathcal{P} \) is the set of statistical models considered: these are mathematical representations of some aspects of the reality under consideration (in particular, the observed data can be represented by elements of \( \mathcal{A} \)). It is important to note that we use the term “model” to indicate a single probability measure (without unknown parameters), while in the statistical literature it is also used to indicate a whole family of probability measures (such a family can be a subset of \( \mathcal{P} \)). It is not assumed that one of the models in \( \mathcal{P} \) is in some sense “true”, but our conclusions will be based on \( \mathcal{P} \) and can thus not be expected to be satisfactory when all models in \( \mathcal{P} \) poorly represent the reality. No structure is imposed on \( \mathcal{P} \); in particular, it is possible to enlarge \( \mathcal{P} \) in order to improve the robustness of the conclusions.
Introduction

\( \mathcal{D} \) is the set of possible decisions: a solution of the statistical decision problem is a subset of \( \mathcal{D} \) consisting of the decisions that are optimal according to some criterion. No structure is imposed on \( \mathcal{D} \); in particular, it is possible to restrict \( \mathcal{D} \) by excluding the dominated decisions.

The loss function \( L \) summarizes all important aspects of the possible decisions: \( L(P, d) \) is the loss we would incur, according to the model \( P \in \mathcal{P} \), by making the decision \( d \in \mathcal{D} \). The loss function is usually finite, but infinite values pose no problems. The decision \( d \) is correct for the model \( P \) if \( L(P, d) = 0 \) (it is not necessary that for each model there is a correct decision, or that each decision is correct for some model), and the unit in which the loss is expressed is of no importance. The loss is thus a relative quantity: that is, for all \( c \in \mathcal{P} \), the loss functions \( L \) and \( cL \) are equivalent, while the loss functions \( L \) and \( L+c \) are not (because of the peculiar meaning of the value 0). For each \( d \in \mathcal{D} \), let \( l_d \) be the function on \( \mathcal{P} \) defined by

\[
l_d(P) = L(P, d) \quad \text{for all } P \in \mathcal{P}.
\]

If \( l_d \leq l_{d'} \), then \( d' \) cannot be preferred to \( d \); and if \( l_d = l_{d'} \), then \( d \) and \( d' \) are equivalent (in the sense that no one can be preferred to the other). A decision \( d \) is said to dominate a decision \( d' \) if \( l_d \leq l_{d'} \) and \( l_d \neq l_{d'} \); in this case, if we have to choose between \( d \) and \( d' \), then we must choose \( d \) (there is no reason for choosing \( d' \)). A decision criterion usually corresponds to a monotonic functional \( V : \mathcal{P}^p \rightarrow \mathcal{P} \) (where monotonic means that \( l \leq l' \) implies \( V(l) \leq V(l') \)), in the sense that it can be expressed as follows: minimize \( V(l_d) \). In this case, an optimal decision (that is, a decision \( d \) minimizing \( V(l_d) \)) can in general be dominated, but only by another optimal decision; if necessary, the criterion can thus be refined by discarding from the set of optimal decisions those that are dominated (by other optimal decisions).

Wald (1939) introduced the statistical decision problem as a generalization of the problems of statistical estimation and testing hypotheses. For example, if \( \mathcal{G} \) is a set, and \( g : \mathcal{P} \rightarrow \mathcal{G} \) is a mapping, then the problem of estimating \( g(P) \) can be described by a loss function \( L \) on \( \mathcal{P} \times \mathcal{G} \) expressing the estimation error: in particular, \( g(P) \) is a correct decision for \( P \). Inference problems (such as estimation problems) can also be stated without reference to a particular loss function, but some expression of the loss incurred is actually needed in order to compare different solutions. Often some standard loss function is used, as in the following simple example, which will be considered throughout the present chapter.
Example 1.1. Let $X$ be the number of successes in $n$ independent binary experiments (Bernoulli trials) with common success probability $p$. Assume that we know $n$, and we want to estimate $p$ with squared error loss. Let $\mathcal{P} = \{P_p : p \in [0, 1]\}$ be a family of probability measures (the choice of the measurable space $(\Omega, \mathcal{A})$ is irrelevant) such that under each model $P_p$ the random variable $X$ has the binomial distribution with parameters $n$ and $p$. The estimation problem is then described by the loss function $L : (P_p, d) \mapsto (p - d)^2$ on $\mathcal{P} \times [0, 1]$. 

1.1.1 Classical Decision Theory

Assume first that we have no information at all about which models are more or less plausible. In this case we only know that, according to the models in $\mathcal{P}$, the loss we would incur by making the decision $d$ lies in the range of $l_d$, and thus in particular between $\inf l_d$ and $\sup l_d$. The functional $V = \sup$ gives us a useful upper bound for the loss incurred, and it corresponds to the most important general decision criterion in the case of complete ignorance about the models in $\mathcal{P}$: the minimax criterion:

$$\text{minimize } \sup l_d.$$ 

By contrast, the functional $V = \inf$ gives us a less interesting lower bound for the loss incurred, and in particular all $d \in \mathcal{D}$ that are correct for some $P_d \in \mathcal{P}$ are optimal according to the corresponding decision criterion (called minimin criterion), which is therefore often useless. A compromise between the minimax and the minimin criteria is given by the Hurwicz criterion with optimism parameter $\lambda \in (0, 1)$:

$$\text{minimize } \lambda \inf l_d + (1 - \lambda) \sup l_d.$$ 

This criterion was introduced by Hurwicz (1951); it reduces to the minimax criterion when each $d \in \mathcal{D}$ is correct for some $P_d \in \mathcal{P}$.

Example 1.2. The minimax criterion applied to the estimation problem of Example 1.1 (without considering the realization of $X$) leads to the estimate $d = \frac{1}{2}$. Since each possible estimate $d \in [0, 1]$ is correct for the model $P_d \in \mathcal{P}$, the minimin criterion is useless, and the Hurwicz criterion reduces to the minimax criterion. \qed
Consider now that an event $A \in \mathcal{A}$ has been observed: this observation gives us some information about the relative plausibility of the models in $\mathcal{P}$; and to be reasonable a decision criterion must be able to use this information. The classical approach consists in considering the observation as a particular realization $A = \{X = x_A\}$ of a random object $X : \Omega \rightarrow \mathcal{X}$, in selecting a whole decision function $\delta : \mathcal{X} \rightarrow \mathcal{D}$ instead of a single decision $d = \delta(x_A)$, and in evaluating the different decision functions without considering that $X = x_A$ has already been observed. In this sense, the selection of the decision function is based on a pre-data evaluation. Let $\Delta$ be a set of decision functions $\delta$ such that for each $P \in \mathcal{P}$ the function $x \mapsto L[P, \delta(x)]$ on $\mathcal{X}$ is measurable: under each model $P$ the loss we would incur by using the decision function $\delta \in \Delta$ is the random variable $L[P, \delta(X)]$. In order to compare different decision functions, the random loss is reduced to a single, representative value: usually the expected value, but other quantities (such as quantiles) can be used instead. That is, for each model $P \in \mathcal{P}$ and each decision function $\delta \in \Delta$ we have a representative value $L'(P, \delta)$ of the random loss: the loss function $L'$ on $\mathcal{P} \times \Delta$ describes the problem of selecting a decision function $\delta \in \Delta$. If $L'(P, \delta)$ is defined as the expected value of $L[P, \delta(X)]$ with respect to $P$, then $L'$ is called risk function, and the minimax criterion applied to the decision problem described by $L'$ is the minimax risk criterion introduced by Wald (1939).

**Example 1.3.** The minimax risk criterion applied to the estimation problem of Example 1.1 (with $\mathcal{X} = \{0, \ldots, n\}$, and $\Delta$ defined as the set of all decision functions on $\mathcal{X}$) leads to the estimate

$$\lambda_n \frac{X}{n} + (1 - \lambda_n) \frac{1}{2},$$

with $\lambda_n = \frac{\sqrt{n}}{\sqrt{n} + 1}$

(see Hodges and Lehmann, 1950). Note that the case with $n = 0$ corresponds to the case without observations (the estimate obtained in Example 1.2), while for large $n$ the estimate is approximately $\frac{X}{n}$.  

The classical approach requires the definition of the random object $X$; that is, it requires the consideration of what could have been observed instead of $A$. In general this is not a problem, since the alternatives to $A$ have already been considered when defining $\mathcal{P}$. More problematic can be the assumption that the same decision problem would have been faced for all possible realizations of $X$. Even if this assumption is reasonable,
the classical approach has two main drawbacks with respect to conditional methods (that is, methods selecting a single decision in the light of the observation \( X = x_\theta \), without considering the other possible realizations of \( X \)). The first one is that selecting a decision function is in general much more complicated than selecting a single decision. The second one is that the selection of the decision function is based on a pre-data evaluation, whose meaning can be questionable once the data have been observed (see for instance Goutis and Casella, 1995).

### 1.1.2 Bayesian Decision Theory

When no data are observed, the Bayesian approach consists in averaging over the models in \( \mathcal{P} \) by means of a probability measure \( \pi \) on \( (\mathcal{P}, \mathcal{C}) \), where \( \mathcal{C} \) is a \( \sigma \)-algebra of subsets of \( \mathcal{P} \) such that \( I_d \) is measurable for all \( d \in \mathcal{D} \): we obtain the **Bayesian criterion**:

\[
\text{minimize } E_\pi(l_d).
\]

If for each \( A \in \mathcal{A} \) the function \( P \mapsto P(A) \) on \( \mathcal{P} \) is measurable, then \( \pi \) can be combined with the models in \( \mathcal{P} \) (interpreted as a stochastic kernel) to obtain a probability measure \( P_\pi \) on \( (\mathcal{P} \times \Omega, \mathcal{E}) \), where \( \mathcal{E} \) is the \( \sigma \)-algebra of subsets of \( \mathcal{P} \times \Omega \) generated by the sets \( C \times A \) with \( C \in \mathcal{C} \) and \( A \in \mathcal{A} \). If we observe an event \( A \in \mathcal{A} \), then we can condition the probability measure \( P_\pi \) on the event \( \mathcal{P} \times A \in \mathcal{E} \), obtaining in particular an updated marginal probability measure on \( (\mathcal{P}, \mathcal{C}) \), which can then be used in the Bayesian criterion. The Bayesian approach consists thus in combining the models in \( \mathcal{P} \) to obtain a single model \( P_\pi \), in updating it by conditioning it on the observed data, and in reducing the decision problem described by the loss function \( L \) on \( \mathcal{P} \times \mathcal{D} \) to the one described by the loss function \( L'' \) on \( \Pi \times \mathcal{D} \), where \( \Pi = \{ P_\pi \} \), and \( L''(P_\pi, d) \) is defined as the expected value of \( L(P, d) \) with respect to the updated version of \( P_\pi \). Since \( \Pi \) is a singleton, there is no concern about the decision criterion to be applied to the decision problem described by \( L'' \).

**Example 1.4.** For the estimation problem of Example 1.1, let \( \pi \) be a probability measure on \( \mathcal{P} \) corresponding to a beta distribution for the parameter \( p \in [0, 1] \). When observing \( X = x \), the updating of \( \pi \) corresponds to increasing the two parameters of the beta distribution by \( x \) and \( n - x \), respectively. The estimate of \( p \) obtained by applying the Bayesian criterion is the mean of the resulting beta distribution: if \( a, b \in \mathbb{P} \) are the two
parameters of the initial beta distribution corresponding to $\pi$, then the estimate is

$$\frac{X + a}{n + a + b} = \lambda_n \frac{X}{n} + (1 - \lambda_n) \frac{a}{a + b}, \quad \text{with} \quad \lambda_n = \frac{n}{n + a + b}.$$  

This estimate depends strongly on the choice of $a, b \in \mathbb{P}$; if $a + b$ is large with respect to $n$, then the estimate is approximately $\frac{a}{a+b}$, while if $a + b$ is small with respect to $n$, then the estimate is approximately $\frac{X}{n}$.

When the only available information about the relative plausibility of the models in $\mathcal{P}$ is provided by the observation $X = x$, the two most popular choices of $\pi$ are the probability measures on $\mathcal{P}$ corresponding to the beta distributions with parameters $a = b = 1$ (that is, the uniform distribution on $[0,1]$, proposed by Bayes, 1763) and $a = b = \frac{1}{2}$ (proposed by Jeffreys, 1946), respectively.

The Bayesian approach is conditional (the selection of a decision is based on a post-data evaluation), and it possesses the following important property of “temporal coherence” between pre-data and post-data evaluations. Let $X : \Omega \to \mathcal{X}$ be a random object, and define the decision function $\delta' : \mathcal{X} \to \mathcal{D}$ as follows: let $\delta'(x)$ be the decision that we would select after having observed $X = x$. If in the pre-data decision problem considered in Subsection 1.1.1 we define $L'$ as the risk function, some regularity conditions are satisfied (see for example Brown and Purves, 1973), and $\Delta$ is sufficiently wide to contain $\delta'$, then $\delta'$ is an optimal decision, according to the Bayesian criterion, for the decision problem described by $L'$. The temporal coherence is useful because it allows us to construct an optimal decision function without having to compare decision functions, but only single decisions. Thanks to the additivity of $E_\pi$, the Bayesian approach possesses also the following important property of “coherence with respect to additive loss”. Let $L_1$ and $L_2$ be two loss functions on $\mathcal{P} \times \mathcal{D}_1$ and $\mathcal{P} \times \mathcal{D}_2$, respectively, let $\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2$, and let $L$ be the loss function on $\mathcal{P} \times \mathcal{D}$ defined by $L[P, (d_1, d_2)] = L_1(P, d_1) + L_2(P, d_2)$ (for all $P \in \mathcal{P}$ and all $(d_1, d_2) \in \mathcal{D}$). If the Bayesian criterion can be applied to the three decision problems (in the sense that the measurability condition stated at the beginning of the present subsection is satisfied), and $d_1$ and $d_2$ are optimal for the decision problems described by $L_1$ and $L_2$, respectively, then $(d_1, d_2)$ is optimal for the decision problem described by $L$. The coherence with respect to additive loss is useful because for some decision problems it allows us to construct an optimal decision by considering only simpler problems.
These coherence properties are possible because in the (strict) Bayesian approach there is no uncertainty about the model: \( \Pi \) is a singleton. In fact, the central problem of the Bayesian approach is the choice of the averaging probability measure \( \pi \), which is usually interpreted either as a statistical model, or as a description of subjective uncertain knowledge. Independently of the interpretation of \( \pi \), it is reasonable to allow some uncertainty about the model \( P_\pi \) by assuming only that \( \pi \) is an element of some set \( \Gamma \) of probability measures on \((\mathcal{P}, \mathcal{C})\); the set \( \Gamma \) can also be interpreted as an imprecise probability measure on \((\mathcal{P}, \mathcal{C})\): see for example Walley (1991) and Weichselberger (2001). Each element of the set \( \Pi = \{ P_\pi : \pi \in \Gamma \} \) can be conditioned on the observed data, and the decision problem described by the loss function \( L \) on \( \mathcal{P} \times \mathcal{D} \) can be reduced to the one described by the loss function \( L'' \) on \( \mathcal{P} \times \mathcal{D} \), where \( L'' \) is defined as above. The minimax criterion applied to the (post-data) decision problem described by \( L'' \) is the \( \Gamma \)-minimax criterion, which was introduced by Hurwicz (1951) for the pre-data decision problem under the name of “generalized Bayes-minimax principle”; but the pre-data and the post-data applications of the \( \Gamma \)-minimax criterion do not coincide, since the temporal coherence of the strict Bayesian approach is lost (see for example Augustin, 2003). A decision criterion similar in spirit to the \( \Gamma \)-minimax is the restricted Bayesian criterion, which was introduced by Hodges and Lehmann (1952) for the pre-data decision problem. It consists in applying the Bayesian criterion (with respect to an averaging probability measure \( \pi \)) to the decision problem described by the risk function \( L' \), but only after having restricted the set \( \Delta \) to those decision functions for which the expected loss is bounded above by a particular constant. The intent of the restriction of \( \Delta \) is to limit the importance of the particular choice of \( \pi \).

**Example 1.5.** The application of the restricted Bayesian criterion is difficult even for the simple estimation problem of Example 1.1: some limited results are reported by Hodges and Lehmann (1952). The application of the \( \Gamma \)-minimax criterion is much simpler, but \( \Gamma \) must be chosen carefully. For instance, if for the estimation problem of Example 1.1 we choose as \( \Gamma \) the family of probability measures on \( \mathcal{P} \) corresponding to the family of beta distributions for the parameter \( p \in [0, 1] \), then the resulting estimate is \( \frac{1}{2} \), independently of \( n \) and of the observation \( X = x \). The family of beta distributions is a natural choice for \( \Gamma \) (it is the so-called conjugate family for this estimation problem, and for instance the two popular choices of \( \pi \)
considered at the end of Example 1.4 belong to it): this example simply shows that the $\Gamma$-minimax criterion is useless when $\Gamma$ is too wide.  

\section{1.2 Likelihood Function}

Let $\mathcal{P}$ be a set of statistical models, and let $A$ represent the observed data ($\mathcal{P}$ is a set of probability measures on a measurable space $(\Omega, \mathcal{A})$, and $A \in \mathcal{A}$). The \textbf{likelihood function} $lik$ on $\mathcal{P}$ induced by the observation of $A$ is defined by

\[ lik(P) = P(A) \quad \text{for all } P \in \mathcal{P}. \]

The likelihood function measures the relative plausibility of the models in the light of the observed data alone. Only ratios of the values of $lik$ for different models have meaning: proportional likelihood functions are equivalent. In particular, if the statistic $s : \mathcal{X} \to \mathcal{S}$ is sufficient for the random object $X : \Omega \to \mathcal{X}$, then the likelihood function induced by the observation $X = x_A$ is equivalent to the one induced by the observation $s(X) = s(x_A)$. That is, (the equivalence class of) the likelihood function depends only on sufficient statistics.

If $X : \Omega \to \mathcal{X}$ is a random object that is continuous under each model in $\mathcal{P}$, then the observation $X = x_A$ would induce a likelihood function with constant value 0. But in reality, because of the finite precision of all measurements, it is only possible to observe $X \in N_A$, for a neighborhood $N_A$ of $x_A$. If $f_P$ is a density of $X$ under the model $P$, then it can be useful to consider the approximation $P\{X \in N_A\} \approx \delta f_P(x_A)$, where $\delta \in \mathbb{P}$. If this holds for all $P \in \mathcal{P}$, then the function $P \mapsto f_P(x_A)$ on $\mathcal{P}$ is approximately equivalent to the likelihood function induced by the observation $X \in N_A$. When all $f_P$ are densities of $X$ with respect to the same measure on $\mathcal{X}$, the function $P \mapsto f_P(x_A)$ on $\mathcal{P}$ is often considered as the likelihood function induced by the observation $X = x_A$, independently of the quality of the above approximation. This alternative definition of likelihood function has some little problems (such as nonuniqueness), but it leads to an elegant mathematical theory, and all the results that we shall obtain for discrete random objects would be valid also for continuous ones if this alternative definition were used.

Under each model $P \in \mathcal{P}$, the likelihood ratio $\frac{lik(P)}{lik(P')}$. of $P$ against a different model $P' \in \mathcal{P}$ almost surely increases without bound when
more and more data are observed, if some weak conditions are satisfied. Consequently, when $\mathcal{P}$ is finite, the likelihood function tends to concentrate around $P$; and the same holds also when $\mathcal{P}$ is infinite, if some regularity conditions are satisfied.

A **parametrization** of $\mathcal{P}$ is a bijection $t : \mathcal{P} \rightarrow \Theta$, where the parameter space $\Theta$ can be any set, but usually $\Theta \subseteq \mathbb{R}^m$. The set $\mathcal{P}$ can be identified with $\Theta$, and therefore a likelihood function $lik$ on $\mathcal{P}$ can be identified with $lik \circ t^{-1}$, which is called likelihood function on $\Theta$ (with respect to $t$). A **pseudo likelihood function** on a set $\mathcal{G}$ with respect to a mapping $g : \mathcal{P} \rightarrow \mathcal{G}$ is a function that, to some extent at least, can be used as if it were a likelihood function on $\mathcal{G}$. For instance, the function $P \mapsto f_p(x_A)$ on $\mathcal{P}$ considered above as an approximation of the likelihood function induced by the observation $X \in N_A$ is a pseudo likelihood function on $\mathcal{P}$ with respect to $id_\mathcal{P}$. Other pseudo likelihood functions on $\mathcal{P}$ with respect to $id_\mathcal{P}$ can be obtained for example by modifying the likelihood function in order to take into account the complexity of the models in $\mathcal{P}$. If $lik$ is a likelihood function on $\mathcal{P}$, and $g : \mathcal{P} \rightarrow \mathcal{G}$ is a mapping, then the **profile likelihood function** $lik_g$ on $\mathcal{G}$ with respect to $g$ is defined by

$$lik_g(\gamma) = \sup_{\{g=\gamma\}} lik \text{ for all } \gamma \in \mathcal{G}.$$ 

In particular, if $t : \mathcal{P} \rightarrow \Theta$ is a parametrization of $\mathcal{P}$, then $lik_t$ is the likelihood function on $\Theta$. The profile likelihood function is a pseudo likelihood function that is defined for all functions $g$ on $\mathcal{P}$, and that usually leads to reasonable results; but better results can sometimes be achieved by using other pseudo likelihood functions. In the literature on likelihood-based statistical inference, many alternative pseudo likelihood functions have been proposed for different situations: see for instance Barndorff-Nielsen (1991) and Severini (2000, Chapters 8 and 9), and the references therein.

**Example 1.6.** Consider the estimation problem of Example 1.1: the mapping $P_p \mapsto p$ is a parametrization of $\mathcal{P}$ with parameter space $[0, 1]$. The likelihood function $lik$ on $[0, 1]$ induced by the observation $X = x$ satisfies

$$lik(p) = \binom{n}{x} p^x (1 - p)^{n-x} \text{ for all } p \in [0, 1].$$

By observing the realization of each one of the $n$ binary experiments, we would have obtained the likelihood function $lik'$ on $[0, 1]$ defined by

$$lik'(p) = p^x (1 - p)^{n-x} \text{ for all } p \in [0, 1],$$
Figure 1.1. Profile likelihood functions from Example 1.6.

where $x$ is the number of successes. The likelihood functions $\text{lik}$ and $\text{lik}'$ are equivalent (that is, $\text{lik} \propto \text{lik}'$), and in fact the number of successes is a sufficient statistic.

Let $g$ be the function on $\mathcal{P}$ assigning to each model $P_p$ the probability of observing at least 3 successes in 5 future independent binary experiments with the same success probability $p$:

$$g(P_p) = \sum_{k=3}^{5} \binom{5}{k} p^k (1-p)^{5-k} = 6 p^5 - 15 p^4 + 10 p^3 \quad \text{for all } p \in [0, 1].$$

Figure 1.1 shows the graphs of the profile likelihood functions $\text{lik}_g$ on $[0, 1]$ (actually, $g$ is a parametrization of $\mathcal{P}$) induced by the three observations $X = 8$ with $n = 10$ (dotted line), $X = 33$ with $n = 50$ (dashed line), and $X = 178$ with $n = 250$ (solid line), respectively (the three functions have been scaled to have the same maximum). The data have been generated using the model $P_{0.7}$, and in fact the profile likelihood function tends to concentrate around $g(P_{0.7}) \approx 0.837$. 

1.2.1 Likelihood-Based Inference

Let $\text{lik}$ be a likelihood function (induced by some observation $A$) on a set $\mathcal{P}$ of statistical models, let $\mathcal{G}$ be a set, and let $g : \mathcal{P} \rightarrow \mathcal{G}$ be a mapping. The maximum likelihood estimate $\hat{\gamma}_{\text{ML}}$ of $g(P)$ is the $\gamma \in \mathcal{G}$ maximizing
1.2 Likelihood Function

**Table 1.1.** Likelihood-based inferences from Example 1.7.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$n$</th>
<th>$\hat{\gamma}_{ML}$</th>
<th>$LR(\mathcal{H})$</th>
<th>${lik_g &gt; \beta \sup lik}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>10</td>
<td>0.942</td>
<td>0.146</td>
<td>(0.321, 1.000)</td>
</tr>
<tr>
<td>33</td>
<td>50</td>
<td>0.780</td>
<td>0.074</td>
<td>(0.461, 0.952)</td>
</tr>
<tr>
<td>178</td>
<td>250</td>
<td>0.852</td>
<td>$10^{-10}$</td>
<td>(0.741, 0.927)</td>
</tr>
</tbody>
</table>

The likelihood ratio test of the null hypothesis $H_0 : P \in \mathcal{H}$ versus the alternative $H_1 : P \in \mathcal{P} \setminus \mathcal{H}$ rejects the null hypothesis when $LR(\mathcal{H}) \leq \beta$, where the critical value $\beta \in (0, 1)$ is chosen in order to obtain a test of a particular level (that is, $\beta$ is chosen on the basis of a pre-data evaluation). The likelihood ratio tests were introduced by Neyman and Pearson (1928), and have a central place in both the hypothesis testing literature and practice; they are closely related to the following confidence regions.

The likelihood-based confidence region for $g(P)$ with cutoff point $\beta \in (0, 1)$ is the set

$$\{\gamma \in \mathcal{G} : LR(g = \gamma) > \beta\} = \{lik_g > \beta \sup lik\}.$$ 

The (pre-data) coverage probability of a likelihood-based confidence region with cutoff point $\beta$ depends on the set $\mathcal{P}$ of statistical models considered (and on the mapping $g$): this is the so-called calibration problem of likelihood-based inference.

**Example 1.1.** The method of maximum likelihood applied to the estimation problem of Example 1.1 leads to the estimate $\hat{p}_{ML} = \frac{X}{n}$ (when $n \geq 1$; while for $n = 0$ the maximum likelihood estimate is undefined). Table 1.1 gives some likelihood-based inferences for the mapping $g$ and the data considered in Example 1.6 (the corresponding profile likelihood functions $lik_g$ are plotted in Figure 1.1). The maximum likelihood estimate $\hat{\gamma}_{ML}$ tends to
the value \( g(P_{0.7}) \approx 0.837 \) of the function \( g \) for the model \( P_{0.7} \) used to generate the data. The likelihood ratio \( LR(\mathcal{H}) \) of the hypothesis \( \mathcal{H} = \{ g \leq \frac{1}{2} \} \) (which states that the probability \( g(P_p) \) is at most \( \frac{1}{2} \)) tends to 0. The likelihood-based confidence region for \( g(P) \) with cutoff point \( \beta \approx 0.036 \) (so that \(-2 \log \beta \) is the 99%-quantile of the \( \chi^2 \) distribution with one degree of freedom) is an interval tending to concentrate around \( g(P_{0.7}) \approx 0.837 \). For large \( n \) the coverage probability of this likelihood-based confidence region is approximately 99%; the exact coverage probability for \( n = 10 \) is plotted in Figure 1.2 as a function of \( p \in [0, 1] \).

The method of maximum likelihood is conditional, while the tests and confidence regions based on the likelihood ratio are fully conditional only if the choice of the critical value or cutoff point \( \beta \) is not based directly on pre-data evaluations (for example, it can be based on experience and on analogies with simple situations). Although they are (at least partially) conditional, in general the likelihood-based inference methods perform well also from the pre-data point of view. Under suitable regularity conditions, they are consistent and even asymptotically efficient. But they also have good pre-data performances in a surprisingly large number of important small sample problems, even though examples with bad performances can be easily constructed.

If we consider a set \( \Gamma \) of averaging probability measures \( \pi \) on \( \mathcal{P} \), as done at the end of Subsection 1.1.2, then we obtain a set \( \Pi \) of probability measures \( P_\pi \) on \( \mathcal{P} \times \Omega \). The observation of an event \( A \in \mathcal{A} \) induces a
likelihood function $lik$ on $\Gamma$, in the sense that the observation $\mathcal{P} \times A$ induces a likelihood function on $\Pi$, and the mapping $P_\pi \mapsto \pi$ is a parametrization of $\Pi$. The likelihood $lik(\pi)$ of $\pi$ is the expected value of $P(A)$ with respect to $\pi$: that is, it is the average with respect to $\pi$ of the likelihood function on $\mathcal{P}$ induced by the observation of $A$. The likelihood-based inference methods can be applied to $lik$ in order to obtain inferences about the averaging probability measures $\pi$ in $\Gamma$, as proposed for example by Good (1965).

1.3 Likelihood-Based Statistical Decisions

Since the statistical decision problems were introduced as a generalization of the problems of statistical estimation and testing hypotheses, and the most appreciated general methods for these inference problems are based directly on the likelihood function, it is natural to consider general decision criteria based directly on the likelihood function. Of particular interest are decision criteria leading to maximum likelihood estimates and likelihood ratio tests when applied to (some standard form of) estimation and hypothesis testing problems, respectively.

1.3.1 MPL Criterion

Consider a statistical decision problem described by a loss function $L$ on $\mathcal{P} \times \mathcal{D}$: let $A$ represent the observed data ($\mathcal{P}$ is a set of probability measures on a measurable space $(\Omega, \mathcal{A})$, and $A \in \mathcal{A}$), and let $lik$ be the likelihood function on $\mathcal{P}$ induced by the observation of $A$. If we use the likelihood of the statistical models to weight the respective losses, before applying the minimax criterion, then we obtain the MPL criterion:

$$\text{minimize } \sup_{\pi \in \mathcal{P}} lik l_d.$$  

The MPL criterion was introduced by Cattaneo (2005): “MPL” means “Minimax Plausibility-weighted Loss”, and for the present situation “plausibility” can simply be read as “likelihood”. A likelihood function with constant value $c \in \mathbb{P}$ is equivalent to the one induced by the observation of $\Omega$, which corresponds to having observed no data; with such a likelihood function, the MPL criterion reduces to the minimax criterion. In general, the likelihood function measures the relative plausibility of the models (in
the light of the observed data alone), and the MPL criterion is a simple, intuitive modification of the minimax criterion, making use of the information provided by the likelihood function: the more plausible a model, the larger its influence on the evaluation of the different decisions. In the extreme case of a likelihood function proportional to $I_P$, for a model $P \in \mathcal{P}$ (that is, $P$ is infinitely more plausible than every other model in $\mathcal{P}$), a decision $d \in \mathcal{D}$ is optimal according to the MPL criterion when it minimizes $L(P, d)$: in particular, if $d$ is correct for $P$, then $d$ is optimal.

**Example 1.8.** The MPL criterion applied to the estimation problem of Example 1.1 leads to the (unique) estimate

$$\arg \min_{d \in [0,1]} \max_{p \in [0,1]} p^X (1 - p)^{n-X} (p - d)^2.$$ 

Except for the simplest cases (with very small $n$, or with $X = \frac{n}{2}$), it is not possible to express this estimate in a simple form, but its numerical evaluation poses no problems. When $n = 0$ or $n = 1$, this estimate coincides with the one obtained by applying the minimax risk criterion (considered in Example 1.3); but as $n$ increases, the present estimate tends to the maximum likelihood estimate $\hat{p}_{ML} = \frac{X}{n}$ (considered in Example 1.7) faster than the one obtained by applying the minimax risk criterion. Both estimates obtained by applying the Bayesian criterion with as initial averaging probability measures $\pi$ on $\mathcal{P}$ the two popular choices considered at the end of Example 1.4 have a behavior (for increasing $n$) similar to the one of the present estimate (which coincides with both of them when $n = 0$, and with the one based on the proposal of Jeffreys when $n = 1$). Figure 1.3 shows, for various $n$, the graphs of the expected losses (as functions of $p \in [0,1]$) for the estimates obtained by applying the MPL criterion (solid line), the minimax risk criterion (dotted horizontal line), the method of maximum likelihood (dotted curved line), and the Bayesian criterion for the two choices of $\pi$ considered above (dashed lines: when $n \geq 1$ and $p = \frac{1}{2}$, the resulting expected loss is larger for the proposal of Jeffreys than for the one of Bayes).

Let $\mathcal{G}$ be a set, and let $g : \mathcal{P} \rightarrow \mathcal{G}$ be a mapping. If the loss function $L$ depends on the models $P$ only through $g(P)$, in the sense that there is a function $L'$ on $\mathcal{G} \times \mathcal{D}$ such that $L(P, d) = L'[g(P), d]$ for all $P \in \mathcal{P}$ and all $d \in \mathcal{D}$, then the MPL criterion can be expressed as follows:

$$\text{minimize } \sup_{g \in \mathcal{G}} \inf_{d \in \mathcal{D}} L'_d,$$
where for each \( d \in D \) the function \( l_d' \) on \( \mathcal{G} \) is defined by \( l_d'(\gamma) = L'(\gamma, d) \) (for all \( \gamma \in \mathcal{G} \)). That is, the MPL criterion automatically employs the profile likelihood function \( \text{lik}_g \) on \( \mathcal{G} \); but other pseudo likelihood functions on \( \mathcal{G} \) (with respect to \( g \)) can be used instead of \( \text{lik}_g \) in the above version of the MPL criterion, if they are expected to give better results.

When \( \mathcal{G} \) is finite and we want to estimate \( g(P) \), a simple way to define an estimation error is to assign a constant error to estimates \( \gamma \neq g(P) \), while assigning no error to the correct estimates \( \gamma = g(P) \). That is, the estimation problem can be described by the loss function \( I_w \) on \( \mathcal{P} \times \mathcal{G} \),
where $W = \{(P, \gamma) \in \mathcal{P} \times \mathcal{G} : g(P) \neq \gamma\}$. If $lik_g$ has a unique maximum at $\gamma = \hat{\gamma}_{ML}$, then the maximum likelihood estimate $\hat{\gamma}_{ML}$ of $g(P)$ is the unique optimal decision, according to the MPL criterion, for the decision problem described by $I_W$. This is in general not true when $\mathcal{G}$ is infinite, and in fact in this case the loss function $I_W$ is usually unreasonable. When the (finite or infinite) set $\mathcal{G}$ possesses a metric $\rho$, we can generalize the finite case by considering the loss function $I_{W_\varepsilon}$ on $\mathcal{P} \times \mathcal{G}$, where $\varepsilon \in \mathbb{P}$, and $W_\varepsilon = \{(P, \gamma) \in \mathcal{P} \times \mathcal{G} : \rho[g(P), \gamma] > \varepsilon\}$. When a maximum likelihood estimate $\hat{\gamma}_{ML}$ of $g(P)$ exists, it can be considered really unique only if $LR\{\rho(g, \hat{\gamma}_{ML}) > \delta\} < 1$ for all $\delta \in \mathbb{P}$. In this case, if $\gamma$ is an optimal decision, according to the MPL criterion, for the decision problem described by $I_{W_\varepsilon}$, then $\rho(\gamma, \hat{\gamma}_{ML}) \leq \varepsilon$. Thus by selecting a sufficiently small $\varepsilon \in \mathbb{P}$, we obtain a situation which in practice is analogous to the finite case: if a unique maximum likelihood estimate exists, then it corresponds practically to the unique optimal decision (according to the MPL criterion) for the estimation problem.

Consider now the problem of testing the null hypothesis $H_0 : P \in \mathcal{H}$ versus the alternative $H_1 : P \in \mathcal{P} \setminus \mathcal{H}$, for a subset $\mathcal{H}$ of $\mathcal{P}$: a simple way to define a loss function is to assign constant losses $c_1, c_2 \in \mathbb{P}$ to errors of the first and of the second kind, respectively, where $c_1 \geq c_2$. That is, the hypothesis testing problem can be described by the loss function $L$ on $\mathcal{P} \times \{r, n\}$ such that $l_r = c_1 I_\mathcal{H}$ and $l_n = c_2 I_{\mathcal{P} \setminus \mathcal{H}}$, where $r$ and $n$ are respectively the decisions of rejecting and of not rejecting the null hypothesis. The MPL criterion applied to the decision problem described by $L$ leads to the rejection of the null hypothesis when $LR(\mathcal{H}) < \frac{c_2}{c_1}$ (when $LR(\mathcal{H}) = \frac{c_2}{c_1}$, both decisions $r$ and $n$ are optimal); that is, practically it leads to the likelihood ratio test with critical value $\frac{c_2}{c_1}$. The level of this test depends on the sets $\mathcal{H}$ and $\mathcal{P}$ (this is the calibration problem of likelihood-based inference): it can be argued that those sets should have been considered when selecting the constants $c_1$ and $c_2$; another way to address the calibration problem is by using some pseudo likelihood function depending on the sets $\mathcal{H}$ and $\mathcal{P}$. The choice of a loss function to describe the problem of selecting a confidence region for $g(P)$ is complicated by the fact that also the extent of the region plays a role; but a simple loss function can be chosen if we consider only a limited set $\mathcal{D} \subset 2^\mathcal{G}$ of possible confidence regions (for example all the intervals whose length does not exceed a particular constant). In this case we can assign a constant error to regions $d \not\supset g(P)$, while assigning no error to regions $d \supset g(P)$; that is,
1.3 Likelihood-Based Statistical Decisions

the problem of selecting a confidence region can be described by the loss function \( L \) on \( P \times D \) such that \( l_d = I_{\{g \notin d\}} \) for all \( d \in D \). The MPL criterion applied to the decision problem described by \( L \) consists in minimizing \( LR\{g \notin d\} \); and for each \( \beta \in (0,1) \) the smallest region \( r \subseteq \mathcal{G} \) satisfying \( LR\{g \notin r\} = \beta \) is the likelihood-based confidence region for \( g(P) \) with cutoff point \( \beta \). That is, if \( D \) is sufficiently wide, then the MPL criterion leads to the selection of a likelihood-based confidence region for \( g(P) \), in the sense that it is the smallest region among the optimal ones.

The MPL criterion leads thus to the usual likelihood-based inference methods, when applied to some standard form of the corresponding decision problems. As the likelihood-based inferences, under suitable regularity conditions the decisions selected on the basis of the MPL criterion are asymptotically optimal and even asymptotically efficient (see Subsections 4.2.2 and 5.1.1). However, asymptotic properties have no practical meaning if they are not supported by some experience with finite sample problems, telling us how many data are necessary to approximate the asymptotic results. In fact, the appreciation of the likelihood-based inference methods rests on the positive experience with a myriad of finite sample problems, and on the property of being general, simple, and intuitive. This property is shared by the MPL criterion: it implies a wide applicability of the methods, and is related with the property of being based directly on the likelihood function (which implies in particular conditionality, parametrization invariance, and dependence only on sufficient statistics).

Besides the analogy with the usual likelihood-based inference methods, there are other interesting analogies concerning the MPL criterion. Consider a decision problem described by a loss function \( L \) on \( P \times D \), and a random object \( X : \Omega \rightarrow \mathcal{X} \); and assume for simplicity that the sets \( P, D, \) and \( \mathcal{X} \) are finite. After having observed \( X = x \), the decision \( \delta(x) \in D \) is optimal according to the MPL criterion if it minimizes

\[
\max_{P \in \mathcal{P}} P\{X = x\} L[P, \delta(x)].
\]

If we start with the uniform distribution on \( \mathcal{P} \) as the averaging probability measure of the Bayesian approach, then after having observed \( X = x \), the decision \( \delta(x) \in D \) is optimal according to the Bayesian criterion if it minimizes

\[
\sum_{P \in \mathcal{P}} P\{X = x\} L[P, \delta(x)].
\]
The decision function $\delta \in \mathcal{D}^X$ is optimal according to the minimax risk criterion if it minimizes

$$\max_{\mathcal{P} \in \mathcal{P}} \sum_{x \in \mathcal{X}} P\{X = x\} L[P, \delta(x)].$$

There is a high similarity between these three expressions, although the term $P\{X = x\}$ appears with different meanings. It can be noted that the MPL criterion corresponds to the minimax risk criterion by assuming that for all the alternative, unobserved realizations of $X$ we would have selected a correct decision. If we apply the Bayesian approach (with the uniform distribution on $\mathcal{P}$ as the averaging probability measure) to the pre-data decision problem described by the risk function, then the decision function $\delta \in \mathcal{D}^X$ is optimal according to the Bayesian criterion if it minimizes

$$\sum_{\mathcal{P} \in \mathcal{P}} \sum_{x \in \mathcal{X}} P\{X = x\} L[P, \delta(x)].$$

Since the summation operators commute, we obtain the temporal coherence of the Bayesian approach. The maximization and summation operators do not commute, and in fact the MPL approach does not in general satisfy this temporal coherence: the MPL criterion for the pre-data decision problem described by the risk function corresponds to the minimax risk criterion (since the likelihood function induced by observing no data is constant). Since the maximization operators commute, a property analogous to the temporal coherence of the Bayesian approach is obtained for the MPL approach if

$$\max_{x \in \mathcal{X}} P\{X = x\} L[P, \delta(x)]$$

is used (instead of the expected value) as the representative value of the random loss of $\delta \in \mathcal{D}^X$, when defining the loss function $L'$ on $\mathcal{P} \times \mathcal{D}^X$ for the pre-data decision problem considered in Subsection 1.1.1. This representative value has the drawback of depending on spurious modifications of the random object $X$ (for example when a possible observation $x \in \mathcal{X}$ is “subdivided” into two equiprobable new observations by considering also the result of flipping a fair coin), but the decision functions obtained by conditional application of the MPL criterion are the only decision functions that are optimal for all such modifications (according to the MPL criterion applied to the decision problem described by $L'$).
Besides temporal coherence, the Bayesian approach possesses also the property of coherence with respect to additive loss: the MPL approach possesses the analogous, important property of “coherence with respect to maxitive loss”. Let $L_1$ and $L_2$ be two loss functions on $\mathcal{P} \times \mathcal{D}_1$ and $\mathcal{P} \times \mathcal{D}_2$, respectively, let $\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2$, and let $L$ be the loss function on $\mathcal{P} \times \mathcal{D}$ defined by $L(P,(d_1,d_2)) = \max\{L_1(P,d_1), L_2(P,d_2)\}$ (for all $P \in \mathcal{P}$ and all $(d_1,d_2) \in \mathcal{D}$). If $d_1$ and $d_2$ are optimal for the decision problems described by $L_1$ and $L_2$, respectively, then $(d_1,d_2)$ is optimal for the decision problem described by $L$. The coherence with respect to maxitive loss is useful because for some decision problems it allows us to construct an optimal decision by considering only simpler problems.

Consider a statistical decision problem described by a loss function $L$ on $\mathcal{P} \times \mathcal{D}$: let $A$ represent the observed data, and let $lik$ be the likelihood function on $\mathcal{P}$ induced by the observation of $A$. If we consider a set $\Gamma$ of averaging probability measures $\pi$ on $\mathcal{P}$, then the MPL criterion for the decision problem described by the loss function $L''$ considered at the end of Subsection 1.1.2 can be expressed as follows:

$$\min \sup_{\pi \in \Gamma} E_{\pi}(lik l_d).$$

In fact, it can be easily proved that for all $\pi \in \Gamma$ and all $d \in \mathcal{D}$

$$P_{\pi}(\mathcal{P} \times A) E_{P_{\pi}}[L(P,d) | \mathcal{P} \times A] = E_{\pi}(lik l_d).$$

Since $E_{\pi}(lik l_d) \leq \sup lik l_d$, if $\Gamma$ can be interpreted as an extension of $\mathcal{P}$, in the sense that $\Gamma$ contains, at least as limits, the Dirac measures $\delta_{\pi}$ on $\mathcal{P}$, for all $P \in \mathcal{P}$, then the application of the MPL criterion to the decision problem described by $L''$ leads to the same results as its application to the decision problem described by $L$. More precisely, if for each $P \in \mathcal{P}$ and each $d \in \mathcal{D}$ there is a sequence $\pi_1, \pi_2, \ldots \in \Gamma$ such that

$$\lim_{n \to \infty} E_{\pi_n}(lik l_d) = lik(P) l_d(P),$$

then according to the MPL criterion, a decision $d \in \mathcal{D}$ is optimal for the decision problem described by $L''$ if and only if it is optimal for the decision problem described by $L$. This property too can be considered as a sort of coherence; its premise is satisfied in particular when $\Gamma$ is the set of all probability measures on $(\mathcal{P}, \mathcal{C})$, and $\mathcal{C}$ contains all singletons of $\mathcal{P}$. 
Example 1.9. If for the estimation problem of Example 1.1 we consider the family \( \Gamma \) of averaging probability measures on \( \mathcal{P} \) corresponding to the family of beta distributions for the parameter \( p \in [0, 1] \), as done in Example 1.5, then the MPL criterion applied to the decision problem described by \( L'' \) leads to the estimate obtained in Example 1.8 without considering the family \( \Gamma \) of averaging probability measures, since \( \Gamma \) contains as limits the Dirac measures on \( \mathcal{P} \). Hence, unlike the \( \Gamma \)-minimax criterion, the MPL criterion can be useful even when \( \Gamma \) is wide enough to be interpreted as an extension of \( \mathcal{P} \) (see also Subsection 3.2.2).

1.3.2 Other Likelihood-Based Decision Criteria

Consider a statistical decision problem described by a loss function \( L \) on \( \mathcal{P} \times \mathcal{D} \), and let \( lik \) be the likelihood function on \( \mathcal{P} \) induced by the observed data. When a (unique) maximum likelihood estimate \( \hat{P}_{ML} \) of \( P \) exists, a likelihood-based decision criterion that is often used in practice (even though perhaps it was never stated in its full generality) is the following one: simply replace \( P \) with its maximum likelihood estimate \( \hat{P}_{ML} \) in the loss function \( L \), and thus minimize \( L(\hat{P}_{ML}, d) \). The asymptotic properties of this criterion (and of all reasonable likelihood-based decision criteria) are similar to those of the MPL one; and since it reduces the set \( \mathcal{P} \) to a single model \( \hat{P}_{ML} \), the present criterion possesses important coherence properties (such as the coherence with respect to additive or maxitive loss), but not the temporal coherence, because the model \( \hat{P}_{ML} \) depends on the data. The above definition of this criterion is not general (since it depends on the existence of a maximum likelihood estimate of \( P \)): we will consider the general definition of an almost equivalent criterion in a moment.

The likelihood function measures the relative plausibility of the models (in the light of the observed data alone), and a simple way to use it in a decision criterion is to reduce the set \( \mathcal{P} \) of considered statistical models by excluding the less plausible ones. If we reduce \( \mathcal{P} \) to the likelihood-based confidence region for \( P \) with cutoff point \( \beta \in (0, 1) \), before applying the minimax criterion, then we obtain the \( LR_{\beta} \) criterion:

\[
\text{minimize } \sup_{\{lik > \beta \sup lik\}} \{L \}
\]

It is possible that this criterion was never stated in such generality before: "LRM" means "Likelihood-based Region Minimax". The \( LR_{\beta} \) criterion is thus simply the minimax criterion applied after having discarded the
models in \( \mathcal{P} \) that are too implausible (when compared to other models in \( \mathcal{P} \)) in the light of the observed data: the larger is \( \beta \), the more plausible are the discarded models. By letting \( \beta \) tend to 1, we obtain the **MLD criterion**:

\[
\text{minimize } \limsup_{\beta \uparrow 1} \{ \sup_{\{ l_d \} \sup_{\beta \uparrow \beta} \} l_d \}.
\]

This criterion is almost equivalent to the one considered above, based on the maximum likelihood estimate \( \hat{P}_{ML} \) of \( P \): “MLD” means “Maximum Likelihood Decision”. In fact, if there is a topology on \( \mathcal{P} \) such that \( l_d \) is continuous, and \( LR(\mathcal{P} \setminus \mathcal{N}) < 1 \) for all neighborhoods \( \mathcal{N} \) of \( \hat{P}_{ML} \), then

\[
\limsup_{\beta \uparrow 1} \{ \sup_{\{ l_d \} \sup_{\beta \uparrow \beta} \} l_d \} = L(\hat{P}_{ML}, d).
\]

If such a topology does not exist, then the use of \( L(\hat{P}_{ML}, d) \) to evaluate the decision \( d \) does not seem very reasonable: we can say that the MLD criterion corresponds to the one considered at the beginning of the present subsection, in all the cases in which this is well-defined and reasonable.

When applied to an estimation problem, the MLD criterion leads to the maximum likelihood estimate if some weak conditions (such as the above existence of a topology) are satisfied, while the \( LRM_\beta \) criterion does not in general lead to the maximum likelihood estimate even in the simple problem described by the loss function \( I_W \) considered in Subsection 1.3.1.

When applied to the hypothesis testing problem considered in Subsection 1.3.1 (with \( c_1 > c_2 \)), the MLD criterion leads to the rejection of the null hypothesis \( H_0 : P \in \mathcal{H} \) if and only if \( LR(\mathcal{H}) < 1 \), while the \( LRM_\beta \) criterion leads to the rejection of \( H_0 \) if and only if \( LR(\mathcal{H}) \leq \beta \). That is, the MLD criterion leads to an unreasonable test, while the \( LRM_\beta \) criterion leads to the likelihood ratio test with critical value \( \beta \) (independently of \( \frac{c_2}{c_1} \)). Analogously, for the problem of selecting a confidence region considered in Subsection 1.3.1, a region \( d \in \mathcal{D} \) is optimal according to the MLD criterion if \( LR\{ g \notin d \} < 1 \), while it is optimal according to the \( LRM_\beta \) criterion if \( LR\{ g \notin d \} \leq \beta \). In conclusion, both these criteria use the likelihood in a very limited way: the MLD criterion divides the values of the likelihood ratio into two categories: equal 1 or not, while the \( LRM_\beta \) criterion divides them into the two categories: larger than \( \beta \) or not. The \( LRM_\beta \) criterion has the advantage of allowing some control through the choice of \( \beta \), but on the other hand this choice is complicated by the calibration problem of likelihood-based inference.
Example 1.10. Consider the estimation problem of Example 1.1: the estimate obtained by applying the LRM\(_\beta\) criterion is the midpoint \(\hat{p}_\beta\) of the interval corresponding to the likelihood-based confidence region for \(p\) with cutoff point \(\beta\). The case with \(n = 0\) corresponds to the case without observations (the midpoint \(\frac{1}{2}\) of the interval \([0,1]\) is the estimate obtained in Example 1.2), while as \(n\) increases, the estimate \(\hat{p}_\beta\) tends to the maximum likelihood estimate \(\hat{p}_{ML} = \frac{X}{n}\) (considered in Example 1.7), if \(\beta\) is held constant. But note that \(\lim_{\beta \to 0} \hat{p}_\beta = \frac{1}{2}\), independently of \(n\) and of the observation \(X = x\). The MLD criterion applied to this estimation problem leads to the estimate \(\lim_{\beta \to 1} \hat{p}_\beta\). When \(n \geq 1\), this is the maximum likelihood estimate \(\hat{p}_{ML} = \frac{X}{n}\), while for \(n = 0\) the maximum likelihood estimate is undefined, and the MLD criterion leads to the estimate \(\hat{p}_\beta = \frac{1}{2}\).

It is important to note that the estimates obtained by applying the LRM\(_\beta\) and MLD criteria would be the same for all symmetric, strictly increasing estimation errors; that is, for all loss functions \(L : (P, d) \mapsto f(|p - d|)\) on \(\mathcal{P} \times [0,1]\), where \(f : [0,1] \to \mathbb{P}\) is strictly increasing. This is not true (when \(n \geq 1\)) for the estimates obtained by applying the minimax risk, Bayesian, and MPL criteria considered in Examples 1.3, 1.4, and 1.8, respectively.

If we consider a set \(\Gamma\) of averaging probability measures \(\pi\) on \(\mathcal{P}\), then applying the MLD criterion to the decision problem described by the loss function \(L''\) considered at the end of Subsection 1.1.2 usually means applying the Bayesian criterion (to the decision problem described by \(L\)) with as averaging probability measure on \(\mathcal{P}\) the maximum likelihood estimate \(\hat{\pi}_{ML}\) (obtained from the likelihood function on \(\Gamma\)). As stated at the end of Subsection 1.2.1, this way of selecting \(\pi\) was proposed for example by Good (1965); but also most instances of the empirical Bayes approach introduced by Robbins (1951) can be considered as applications of the MLD criterion (to the decision problem described by \(L''\), after having selected a suitable set \(\Gamma\) of averaging probability measures). DasGupta and Studden (1989) applied a criterion similar to the LRM\(_\beta\) one to the decision problem (described by \(L''\)) obtained from a particular estimation problem (described by \(L\)) by using a particular set \(\Gamma\) of averaging probability measures.

The choice of the way of using the likelihood function \(lik\) to weight the function \(l_d\) in the MPL criterion is intuitive, but arbitrary. In particular, reasonable alternative criteria can be easily obtained by transforming \(lik\) before multiplying it with \(l_d\), or by changing the way in which (the transformed version of) \(lik\) and \(l_d\) are combined. For example, the LRM\(_\beta\)
1.3 Likelihood-Based Statistical Decisions

criterion can be obtained by transforming \( lik \) into \( I_{\beta \sup \{lik, \infty\}} \circ lik \) before multiplying it with \( l_d \) (the MLD criterion can also be obtained, but in a slightly more complicated way), while an alternative way of combining \( lik \) and \( l_d \) was proposed by Cattaneo (2005) under the name of “MPL* criterion”. Such modifications of the MPL criterion are special cases of the general definition of likelihood-based decision criterion that will be used in Chapters 4 and 5. The main advantage of the way in which the MPL criterion combines \( lik \) and \( l_d \) is its simplicity (and thus its applicability in difficult problems; see also Subsection 4.1.2 and Section 5.2), and the best reasons for not transforming \( lik \) before multiplying it with \( l_d \) are the analogies and the coherence properties considered at the end of Subsection 1.3.1 (some transformed versions of \( lik \) can simply be interpreted as pseudo likelihood functions).

Lehmann (1959, Section 1.7, which remained substantially unchanged in the following two editions of the book) proposed a likelihood-based decision criterion for decision problems that involve at most countably many possible decisions, and that are formulated as follows: for each model \( P \in \mathcal{P} \) there is exactly one correct decision \( \psi(P) \in \mathcal{D} \), which would result in a positive gain \( G(P) \), while every other decision in \( \mathcal{D} \) would yield no gain (\( \psi \) and \( G \) are functions on \( \mathcal{P} \)). The proposed criterion consists in selecting the decision \( \psi(P') \), where \( P' \) is the \( P \in \mathcal{P} \) maximizing \( likG \), when such a \( P \) exists and is unique. That is, this criterion modifies the one based on the maximum likelihood estimate \( P_{ML} \) of \( P \) (considered at the beginning of the present subsection) by using the possible gain \( G \) to weight the likelihood function \( lik \) before maximizing it (these two criteria correspond if \( G \) is constant). The above decision problem can be described by the loss function \( L \) on \( \mathcal{P} \times \mathcal{D} \) such that \( l_d = I_{\{\psi \neq d\}} G \) for all \( d \in \mathcal{D} \); then \( L(P, d) \) is the additional gain that would have resulted from the selection of the correct decision \( \psi(P) \) instead of \( d \) (in particular, \( L[P, \psi(P)] = 0 \) for all \( P \in \mathcal{P} \)). When a \( P' \) maximizing \( likG \) exists, it can be considered really unique, and the decision \( d' = \psi(P') \) clearly determined, only if \( \sup_{\psi \neq d'} likG < lik(P')G(P') \) (this condition could also be formulated as the existence of a particular topology or metric, in analogy with other conditions considered above). In this case, \( d' \) is the unique optimal decision, according to the MPL criterion, for the decision problem described by \( L \). That is, the criterion based on \( P' \) can be considered as a special case of the MPL criterion, applicable to the decision problems that involve at most countably many possible decisions, and that can be formulated as above in terms of the functions \( \psi \) and \( G \) on \( \mathcal{P} \).
Giang and Shenoy (2002) proposed a likelihood-based decision criterion for decision problems that are described by a binary utility function $U : \mathcal{P} \times \mathcal{D} \to \mathcal{U}$, where $\mathcal{U} = \{(a, b) \in [0, 1]^2 : \max\{a, b\} = 1\}$ is the set of possible values for the binary utility: if $c > c'$, then $(c, 1)$ is better than $(c', 1)$, while $(1, c)$ is worse than $(1, c')$. That is, $(0, 1)$ is the worst value for the binary utility, $(1, 0)$ is the best one, while $(1, 1)$ is an intermediate value. Each decision $d \in \mathcal{D}$ is evaluated by the binary utility $(\sup \text{lik} u_1, \sup \text{lik} u_2)$, where $u_1$ and $u_2$ are the functions on $\mathcal{P}$ such that $U(P, d) = (u_1(P) \sup \text{lik}, u_2(P) \sup \text{lik})$ for all $P \in \mathcal{P}$; the criterion consists in selecting the decision with the best evaluation in terms of binary utility. If a maximum likelihood estimate $\hat{P}_{ML}$ of $P$ exists, then the evaluation of a decision $d$ lies somewhere (on the binary utility scale) between $U(\hat{P}_{ML}, d)$ and $(1, 1)$; that is, the intermediate value $(1, 1)$ plays a peculiar role in the evaluation of decisions (see also the remarks at the end of Subsection 4.1.2). This criterion is based on a modification of the axiomatic approach to qualitative decision making with possibility theory studied in Giang and Shenoy (2001); a direct application of the criterion to a statistical decision problem described by a loss function $L$ on $\mathcal{P} \times \mathcal{D}$ is not possible, because the values of $L$ must first be translated into (subjective) binary utility values. If the loss function $L$ is bounded above by $c \in \mathbb{R}$, then by defining $U = (1, \frac{L}{c})$, we obtain that the present criterion corresponds to the MPL one (independently of the choice of the upper bound $c$). However, this definition of $U$ does not really conform with the above definition of $\mathcal{U}$, since $\sup L$ should in fact be translated into $(0, 1)$, but then we would be faced with the problematic choice of a loss value to be translated into $(1, 1)$. 
This chapter is an introduction to nonadditive measures and integrals, which will be used in the following chapters. The literature on this topic is extensive and heterogeneous, ranging from pure mathematics to decision theory and artificial intelligence (a partial survey is given by Grabisch, Murofushi, and Sugeno, 2000). In particular, many concepts appear in the literature under different names: the names used in this chapter have been selected in order to avoid possible confusion. The present introduction focuses on maxitive measures and regular integrals, because these topics are particularly important for the following chapters. Many definitions are new (for example the one of regular integral), and as a consequence many results are new too, but in general they are rather simple.

2.1 Nonadditive Measures

Usually, a measure is a nonnegative, extended real-valued set function satisfying countable additivity; while for nonadditive measures the requirement of additivity is dropped. Thus, in general a nonadditive measure is not a measure, and an additive measure is a nonadditive measure. To avoid such contradictions, we will use the term “measure” to denote the general concept of nonnegative, extended real-valued set function, which can then be specialized by attributes such as “countably additive”.

Nonadditive Measures and Integrals
2 Nonadditive Measures and Integrals

2.1.1 Types of Measures

A measure on a set $\Omega$ is a function

$$\mu : 2^\Omega \to \mathbb{R}$$

such that $\mu(\emptyset) = 0$.

This definition could be generalized by allowing $\mu$ to be defined only on some subset of $2^\Omega$, but this generality will not be really needed in the following. The requirement $\mu(\emptyset) = 0$ is not very restrictive, since we will consider monotone measures only. A measure $\mu$ on $\Omega$ is said to be monotone if

$$A \subseteq B \Rightarrow \mu(A) \leq \mu(B) \quad \text{for all } A, B \subseteq \Omega.$$

In this case, if $\mu(\emptyset) = 0$ were not required, we would have $\mu(\emptyset) = \min \mu$, and thus $\mu' = \mu - \mu(\emptyset)$ would be a monotone measure satisfying $\mu'(\emptyset) = 0$ (except for the uninteresting case with $\mu = \infty$). That is, the requirement $\mu(\emptyset) = 0$ simply spares us the subtraction of $\mu(\emptyset)$. If $\mu$ is monotone, then $\mu(\emptyset) = \max \mu$, and therefore $\mu$ is bounded when it is finite. A monotone measure $\mu$ on $\Omega$ is said to be nonzero if $\mu(\emptyset) > 0$; otherwise it is the zero measure, which has constant value 0. A monotone measure $\mu$ on $\Omega$ is said to be normalized if $\mu(\emptyset) = 1$.

A measure $\mu$ on $\Omega$ is said to be subadditive if

$$\mu(A \cup B) \leq \mu(A) + \mu(B) \quad \text{for all disjoint } A, B \subseteq \Omega.$$

A monotone measure is not necessarily subadditive, and a subadditive measure is not necessarily monotone. When combined, monotonicity and subadditivity have important consequences: in particular, it can be easily proved that if $\mu$ is a monotone, subadditive measure on $\Omega$, and $A \subseteq \Omega$ is a set such that $\mu(A) = 0$, then

$$\mu(B) = \mu(B \setminus A) \quad \text{for all } B \subseteq \Omega.$$

Thanks to this property, it makes sense to consider a statement whose truth depends on the elements of $\Omega$ as true almost everywhere with respect to $\mu$ (written a.e. $[\mu]$) when it is false only on a set $A \subseteq \Omega$ such that $\mu(A) = 0$.

A measure $\mu$ on $\Omega$ is said to be 2-alternating if it is monotone and

$$\mu(A \cup B) + \mu(A \cap B) \leq \mu(A) + \mu(B) \quad \text{for all } A, B \subseteq \Omega.$$
A 2-alternating measure is thus monotonic and subadditive, and has other important properties: see for instance Choquet (1954) and Huber and Strassen (1973).

If a measure $\mu$ on $Q$ is monotonic and subadditive, then
$$\max\{\mu(A), \mu(B)\} \leq \mu(A \cup B) \leq \mu(A) + \mu(B)$$
for all disjoint $A, B \subseteq Q$.

In this sense, at one extreme of the class of monotonic, subadditive measures we have the finitely additive ones, satisfying
$$\mu(A \cup B) = \mu(A) + \mu(B)$$
for all disjoint $A, B \subseteq Q$.

while at the other extreme we have the finitely maxitive ones. A measure $\mu$ on $Q$ is said to be **finitely maxitive** if
$$\mu(A \cup B) = \max\{\mu(A), \mu(B)\}$$
for all disjoint $A, B \subseteq Q$.

It can be easily proved that a finitely maxitive measure $\mu$ on $Q$ is 2-alternating and satisfies
$$\mu\left(\bigcup A\right) = \max_A \mu$$
for all finite, nonempty $A \subseteq 2^Q$.

Unlike additivity, this property can be extended without problems to all sets $A$ (not only to the countable ones). A measure $\mu$ on $Q$ is said to be **completely maxitive** if
$$\mu\left(\bigcup A\right) = \sup_A \mu$$
for all $A \subseteq 2^Q$.

In this case, $\mu$ is uniquely determined by its **density function** $\mu^\downarrow$ on $Q$, defined by $\mu^\downarrow(q) = \mu\{q\}$ (for all $q \in Q$); in fact,
$$\mu(A) = \sup_A \mu^\downarrow$$
for all $A \subseteq Q$.

Density functions do not satisfy particular requirements: each nonnegative, extended real-valued function on $Q$ is the density function of a completely maxitive measure on $Q$. Maxitive measures were introduced by Shilkret (1971), who focused on countable maxitivity in analogy with additive measures.

**Example 2.1.** The measure $\mu$ on $[0, 1]$ defined by $\mu(A) = \sup A$ (for all $A \subseteq [0, 1]$) is normalized and completely maxitive, with density function $\mu^\downarrow = id_{[0,1]}$. \(\Diamond\)
A possibility measure on a set $\mathcal{Q}$ is a completely maxitive measure $\mu$ on $\mathcal{Q}$ such that $\mu(\mathcal{Q}) \leq 1$. The density function of a possibility measure is often called “possibility distribution function”, but we will use the expression “distribution function” with another meaning. Possibility measures were introduced by Zadeh (1978) in the context of the theory of fuzzy sets. Other authors use different definitions: for example, for Dubois and Prade (1988) a possibility measure is a normalized, finitely maxitive measure.

A monotonic measure $\mu$ on $\mathcal{Q}$ is said to be continuous from below if

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} \mu(A_n)$$

for all sequences $A_1 \subseteq A_2 \subseteq \ldots \subseteq \mathcal{Q}$.

A completely maxitive measure is continuous from below, while this is in general not true if the maxitivity is only finite. It can be easily proved that, as in the case of additivity, a finitely maxitive measure is continuous from below if and only if it is countably maxitive. But even complete maxitivity is not enough to assure continuity from above.

### 2.1.2 Derived Measures

The dual of a finite, monotonic measure $\mu$ on $\mathcal{Q}$ is the finite, monotonic measure $\overline{\mu}$ on $\mathcal{Q}$ defined by

$$\overline{\mu}(A) = \mu(\mathcal{Q}) - \mu(\mathcal{Q} \setminus A)$$

for all $A \subseteq \mathcal{Q}$.

The equality $\overline{\mu} = \mu$ justifies the use of the term “dual”. If $\mu$ is subadditive, then $\overline{\mu} \leq \mu$, and therefore $\overline{\mu}$ is also subadditive only if $\overline{\mu} = \mu$; in this case $\mu$ is 2-alternating only if it is finitely additive.

It can be easily proved that if $\mathcal{Q}$ and $\mathcal{T}$ are sets, $\mu$ is a measure on $\mathcal{Q}$, and $t: \mathcal{Q} \to \mathcal{T}$ is a function, then $\mu \circ t^{-1}$ is a measure on $\mathcal{T}$; and if $\mu$ satisfies one of the properties considered in Subsection 2.1.1, then so does $\mu \circ t^{-1}$. A class $\mathcal{M}$ of measures (that is, each $\mu \in \mathcal{M}$ is a measure on some set $\mathcal{Q}_\mu$) is said to be closed under transformations if $\mu \circ t^{-1} \in \mathcal{M}$ for all $\mu \in \mathcal{M}$, all sets $\mathcal{T}$, and all functions $t: \mathcal{Q}_\mu \to \mathcal{T}$. For instance, the class of all normalized, completely maxitive measures is closed under transformations.
If \( \mu \) is a measure on \( \mathcal{Q} \), and \( \delta : \mathcal{P} \to \mathcal{P} \) is a nondecreasing function such that \( \delta(0) = 0 \), then \( \delta \circ \mu \) is a measure on \( \mathcal{Q} \). Among the possible properties of \( \mu \) considered in Subsection 2.1.1, the only ones that are certainly maintained by \( \delta \circ \mu \) are monotonicity and finite maxitivity; but if \( \delta \) is left-continuous, then also complete maxitivity and continuity from below are maintained. If \( c \in \mathcal{P} \), and \( \delta \) is the function \( x \mapsto cx \) on \( \mathcal{P} \), then \( \delta \circ \mu = c \mu \) maintains all the properties of \( \mu \), except for the properties of being normalized and of being a possibility measure. A class \( \mathcal{M} \) of measures is said to be regular if \( c \mu \in \mathcal{M} \) for all \( \mu \in \mathcal{M} \) and all \( c \in \mathcal{P} \), and \( \mathcal{M} \) is closed under transformations. For instance, the class of all finite, nonzero, completely maxitive measures is regular.

**Example 2.2.** Let \( \mu \) be the completely maxitive measure on \([0, 1]\) considered in Example 2.1, and let \( \delta \) be the function on \( \mathcal{P} \) defined by

\[
\delta(x) = \begin{cases} 
\frac{x}{2} & \text{if } 0 \leq x < 1, \\
x & \text{if } 1 \leq x \leq \infty.
\end{cases}
\]

The function \( \delta \) is not left-continuous at 1, and in fact the measure \( \nu = \delta \circ \mu \) on \([0, 1]\) is finitely maxitive, but not continuous from below (and thus not completely maxitive), since for example

\[
\lim_{x \uparrow 1} \nu([0, x)) = \lim_{x \uparrow 1} \delta(x) = \frac{1}{2} < 1 = \delta(1) = \nu([0, 1]).
\]

The restriction \( \mu|_A \) to \( A \subseteq \mathcal{Q} \) of the measure \( \mu \) on \( \mathcal{Q} \) is the measure on \( A \) defined by \( \mu|_A(B) = \mu(B) \) (for all \( B \subseteq A \)). Since the function \( \mu \) on \( 2^\mathcal{Q} \) is restricted to \( 2^A \), we should write \( \mu|_{2^A} \) instead of \( \mu|_A \), but the abuse of notation in writing \( \mu|_A \) is coherent with the abuse of terminology committed by saying that \( \mu \) is a measure on \( \mathcal{Q} \) when in fact it is a function on \( 2^\mathcal{Q} \). If \( \mu \) satisfies one of the properties considered in Subsection 2.1.1, then so does \( \mu|_A \), except for the properties of being normalized and of being nonzero. A class \( \mathcal{M} \) of measures (each \( \mu \in \mathcal{M} \) is a measure on some set \( \mathcal{Q}_\mu \)) is said to be **closed under restrictions** if \( \mu|_A \in \mathcal{M} \) for all \( \mu \in \mathcal{M} \) and all \( A \subseteq \mathcal{Q}_\mu \). For instance, the class of all finite, completely maxitive measures is closed under restrictions.
2.2 Nonadditive Integrals

As with nonadditive measures, to avoid contradictions, we will use the term "integral" to denote a very general concept, which can then be specialized by attributes such as "additive".

Let \( \mathcal{M} \) be a class of measures (each \( \mu \in \mathcal{M} \) is a measure on some set \( Q_\mu \)). An integral on \( \mathcal{M} \) is a function that associates a value \( \int f \, d\mu \in \mathbb{P} \) (the integral of \( f \) with respect to \( \mu \)) to each pair consisting of a measure \( \mu \in \mathcal{M} \) and of a function \( f : Q_\mu \to \mathbb{P} \), and that satisfies the following indicator property:

\[
\int I_A \, d\mu = \mu(A) \quad \text{for all } \mu \in \mathcal{M} \text{ and all } A \subseteq Q_\mu.
\]

That is, integrals can be seen as extensions of measures: from the indicator functions to all nonnegative, extended real-valued functions. The definition of integrals for functions taking also negative values would pose some annoying problems (since \( -\infty \) is undefined) and will not be really needed in the following.

Usually, a basic property of integrals is linearity, which consists of (finite) additivity:

\[
\int (f + g) \, d\mu = \int f \, d\mu + \int g \, d\mu \quad \text{for all } \mu \in \mathcal{M} \text{ and all } f, g \in \mathbb{P}^{Q_\mu};
\]

and (positive) homogeneity. An integral on \( \mathcal{M} \) is said to be (positively) homogeneous if

\[
\int c \, f \, d\mu = c \int f \, d\mu \quad \text{for all } \mu \in \mathcal{M}, \text{ all } c \in \mathbb{P}, \text{ and all } f \in \mathbb{P}^{Q_\mu}.
\]

Note that the case with \( c = 0 \) is already implied by the indicator property: \( 0 \, d\mu = 0 \), since \( \mu(\emptyset) = 0 \) by definition. When combined with the indicator property, homogeneity places no restrictions on the measures \( \mu \in \mathcal{M} \). By contrast, additivity implies that they are finitely additive, since \( I_{A \cup B} = I_A + I_B \) for all disjoint \( A, B \subseteq Q_\mu \). Thus, additivity cannot be required for an integral on nonadditive measures. But if the measures are at least monotonic, then the following fundamental consequence of ad-
ditivity can still be required. An integral on $\mathcal{M}$ is said to be **monotonic** if
\[ f \leq g \Rightarrow \int f \, d\mu \leq \int g \, d\mu \quad \text{for all } \mu \in \mathcal{M} \text{ and all } f, g \in \mathbb{P}^Q. \]

When combined with the indicator property, monotonicity implies that the measures $\mu \in \mathcal{M}$ are monotonic, since $I_A \leq I_B$ for all $A \subseteq B \subseteq Q_\mu$.

### 2.2.1 Shilkret Integral

The **Shilkret integral** associates the value
\[
\int_S f \, d\mu = \sup_{x \in \mathbb{P}} x \mu\{f \geq x\}
\]
to each pair consisting of a monotonic measure $\mu$ on a set $Q$ and of a function $f : Q \rightarrow \mathbb{P}$. It was introduced by Shilkret (1971) as the unique homogeneous, countably maxitive integral on countably maxitive measures. Before examining this aspect, we consider two other characterizations of the Shilkret integral.

**Theorem 2.3.** The Shilkret integral is a homogeneous, monotonic integral on the class of all monotonic measures.

If $\mu$ is a monotonic measure on a set $Q$, and $f : Q \rightarrow \mathbb{P}$ is a function, then
\[
\int_S f \, d\mu \leq \int f \, d\mu
\]
for all homogeneous, monotonic integrals on classes $\mathcal{M} \ni \mu$. That is, the Shilkret integral is the minimum of all homogeneous, monotonic integrals.

**Proof.** The indicator property for the Shilkret integral follows from
\[
\mu\{I_A \geq x\} = \begin{cases} 
\mu(A) & \text{if } 0 < x \leq 1, \\
0 & \text{if } 1 < x < \infty.
\end{cases}
\]

Since for $c \in \mathbb{P}$ we have $x \mu\{cf \geq x\} = cx' \mu\{f \geq x'\}$ with $x' = \frac{x}{c}$, the Shilkret integral is homogeneous. It is also monotonic, since $f \leq g$ implies $\{f \geq x\} \subseteq \{g \geq x\}$, and $\mu$ is monotonic.
For all \( x \in \mathbb{P} \) we have \( f \geq x \mathbb{I}_{\{f \geq x\}} \). Hence, a homogeneous, monotonic integral satisfies

\[
\int f \, d\mu \geq \int x \mathbb{I}_{\{f \geq x\}} \, d\mu = x \int \mathbb{I}_{\{f \geq x\}} \, d\mu = x \mu \{ f \geq x \}.
\]

The desired result is obtained by taking the supremum (over all \( x \in \mathbb{P} \)) of the right-hand side of the inequality. \( \Box \)

A well-known nonadditive integral was introduced by Sugeno (1974): the following generalization was considered for instance by Weber (1984), Mesiar (1995), and de Cooman (2000). A generalized Sugeno integral associates the value

\[
\int f \, d\mu = \sup_{x \in \mathbb{P}} m(x, \mu \{ f \geq x \})
\]

to each pair consisting of a monotonic measure \( \mu \) on a set \( \mathcal{Q} \) and of a function \( f : \mathcal{Q} \rightarrow \overline{\mathbb{P}} \), where \( m \) is a nonnegative, extended real-valued function on \( \mathbb{P} \times \overline{\mathbb{P}} \) that is nondecreasing in both arguments. In the Sugeno integral, the function \( m \) returns the minimum of its two arguments; if it returns the product, we obtain the Shilkret integral.

**Theorem 2.4.** A generalized Sugeno integral satisfies the indicator property if and only if the function \( m \) fulfills the following two conditions:

\[
m(x, 0) = 0 \quad \text{for all } x \in \mathbb{P}, \quad (2.2)
\]

\[
m(1, y) = y \quad \text{for all } y \in \overline{\mathbb{P}}. \quad (2.3)
\]

In this case it is a monotonic integral on the class of all monotonic measures.

It is also homogeneous if and only if the function \( m \) returns the product of its two arguments. That is, the Shilkret integral is the homogeneous version of the generalized Sugeno integral.

**Proof.** The indicator property implies \( \int 0 \, d\mu = 0 \), and thus also condition (2.2). If condition (2.2) holds, then from the equality (2.1) we obtain

\[
\int \mathbb{I}_A \, d\mu = m[1, \mu(A)] \quad (\text{since } m \text{ is nondecreasing in the first argument}),
\]

and therefore the indicator property is equivalent to condition (2.3). The monotonicity of the integral can be proved in the same way as for the Shilkret integral, since \( m \) is nondecreasing in the second argument.
Reasoning as above, we obtain \( \int c I_A \, d\mu = m[c, \mu(A)] \) for all \( c \in \mathbb{P} \). Hence, the homogeneity of the integral implies that \( m \) returns the product of its two arguments. The converse was proved in Theorem 2.3. \( \square \)

Let \( \mathcal{M} \) be a class of measures (each \( \mu \in \mathcal{M} \) is a measure on some set \( \mathcal{Q}_\mu \)). An integral on \( \mathcal{M} \) is said to be (finitely) **maxitive** if

\[
\int \max\{f, g\} \, d\mu = \max\left\{ \int f \, d\mu, \int g \, d\mu \right\} \quad \text{for all } \mu \in \mathcal{M} \text{ and all } f, g \in \mathbb{P}^{\mathcal{Q}_\mu}.
\]

When combined with the indicator property, maxitivity implies that the measures \( \mu \in \mathcal{M} \) are finitely maxitive, since \( I_{A \cup B} = \max\{I_A, I_B\} \) for all \( A, B \subseteq \mathcal{Q}_\mu \). It can be easily proved that for an integral on a class of finitely maxitive measures, maxitivity implies monotonicity, while the converse is not true.

**Theorem 2.5.** The Shilkret integral is maxitive on the class of all finitely maxitive measures.

**Proof.** The maxitivity of the Shilkret integral follows from

\[
\int \max\{f, g\} \, d\mu = \max\left\{ \int f \, d\mu, \int g \, d\mu \right\} \quad \text{for all } \mu \in \mathcal{M} \text{ and all } f, g \in \mathbb{P}^{\mathcal{Q}_\mu}.
\]

Shilkret (1971) noted that homogeneity and maxitivity do not characterize his integral: maxitivity must be at least countable. We are not interested in countable maxitivity, so we proceed directly to the strongest version of maxitivity. An integral on \( \mathcal{M} \) is said to be **completely maxitive** if

\[
\int \sup f \, d\mu = \sup_{f \in \mathcal{F}} \int f \, d\mu \quad \text{for all } \mu \in \mathcal{M} \text{ and all } \mathcal{F} \subseteq \mathbb{P}^{\mathcal{Q}_\mu}.
\]

When combined with the indicator property, complete maxitivity implies that the measures \( \mu \in \mathcal{M} \) are completely maxitive, since \( I_{\bigcup A} = \sup_{A \in \mathcal{A}} I_A \) for all \( A \subseteq 2^{\mathcal{Q}_\mu} \).

**Theorem 2.6.** The Shilkret integral is completely maxitive on the class of all completely maxitive measures.
If $\mu$ is a completely maxitive measure on a set $\mathcal{Q}$, and $f : \mathcal{Q} \to \bar{R}$ is a function, then
$$\int f \, d\mu = \int f \, d\mu = \sup f \, \mu^\perp$$
for all homogeneous, completely maxitive integrals on classes $\mathcal{M} \ni \mu$. That is, the Shilkret integral on completely maxitive measures is characterized by homogeneity and complete maxitivity.

Proof. By definition of $\mu^\perp$, we have $\mu\{f \geq x\} = \sup\{f \geq x\} \, \mu^\perp$, and therefore
$$\int f \, d\mu = \sup \sup x \, \mu^\perp(q) = \sup \sup x \, \mu^\perp(q) = \sup \sup f(q) \, \mu^\perp(q) = \sup f \, \mu^\perp.$$  

This expression implies the complete maxitivity of the Shilkret integral, since $\sup_{q \in \mathcal{Q}}$ and $\sup_{f \in \mathcal{F}}$ commute.

If an integral is completely maxitive and homogeneous, then it satisfies
$$\int f \, d\mu = \int \sup \sup c \, I(q) \, d\mu = \sup \sup f(q) \, \mu^\perp(q) = \sup f \, \mu^\perp.$$  

Example 2.7. Let $\mu$ be the completely maxitive measure on $[0, 1]$ considered in Example 2.1, and for each $d \in [0, 1]$ let $l_d$ be the function $x \mapsto (x - d)^2$ on $[0, 1]$. Theorem 2.6 implies that
$$\int l_d \, d\mu = \sup x \cdot (x - d)^2 = \begin{cases} (1 - d)^2 & \text{if } 0 \leq d \leq \frac{3}{4}, \\ \frac{4}{27} d^3 & \text{if } \frac{3}{4} \leq d \leq 1; \end{cases}$$
this integral is plotted (as a function of $d \in [0, 1]$) in the first diagram of Figure 2.1 (dashed line).

A comparison of the above expression with the results of Example 1.8 shows that the value $d = \frac{3}{4}$ minimizing $\int l_d \, d\mu$ is the estimate obtained by applying the MPL criterion to the estimation problem of Example 1.1, when $n = 1$ and $X = 1$; in fact, in this case $\mu^\perp$ is equivalent to the likelihood function $lik$ on $[0, 1]$ considered in Example 1.6 (see Section 3.1).  

\top
2.2 Invariance Properties

The properties considered so far for integrals on classes of measures apply to each measure separately, while the following fundamental invariance involves integrals with respect to different measures. Let $\mathcal{M}$ be a class of measures (each $\mu \in \mathcal{M}$ is a measure on some set $\mathcal{Q}_\mu$) that is closed under transformations. An integral on $\mathcal{M}$ is said to be transformation invariant if

$$\int (f \circ t) \, d\mu = \int f \, d(\mu \circ t^{-1})$$

for all $\mu \in \mathcal{M}$, all sets $\mathcal{T}$, all $f \in \mathcal{P}^T$, and all $t \in \mathcal{T}^{\mathcal{Q}_\mu}$.

When combined with the indicator property, transformation invariance places no restrictions on the measures in $\mathcal{M}$. But it is a powerful property, since the integral of $f$ with respect to $\mu$ can be written as

$$\int f \, d\mu = \int (id_{\mathcal{P}} \circ f) \, d\mu = \int id_{\mathcal{P}} \, d(\mu \circ f^{-1}).$$

That is, a transformation invariant integral on $\mathcal{M}$ can be seen as a function on the subset of $\mathcal{M}$ consisting of the measures on $\mathcal{P}$.

Theorem 2.8. All transformation invariant integrals on classes $\mathcal{M}$ of measures ($\mathcal{M}$ closed under transformations) have the following property: if $\mu \in \mathcal{M}$ is a monotonic, subadditive measure on a set $\mathcal{Q}$, the nonempty
set $A \subseteq \mathcal{Q}$ satisfies $\mu(\mathcal{Q} \setminus A) = 0$, and $f, g : \mathcal{Q} \to \bar{P}$ are functions, then $\mu|_A \in \mathcal{M}$,

$$\int f \, d\mu = \int f|_A \, d\mu|_A,$$

and in particular

$$f = g \text{ a.e. } [\mu] \implies \int f \, d\mu = \int g \, d\mu.$$

**Proof.** As noted in Subsection 2.1.1, we have $\mu(B) = \mu(B \cap A)$ for all $B \subseteq \mathcal{Q}$ (since $\mu$ is monotone and subadditive, and $\mu(\mathcal{Q} \setminus A) = 0$). Hence, if $T$ is a set, and $t : \mathcal{Q} \to T$ is a function, then $\mu \circ t^{-1} = \mu|_A \circ (t|_A)^{-1}$, since for all $C \subseteq T$ we have

$$\mu[t^{-1}(C)] = \mu[t^{-1}(C) \cap A] = \mu[(t|_A)^{-1}(C)] = \mu|_A[(t|_A)^{-1}(C)].$$

This result implies the first two statements of the theorem. To conclude that $\mu|_A \in \mathcal{M}$, it suffices to choose a function $t : \mathcal{Q} \to A$ such that $t|_A = \text{id}_A$; we obtain $\mu \circ t^{-1} = \mu|_A$, and $\mathcal{M}$ is closed under transformations.

To conclude that $\int f \, d\mu = \int f|_A \, d\mu|_A$, it suffices to choose $t = f$: we obtain $\mu \circ f^{-1} = \mu|_A \circ (f|_A)^{-1}$, and the integral is transformation invariant.

The last statement of the theorem follows from the first two, except for the case with $\{f = g\} = \emptyset$. In this case $\mu$ is the zero measure on $\mathcal{Q}$, and therefore $\int f \, d\mu = 0$ for all $f$, since $\mu \circ f^{-1} = \mu \circ (I_\emptyset)^{-1}$.

Let $\mu$ be a monotone measure on a set $\mathcal{Q}$, and let $f : \mathcal{Q} \to \bar{P}$ be a function. The essential supremum of $f$ with respect to $\mu$ is the value

$$\text{ess}_\mu \sup f = \inf \{x \in \bar{P} : \mu\{f \geq x\} = 0\}.$$

When $\mu$ is also subadditive, we can say that $\text{ess}_\mu \sup f$ is the infimum of all $x \in \bar{P}$ such that $f < x$ a.e. $[\mu]$. In particular, it can be easily proved that if $\mu$ is completely maxitive, then $\text{ess}_\mu \sup f = \sup_{\mu|_A > 0} f$.

The **rectangular integral** associates the value

$$\int f \, d\mu = (\text{ess}_\mu \sup f) \mu\{f > 0\}$$

to each pair consisting of a monotone measure $\mu$ on a set $\mathcal{Q}$ and of a function $f : \mathcal{Q} \to \bar{P}$. When $\mu$ is completely maxitive, the rectangular integral of $f$ with respect to $\mu$ can be written as $\int f \, d\mu = (\sup_{\mu|_A > 0} f) (\sup_{f > 0} \mu|_A)$. 

Theorem 2.9. The rectangular integral is a transformation invariant, homogeneous, monotonic integral on the class of all monotonic measures.

If \( \mu \) is a monotonic, subadditive measure on a set \( Q \), and \( f : Q \to \mathbb{P} \) is a function, then

\[
\int f \, d\mu \leq \int f \, d\mu
\]

for all transformation invariant, homogeneous, monotonic integrals on classes \( M \supseteq \mu \). That is, the rectangular integral on monotonic, subadditive measures is the maximum of all transformation invariant, homogeneous, monotonic integrals.

Proof. The rectangular integral on monotonic measures satisfies the indicator property, monotonicity, and homogeneity, since the essential supremum satisfies

\[
\text{ess}_\mu \sup I_A = \begin{cases} 
0 & \text{if } \mu(A) = 0, \\
1 & \text{if } \mu(A) > 0,
\end{cases}
\]

\( f \leq g \implies \text{ess}_\mu \sup f \leq \text{ess}_\mu \sup g \),

\( \text{ess}_\mu \sup c f = c \text{ess}_\mu \sup f \) for all \( c \in \mathbb{P} \).

The rectangular integral is transformation invariant, since for all \( A \subseteq \mathbb{P} \)

\[
\mu \{ f \circ t \in A \} = [\mu \circ (f \circ t)^{-1}](A) = [\mu \circ (f^{-1}) \circ t^{-1}](A) = (\mu \circ t^{-1}) \{ f \in A \}.
\]

Let \( \mu \) be a monotonic, subadditive measure. If \( \text{ess}_\mu \sup f \) is infinite, then \( \int f \, d\mu = \infty \). If \( \text{ess}_\mu \sup f \) is finite, then for all \( x \in (\text{ess}_\mu \sup f, \infty) \) we have \( f = \min\{f, x\} \) a.e. \( [\mu] \). Hence, a transformation invariant, homogeneous, monotonic integral satisfies

\[
\int f \, d\mu = \int \min\{f, x\} \, d\mu \leq \int x I_{\{f > 0\}} \, d\mu = x \mu \{ f > 0 \}.
\]

The desired result is obtained by letting \( x \) tend to \( \text{ess}_\mu \sup f \). \( \square \)

Example 2.10. Let \( \mu \) and \( l_d \) be the measure and the functions on \([0, 1]\) defined in Examples 2.1 and 2.7, respectively. Since \( \mu \) is completely maxitive,

\[
\int l_d \, d\mu = (\sup_{(0,1]} l_d)(\sup_{[0,1]\setminus\{d\}} id_{[0,1]}) = \begin{cases} 
(1-d)^2 & \text{if } 0 \leq d \leq \frac{1}{2}, \\
d^2 & \text{if } \frac{1}{2} \leq d \leq 1;
\end{cases}
\]

this integral is plotted (as a function of \( d \in [0, 1] \)) in the first diagram of Figure 2.1 (dotted line). Since \( \int S l_d \, d\mu = \int l_d \, d\mu \) for all \( d \in [0, \frac{1}{2}] \) (see
Example 2.7), Theorems 2.3 and 2.9 imply that \( \int l_d \, d\mu = (1 - d)^2 \) for all \( d \in [0, \frac{1}{2}] \) and all transformation invariant, homogeneous, monotone integrals on classes \( M \ni \mu \).

Theorem 2.8 states that for a transformation invariant integral on \( M \), if \( \mu \in M \) is a monotone, subadditive measure on a set \( Q \), and \( f : Q \to \mathbb{P} \) is a function, then \( \int f \, d\mu = \int f|_A \, d\mu|_A \) when \( A \subseteq Q \) is a nonempty set such that \( \mu|_{Q \setminus A} = 0 \). The reciprocal invariance property is that \( \int f \, d\mu = \int f|_A \, d\mu|_A \) when \( A \subseteq Q \) is a nonempty set such that \( f|_{Q \setminus A} = 0 \). But there is a difference: if \( A = \{ A \subseteq Q : \mu|_{Q \setminus A} = 0 \} \), then in general \( \mu|_{Q \cap \mathbb{A}} \neq 0 \) (we are sure that \( \mu|_{Q \cap \mathbb{A}} = 0 \) if \( \mu \) is also continuous from below); while if \( A = \{ A \subseteq Q : f|_{Q \setminus A} = 0 \} \), then \( f|_{Q \cap \mathbb{A}} = 0 \), since \( \mathbb{A} = \{ f > 0 \} \), the support of \( f \). That is, the reciprocal invariance property can simply be stated as follows.

Let \( M \) be a class of measures (each \( \mu \in M \) is a measure on some set \( Q_\mu \) that is closed under restrictions. An integral on \( M \) is said to be support-based if

\[
\int f \, d\mu = \int f|_{\{f > 0\}} \, d\mu|_{\{f > 0\}} \quad \text{for all } \mu \in M \text{ and all } f \in \mathbb{P} Q_\mu.
\]

This property places no restrictions on the measures in \( M \), since for indicator functions \( f = I_A \) it is implied by the indicator property. For instance, the rectangular integral is support-based on the class of all monotone measures, since it depends only on \( \mu \{ f > 0 \} \) and on \( \mu \{ f \geq x \} \) for \( x \in \mathbb{P} \).

2.2.3 Regular Integrals

Let \( X \) and \( Y \) be nonnegative random variables on a probability space with probability measure \( P \). Usually, \( Y \) is said to (first-order) stochastically dominate \( X \) (with respect to \( P \)) if

\[
P\{ X \leq x \} \geq P\{ Y \leq x \} \quad \text{for all } x \in \mathbb{P}.
\]

Since a probability measure is finitely additive, finite, and continuous (from both above and below), this condition is equivalent to the following one:

\[
P\{ X \geq x \} \leq P\{ Y \geq x \} \quad \text{for all } x \in \mathbb{P}.
\]

Let \( \mu \) be a monotone measure on a set \( Q \), and let \( f : Q \to \mathbb{P} \) be a function. The (decreasing) distribution function of \( f \) with respect to \( \mu \)
is the function \( x \mapsto \mu\{f \geq x\} \) on \( \overline{\mathbb{R}} \); it shall be denoted by \( \mu\{f \geq \cdot\} \). Let \( M \) be a class of monotonic measures (each \( \mu \in M \) is a measure on some set \( Q_\mu \)). An integral on \( M \) is said to respect **distributional dominance** if

\[
\mu\{f \geq \cdot\} \leq \nu\{g \geq \cdot\} \Rightarrow \int f \, d\mu \leq \int g \, d\nu \quad \text{for all } \mu, \nu \in M, \quad \text{all } f \in \overline{P}_\mu, \text{ and all } g \in \overline{P}_\nu.
\]

If \( \mu \) and \( \nu \) are probability measures, then this property corresponds to respecting (first-order) stochastic dominance with respect to the product probability measure \( \mu \times \nu \).

**Example 2.11.** Let \( \mu \) and \( l_d \) be the measure and the functions on \([0, 1]\) defined in Examples 2.1 and 2.7, respectively. For each \( d \in [0, 1] \), the distribution function of \( l_d \) with respect to \( \mu \) satisfies

\[
\mu\{l_d \geq x\} = \begin{cases} 
1 & \text{if } 0 \leq x \leq (1 - d)^2, \\
\frac{d}{d - \sqrt{x}} & \text{if } (1 - d)^2 < x \leq d^2, \\
0 & \text{if } \max\{d^2, (1 - d)^2\} < x \leq \infty;
\end{cases}
\]

the function \( x \mapsto \mu\{l_{\frac{1}{2}} \geq x\} \) is plotted in the second diagram of Figure 2.1 for \( x \in [0, 1] \) (solid line).

Let \( \nu = \delta \circ \mu \) be the finitely maxitive measure on \([0, 1]\) considered in Example 2.2. Since

\[
\nu\{l_1 \geq \cdot\} = \delta \circ \mu\{l_1 \geq \cdot\} \leq I\{0\} + \frac{1}{2} I\{0,1\} = \mu\{I_{\{\frac{1}{2}\}} \geq \cdot\},
\]

all integrals respecting distributional dominance on classes \( M \ni \mu, \nu \) satisfy \( \int l_1 \, d\nu \leq \int I_{\{\frac{1}{2}\}} \, d\mu = \frac{1}{4} \). This is not the case for the rectangular integral: in fact, for all \( d \in [0, 1] \)

\[
\int l_d \, d\nu = (\text{ess}_{\delta \circ \mu} \sup l_d) \delta(\mu\{l_d > 0\}) = \int l_d \, d\mu = \max\{d^2, (1 - d)^2\},
\]

since \( \delta^{-1}\{0\} = \{0\} \) and \( \delta(1) = 1 \) (see Example 2.10).

The property of respecting distributional dominance strengthens monotonicity, implying in particular also monotonicity with respect to measures; but the next theorem states that at least for integrals on the class of all completely maxitive measures, it is equivalent to monotonicity, provided that the invariance properties considered in Subsection 2.2.2 are satisfied. The term "bimonotonicity" shall designate monotonicity with respect to
both functions and measures, and analogously, the term “bihomogeneity” shall designate the double homogeneity. An integral on $\mathcal{M}$ is said to be \textbf{bimonotonic} if it is monotonic and

$$\mu \leq \nu \Rightarrow \int f \, d\mu \leq \int f \, d\nu$$

for all $\mu, \nu \in \mathcal{M}$ such that $\mathcal{Q}_\mu = \mathcal{Q}_\nu$, and all $f \in \overline{\mathbb{F}}^\mathcal{U}$.

**Theorem 2.12.** Let $\mathcal{M}$ be a class of monotonic measures. All integrals on $\mathcal{M}$ respecting distributional dominance are bimonotonic, and they are also transformation invariant if $\mathcal{M}$ is closed under transformations.

All transformation invariant, bimonotonic integrals on the class of all completely maxitive measures respect distributional dominance.

All support-based, transformation invariant, monotonic integrals on the class of all completely maxitive measures respect distributional dominance.

**Proof.** An integral respecting distributional dominance is bimonotonic and transformation invariant, since for all $x \in \overline{\mathbb{F}}$ we have

$$f \leq g \Rightarrow \mu\{f \geq x\} \leq \mu\{g \geq x\},$$

$$\mu \leq \nu \Rightarrow \mu\{f \geq x\} \leq \nu\{f \geq x\},$$

$$\mu\{f \circ t \geq x\} = (\mu \circ t^{-1})\{f \geq x\}.$$

If $\mu$ is a completely maxitive measure and the integral is transformation invariant, then $\int f \, d\mu$ depends only on the function $(\mu \circ f^{-1})^\perp$ on $\overline{\mathbb{F}}$. Let $f'$ and $\mu'$ be respectively the function and the completely maxitive measure on $Q = \overline{\mathbb{F}}^2 \times \{0, 1\}$ defined by

$$f'(x, y, z) = x z \quad \text{and} \quad (\mu')^\perp(x, y, z) = \begin{cases} y z & \text{if } y < (\mu \circ f^{-1})^\perp(x), \\ 0 & \text{otherwise.} \end{cases}$$

Then $\int f \, d\mu = \int f' \, d\mu'$, because for all $x \in \overline{\mathbb{F}}$

$$[\mu' \circ (f')^{-1}]^\perp(x) = \sup_{(f')^{-1}(x)} (\mu')^\perp = (\mu \circ f^{-1})^\perp(x).$$

If $\nu$ is a completely maxitive measure, then for all $x \in \overline{\mathbb{F}}$

$$\mu\{f \geq x\} \leq \nu\{g \geq x\} \iff \sup_{[x, \infty)} (\mu \circ f^{-1})^\perp \leq \sup_{[x, \infty)} (\nu \circ g^{-1})^\perp.$$ 

In this case, for all $(x, y) \in \overline{\mathbb{F}}^2$, if $y < (\mu \circ f^{-1})^\perp(x)$, then there is an $x' \geq x$ such that $y < (\nu \circ g^{-1})^\perp(x')$: define $g'(x, y, 1) = x'$ and $g'(x, y, 0) = x$. We
obtain a function \( g' \) on \( Q \) such that \( f' \leq g' \), and therefore, if the integral is monotonic, then \( \int f' \, d\mu' \leq \int g' \, d\mu' \). Now, let \( \nu' \) be the completely maxitive measure on \( Q \) defined by

\[
(v')^\dagger(x, y, z) = \begin{cases} 
  y & \text{if } y < (\mu \circ f^{-1})^\dagger(x) \text{ and } z = 1, \\
  y & \text{if } y < (\nu \circ g^{-1})^\dagger(x) \text{ and } z = 0, \\
  0 & \text{otherwise}.
\end{cases}
\]

Since \( \mu' \leq \nu' \), if the integral is bimonotonic, then \( \int g' \, d\mu' \leq \int g' \, d\nu' \).

To obtain the second statement of the theorem, it suffices to show that \( \int g' \, d\nu' = \int g \, d\nu \); that is, it suffices to show that for all \( x' \in \overline{P} \)

\[
[
u' \circ (g')^{-1}]^\dagger(x') = \sup_{(g')^{-1}(x')} (v')^\dagger = (\nu \circ g^{-1})^\dagger(x').
\]

The second equality holds, since if \( g'(x, y, 1) = x' \) and \( y < (\mu \circ f^{-1})^\dagger(x) \), then \( y < (\nu \circ g^{-1})^\dagger(x') \).

To obtain the third statement of the theorem, it suffices to show that if the integral is support-based, then the second equality of \( \int f' \, d\mu' = \int g' \, d\nu' \leq \int g' \, d\nu' = \int g \, d\nu \)

holds, since the other two equalities have already been proved, and the inequality is implied by monotonicity. But if the integral is support-based, then the second equality is implied by \( (\mu')|_{\{f' > 0\}} = (\nu')|_{\{f' > 0\}} \), which holds since it is equivalent to \( [(\mu')^\dagger]|_{\{f' > 0\}} = [(\nu')^\dagger]|_{\{f' > 0\}} \).

Let \( \mathcal{M} \) be a regular class of monotonic measures (each \( \mu \in \mathcal{M} \) is a measure on some set \( Q_{\mu} \)). An integral on \( \mathcal{M} \) is said to be bihomogeneous if it is homogeneous and

\[
\int f \, d(c \mu) = c \int f \, d\mu \quad \text{for all } \mu \in \mathcal{M}, \text{ all } c \in \mathbb{P}, \text{ and all } f \in \overline{P}_{Q_{\mu}}.
\]

An integral on \( \mathcal{M} \) is said to be regular if it is bihomogeneous and it respects distributional dominance.

The next theorem states that a regular integral on \( \mathcal{M} \) corresponds to a functional on the set

\[
\mathcal{D}(\mathcal{M}) = \{ \mu \{ f \geq \cdot \} : \mu \in \mathcal{M}, \, f \in \overline{P}_{Q_{\mu}} \}
\]
of all distribution functions generated by the measures in \( \mathcal{M} \). Since these measures are monotonic, \( \mathcal{DF}(\mathcal{M}) \) is a subset of
\[
\mathcal{NI}_\overline{\mathbb{P}} = \{ \varphi \in \overline{\mathbb{P}}^\mathbb{P} : \varphi \text{ is nonincreasing} \}.
\]
In general \( \mathcal{DF}(\mathcal{M}) \) is a proper subset of \( \mathcal{NI}_\overline{\mathbb{P}} \): the class \( \mathcal{M} \) is said to be **exhaustive** if the two sets are equal; for example, the class of all completely maxitive measures is exhaustive. In order to state the theorem, we need some additional definitions.

Let \( \mathcal{Q} \) be a set, and let \( \mathcal{F} \) be a subset of \( \overline{\mathbb{P}}^\mathcal{Q} \). A functional \( V : \mathcal{F} \to \overline{\mathbb{P}} \) is said to be **monotonic** if
\[
f \leq g \Rightarrow V(f) \leq V(g) \quad \text{for all } f, g \in \mathcal{F}.
\]
A functional \( V : \mathcal{F} \to \overline{\mathbb{P}} \) is said to be **homogeneous** if
\[
V(cf) = cV(f) \quad \text{for all } f \in \mathcal{F} \text{ and all } c \in \mathbb{P} \text{ such that } cf \in \mathcal{F}.
\]
For example, if an integral on \( \mathcal{M} \) is monotonic and homogeneous, then for each measure \( \mu \in \mathcal{M} \) the functional \( f \mapsto \int f \, d\mu \) on \( \overline{\mathbb{P}}^\mathbb{P} \) is monotonic and homogeneous. If \( \varphi : \overline{\mathbb{P}} \to \overline{\mathbb{P}} \) is a function, and \( c \in \mathbb{P} \), then \( \varphi\left(\frac{\cdot}{c}\right) \) denotes the function \( x \mapsto \varphi\left(\frac{x}{c}\right) \) on \( \overline{\mathbb{P}} \). Let \( \mathcal{Y} \) be a subset of \( \overline{\mathbb{P}}^\mathbb{P} \). A functional \( F : \mathcal{Y} \to \overline{\mathbb{P}} \) is said to be **bihomogeneous** if it is homogeneous and
\[
F[\varphi\left(\frac{\cdot}{c}\right)] = cF(\varphi) \quad \text{for all } \varphi \in \mathcal{Y} \text{ and all } c \in \mathbb{P} \text{ such that } \varphi\left(\frac{\cdot}{c}\right) \in \mathcal{Y}.
\]
Note that if \( \varphi \in \mathcal{DF}(\mathcal{M}) \) and \( c \in \mathbb{P} \), then \( c\varphi, \varphi\left(\frac{\cdot}{c}\right) \in \mathcal{DF}(\mathcal{M}) \). A functional \( F : \mathcal{Y} \to \overline{\mathbb{P}} \) is said to be **\( x \)-independent**, for an \( x \in \overline{\mathbb{P}} \), if
\[
\varphi|_{\overline{\mathbb{P}} \setminus \{x\}} = \psi|_{\overline{\mathbb{P}} \setminus \{x\}} \quad \Rightarrow \quad F(\varphi) = F(\psi) \quad \text{for all } \varphi, \psi \in \mathcal{Y}.
\]
A functional \( F : \mathcal{NI}_\overline{\mathbb{P}} \to \overline{\mathbb{P}} \) is said to be **calibrated** if
\[
F(I_{[0,1]} + y I_{\{0\}}) = 1 \quad \text{for all } y \in \overline{\mathbb{P}}.
\]
Note that a \( 0 \)-independent functional \( F \) on \( \mathcal{NI}_\overline{\mathbb{P}} \) is calibrated if and only if \( F(I_{[0,1]}) = 1 \). The left-continuous version of a function \( \varphi \in \mathcal{NI}_\overline{\mathbb{P}} \) is the function \( \varphi^- \in \mathcal{NI}_\overline{\mathbb{P}} \) defined by
\[
\varphi^-(x) = \inf_{[0,x]} \varphi \quad \text{for all } x \in \overline{\mathbb{P}}.
\]
It can be easily proved that \( \varphi^- \) is left-continuous, and that \( (\varphi^-)^- = \varphi^- \). Note that \( \varphi^-(0) = \infty \).
Theorem 2.13. Let $\mathcal{M}$ be a regular class of monotonic measures. An integral on $\mathcal{M}$ is regular if and only if the functional

$$F : \mu \{ f \geq \cdot \} \mapsto \int f \, d\mu$$

on $\mathcal{D}(\mathcal{M})$ is well-defined, monotonic, and bihomogeneous. In this case, the functional $F$ is $x$-independent for all $x \in (0, \infty]$; if it is also 0-independent and $\mathcal{M}$ is closed under restrictions, then the integral is support-based.

Conversely, if $F : \mathcal{N} \rightarrow [0, \infty]$ is a calibrated, bihomogeneous, monotonic functional, then the equality

$$\int f \, d\mu = F(\mu \{ f \geq \cdot \})$$

defines a regular integral on the class of all monotonic measures. The integral is support-based if and only if $F$ is 0-independent. In this case, $F(\varphi) = F(\varphi^-)$ for all $\varphi \in \mathcal{N}$, and in particular, denoting by $\mu \{ f > \cdot \}$ the function $x \mapsto \mu \{ f > x \}$ on $\mathbb{P}$, the integral can be written also as

$$\int f \, d\mu = F(\mu \{ f > \cdot \}).$$

Proof. An integral on $\mathcal{M}$ respects distributional dominance if and only if the functional $F$ is well-defined and monotonic. In this case, the integral is bihomogeneous if and only if $F$ is bihomogeneous. If $\varphi \mid \mathbb{P} \{ x \} = \psi \mid \mathbb{P} \{ x \}$ for an $x \in \mathbb{P}$, then for all $x' \in \mathbb{P} \{ x \}$ and all $c \in (1, \infty)$ we have

$$\psi(x) \leq \psi(\frac{x}{c}) = \varphi(\frac{x}{c}) \quad \text{and} \quad \psi(x') = \varphi(x') \leq \varphi(\frac{x'}{c}).$$

Hence $F(\psi) \leq F(\varphi(\cdot)) = c F(\varphi)$ holds when $F$ is monotonic and bihomogeneous: by letting $c$ tend to 1 we obtain the inequality $F(\psi) \leq F(\varphi)$, and thus, by symmetry, the $x$-independence. To prove the $\infty$-independence, consider the following consequence of Theorem 2.3:

$$\lim_{x \uparrow \infty} \mu \{ f \geq x \} > 0 \Rightarrow \int f \, d\mu \geq \int f \, d\mu = \infty.$$
The integral defined by a calibrated, bihomogeneous, monotonic functional $F : \mathcal{N}_\mathcal{P} \to \mathcal{P}$ satisfies the indicator property, since if $\mu(A) \in \mathcal{P}$, then using bihomogeneity and calibration we obtain

$$\int I_A \, d\mu = F(\mu\{I_A \geq \cdot\}) = F[\mu(A) I_{(0,1]} + \mu(Q_\mu) I_{\{0,1\}}] = \mu(A),$$

and using limits and monotonicity we obtain the same result also for $\mu(A) \in \{0, \infty\}$. We have already proved that the integral is regular, that $F$ is $\infty$-independent, and that the integral is support-based when $F$ is $0$-independent. Now assume that the integral is support-based, and consider a function $\varphi \in \mathcal{N}_\mathcal{P}$. To prove the $0$-independence of $F$, it suffices to show that defining $\varphi' = \varphi I_{(0,\infty]} + (\sup_{(0,\infty]} \varphi) I_{\{0,1\}}$, we obtain $F(\varphi) = F(\varphi')$. Let $\mu$ be the completely maxitive measure on $\mathcal{P}$ defined by $\mu^\perp = \varphi$. Since $\varphi$ is nonincreasing, we have $\mu\{id_{\mathcal{P}} \geq \cdot\} = \varphi$, and therefore, since the integral is support-based and $\{id_{\mathcal{P}} > 0\} = (0, \infty]$, to obtain the desired result it suffices to show that $\mu\{id_{\mathcal{P}}|_{(0,\infty]} \geq x\} = \varphi'(x)$ for all $x \in \mathcal{P}$. This equality clearly holds for positive $x$, and for $x = 0$ we have

$$\mu\{0,\infty\}\{id_{\mathcal{P}}|_{(0,\infty]} \geq 0\} = \mu(0, \infty] = \sup_{(0,\infty]} \varphi = \varphi'(0).$$

Defining $\varphi'' = \varphi I_{(0,\infty]} + \varphi^- I_{(0,\infty]}$, we obtain $\varphi^-(x) \leq \varphi''(x)$ for all $x \in \mathcal{P}$ and all $c \in (1, \infty)$, and therefore

$$F(\varphi'') = F(\varphi) \leq F(\varphi^-) \leq c F(\varphi'').$$

The equality $F(\varphi) = F(\varphi^-)$ is obtained by letting $c$ tend to $1$. The last statement of the theorem follows from this result, because the equality $(\mu\{f \geq \cdot\})^- = (\mu\{f > \cdot\})^-$ is obtained from the inequalities

$$\mu\{f > x\} \leq \mu\{f \geq x\} \leq \inf_{[0,x]} \mu\{f > \cdot\}$$

by considering the left-continuous versions of the corresponding functions of $x \in \mathcal{P}$:

$$(\mu\{f > \cdot\})^- \leq (\mu\{f \geq \cdot\})^- \leq [(\mu\{f > \cdot\})^-]^- = (\mu\{f > \cdot\})^-.$$
Many authors (for example Denneberg, 1994) define the distribution function as \( \mu\{f > \cdot\} \). Theorem 2.13 assures us that if a support-based, regular integral on an exhaustive class of monotonic measures is defined by a functional \( F \) on \( \mathcal{N}\mathcal{I}\), then it is not affected by the choice between \( \mu\{f \geq \cdot\} \) and \( \mu\{f > \cdot\} \).

Theorem 2.13 implies that the Shilkret integral is regular and support-based on the class of all monotonic measures, since it corresponds to the 0-independent, bihomogeneous, monotonic functional \( F_S : \mathcal{N}\mathcal{I} \rightarrow \mathbb{P} \) defined by
\[
F_S(\varphi) = \sup_{x \in \mathbb{P}} x \varphi(x) \quad \text{for all } \varphi \in \mathcal{N}\mathcal{I}.
\]
By contrast, the rectangular integral is not regular on the class of all monotonic measures, because the functional \( F \) is not well-defined, since in general \( \mu\{f > 0\} \) is not determined by \( \mu\{f \geq \cdot\} \). But if \( \mu \) is continuous from below, then \( \mu\{f > 0\} = \sup_{\mathbb{P}} \mu\{f \geq \cdot\} \), and therefore the rectangular integral is regular and support-based on the class of all monotonic measures that are continuous from below, since it correspond to the 0-independent, bihomogeneous, monotonic functional \( F_r : \mathcal{N}\mathcal{I} \rightarrow \mathbb{P} \) defined by
\[
F_r(\varphi) = \inf\{\varphi = 0\} \sup_{\mathbb{P}} \varphi \quad \text{for all } \varphi \in \mathcal{N}\mathcal{I}.
\]

The **regularized rectangular integral** is the integral defined by \( F_r \) on the class of all monotonic measures: it associates the value
\[
\int_{rr} f \, d\mu = (\text{ess}_\mu \sup f) \sup_{\mathbb{P}} \mu\{f \geq \cdot\}
\]
to each pair consisting of a monotonic measure \( \mu \) on a set \( Q \) and of a function \( f : Q \rightarrow \mathbb{P} \). The regularized rectangular integral is thus regular and support-based on the class of all monotonic measures, and it corresponds to the rectangular integral on the class of all monotonic measures that are continuous from below.

**Example 2.14.** Let \( \mu, \nu, \delta, \) and \( l_d \) be the measures and the functions defined in Examples 2.1, 2.2, and 2.7. For all \( d \in [0, 1) \)
\[
\int_{rr}^{rr} l_d \, d\mu = \int_{r}^{r} l_d \, d\mu = \max\{d^2, (1 - d)^2\} = \int_{r}^{r} l_d \, d\nu = \int_{rr}^{rr} l_d \, d\nu;
\]
the first equality holds because \( \mu \) is completely maxitive (and thus continuous from below), the second and third ones were proved in Examples 2.10
and 2.11, and the last one is implied by \( \sup_{\mathbb{P}} \nu\{l_d \geq \cdot\} = 1 = \nu\{l_d > 0\} \).

When \( d = 1 \), the first three equalities still hold, but the last one does not: \( \int \nu_{1} \, d\nu = \frac{1}{2} \), since \( \sup_{\mathbb{P}} \nu\{l_1 \geq \cdot\} = \frac{1}{2} \). In fact, Theorem 2.13 implies that all support-based, regular integrals on exhaustive classes \( \mathcal{M} \ni \mu, \nu \) satisfy \( \int \nu_{1} \, d\nu = \frac{1}{2} \int \nu_{1} \, d\mu \), because \( \nu\{l_1 \geq \cdot\}|(0,\infty] = \frac{1}{2} \mu\{l_1 \geq \cdot\}|(0,\infty] \).

\[ \begin{align*}
\text{Theorem 2.15.} & \quad \text{If} \ \mu \text{ is a monotonic measure on a set} \ Q, \ \text{and} \ f : Q \to \mathbb{P} \ \\
& \text{is a function, then} \int_{S}^{rr} f \, d\mu \leq \int f \, d\mu \leq \int f \, d\mu.
\end{align*} \]

for all regular integrals on classes \( \mathcal{M} \ni \mu \). That is, the Shilkret integral and the regularized rectangular integral are respectively the minimum and the maximum of all regular integrals.

In particular, if \( A \subseteq Q \) is a nonempty set, and \( \mu \) is the completely maxitive measure on \( Q \) defined by \( \mu^\perp = I_A \), then

\[ \begin{align*}
\int f \, d\mu = \sup_A f & \quad \text{and} \quad \int f \, d\overline{\mu} = \inf_A f
\end{align*} \]

for all regular integrals on classes \( \mathcal{M} \ni \mu, \overline{\mu} \).

\[ \begin{align*}
\text{Proof.} & \quad \text{Since a regular integral is monotonic and homogeneous, the first} \\
& \text{inequality was proved in Theorem 2.3. If} \ y = \sup_{\mathbb{P}} \mu\{f \geq \cdot\} \in \{0,\infty\}, \\
& \text{then the second inequality clearly holds. If} \ y \in \mathbb{P}, \ \text{then there is an} \ x \in \mathbb{P} \\
& \text{such that} \ y' = \mu\{f \geq x\} \in (0,y]. \ \text{Let} \ g = (\text{ess}_\mu \sup f) I_{\{f \geq x\}} \ \text{and} \ \nu = \frac{y}{y} \mu. \\
& \text{Since} \ \mu\{f \geq \cdot\} \leq \mu(Q) I_{\{0\}} + y I_{[0,\text{ess}_\mu \sup f]} \leq \nu\{g \geq \cdot\}, \ \text{we have} \\
& \int f \, d\mu \leq \int g \, d\nu = (\text{ess}_\mu \sup f) \nu\{f \geq x\} = \int f \, d\mu.
\end{align*} \]

Consider now the case with \( \mu^\perp = I_A \). If \( x < \sup_A f \), then \( \mu\{f \geq x\} = 1 \); while if \( x > \sup_A f \), then \( \mu\{f \geq x\} = 0 \). Hence we have

\[ \int f \, d\mu = \int f \, d\mu = \int f \, d\mu = \sup_A f. \]

The last equality of the theorem can be obtained in the same way: if \( x < \inf_A f \), then \( \overline{\mu}\{f \geq x\} = 1 - \mu\{f < x\} = 1 \); while if \( x > \inf_A f \), then \( \overline{\mu}\{f \geq x\} = 1 - \mu\{f < x\} = 0 \). \( \square \)
2.2 Nonadditive Integrals 47

The second part of Theorem 2.15 justifies the use of the expression “density function” for \( \mu^\perp \) (where \( \mu \) is a completely maxitive measure on a set \( Q \)) when regular integrals are considered, since \( \mu^\perp \) is the “Radon-Nikodým derivative” of \( \mu \) with respect to the completely maxitive measure \( \nu \) on \( Q \) defined by \( \nu^\perp = 1 \). In fact, for all regular integrals on classes \( \mathcal{M} \supseteq \nu \) we have

\[
\int_A \mu^\perp \, d\nu = \int A \mu^\perp I_A \, d\nu = \sup \mu^\perp I_A = \mu(A) \quad \text{for all } A \subseteq Q,
\]

where the left-hand side of the first equality is the usual abbreviation for the right-hand side.

Geometrically, if we consider the rectangles lying in the first quadrant of the (extended) Cartesian plane and having two sides on the coordinate axes, then \( F_S(\varphi) \) is the area of the largest rectangle lying under the graph of \( \varphi \) (if such a rectangle exists), while \( F_r(\varphi) \) is the area of the smallest rectangle containing the portion of the graph of \( \varphi \) not lying on the coordinate axes. From this geometrical standpoint, the most natural calibrated, bihomogeneous, monotone functional on \( \mathcal{N}_\mathbb{P} \) is probably the one assigning to each function the area under its graph. That is, the functional \( F_C : \mathcal{N}_\mathbb{P} \to \mathbb{P} \) defined by

\[
F_C(\varphi) = \int_0^\infty \varphi(x) \, dx \quad \text{for all } \varphi \in \mathcal{N}_\mathbb{P},
\]

where the right-hand side of the equality is the Lebesgue integral of \( \varphi|_\mathbb{P} \) (with respect to the Lebesgue measure on \( \mathbb{P} \): since \( \varphi \) is nonincreasing, it is Borel measurable), which corresponds to the improper Riemann integral when \( \varphi|_\mathbb{P} \) is finite (in this case, \( \varphi \) is Riemann integrable on every bounded, closed interval \( I \subseteq \mathbb{P} \), since \( \varphi|_I \) is bounded and nonincreasing), while the integral is infinite when \( \varphi|_\mathbb{P} \) is not finite.

The Choquet integral is the integral defined by \( F_C \) on the class of all monotone measures: it associates the value

\[
\int f \, d\mu = \int_0^\infty \mu\{f \geq x\} \, dx
\]

to each pair consisting of a monotone measure \( \mu \) on a set \( Q \) and of a function \( f : Q \to \mathbb{P} \). The Choquet integral is a well-known nonadditive integral, which was introduced by Choquet (1954) and then developed by
many authors (see for example Denneberg, 1994); it is regular and support-based on the class of all monotonic measures, since $F_C$ is 0-independent, bihomogeneous, and monotonic.

Example 2.16. Let $\mu$ and $l_d$ be the measure and the functions on $[0, 1]$ defined in Examples 2.1 and 2.7, respectively. The results of Example 2.11 imply that

$$\int_0^\infty \mu\{l_d \geq x\} \, dx = \begin{cases} (1-d)^2 & \text{if } 0 \leq d \leq \frac{1}{2}, \\ \frac{1}{3} d^3 + \frac{5}{3} (1-d)^3 & \text{if } \frac{1}{2} \leq d \leq 1; \end{cases}$$

this integral is plotted (as a function of $d \in [0, 1]$) in the first diagram of Figure 2.1 (solid line). The value $d \approx 0.691$ minimizing $\int_0^\infty l_d \, d\mu$ is the estimate obtained by applying the “MPL* criterion” of Cattaneo (2005) to the estimation problem of Example 1.1, when $n = 1$ and $X = 1$ (compare with Example 2.7).

In the second diagram of Figure 2.1, the area under the solid line (that is, under the graph of $\mu\{l_{\frac{d}{2}} \geq \cdot\}$) corresponds to the Choquet integral of $l_{\frac{d}{2}}$ with respect to $\mu$, while the areas of the dashed and dotted rectangles correspond to the Shilkret integral and to the (regularized) rectangular integral, respectively.

All the regular integrals (on regular classes $\mathcal{M}$ of monotonic measures) that we have considered so far are support-based (when $\mathcal{M}$ is closed under restrictions). The next example shows that there are regular integrals (on classes $\mathcal{M}$ closed under restrictions) that are not support-based.

Example 2.17. Consider the functional $F : \mathcal{N}\mathcal{I}_{\mathbb{P}} \to \mathbb{P}$ defined by

$$F(\varphi) = \max \left\{ F_S(\varphi), \inf \{ \varphi = 0 \} \min \left\{ \frac{1}{2} \varphi(0), \sup_{\mathbb{P}} \varphi \right\} \right\} \quad \text{for all } \varphi \in \mathcal{N}\mathcal{I}_{\mathbb{P}}.$$  

It can be easily proved that $F$ is calibrated, bihomogeneous, and monotonic, but not 0-independent. Therefore the integral defined by $F$ on the class of all monotonic measures is regular, but not support-based.

The hypograph of a function $\varphi : \mathbb{P} \to \mathbb{P}$ is the set

$$H(\varphi) = \{(x,y) \in \mathbb{P}^2 : y \leq \varphi(x)\}.$$  

The functional $F_C$ assigns to each $\varphi \in \mathcal{N}\mathcal{I}_{\mathbb{P}}$ the value $\nu_C[H(\varphi)]$, where
\(\nu_C\) is the countably additive measure on (the Borel sets of) \(\mathbb{F}^2\) such that 
\(\nu_C(\mathbb{F}^2 \setminus \mathbb{F}^2) = 0\), and \(\nu_C|_{\mathbb{F}^2}\) is the Lebesgue measure on \(\mathbb{F}^2\). It can be easily proved that, in analogy with the Choquet integral, each integral respecting distributional dominance on some class of monotone measures can be represented by some (not unique) monotone measure \(\nu\) on \(\mathbb{F}^2\), in the sense that it can be written as

\[
\int f \, d\mu = \nu[H(\mu \{ f \geq \cdot \})]. \tag{2.4}
\]

For example, the Shilkret integral and the regularized rectangular integral are respectively represented by the measures \(\nu_S\) and \(\nu_r\) on \(\mathbb{F}^2\) defined by

\[
\nu_S(A) = \sup_{(x,y) \in A} xy \quad \text{and} \quad \nu_r(A) = \left( \sup_{(x,y) \in A \cap \mathbb{F}^2} x \right) \left( \sup_{(x,y) \in A \cap \mathbb{F}^2} y \right),
\]

for all \(A \subseteq \mathbb{F}^2\).

Conversely, if a monotone measure \(\nu\) on (the Borel sets of) \(\mathbb{F}^2\) satisfies

\[
\nu[H(y I_{[0,1]} + y' I_{\{0\}})] = y \quad \text{for all } y, y' \in \mathbb{F}, \tag{2.5}
\]

then the equality (2.4) defines an integral respecting distributional dominance on the class of all monotone measures, since the indicator property is implied by the condition (2.5). For instance, with the measure \(\nu\) on \(\mathbb{F}^2\) defined by

\[
\nu(A) = \sup_{A \cap (\mathbb{F} \times \mathbb{F})} m \quad \text{for all } A \subseteq \mathbb{F}^2,
\]

where \(m\) is a nonnegative, extended real-valued function on \(\mathbb{F} \times \mathbb{F}\) that is nondecreasing in both arguments and that fulfills the conditions (2.2) and (2.3), we obtain the generalized Sugeno integral corresponding to \(m\). Imaoka (1997) introduced a class of integrals obtained in this way using a particular kind of countably additive measures. Thanks to the Carathéodory extension theorem, it can be proved that the Choquet integral is the only homogeneous integral that can be represented by a countably additive measure on (the Borel sets of) \(\mathbb{F}^2\), although this measure is not unique. That is, if we call “generalized Imaoka integrals” the integrals representable by countably additive measures on (the Borel sets of) \(\mathbb{F}^2\), then what distinguishes the Choquet integral among the generalized Imaoka integrals is the homogeneity, which is also the distinguishing feature of the Shilkret integral among the generalized Sugeno integrals.
The left-continuous pseudo-inverse of a function \( \varphi \in \mathcal{N}_\mathbb{P} \) is the function \( \varphi^\sim \in \mathcal{N}_\mathbb{P} \) defined by

\[
\varphi^\sim(x) = \sup\{\varphi \geq x\} = \inf\{\varphi < x\} \quad \text{for all } x \in \mathbb{P}.
\]

It can be easily proved that \((\varphi^\sim)^\sim = \varphi^\sim\), and that \(H(\varphi^\sim)\) is the mirror image of \(H(\varphi^-)\) with respect to the diagonal \(y = x\) of \(\mathbb{P}^2\). Hence, if a support-based, regular integral on the class of all monotonic measures is represented by a measure on (the Borel sets of) \(\mathbb{P}^2\) that is invariant with respect to the reflection about the diagonal \(y = x\), then the corresponding functional \(F\) on \(\mathcal{N}_\mathbb{P}\) satisfies

\[
F(\varphi) = F(\varphi^-) = \nu[H(\varphi^-)] = \nu[H(\varphi^\sim)] = F(\varphi^\sim) \quad \text{for all } \varphi \in \mathcal{N}_\mathbb{P}.
\]

A regular integral on an exhaustive, regular class of monotonic measures is said to be symmetric if the corresponding functional \(F : \mathcal{N}_\mathbb{P} \to \mathbb{P}\) satisfies

\[
F(\varphi) = F(\varphi^\sim) \quad \text{for all } \varphi \in \mathcal{N}_\mathbb{P}.
\]

The Choquet integral, the Shilkret integral, and the regularized rectangular integral are symmetric, since the measures \(\nu_C\), \(\nu_S\), and \(\nu_r\) on (the Borel sets of) \(\mathbb{P}^2\) are invariant with respect to the reflection about the diagonal \(y = x\).

The next theorem illustrates an interesting consequence of the symmetry of an integral.

**Theorem 2.18.** All symmetric, regular integrals on exhaustive, regular classes \(\mathcal{M}\) of monotonic measures are support-based (if \(\mathcal{M}\) is closed under restrictions) and have the following property: if \(\mu \in \mathcal{M}\) is a completely maximal measure on a set \(Q\), and \(f : Q \to \mathbb{P}\) is a function, then, denoting by \(F\) the functional on \(\mathcal{N}_\mathbb{P}\) corresponding to the integral, and by \(\sup_{\{\mu^\downarrow \geq \cdot\}} f\) the function \(x \mapsto \sup_{\{\mu^\downarrow \geq x\}} f\) on \(\mathbb{P}\), we have

\[
\int f \, d\mu = F(\sup_{\{\mu^\downarrow \geq \cdot\}} f);
\]

moreover, if \(\mu\) is finite, then denoting by \(\inf_{\{\mu^\downarrow \geq \mu(Q) - \cdot\}} f\) the function \(x \mapsto \inf_{\{\mu^\downarrow \geq \mu(Q) - x\}} f\) on \(\mathbb{P}\), we have also

\[
\int f \, d\mu = F(I_{[0, \mu(Q)]} \inf_{\{\mu^\downarrow \geq \mu(Q) - \cdot\}} f).
\]
2.2 Nonadditive Integrals

Proof. The integral is support-based, because $\varphi|_{[0,\infty]} = \psi|_{[0,\infty]}$ implies $\varphi^- = \psi^-$, and therefore $F$ is 0-independent. If we define $\varphi = \mu\{f > \cdot\}$ and $\psi = \sup_{\mu \geq \cdot} f$, and we prove that $\varphi^- = \psi^-$, then we can use Theorem 2.13 to obtain the desired result

$$\int f \, d\mu = F(\varphi) = F(\varphi^-) = F(\psi^-) = F(\psi).$$

The equality $\varphi^- = \psi^-$ can be proved as follows: for all $x, y \in \overline{\mathbb{P}}$ we have

$$\varphi^-(x) \leq y \iff \forall y' \in (y, \infty] \quad \mu\{f > y'\} < x$$
$$\iff \forall y' \in (y, \infty] \quad \exists x' \in [0, x) \quad \{f > y'\} \cap \{\mu \geq x'\} = \emptyset$$
$$\iff \forall y' \in (y, \infty] \quad \exists x' \in [0, x) \quad \sup_{\mu \geq x'} f \leq y'$$
$$\iff \psi^-(x) \leq y.$$

The last statement of the theorem can be proved in a similar way: defining $\varphi = \mu\{f \geq \cdot\}$ and $\psi = \inf_{\mu \geq \mu(Q) -} f$, it suffices to show that $\varphi^- = \psi^-$. This can be done as follows: for all $x, y \in \overline{\mathbb{P}}$ we have

$$\varphi^-(x) > y \iff \exists y' \in (y, \infty] \quad \mu\{f \geq y'\} \geq x$$
$$\iff \exists y' \in (y, \infty] \quad \mu\{f < y'\} \leq \mu(Q) - x$$
$$\iff \exists y' \in (y, \infty] \quad \exists x' \in [0, x) \quad \{f < y'\} \cap \{\mu \geq \mu(Q) - x'\} = \emptyset \text{ and } x \leq \mu(Q)$$
$$\iff \exists y' \in (y, \infty] \quad \forall x' \in [0, x) \quad I_{[0, \mu(Q)]}(x') \inf_{\mu \geq \mu(Q) - x'} f \geq y'$$
$$\iff \psi^-(x) > y.$$

2.2.4 Subadditivity

Let $\mathcal{M}$ be a class of measures (each $\mu \in \mathcal{M}$ is a measure on some set $Q_\mu$). We have seen that if some measures in $\mathcal{M}$ are not finitely additive, then an integral on $\mathcal{M}$ cannot be additive; but if the measures in $\mathcal{M}$ are subadditive, then we have at least

$$\int (I_A + I_B) \, d\mu \leq \int I_A \, d\mu + \int I_B \, d\mu \quad \text{for all } \mu \in \mathcal{M}$$
$$\text{and all disjoint } A, B \subseteq Q_\mu.$$
A quasi-subadditive integral extends this property to all pairs of functions $f, g \in \overline{\mathcal{P}}$ with disjoint supports (that is, $\{f > 0\} \cap \{g > 0\} = \emptyset$). An integral on $\mathcal{M}$ is said to be **quasi-subadditive** if

$$
\int (f+g) \, d\mu \leq \int f \, d\mu + \int g \, d\mu \quad \text{for all } \mu \in \mathcal{M}
$$

and all $f, g \in \overline{\mathcal{P}}$ with disjoint supports.

When combined with the indicator property, quasi-subadditivity implies that the measures $\mu \in \mathcal{M}$ are subadditive, since $I_{A \cup B} = I_A + I_B$ for all disjoint $A, B \subseteq \Omega$. A functional $F : \mathcal{N}_\mathcal{P} \to \overline{\mathcal{P}}$ is said to be **subadditive** if

$$
F(\psi + \varphi) \leq F(\varphi) + F(\psi) \quad \text{for all } \varphi, \psi \in \mathcal{N}_\mathcal{P}.
$$

**Theorem 2.19.** All regular integrals on exhaustive, regular classes $\mathcal{M}$ of subadditive, monotonic measures have the following property: if the corresponding functional on $\mathcal{N}_\mathcal{P}$ is subadditive, then the integral is support-based (if $\mathcal{M}$ is closed under restrictions) and quasi-subadditive.

**Proof.** The integral is support-based, since the corresponding functional $F$ on $\mathcal{N}_\mathcal{P}$ is 0-independent. In fact, if $\varphi|_{[0,\infty]} = \psi|_{[0,\infty]}$ and $\varphi(0) < \psi(0)$, then $(\psi - \varphi) \in \mathcal{N}_\mathcal{P}$, and so there are a measure $\mu \in \mathcal{M}$ and a function $f$ on $\Omega$ such that $(\psi - \varphi) = \mu \{ f \geq \cdot \}$; but this implies that $(\psi - \varphi) = \mu \{ I_\varphi \geq \cdot \}$, and therefore

$$
F(\varphi) \leq F(\psi) \leq F(\varphi) + F(\psi - \varphi) = F(\varphi) + \int I_\varphi \, d\mu = F(\varphi).
$$

The integral is quasi-subadditive, because if $f$ and $g$ have disjoint supports, then $\{ f + g \geq x \} = \{ f \geq x \} \cup \{ g \geq x \}$, and therefore

$$
\int (f + g) \, d\mu \leq F(\mu \{ f \geq x \} \cup \mu \{ g \geq x \}) \leq \int f \, d\mu + \int g \, d\mu,
$$

since $\mu$ is subadditive.

Theorem 2.19 implies that the Shilkret integral and the Choquet integral are quasi-subadditive on the class of all monotonic, subadditive measures, since the functionals $F_S$ and $F_C$ are subadditive. By contrast, the next example shows that the regularized rectangular integral can be quasi-subadditive only on very limited classes of monotonic, subadditive measures.
Example 2.20. Let \( \mu \) be a monotonic, subadditive measure on a set \( Q \). If there are two disjoint sets \( A, B \subset Q \) such that \( 0 < \mu(A) < \frac{1}{2} \mu(B) < \infty \), then the two functions \( f = \mu(A) I_B \) and \( g = \mu(B) I_A \) on \( Q \) have disjoint supports, but

\[
\int (f + g) \, d\mu = \mu(A \cup B) \mu(B) > 2 \mu(A) \mu(B) = \int f \, d\mu + \int g \, d\mu. 
\]

In fact, as stated by the next two theorems, the Shilkret integral and the Choquet integral satisfy the following stronger property on suitable classes of monotonic, subadditive measures. An integral on \( M \) is said to be subadditive if

\[
\int (f + g) \, d\mu \leq \int f \, d\mu + \int g \, d\mu \quad \text{for all } \mu \in M \text{ and all } f, g \in \mathbb{P}^Q \mu.
\]

**Theorem 2.21.** The Shilkret integral is subadditive on a class of monotonic measures if and only if these are finitely maxitive.

**Proof.** For the “if” part, we can adapt the proof of Shilkret (1971) to the finitely maxitive measures. It suffices to show

\[
x \mu\{f + g \geq x\} \leq \int f \, d\mu + \int g \, d\mu \quad \text{for all } x \in \mathbb{P},
\]

since then the desired result is obtained by taking the supremum (over all \( x \in \mathbb{P} \)) of the left-hand side of the inequality. If one of the two integrals of the right-hand side is infinite, the inequality clearly holds. If both are finite, observe that

\[
\{f + g \geq x\} \subseteq \{f \geq \lambda x\} \cup \{g \geq (1 - \lambda) x\} \quad \text{for all } \lambda \in (0, 1).
\]

If one of the two integrals is 0, say \( \int g \, d\mu = 0 \), then \( \mu\{g \geq x' \} = 0 \) for all \( x' \in \mathbb{P} \), and therefore

\[
x \mu\{f + g \geq x\} = \sup_{\lambda \in (0, 1)} \lambda x \mu\{f + g \geq x\} \leq \sup_{\lambda \in (0, 1)} \lambda x \mu\{f \geq \lambda x\} \leq \int f \, d\mu = \int f \, d\mu + \int g \, d\mu.
\]
If both the integrals are positive and finite, then setting
\[ \lambda = \frac{\int f \, d\mu}{\int f \, d\mu + \int g \, d\mu} \]
we obtain
\[ x \mu \{ f + g \geq x \} \leq \max \{ x \mu \{ f \geq \lambda x \}, x \mu \{ g \geq (1 - \lambda) x \} \} \leq \]
\[ \leq \max \left\{ \frac{1}{\lambda} \int f \, d\mu, \frac{1}{1 - \lambda} \int g \, d\mu \right\} = \int f \, d\mu + \int g \, d\mu. \]

For the “only if” part, assume that the Shilkret integral is subadditive on a class \( \mathcal{M} \) of monotonic measures, let \( \mu \in \mathcal{M} \) be a measure on \( \mathcal{Q} \), and let \( A, B \subseteq \mathcal{Q} \) be two disjoint sets. We have
\[ \mu \{ \mu(A \cup B) I_A + \mu(A) I_B \geq x \} = \begin{cases} \mu(A \cup B) & \text{if } 0 < x \leq \mu(A), \\ \mu(A) & \text{if } \mu(A) < x \leq \mu(A \cup B), \\ 0 & \text{if } \mu(A \cup B) < x < \infty; \end{cases} \]
and therefore
\[ \int [\mu(A \cup B) I_A + \mu(A) I_B] \, d\mu = \mu(A \cup B) \mu(A). \]

From this result and the equality
\[ \mu(A \cup B) I_{A \cup B} = [\mu(A \cup B) I_A + \mu(A) I_B] + [\mu(A \cup B) I_B - \mu(A) I_B], \]
using the subadditivity of the Shilkret integral and the property
\[ \int c I_C \, d\mu = c \mu(C) \quad \text{for all } c \in \overline{\mathbb{P}} \text{ and all } C \subseteq \mathcal{Q}, \]
we obtain the inequality
\[ \mu(A \cup B)^2 \leq \mu(A \cup B) \mu(A) + [\mu(A \cup B) - \mu(A)] \mu(B), \]
which is equivalent to
\[ [\mu(A \cup B) - \mu(A)] [\mu(A \cup B) - \mu(B)] \leq 0. \]
Since both factors of the left-hand side of this inequality are nonnegative, one of them must vanish; that is, \( \mu(A \cup B) = \max \{ \mu(A), \mu(B) \} \).

\( \square \)
The next theorem is proved for example in Denneberg (1994, Chapter 6).

**Theorem 2.22.** The Choquet integral is subadditive on a class of monotonic measures if and only if these are 2-alternating.

### 2.2.5 Continuity

There is a correspondence between convergence theorems for regular integrals on regular classes $\mathcal{M}$ of monotonic measures, and continuity properties of the corresponding functionals on $DF(\mathcal{M})$. The next one is the convergence theorem corresponding to the continuity property implied by monotonicity and bihomogeneity.

**Theorem 2.23.** All regular integrals on regular classes $\mathcal{M}$ of monotonic measures have the following property: if the sequence $\mu_1, \mu_2, \ldots \in \mathcal{M}$ of measures on a set $\Omega$ converges uniformly to the measure $\mu \in \mathcal{M}$ on $\Omega$, the sequence $f_1, f_2, \ldots : \Omega \to \mathbb{P}$ of functions converges uniformly to the function $f : \Omega \to \mathbb{P}$, and one of the following two conditions is fulfilled

\[
\lim_{n \to \infty} \sup_{\mathbb{P}} \mu_n \{ f_n \geq \cdot \} = \sup_{\mathbb{P}} \mu \{ f \geq \cdot \} < \infty
\]

and

\[
\lim_{n \to \infty} \text{ess} \sup_{\mathbb{P}} f_n = \text{ess} \sup_{\mathbb{P}} f < \infty,
\]

(2.6)

then

\[
\lim_{n \to \infty} \int f_n \, d\mu_n = \int f \, d\mu.
\]

**Proof.** Define

\[
\varphi = \mu \{ f \geq \cdot \}, \quad \varphi_n = \mu_n \{ f_n \geq \cdot \}, \quad a = \sup_{\mathbb{P}} \varphi, \quad a_n = \sup_{\mathbb{P}} \varphi_n,
\]

\[
b = \text{ess} \sup_{\mathbb{P}} f = \sup_{\mathbb{P}} \varphi^\sim, \quad \text{and} \quad b_n = \text{ess} \sup_{\mathbb{P}} f_n = \sup_{\mathbb{P}} (\varphi_n)^\sim.
\]

First assume that condition (2.6) is fulfilled: then

\[
\varepsilon_n = \max \{ \| f_n - f \|, \| \mu_n - \mu \|, |a_n - a|, |b_n - b|, \frac{1}{n} \}
\]

is positive, and $\lim_{n \to \infty} \varepsilon_n = 0$. If $a = 0$ or $b = 0$, then $\int f \, d\mu = 0$, and
the desired result follows from

$$\int f_n \, d\mu_n \leq \int f_n \, d\mu_n = a_n \, b_n.$$ 

If $a$ and $b$ are positive, and $n$ is sufficiently large, then

$$c_n = \min \left\{ 1 - \sqrt{\varepsilon_n}, \frac{\varphi_n(\sqrt{\varepsilon_n})}{a_n}, \frac{\varphi(\sqrt{\varepsilon_n})}{a_n}, \frac{(\varphi_n)^-(\sqrt{\varepsilon_n})}{b + \varepsilon_n}, \frac{\varphi^-(\sqrt{\varepsilon_n})}{b_n + \varepsilon_n} \right\}$$

is well-defined and $c_n \in (0, 1)$. We have $\lim_{n \to \infty} c_n = 1$, since

$$\liminf_{n \to \infty} \varphi_n(\sqrt{\varepsilon_n}) \geq \lim_{n \to \infty} \left( \mu \{ f - \varepsilon_n \geq \sqrt{\varepsilon_n} \} - \varepsilon_n \right) = a,$$

$$\liminf_{n \to \infty} (\varphi_n)^-(\sqrt{\varepsilon_n}) \geq \lim_{n \to \infty} \sup \{ x \in \mathbb{P} : \mu \{ f - \varepsilon_n \geq x \} - \varepsilon_n \geq \sqrt{\varepsilon_n} \} = b.$$

If we show that for $n$ sufficiently large and all $x \in \mathbb{P}$

$$c_n \varphi(\frac{x}{c_n}) \leq \varphi_n(x) \leq \frac{1}{c_n} \varphi(c_n x), \quad (2.8)$$

then the desired result is obtained by using Theorem 2.13 and letting $n$ tend to infinity. If $x = 0$, then (2.8) reduces to $c_n \mu(Q) \leq \mu_n(Q) \leq \frac{1}{c_n} \mu(Q)$, which is valid when $\varepsilon_n \leq \mu(Q)^2$, since then

$$c_n \leq 1 - \sqrt{\varepsilon_n} \leq 1 - \frac{\varepsilon_n}{\mu(Q)} \quad \text{and} \quad \frac{1}{c_n} \geq \frac{1}{1 - \sqrt{\varepsilon_n}} \geq 1 + \sqrt{\varepsilon_n} \geq 1 + \frac{\varepsilon_n}{\mu(Q)}.$$

If $x \in (0, \sqrt{\varepsilon_n}]$, then (2.8) can be obtained as follows:

$$c_n \varphi(\frac{x}{c_n}) \leq c_n a \leq \varphi_n(\sqrt{\varepsilon_n}) \leq \varphi_n(x) \leq a_n \leq \frac{1}{c_n} \varphi(\sqrt{\varepsilon_n}) \leq \frac{1}{c_n} \varphi(c_n x).$$

If $x \in (\sqrt{\varepsilon_n}, (\varphi_n)^-(\sqrt{\varepsilon_n}))$ and $\varepsilon_n < 1$, then $\varphi_n(x) \geq \sqrt{\varepsilon_n}$, and (2.8) follows from

$$\varphi(x + \varepsilon_n) - \varepsilon_n \leq \varphi_n(x) \leq \varphi(x - \varepsilon_n) + \varepsilon_n,$$

because for $y \geq \sqrt{\varepsilon_n}$ the inequalities $c_n y \leq y - \varepsilon_n$ and $\frac{1}{c_n} y \geq y + \varepsilon_n$ are a consequence of $c_n \leq 1 - \sqrt{\varepsilon_n} \leq \frac{1}{1 + \sqrt{\varepsilon_n}}$. In fact, we have

$$c_n \varphi(\frac{x}{c_n}) \leq c_n \varphi(x + \varepsilon_n) \leq c_n [\varphi_n(x) + \varepsilon_n] \leq \varphi_n(x),$$

$$\frac{1}{c_n} \varphi(c_n x) \geq \frac{1}{c_n} \varphi(x - \varepsilon_n) \geq \frac{1}{c_n} [\varphi_n(x) - \varepsilon_n] \geq \varphi_n(x).$$

If $x \in [(\varphi_n)^-(\sqrt{\varepsilon_n}), \infty]$, then the first inequality of (2.8) is a direct consequence of $\frac{x}{c_n} > b$, while the second one is obvious only if $x > b_n$, but it follows from $\varphi(c_n b_n) \geq \sqrt{\varepsilon_n} \geq \varphi_n(x)$ if $x \leq b_n$. 
Now assume that condition (2.7) is fulfilled: then

$$\varepsilon_n = \max\{\|f_n - f\|, \|\mu_n - \mu\|, \frac{1}{n}\}$$

is positive, and \(\lim_{n \to \infty} \varepsilon_n = 0\). Define

$$f' = \min\{f, b\} \quad \text{and} \quad f'_n = \min\{f_n, b + 2\varepsilon_n\} I_{\{f \geq \varepsilon_n\}}.$$ 

Since \(\mu\{f' \geq \cdot\} = \varphi\), the part of the theorem already proved implies

$$\lim_{n \to \infty} \int f'_n \, d\mu_n = \int f \, d\mu,$$

because \(\|f'_n - f'\| \leq 2\varepsilon_n\),

$$\sup_P \mu_n\{f'_n \geq \cdot\} \leq \mu_n\{f'_n > 0\} \leq \mu_n\{f \geq \varepsilon_n\} \leq \varphi(\varepsilon_n) + \varepsilon_n,$$

$$\sup_P \mu_n\{f'_n \geq \cdot\} \geq \mu_n\{f'_n \geq \varepsilon_n\} \geq \mu_n\{f \geq 2\varepsilon_n\} \geq \varphi(2\varepsilon_n) - \varepsilon_n,$$

$$\text{ess} \mu_n \sup f'_n \leq b + 2\varepsilon_n,$$

$$\text{ess} \mu_n \sup f'_n \geq \sup\{x \in P : \mu_n\{f'_n \geq x\} \geq \varepsilon_n\} \geq$$

$$\geq \sup\{x \in P : \varphi(x + \varepsilon_n) \geq 2\varepsilon_n\} \geq \varphi^{-1}(2\varepsilon_n) - \varepsilon_n.$$ 

So, defining \(A = \{f \geq \varepsilon_n\}\) and \(B = \{f_n \leq b + 2\varepsilon_n\}\), the desired result follows from

$$\int f'_n \, d\mu_n \leq \int f_n \, d\mu_n \leq$$

$$\leq \int f_n I_{A \cap B} \, d\mu_n + \int f_n I_{Q \setminus A} \, d\mu_n + \int f_n I_{A \setminus B} \, d\mu_n \leq$$

$$\leq \int f'_n I_{A \cap B} \, d\mu_n + \mu_n(Q) \sup_{Q \setminus A} f_n + \mu_n(Q \setminus B) \sup f_n \leq$$

$$\leq \int f'_n \, d\mu_n + \mu(Q) + \varepsilon_n \leq 2\varepsilon_n + \varepsilon_n \sup f + \varepsilon_n$$

by letting \(n\) tend to infinity. \(\square\)

**Example 2.24.** Let \(\mu, \nu, \) and \(l_d\) be the measures and the functions on \([0, 1]\) defined in Examples 2.1, 2.2, and 2.7, respectively. Since the mapping \(d \mapsto l_d\) on \([0, 1]\) is continuous (with respect to the supremum norm), and \(\sup_P \mu\{l_d \geq \cdot\} = 1\) and \(\text{ess} \mu \sup l_d = \max\{d, (1 - d)^2\}\) for all \(d \in [0, 1]\), Theorem 2.23 implies that the function \(d \mapsto \int l_d \, d\mu\) on \([0, 1]\) is continuous.
for all regular integrals on classes $\mathcal{M} \ni \mu$. It also implies that the function $d \mapsto \int l_d \, d\nu$ on $[0, 1]$ is continuous for all quasi-subadditive, regular integrals on classes $\mathcal{M} \ni \nu$. But since $\lim_{d \uparrow 1} \sup \nu\{l_d \geq \cdot\} \neq \sup \nu\{l_1 \geq \cdot\}$, Theorem 2.23 does not imply that $d \mapsto \int l_d \, d\nu$ is continuous at 1 for all regular integrals on classes $\mathcal{M} \ni \nu$; in fact, the results of Example 2.14 show that $d \mapsto \int^r l_d \, d\nu$ is not continuous at 1. 

\[ \Box \]

### 2.2.6 Choquet Integral

An important property of the Choquet integral is the additivity of the corresponding functional $F_C$ on $\mathcal{N} \mathcal{I}_{\mathbb{P}}$, implying in particular the additivity of the integral with respect to measures:

\[
\int C f \, d(\mu + \nu) = \int C f \, d\mu + \int C f \, d\nu,
\]

for all pairs of monotone measures $\mu, \nu$ on a set $\mathcal{Q}$, and all functions $f : \mathcal{Q} \to \mathbb{P}$. The next theorem states another important property of the Choquet integral, but first we need some definitions.

Two functions $f, g : \mathcal{Q} \to \mathbb{P}$ on a set $\mathcal{Q}$ are said to be comonotonic if $f(q) < f(q') \Rightarrow g(q) \leq g(q')$ for all $q, q' \in \mathcal{Q}$.

Let $\mathcal{M}$ be a class of measures (each $\mu \in \mathcal{M}$ is a measure on some set $\mathcal{Q}_\mu$). An integral on $\mathcal{M}$ is said to be **comonotonic additive** if

\[
\int (f + g) \, d\mu = \int f \, d\mu + \int g \, d\mu \quad \text{for all } \mu \in \mathcal{M}
\]

and all comonotonic $f, g \in \mathbb{P}_{\mathcal{Q}_\mu}$.

**Theorem 2.25.** The Choquet integral is comonotonic additive on the class of all monotone measures.

**Proof.** The additivity of the Choquet integral for comonotonic, finite functions is proved for instance in Denneberg (1994, Proposition 5.1); we can extend this result as follows. If $f$ and $g$ are comonotonic, then there are no $q, q'$ such that $q \in \{f < \infty\} \cap \{g = \infty\}$ and $q' \in \{f = \infty\} \cap \{g < \infty\}$, and therefore the inclusion $A = \{f < \infty\} \subseteq \{g < \infty\}$ can be assumed without loss of generality. If $\mu(\mathcal{Q}_\mu \setminus A) > 0$, then the desired result is obvious, since both the integrals of $f$ and of $(f + g)$ are infinite. If $\mu(\mathcal{Q}_\mu \setminus A) = 0$, then
define \( b = \sup_A (f + g) \), and note that \( \sup_A g \leq \inf_{Q \setminus A} g \), since \( f \) and \( g \) are comonotonic. Proposition 4.5 of Denneberg (1994) implies that there are two continuous, nondecreasing functions \( u, v : [0, b] \rightarrow \mathbb{P} \) such that

\[
  u + v = \text{id}_{[0,b]}, \quad f_A | = u \circ (f_A + g_A), \quad \text{and} \quad g_A | = v \circ (f_A + g_A).
\]

Define \( h = \min\{f + g, b\} \): Proposition 4.1 of the same book implies that

\[
  \int (u \circ h) \, d\mu + \int (v \circ h) \, d\mu = \int h \, d\mu.
\]

The desired result follows from this equality, since \( u \circ h = \min\{f, \sup_A f\} \) and \( v \circ h = \min\{g, \sup_A g\} \), and for all functions \( f' : Q \mu \rightarrow \mathbb{P} \)

\[
  \int \min\{f', \sup_A f'\} \, d\mu = \int f' \, d\mu,
\]

because \( \mu(Q \setminus A) = 0 \), and therefore the distribution functions of \( f' \) and \( \min\{f', \sup_A f'\} \) are equal.

Schmeidler (1986) proved that for finite, monotonic measures and bounded functions, the Choquet integral is characterized by monotonicity and comonotonic additivity. But the next example shows that this characterization fails when the measures or the functions are unbounded (see also Wakker, 1993).

**Example 2.26.** Consider the integral that associates the value

\[
  \| f \|_\mu = \left\{ \begin{array}{ll}
  \int f \, d\mu & \text{if } \int f \, d\mu < \infty, \\
  \infty & \text{if } \int f \, d\mu = \infty,
  \end{array} \right.
\]

(2.9)

to each pair consisting of a monotonic measure \( \mu \) on a set \( Q \) and of a function \( f : Q \rightarrow \mathbb{P} \). It can be easily proved that this integral is regular and symmetric on the class of all monotonic measures, and subadditive on the class of all 2-alternating measures. If the functions \( f, g : Q \rightarrow \mathbb{P} \) are comonotonic, then for all \( x \in \mathbb{P} \) one of the two sets \( \{f \geq x\}, \{g \geq x\} \) is a subset of the other, and therefore

\[
  \mu\{f + g \geq x\} \leq \mu\{\{f \geq \frac{x}{2}\} \cup \{g \geq \frac{x}{2}\}\} = \max\{\mu\{f \geq \frac{x}{2}\}, \mu\{g \geq \frac{x}{2}\}\}.
\]

Hence, if \( \int f \, d\mu \) and \( \int g \, d\mu \) are finite, then so is \( \int (f + g) \, d\mu \); and thus the integral defined by (2.9) is comonotonic additive on the class of all monotonic measures.
It is interesting to note that this integral can be represented (in the sense considered at the end of Subsection 2.2.3) by a finitely additive measure on (the Borel sets of) $\mathbb{F}^2$. That is, homogeneity does not suffice to characterize the Choquet integral among the ones representable by finitely additive measures on (the Borel sets of) $\mathbb{F}^2$; and the countable additivity of the measure is not necessary to obtain the comonotonic additivity of the integral (see also Klement, Mesiar, and Pap, 2004).

An integral on $\mathcal{M}$ is said to be **translation equivariant** if

$$\int (f + c) \, d\mu = \int f \, d\mu + c \, \mu(\mathcal{Q}_\mu) \quad \text{for all } \mu \in \mathcal{M}, \text{ all } c \in \mathbb{P},$$

and all $f \in \mathbb{F}^{\mathcal{Q}_\mu}$.

Since two functions are certainly comonotonic when one is constant, all comonotonic additive, homogeneous integrals are translation equivariant. By contrast, the next example shows that a maxitive, homogeneous integral can be translation equivariant only on limited classes of finitely maxitive measures.

**Example 2.21.** Let $\mu$ be a finitely maxitive, finite measure on a set $\mathcal{Q}$. If there is a set $A \subset \mathcal{Q}$ such that $0 < \mu(A) < \mu(\mathcal{Q})$, then all maxitive, homogeneous integrals on classes $\mathcal{M} \ni \mu$ satisfy

$$\int (I_A + 1) \, d\mu = \max\{2 \mu(A), \mu(\mathcal{Q} \setminus A)\} < \mu(A) + \mu(\mathcal{Q}) = \int I_A \, d\mu + \mu(\mathcal{Q}).$$

The equality satisfied by a translation equivariant integral on $\mathcal{M}$ can also be considered as a way to define the integral of functions $f : \mathcal{Q}_\mu \to \mathbb{R}$ with respect to measures $\mu \in \mathcal{M}$, when $\inf f$ and $\mu(Q_\mu)$ are finite. We obtain

$$\int f \, d\mu = \int (f - \inf f) \, d\mu + \mu(Q_\mu) \inf f; \quad (2.10)$$

and if the translation equivariant integral on $\mathcal{M}$ is regular, subadditive, or comonotonic additive, then so is the extended integral (when the definition of distribution function is extended from $\mathbb{P}$ to $\mathbb{R}$). It is important to note that in general the resulting extended integral does not satisfy the equality $\int (-f) \, d\mu = -\int f \, d\mu$: for example, if $A \subseteq \mathcal{Q}_\mu$ is a set, then we have

$$\int (-I_A) \, d\mu = \int (1 - I_A) \, d\mu - \mu(Q_\mu) = \int I_{\mathcal{Q}_\mu \setminus A} \, d\mu - \mu(Q_\mu) = -\int I_A \, d\mu.$$
The next theorem (proved for instance in Denneberg, 1994, Proposition 5.1) states that in the case of the Choquet integral this relation holds for all bounded functions (not only for the indicator functions).

**Theorem 2.28.** If \( \mu \) is a finite, monotonic measure on a set \( \mathcal{Q} \), and the function \( f : \mathcal{Q} \to \mathbb{R} \) is bounded, then the extension of the Choquet integral by means of equality (2.10) satisfies

\[
\int^C (-f) \, d\mu = - \int^C f \, d\bar{\mu}.
\]

There is another way to extend a monotonic, translation equivariant (or homogeneous) integral on \( \mathcal{M} \) to functions \( f : \mathcal{Q}_\mu \to \mathbb{R} \), when \( \inf f \) and \( \mu(\mathcal{Q}_\mu) \) are finite: it is to define

\[
\int f \, d\mu = \int \max\{f, 0\} \, d\mu - \int \max\{-f, 0\} \, d\mu. \tag{2.11}
\]

The right-hand side is certainly well-defined, and the resulting extended integral satisfies the equality \( \int (-f) \, d\mu = - \int f \, d\mu \), but in general it does not maintain the translation equivariance and other possible properties of the original integral, such as subadditivity or maxitivity.

The extension of the Choquet integral by means of equality (2.10) usually keeps the name “Choquet integral”, while the extension by means of equality (2.11) is often called “Šipoš integral” (it was introduced by Šipoš, 1979). Thanks to Fubini’s theorem, it can be proved that if \( \mu \) is countably additive, then both extended versions of the Choquet integral reduce to the Lebesgue integral (when either \( f \geq 0 \), or both \( \inf f \) and \( \mu(\mathcal{Q}_\mu) \) are finite). So the Choquet integral generalizes the Lebesgue integral to nonadditive measures, and has strong properties (subadditivity and comonotonic additivity) on a broad class of measures (the class of all 2-alternating measures), containing all finitely additive and all finitely maxitive measures. By contrast, the Shilkret integral does not generalize the Lebesgue integral, and has strong properties (subadditivity and maxitivity) only on the class of all finitely maxitive measures. The Choquet integral is thus a very general and powerful integral, while the Shilkret integral is specialized on maxitive measures: in particular, on completely maxitive measures it reduces to the very simple form stated in Theorem 2.6. The main drawback of the Shilkret integral is the lack of translation equivariance; this implies in particular that it cannot be extended in a satisfactory way to functions taking also negative values.
The likelihood function can be considered as a description of uncertain knowledge about the statistical models: in the present chapter this kind of description is studied, under the name of “relative plausibility”, from both the static and the dynamic points of view. In particular, it is compared with the Bayesian, probabilistic description of uncertain knowledge about the models, and it is connected to the MPL criterion by a representation theorem for preferences between randomized decisions. The statistical models and their relative plausibility build a hierarchical model with two levels describing two different kinds of uncertain knowledge: this hierarchical model is briefly compared with the classical, Bayesian, and imprecise probability models.

3.1 Description of Uncertain Knowledge

A relative plausibility measure $rp$ on a set $Q$ is an equivalence class of completely maxitive measures on $Q$ with respect to the equivalence relation $\alpha$. Thus, to be precise, $rp$ is not a measure (in the sense introduced in the preceding chapter) but a class of proportional measures; however, this slight abuse of terminology is coherent with the abuse of notation we commit by using $rp$ to indicate one of its representatives $\mu$, when the choice of $\mu \in rp$ is irrelevant. For example, we can say that for each function $f : Q \to \mathbb{P}$ there is a unique relative plausibility measure $f^\dagger$ on $Q$ such that $(f^\dagger)^\dagger \propto f$. On the other hand, when there is no possibility of confusion, we use $rp \circ t^{-1}$ and $rp|_A$ as abbreviations for $[(rp \circ t^{-1})^\dagger]^\dagger$ and $[(rp|_A)^\dagger]^\dagger$, respectively (where $t$ is a function on $Q$, and $A$ is a subset of $Q$). A relative plausibility measure $rp$ on $Q$ is said to be nondegenerate
if \( rp(Q) \in \mathbb{P} \) (that is, if \( rp \) is finite and nonzero); in this case, \( rp \) denotes the (unique) normalized representative of \( rp \):

\[
 rp(A) = \frac{rp(A)}{rp(Q)} \quad \text{for all } A \subseteq Q.
\]

Consider a statistical decision problem described by a loss function \( L \) on \( \mathcal{P} \times \mathcal{D} \), and let \( lik \) be the likelihood function on \( \mathcal{P} \) induced by the observed data; Theorem 2.6 implies that the MPL criterion can be expressed as follows:

\[
 \text{minimize } \int_S l_d \, d\,lik^\uparrow.
\]

The likelihood function \( lik \) measures the relative plausibility of the models in \( \mathcal{P} \) in the light of the observed data alone, and \( lik^\uparrow \) extends this relative measure to the subsets of \( \mathcal{P} \); it is the relative plausibility measure on \( \mathcal{P} \) induced by the observed data. This extension of \( lik \) to the subsets of \( \mathcal{P} \) by means of the supremum agrees with the profile likelihood: if \( \mathcal{G} \) is a set, and \( g : \mathcal{P} \rightarrow \mathcal{G} \) is a mapping, then \( lik^\uparrow \circ g^{-1} = (lik_g)^\uparrow \); the extension is thus in accordance with the likelihood-based inference methods: in particular, if \( lik^\uparrow \) is nondegenerate, then \( lik^\uparrow(\mathcal{H}) = LR(\mathcal{H}) \) for all \( \mathcal{H} \subseteq \mathcal{P} \).

Hence a relative plausibility measure \( rp \) on \( \mathcal{P} \) can be interpreted as a description of uncertain knowledge about the models in \( \mathcal{P} \), based on the description \( rp^\downarrow \) of the relative plausibility of the elements of \( \mathcal{P} \). It is strictly related to other descriptions of uncertain knowledge appearing in the literature: "degrees of potential surprise" (Shackle, 1949), "support by eliminative induction" (Cohen, 1966), "consonant plausibility functions" (Shafer, 1976), and possibility measures (if \( rp \) is nondegenerate, then \( rp \) is a normalized possibility measure on \( \mathcal{P} \)). The models in \( \mathcal{P} \) are mathematical representations of some aspects of the reality under consideration; no model is true, but some models represent the observed reality better than others. The relative plausibility of a model is its relative ability to represent the observed reality; the ability of a set of models is interpreted as the ability of its best element. In the Bayesian approach, the uncertain knowledge about the models in \( \mathcal{P} \) is described by an averaging probability measure \( \pi \) on \( (\mathcal{P}, \mathcal{C}) \) (or by a set \( \Gamma \) of such probability measures), where \( \mathcal{C} \) is a \( \sigma \)-algebra of subsets of \( \mathcal{P} \). The interpretation of a probability measure on \( (\mathcal{P}, \mathcal{C}) \) is based on the assumption that exactly one of the models in \( \mathcal{P} \) is "true": \( \mathcal{P} \) is an exhaustive set of mutually exclusive possible "states of the world" (at least when \( \mathcal{C} \) contains all singletons of \( \mathcal{P} \)). The exhaustivity of \( \mathcal{P} \) is not implied by the interpretation of a relative plausibility
measure \( rp \) on \( P \) (because \( rp \) contains no evaluation of the “absolute plausibility” of \( P \)), although some sort of “actual exhaustivity” is implied by the fact that we do not consider models that are not in \( P \). The same situation can be obtained for the Bayesian approach by considering a “relative probability measure” on \( P \), in analogy with \( rp \): the Bayesian criterion for statistical decision problems is unaffected by the “absolute probability” of \( P \) (Hartigan, 1983, developed a Bayesian approach with non-normalized probability measures). The basic difference between the interpretations of relative plausibility and probability is that besides the exhaustivity, neither the mutual exclusivity of the models in \( P \) is implied by the interpretation of a relative plausibility measure on \( P \), in the following sense. The probability of a model is the probability that it is the “true” one, and truth is exclusive; but the ability to represent the observed reality is not exclusive: different models can coexist as satisfactory representations of the reality. Models are mental constructions, not “states of the world”, even though in particular situations (such as drawing from an urn containing colored balls) a model can be so strongly associated with a state of the world that they can almost be identified. The fundamental qualitative difference between relative plausibility and probability (see for example Dubois, 1988) is that if \( H_1 \), \( H_2 \), and \( H \) are subsets of \( P \) (in the case of probability they must be elements of \( C \)) such that \( rp(H_1) > rp(H_2) \), and \( H \) is disjoint from both \( H_1 \) and \( H_2 \), then the inequality \( rp(H_1 \cup H) > rp(H_2 \cup H) \) holds if and only if \( rp(H) < rp(H_1) \) (while in the case of probability it always holds). In fact, the ability of a set \( H \) of models to represent the observed reality does not increase when it is enlarged by models that are not more satisfactory than the ones in \( H \): if \( rp(H_1) \leq rp(H) \), then \( rp(H_1 \cup H) = rp(H) \).

It is important to note that the identification of the set theoretic operations on subsets of \( P \) with the connectives of propositional logic is based on the assumption that exactly one of the models in \( P \) is “true”. Without this assumption the identification fails; for example, \( rp\{P, P'\} \) is not the (relative) plausibility of “\( P \) or \( P' \)” (in fact, “\( P \) or \( P' \)” means “\( P \) is the true one or \( P' \) is the true one”): it is simply the (relative) plausibility of the set \( \{P, P'\} \). The relative plausibility measure \( lik^{\uparrow} \) on \( P \) is a convenient extension of the (relative) likelihood \( lik \) from the elements of \( P \) to the subsets of \( P \) (in accordance with the likelihood-based inference methods), and since we do not assume that one of the models in \( P \) is “true”, the use of \( lik^{\uparrow} \) does not contradict the assertion of Fisher (1930) that in general “the likelihood of \( P \) or \( P' \)” has no meaning (“you do not know what it is until you know which is meant”).
3.1.1 Representation Theorem

Let $\preceq_2$ be a binary relation on a set $S$. The binary relations $\sim$ and $\prec$ on $S$ are defined as follows on the basis of $\preceq_2$:

\begin{align*}
  a \sim b & \iff a \preceq_2 b \text{ and } b \preceq_2 a \quad \text{for all } a, b \in S, \\
  a \prec b & \iff a \preceq_2 b \text{ and not } b \preceq_2 a \quad \text{for all } a, b \in S.
\end{align*}

The relation $\preceq_2$ is said to be a **total preorder** on $S$ if it fulfills the following two conditions:

\begin{align*}
  &a \preceq_2 b \text{ or } b \preceq_2 a \quad \text{for all } a, b \in S, \\
  &a \preceq_2 b \text{ and } b \preceq_2 c \implies a \preceq_2 c \quad \text{for all } a, b, c \in S.
\end{align*}

If $\preceq_2$ is a total preorder on $S$, then $\sim$ is an equivalence relation on $S$, and for all $a, b \in S$ exactly one of the following three cases applies: $a \prec b$, $b \prec a$, or $a \sim b$. A function $f : S \to \mathbb{R}$ is said to **represent** the relation $\preceq_2$ on $S$ if

\[ a \preceq_2 b \iff f(a) \leq f(b) \quad \text{for all } a, b \in S.\]

If $\preceq_2$ is represented by some function, then it is a total preorder.

In the Bayesian approach, the use of probability measures as a description of uncertain knowledge is usually justified in terms of representation theorems for qualitative preference relations. The most famous representation theorem is the one of Savage (1954): it gives necessary and sufficient conditions for a preference relation $\preceq_2$ on $C^Q$ (where $Q$ is a set of “states of the world” and $C$ is a set of consequences) to be represented by the mapping $l \mapsto E_\pi(v \circ l)$ for a normalized, finitely additive measure $\pi$ on $Q$ (the countable additivity is given up, and so measurability problems are avoided) and a bounded function $v : C \to \mathbb{R}$ (a quantitative evaluation of the consequences). The elements of $C^Q$ are functions $l : Q \to C$ associating a consequence to each state of the world; that is, each $l$ can be considered as the uncertain consequence of a particular decision. By assuming that the uncertain consequence summarizes all important aspects of a decision, we identify the decisions with their uncertain consequences, interpreting the elements of $C^Q$ as decisions, and the relation $\preceq_2$ as a description of preferences between them.

Consider a preference relation $\preceq_2$ on $D \subseteq C^Q$; a representation theorem gives necessary and sufficient conditions for the preferences between
decisions to be represented by a particular kind of functional on $\mathcal{D}$. We shall obtain a representation theorem where the functional is of the form $l \mapsto \int Q(v \circ l) \, drp$ for a nondegenerate relative plausibility measure $rp$ on $Q$ and a function $v : C \to [0, \infty)$. We interpret $v \circ l$ as a quantitative evaluation of the uncertain loss incurred by making the decision $l \in \mathcal{D}$, and accordingly we interpret the expression $l \preceq l'$ (where $l, l' \in \mathcal{D}$) as "$l'$ is not preferred to $l$". For the sake of generality, we formulate the representation theorem with $Q$ as a generic set of alternatives determining the consequences of the decisions; but when interpreting the conditions, $Q$ can be regarded as a set of statistical models. The theorem describes the kind of preference relation on $\mathcal{D}$ induced by the MPL criterion, when applied (with respect to some nondegenerate relative plausibility measure $rp$ on $Q$) to a decision problem described (for some function $v : C \to [0, \infty)$) by the loss function $L_v : (q, l) \mapsto v[l(q)]$ on $Q \times \mathcal{D}$.

Representation theorems are often considered as justifications for particular approaches (such as the Bayesian one), but even if the postulated conditions for the preferences between decisions are reasonable, the conclusion that the only reasonable preference relations are represented by a particular kind of functional is based on the wrong assumption that reasonable conditions are noncontradictory (compare with Chernoff, 1954). Moreover, the postulated conditions are usually motivated only by abstract consideration of simple examples; hence their application beyond those simple examples is in fact unmotivated, and can have unexpected consequences (see for example Ellsberg, 1961). For these reasons, the representation theorem that we shall obtain should be interpreted as a mathematical theorem showing the distinguishing properties of the preferences based on the MPL approach, not as a justification for this approach. The same considerations apply to the axiomatic justifications for the use of the likelihood function as a description of the information content of statistical data: see Birnbaum (1962, 1969) and Evans, Fraser, and Monette (1986).

The first condition postulated by Savage for the preferences between decisions is that the preference relation is a total preorder on $C^Q$. This is a necessary condition for the preferences to be represented by a functional, but it is certainly too strong for the real-world preferences of an individual, also because it is assumed that $\mathcal{D} = C^Q$ (that is, $\mathcal{D}$ is sufficiently wide to encompass all possible uncertain consequences). Since we do not consider the representation theorem as a justification for the MPL approach, we can for simplicity assume that $\mathcal{D} = C^Q$, but we will later consider on
which subsets of $C^Q$ the preference relation suffices to determine its representation by the mapping $l \mapsto \int S(v \circ l) \, dp$ for a nondegenerate relative plausibility measure $rp$ on $Q$ and a function $v : C \rightarrow [0, \infty)$.

If we assume that the preferences between the consequences in $C$ are not affected by the particular case $q \in Q$ considered, and correspond to the preferences between the constant (uncertain) consequences in $C^Q$, then the following condition of monotonicity is hardly debatable. A binary relation $\succeq$ on $C^Q$ is said to be **monotonic** if

$$l(q) \succeq l'(q) \text{ for all } q \in Q \Rightarrow l \succeq l' \text{ for all } l, l' \in C^Q.$$  

If the preference relation $\succeq$ on $C^Q$ is induced by the MPL criterion (applied to the decision problem described by the loss function $L_v$ on $Q \times C^Q$), then the preference relation on $C$ is represented by $v$, and $\succeq$ is monotonic, since the Shilkret integral with respect to completely maxitive measures is homogeneous and monotonic (Theorem 2.3). If $l$ and $l'$ satisfy the left-hand side of the above condition of monotonicity, and $\{l \prec l'\}$ is not empty, then we can say that $l$ dominates $l'$; the monotonicity of the preference relation assures that if $l'$ is dominated by $l$, then $l'$ is not preferred to $l$. The monotonicity is part of Savage’s “sure-thing principle” (at least when $l$ and $l'$ have finite range; see Subsection 4.1.2); but this principle states also that if $l$ dominates $l'$, then $l$ must be preferred to $l'$, at least when $\{l \prec l'\}$ is not impossible (that is, when it has positive “probability”). This condition is satisfied by the preferences induced by the MPL criterion if we discard the dominated decisions: the resulting preference relation $\sim$ on $C^Q$ is such that $l \prec l'$ if and only if either $v \circ l \leq v \circ l'$ and $v \circ l \neq v \circ l'$, or $l$ is preferred to $l'$ on the basis of the MPL criterion applied to the decision problem described by $L_v$. It can be easily proved that in general this preference relation is not a total preorder on $C^Q$, since it is not transitive (but it is important to note that the strict preference relation $\prec$ is transitive); this shows that to be reasonable and useful, a preference relation does not need to be a total preorder.

The problem of the “sure-thing principle” of Savage is that the premise “when $\{l \prec l'\}$ is not impossible” cannot be expressed directly in terms of preferences: therefore the principle was replaced with other conditions. In particular, the second condition postulated by Savage for the preferences between decisions states that the preference between two functions $l, l' \in C^Q$ does not depend on the values taken by the functions on the set $\{l = l'\}$, in the sense that the preference is left unchanged when the two
functions are modified in the same way on \( \{l = l'\} \). This condition is in
general not satisfied by the preferences induced by the MPL criterion, also
if we discard the dominated decisions: it is possible that by modifying the
functions on \( \{l = l'\} \) a strict preference becomes an equivalence, or vice
versa (but it is impossible that a strict preference is reversed). This dis¬
cord between the preferences induced by the MPL criterion and the second
condition postulated by Savage is strictly related to the fundamental qual¬
itative difference between relative plausibility and probability, considered
at the beginning of the present section: for a relative plausibility measure
\( r_p \), the additional consideration of \( \mathcal{H} \) can neutralize the preference between
\( \mathcal{H}_1 \) and \( \mathcal{H}_2 \).

The preferences induced by the MPL criterion present the same kind
of discord with the fourth condition postulated by Savage as with the
second one, while they fulfill the third and fifth ones (the first five pos¬
tulated conditions imply the existence of a qualitative probability on \( \mathcal{Q} \)).
To obtain preferences satisfying the second and fourth conditions postu¬
lated by Savage, it suffices to restrict attention to the subset \( \{l \neq l'\} \) of
\( \mathcal{Q} \) when applying the MPL criterion for discriminating between the deci¬
sions \( l, l' \in \mathcal{C} \mathcal{Q} \), but it can be easily proved that in general the resulting
preference relation is not a total preorder on \( \mathcal{C} \mathcal{Q} \), since it is not transitive
(however, in this case too the strict preference relation is transitive). It
is interesting to note that on the end pages of Savage (1954) the author
restated in a slightly different form the seven postulated conditions that
are scattered through the book: even though he asserted that “the logical
content of each postulate is left unaltered”, the preferences induced by the
MPL criterion satisfy all the conditions appearing on the end pages, apart
from the (highly debatable) sixth one. This shows that minimal changes in
the postulated conditions can have important consequences on the implied
representations, and thus points out how unsuitable the representation the¬
orems are as justifications for particular approaches (in the second edition
of the book, published in 1972, the second condition appearing on the end
pages has been corrected, but the fourth one is still flawed).

Ellsberg (1961) showed that the second condition postulated by Savage
is at variance with the propensity toward hedging bets, which reveals strict
uncertainty aversion. Schmeidler (1989) proved a representation theorem
where the functional is of the form \( l \mapsto \int_{\mathcal{C}} (v \circ l) \, d\mu \) for a normalized,
monotonic measure \( \mu \) on \( \mathcal{Q} \) (a “nonadditive probability” measure) and
a bounded function \( v : \mathcal{C} \to \mathbb{R} \) (a quantitative evaluation of the conse-
quences): the preferences reveal uncertainty aversion if and only if \( \mu \) is 2-alternating (they reveal uncertainty neutrality when \( \mu \) is finitely additive). A strictly related representation theorem for the preferences induced by the \( \Gamma \)-minimax criterion was proved by Gilboa and Schmeidler (1989). Both these theorems use the framework of the representation theorem for the preferences induced by the Bayesian criterion proved by Anscombe and Aumann (1963) and generalized by Fishburn (1970), in which the set \( C \) is enlarged by including randomized consequences (that is, the existence of objective probabilities is assumed). We will use this framework too, but with an important difference: these representation theorems are based on the one of von Neumann and Morgenstern (1944), which considers the consequences of economic decisions, while we are interested in statistical decisions. The main difference is that statistical decision problems possess an additional element: the concept of correct decision; we assume thus the existence of the consequence \( c_0 \in C \) of a correct decision.

For a set \( C \ni c_0 \) of consequences, we define

\[
R(C, c_0) = \{ r \in [0, 1]^C : |\{ r > 0 \}| \leq 1, \quad r(c_0) = 0 \},
\]

and we interpret each element \( r = p I_{\{c\}} \) of \( R(C, c_0) \) as the following randomized consequence: \( c \) with probability \( p \), and \( c_0 \) with probability \( 1 - p \). The set \( C \) of non-randomized consequences can thus be identified with a subset of \( R(C, c_0) \): each \( c \in C \setminus \{ c_0 \} \) can be identified with \( \hat{c} = I_{\{c\}} \), and \( c_0 \) can be identified with \( \hat{c}_0 = 0 \). In our representation theorem, a total preorder \( \preceq \) on the set \( R(C, c_0) \) is considered; that is, the set \( C \) is enlarged by including the randomized consequences of the above form (for all \( p \in (0, 1) \) and all \( c \in C \)). The first condition of our representation theorem states the peculiar role of the consequence \( c_0 \) of a correct decision (that is, no consequence can be preferred to \( c_0 \)):

\[
\hat{c}_0 \preceq \hat{c} \quad \text{for all} \quad c \in C. \tag{3.1}
\]

In the framework of Anscombe and Aumann, preferences between decisions in \( R(C) \) are considered, where

\[
R(C) = \{ r \in [0, 1]^C : |\{ r > 0 \}| < \infty, \quad \sum_{c \in C} r(c) = 1 \}
\]

is the set of all the randomized consequences that are mixtures of a finite number of consequences. The use of \( R(C) \) instead of \( R(C, c_0) \) in our theorem would pose no problem (a condition should be slightly modified), but
thanks to the peculiar role of $c_0$, it suffices to consider the preferences between decisions in the set $R(C, c_0) \subset R(C, c_0)$, which can be identified with a subset of $R(C) \subset R(C)$. Moreover, $C \subset C$ can be identified with a subset of $R(C, c_0) \subset R(C, c_0)$, and if $l'' \in C$ is identified with $l' \in R(C, c_0)$, and $p \in [0, 1]$, then $p l'$ can be interpreted as the following randomized decision: $l''$ with probability $p$, and $c_0$ with probability $1 - p$. We shall see that the preference relation on the set of randomized decisions of the form $p l'$ (where $l'$ can be identified with an element of $C \subset C$) suffices to determine its representation by the mapping $l \mapsto \int_{S} (v \circ l) \, drp$ for a nondegenerate relative plausibility measure $r \circ l \subset [0, \infty)$, where $v : R(C, c_0) \to [0, \infty)$ is the homogeneous extension of $v'$, in the sense that $v(p \circ c) = pv'(c)$ for all $p \in (0, 1)$ and all $c \in C$.

Roughly speaking, each one of the three representation theorems of Anscombe and Aumann, Schmeidler, and Gilboa and Schmeidler imposes three fundamental conditions on the total preorder $\preceq$: independence, continuity, and monotonicity. The independence condition postulated by Anscombe and Aumann can be formulated as follows: $l \preceq l'$ if and only if $p l + (1 - p) l'' \preceq p l' + (1 - p) l''$, where $p \in (0, 1)$; it can be interpreted as stating that the preference between two decisions $l$ and $l'$ does not change if these are randomized by mixing them in the same way with a third decision $l''$: the influence of $l''$ on the evaluation of the mixed decisions is independent of $l$ or $l'$. This is contradicted by the propensity toward hedging bets (when $l \sim l'$, the preference $p l + (1 - p) l' \prec l$ can be explained by strict uncertainty aversion: there is a preference for risk differentiation); therefore Schmeidler relaxed the independence condition by assuming that $l$, $l'$, and $l''$ are pairwise comonotonic ($l$ and $l'$ are comonotonic if there are no $q, q' \in Q$ such that $l(q) \prec l(q')$ and $l'(q') \prec l'(q)$). Gilboa and Schmeidler relaxed the independence condition even more by assuming instead that $l''$ is constant; they called this condition "certainty-independence", and needed to complement it with the one of "uncertainty aversion", which states that $l \sim l'$ implies $p l + (1 - p) l' \preceq l$. The second condition of our representation theorem corresponds to the certainty-independence for the only case in which it is defined in our framework (that is, for $l'' = c_0$):

$$l \preceq l' \iff p l \preceq p l' \quad \text{for all } p \in (0, 1) \text{ and all } l, l' \in R(C, c_0).$$

(3.2)

This condition can be interpreted as stating that if we shall have to decide with probability $p$ (while with probability $1 - p$ we shall know the correct decision), then our preferences must be the same as if we certainly had to decide; the condition can also be considered as expressing the "scale invari-
The continuity condition is the same for the three representation theorems cited above, and can be formulated as follows: if \( l < l' < l'' \), then there are \( p, p' \in (0, 1) \) such that \( pl + (1 - p) l'' < p'l + (1 - p') l'' \). This condition can be interpreted as stating that there are randomized decisions obtained by mixing \( l \) and \( l'' \) that are as close as we want (on the preference scale) to \( l \) or \( l'' \); it is reasonable only if the uncertain consequences are in some sense “bounded” (see also Wakker, 1993). In our framework, when the monotonicity and the conditions (3.1) and (3.2) are satisfied, the uncertain consequences can not be preferred to \( \hat{c}_0 \), and the only case in which the above continuity condition is defined and not vacuous is thus for \( Z = \infty \). In this case, the second implied preference of this condition has no problems related to unbounded uncertain consequences, and the implication is equivalent to the following one:

\[
p l \not\sim l' \quad \text{for all} \quad p \in (0, 1) \quad \Rightarrow \quad l \not\sim l' \quad \text{for all} \quad l, l' \in R(C, c_0). \quad (3.3)
\]

In our representation theorem, we do not need to assume that the uncertain consequences are bounded above, but for simplicity we can assume that there are no “infinitely unfavorable” consequences (otherwise we should impose an additional condition stating their equivalence, to avoid “transfinite evaluations”): the uncertain consequences can be unbounded only if \( C \) is infinite. The following condition is strictly related to the first implied preference of the above continuity condition when \( l'' = \hat{c} \) (and \( l = \hat{c}_0 \)); it assures that the consequences \( c \in C \) are bounded and that there are no “infinitesimal evaluations”:

\[
l \not\sim p \hat{c} \quad \text{for all} \quad p \in S, \quad \text{and} \quad \inf S = 0 \quad \Rightarrow \quad l \sim \hat{c}_0
\]

\[
\text{for all} \quad c \in C, \quad \text{all} \quad S \subseteq (0, 1), \quad \text{and all} \quad l \in R(C, c_0). \quad (3.4)
\]

In the three representation theorems cited above, the monotonicity of the preference relation is assumed; in our theorem it is replaced by the following stronger condition (the monotonicity corresponds to the special
3.1 Description of Uncertain Knowledge

Let $C$ and $Q$ be two nonempty sets, and let $c_0 \in C$. A total preorder $\succeq$ on $R(C,c_0) \otimes$ fulfills the conditions (3.1), (3.2), (3.3), (3.4), and (3.5) if and only if there are a nondegenerate relative plausibility measure $\rho p$ on $Q$ and a homogeneous function $v : R(C,c_0) \rightarrow [0, \infty)$ such that $\succeq$ is represented by the mapping $l \mapsto \int_{S} (v \circ l) \, d\rho p$; the function $v$ is unique up to a (positive) multiplicative constant, and $\rho p$ is unique when $v$ is not constant.

**Proof.** The "if" part is simple: since $v$ is finite and homogeneous, and the Shilkret integral is homogeneous (Theorem 2.3), we have

$$\int_{S} [v \circ (pl)] \, d\rho p = p \int_{S} (v \circ l) \, d\rho p \quad \text{and} \quad \int_{S} (v \circ \hat{c}) \, d\rho p = v(\hat{c}) \rho p(Q).$$
The first four conditions follow easily from these equalities, and for the fifth one we obtain \( v \circ l \leq \sup_{q \in Q} v \circ l_q \): now it suffices to exploit the monotonicity and the complete maxitivity of the Shilkret integral with respect to completely maxitive measures (Theorems 2.3 and 2.6).

For the “only if” part, consider first the trivial case with \( c \sim c_0 \) for all \( c \in C \): condition (3.2) and monotonicity imply that \( l \sim c_0 \) for all \( l \in R(C, c_0) \), and therefore the desired representation is valid if and only if \( v = 0 \). When \( c_1 \in C \) does not satisfy \( c_1 \sim c_0 \), condition (3.1) implies that \( c_0 \prec c_1 \): in this case, let \( V \) be the functional on \( R(C, c_0) \) defined by

\[
V(l) = \begin{cases} 
\sup\{p \in [0, 1] : p \cdot c_1 \preceq l\} & \text{if } l \preceq c_1, \\
\inf\{x \in (1, \infty) : \frac{1}{x} l \preceq c_1\} & \text{if } c_1 \prec l.
\end{cases}
\]

This definition implies that if \( l \preceq l' \), then \( V(l) \leq V(l') \); to prove the converse we need first the following simple result: if \( p' < p \), then \( p' l \preceq pl \), and \( p' l \sim pl \) only if \( pl \sim l \) for all \( p \in (0, 1] \). Thanks to monotonicity, it suffices to show \( p' l \preceq pl \) for \( l = c_1 \); if \( p \cdot c \prec p' \cdot c = p \cdot c \) were true (where \( \gamma = \frac{p'}{p} \)), then by induction and condition (3.2) we would obtain \( p \cdot c \prec \gamma^n p \cdot c \) for all positive integers \( n \), and therefore \( p \cdot c \sim c_0 \) by condition (3.4), but this would contradict \( p \cdot c \prec p' \cdot c \), since \( p' \cdot c = \gamma p \cdot c \sim \gamma c_0 = c_0 \). If \( p' l \sim pl \), then by the same technique we obtain \( pl \sim \gamma^n pl \), and thus \( p' l \sim pl \) for all \( p' \in (0, p] \); but \( p^2 l \sim pl \) and condition (3.2) imply \( p \sim l \), and thus \( p' l \sim l \) for all \( p' \in (0, 1] \). In particular, \( p' < p \) implies \( p' \cdot c_1 \prec p \cdot c_1 \), because otherwise \( c_1 \sim c_0 \) would follow from condition (3.4). Using these results and conditions (3.2) and (3.3), the following equivalences can be easily proved:

\[
V(l) = p \in [0, 1] \iff l \sim p \cdot c_1, \\
V(l) = \frac{1}{p} \in [1, \infty) \iff pl \sim c_1, \\
V(l) = \infty \iff c_1 \prec pl \text{ for all } p \in (0, 1].
\]

These expressions imply that \( V \) is homogeneous (that is, \( V(p l) = p V(l) \)) and it represents \( \preceq \): the only difficulty is to show that if \( V(l) = V(l') = \infty \), then \( l \preceq l' \). To prove it, consider first that \( V(c) < \infty \) (for all \( c \in C \)) follows from condition (3.4); hence \( l(q) \preceq l' \) for all \( q \in Q \), but then condition (3.5) implies \( l \preceq l' \) (with \( l_q = l(q) \)). Let \( \mu \) be the completely maxitive measure on \( Q \) defined by \( \mu(q) = V(c_1 I_{\{q\}}) \); we prove now that \( \mu \) is normalized, and \( V(l) = \int Q (v \circ l) \, d\mu \), where \( v \) is the homogeneous function on \( R(C, c_0) \) defined by \( v(r) = V(r) \). Let \( s_l = \sup_{q \in Q} V[l I_{\{q\}}] \); the monotonicity of \( \preceq \) implies that \( V(l) \geq s_l \). If \( V(l) > s_l \) were true, then there would be
3.1 Description of Uncertain Knowledge

Let $D$ be a subset of $R(C, c_0)$ that contains $c$ and $c I_{\{q\}}$ for all $c \in C$ and all $q \in Q$, and that is closed with respect to randomization, in the sense that if $l \in D$, then $pl \in D$ for all $p \in (0, 1)$. It can be easily proved that Theorem 3.1 is still valid if we substitute $D$ for $R(C, c_0)$ in its statement (and in the five conditions); that is, the preference relation on $D$ suffices to determine its representation by the mapping $l \mapsto \int^S (v \circ l) \, dr_p$ (up to the slight nonuniqueness of $v$ and $rp$). In particular, as noted above, $D$ can be a set of randomized decisions of the form $pl$, where $l$ can be identified with an element of $C^Q$, and $p \in [0, 1]$. The smallest set $D$ satisfying the above requirements consists of the decisions $r$ and $r I_{\{q\}}$ for all $r \in R(C, c_0)$ and all $q \in Q$; by slightly modifying the postulated conditions, the preferences between decisions of the form $r I_{\{q\}}$ would suffice to determine the representation.

Apart from the fact that the evaluations are not necessarily bounded above, the main differences of our representation theorem with respect to the one of Anscombe and Aumann are the peculiar role of $c_0$, the weakening of the independence condition, and the strengthening of the monotonicity condition. The monotonicity is strengthened through worst-case evaluation: this is strictly related to the description of uncertain knowledge by means of relative plausibility; in fact, this description is based on a “best-case evaluation”, in the sense that the ability of a set of models to represent the observed reality is interpreted as the ability of its best element. The apparent conflict between “worst-case” and “best-case” in the previous period is due to the fact that we consider an evaluation of the decisions in
terms of loss; that is, the negativity of the uncertain consequences is evaluated. This is the reason for interpreting the expression \( l \prec l' \) as “\( l' \) is not preferred to \( l \)”, but in the cited literature about representation theorems the above expression is interpreted as “\( l \) is not preferred to \( l' \)”, because the decisions are evaluated in terms of utility: to avoid confusion, all the cited conditions and results have been modified (if necessary) in order to comply with our interpretation of the preference relation (in fact, the only ones needing changes are those concerning uncertainty aversion). The peculiar role of \( c_0 \) and the weakening of the independence condition in our representation theorem with respect to the one of Anscombe and Aumann are strictly related to the differences between the evaluations of decisions in terms of loss and in terms of utility.

Usually, a statistical decision problem is described by a loss function \( L \) on \( \mathcal{P} \times \mathcal{D} \), and there is a clear idea of what constitutes a correct decision for a model \( P \in \mathcal{P} \), also when this decision is not an element of \( \mathcal{D} \). The loss functions \( L \) and \( \alpha L \) are considered equivalent (for all \( \alpha \in \mathcal{P} \)), and therefore, in order to be representable by an evaluation in terms of loss, the preferences between decisions must fulfill the scale invariance expressed by the weak independence condition (3.2), which can be stated in the framework of Anscombe and Aumann only if the existence of the consequence \( c_0 \) of a correct decision has been assumed. A general decision problem is often described by a bounded utility function \( U : \mathcal{Q} \times \mathcal{D} \to \mathbb{R} \), and in general the concept of correct decision does not make sense; in this regard it is important to distinguish between the concept of optimal decision (\( d \in \mathcal{D} \) is optimal for \( q \) if it maximizes \( U(q, d) \) among the elements of \( \mathcal{D} \); the optimality depends thus strongly on \( \mathcal{D} \)) and the one of correct decision (a correct decision for \( q \) is not necessarily an element of \( \mathcal{D} \), and its correctness does not depend on the other elements of \( \mathcal{D} \)). The utility functions \( U \) and \( \alpha U + \beta \) are considered equivalent (for all \( \alpha \in \mathcal{P} \) and all \( \beta \in \mathbb{R} \)), and therefore, in order to be representable by an evaluation in terms of utility, the preferences between decisions must fulfill the certainty-independence condition postulated by Gilboa and Schmeidler. If the loss function \( L \) and the utility function \( U \) on \( \mathcal{Q} \times \mathcal{D} \) describe the same decision problem and are expressed in the same scale, then \( u_0(q) = U(q, d) + L(q, d) \) is the utility of a correct decision for \( q \in \mathcal{Q} \) (and thus in particular it does not depend on \( d \)). Hence, when a decision problem is described by a utility function \( U \) on \( \mathcal{Q} \times \mathcal{D} \), to obtain a description of the problem by a loss function we must determine the function \( u_0 \) on \( \mathcal{Q} \); but this determination is problematic for general decision problems, since there is usually no clear
idea of what constitutes a correct decision. In order to be representable by an evaluation in terms of utility or loss avoiding the problems related to the determination of $u_0$, the preferences between decisions must fulfill the strong independence condition postulated by Anscombe and Aumann: in fact, in this case the utility functions $U$ and $\alpha U + u$ are equivalent for all $\alpha \in \mathbb{P}$ and all bounded functions $u : \mathcal{Q} \to \mathbb{R}$ (interpreted as functions on $\mathcal{Q} \times \mathcal{D}$ such that $u(q, d)$ does not depend on $d$). This points out a fundamental difference between the MPL approach to statistical decision problems and the Bayesian one: the former is specialized on problems described by loss functions, while the latter is able to cope indifferently with loss and utility functions; but this ability comes with a price: it implies uncertainty neutrality, while uncertainty aversion can be considered as the basic attitude in statistics.

Hence, if the evaluation method allows strict uncertainty aversion, then the choice of $u_0$ matters when translating $U$ into a loss function: two alternative assumptions about $u_0$ are often made when its determination is problematic. One of them is that for each $q \in \mathcal{Q}$ a correct decision is contained in $\mathcal{D}$ (at least as a limit): we obtain the loss function $L$ on $\mathcal{Q} \times \mathcal{D}$ defined by $L(q, d) = \sup_{d' \in \mathcal{D}} U(q, d') - U(q, d)$ for all $(q, d) \in \mathcal{Q} \times \mathcal{D}$. That is, the optimality is substituted for the correctness of a decision; this kind of loss function was introduced explicitly by Savage (1951) and is often called “regret function”: it gives the amount of utility lost for not having selected the best available decision (that is, the optimal decision). If a method consists in evaluating the decisions on the basis of the regret function obtained from a utility function, then the induced preferences satisfy the strong independence condition postulated by Anscombe and Aumann (and thus reveal uncertainty neutrality), but the preference between two decisions can depend on the whole set $\mathcal{D}$ (since the optimality depends strongly on $\mathcal{D}$). The other frequent assumption about $u_0$ is that the utility of a correct decision does not depend on $q$ (that is, $u_0$ is constant): we obtain the loss function $c - U$ on $\mathcal{Q} \times \mathcal{D}$, where $c \geq \sup U$. To be reasonable, an evaluation of decisions on the basis of $c - U$ must be independent of $c$; that is, it must be based on an evaluation method such that the induced preferences fulfill the certainty-independence condition postulated by Gilboa and Schmeidler. Hence, the application of the MPL criterion to the decision problem described by the loss function $c - U$ is not reasonable: to obtain preferences independent of $c$ we can replace the Shilkret integral with a translation equivariant integral, and the natural choice is the Choquet integral.
A representation theorem analogous to Theorem 3.1, but with the Choquet integral instead of the Shilkret integral (that is, a representation theorem for the preferences induced by the "MPL* criterion" of Cattaneo, 2005), can be easily obtained by imposing an additional condition to the preferences considered in the representation theorem of Schmeidler. The additional condition must force the complete maxitivity of the "nonadditive probability" measure, as does for example the following one: if \( l|_A \) is constant, and \( l' \prec l \), then there is a \( q \in A \) such that \( l'' \prec l_{(Q \setminus A) \cup \{q\}} \) (for all nonempty \( A \subseteq Q \) and all \( l, l' \in R(C)^Q \)). This condition can be interpreted as stating that the preferences are based on a worst-case evaluation limited to cases implying the same consequence; it could also be expressed in a form analogous to the one of condition (3.5), and in this case it would imply the monotonicity of the preference relation on the set of functions \( l \in R(C)^Q \) having finite range (the preference relation on this set suffices to determine the representation), but anyway this enforcement of complete maxitivity is rather artificial.

3.1.2 Updating

Let \( \mathcal{P} \) be a set of statistical models (\( \mathcal{P} \) is a set of probability measures on a measurable space \((\Omega, A)\)): a relative plausibility measure on \( \mathcal{P} \) is a description of uncertain knowledge about those models. In the previous subsection we considered the static aspect of this description, in relation to the preferences induced by the MPL criterion; in the present subsection we consider the dynamic aspect: the evolution of the description when new information about the models is acquired.

When we observe an event \( A \in A \), we obtain a likelihood function \( lik \) on \( \mathcal{P} \), and we must condition on \( A \) each model in \( \mathcal{P} \); if subsequently we observe a second event \( B \in A \), then we obtain the likelihood function \( lik' \) on \( \mathcal{P} \) defined by \( lik'(P) = P(B|A) \) (for all \( P \in \mathcal{P} \)). To be precise, when we observe \( B \), we obtain a likelihood function \( lik'' \) on the set \( \mathcal{P}' \) of the models in \( \mathcal{P} \) conditioned on \( A \), but it is convenient to avoid explicit reference to \( \mathcal{P}' \) by considering \( lik' = lik'' \circ c \) instead of \( lik'' \), where \( c : \mathcal{P} \rightarrow \mathcal{P}' \) is the surjection representing the conditioning on \( A \) of the models in \( \mathcal{P} \). Since for each \( P \in \mathcal{P} \) we have \( P(A \cap B) = P(A) P(B|A) \) (independently of the definition of \( P(B|A) \) when \( P(A) = 0 \)), the likelihood function on \( \mathcal{P} \) induced by the observation of both events \( A \) and \( B \) is \( lik lik' \); that is, the likelihood function \( lik \) is updated through multiplication with the likelihood function
3.1 Description of Uncertain Knowledge

$lik'$ induced by the new observation. We will use the symbol $\otimes$ to denote the binary operation on relative plausibility measures corresponding to the multiplication of their density functions: $rp \otimes rp' = [rp \cdot (rp')^{-1}]$ for all pairs of relative plausibility measures $rp, rp'$ on the same set $Q$. Hence the relative plausibility measure $lik^\uparrow$ is updated through combination with $(lik')^\uparrow$ by means of the operation $\otimes$.

The dynamic aspect of the description of uncertain knowledge about statistical models by means of relative plausibility can be incorporated in our representation theorem (Theorem 3.1, with $Q = \mathcal{P}$) by imposing an additional condition to the preferences between decisions. This condition can be stated as follows (in a slightly informal way): if the relation $\preceq$ on $R(C, c_0)^\mathcal{P}$ expresses our preferences before observing the event $A$ (the models in $\mathcal{P}$ are assumed to be conditional on past observations), then our preferences after observing $A$ shall be expressed by the relation $\preceq_A$ on $R(C, c_0)^\mathcal{P}$ defined by: $l \preceq_A l'$ if and only if $l_A \preceq l'_A$, where $l_A$ is the decision whose uncertain consequence will be $l$ when $A$ will occur, and $c_0$ otherwise ($l'_A$ is defined analogously, for all $l, l' \in R(C, c_0)^\mathcal{P}$). This condition can be interpreted similarly to condition (3.2), as stating that if we shall have to decide only when $A$ occurs (while otherwise we shall know the correct decision), then our preferences must be the same as if $A$ has already occurred and we certainly have to decide. Under each model $P \in \mathcal{P}$, the decision $l_A$ has the following randomized consequence: $l(P)$ with probability $P(A)$, and $c_0$ with probability $1 - P(A)$; that is, $l_A = lik l$. Therefore, if the preferences before observing $A$ are represented by the mapping $l \mapsto \int^S (\nu \circ l) \, drp$, as stated in Theorem 3.1, then the preferences after observing $A$ will be represented by the same mapping after the updating of the relative plausibility measure on $\mathcal{P}$:

$$l \mapsto \int^S [\nu \circ (lik l)] \, drp \propto \int^S (\nu \circ l) \, d(rp \otimes lik^\uparrow).$$

The relative plausibility measure $1^\uparrow$ on $\mathcal{P}$ describes the complete absence of information for discrimination between the models in $\mathcal{P}$ (in the sense of Kullback and Leibler, 1951): it is the neutral element of the operation $\otimes$, and it is induced for example by the observation of $\Omega$, which corresponds to having observed no data. Basing decisions directly on the likelihood function $lik$ implies in particular using no information about $\mathcal{P}$ that was available prior to the observation of $A$: this means using $1^\uparrow$ as a description of the uncertain knowledge about the models in $\mathcal{P}$ prior to
the observation of $A$, and in fact $\text{lik}^\dagger = 1^\dagger \circ \text{lik}^\dagger$ is then the updated description. On the basis of these considerations, it is natural to allow the possibility of using information about $\mathcal{P}$ that was available prior to the observation of events in $A$, when this is described by a (prior) relative plausibility measure $rP$ on $\mathcal{P}$ (or equivalently, by a prior likelihood function $rP^\dagger$ on $\mathcal{P}$), and the models in $\mathcal{P}$ are assumed to be conditional on the observations that have induced $rP$: in this case, after observing $A$ we obtain the updated relative plausibility measure $rP \circ \text{lik}^\dagger$, which can be used for decision making. The choice of the prior $rP$ can be based on analogies with the relative plausibility measures encountered in past experience, or with those induced by hypothetical data in the situation at hand. It is also possible to interpret the prior $rP$ operationally: if we observed no data, then we would base our decisions on $rP$, and thus the preferences between decisions can shed light on the form of $rP$. In particular, if we decide according to the MPL criterion, then we can choose the prior $rP$ by considering the "ideal" form of the uncertain (relative) loss $l \in \mathbb{F}^P$: the prior likelihood function can simply be defined as proportional to $\frac{1}{l}$.

Hence in general the description of uncertain knowledge about the models in $\mathcal{P}$ by means of a relative plausibility measure on $\mathcal{P}$ evolves from a prior $rP$ (which can also describe the complete absence of information for discrimination between the models) through combination (by means of the operation $\circ$) with the relative plausibility measures on $\mathcal{P}$ induced by the observations of new data, and the description can be used in any moment for decision making. There is a strong similarity with the Bayesian approach, in which the prior probability measure $\pi$ on $\mathcal{P}$ evolves through conditioning of the resulting probability measure $P_\pi$ on $\mathcal{P} \times \Omega$, and can be used in any moment for decision making: in particular, after observing $A$, the MPL and Bayesian criteria applied to a decision problem described by a loss function $L$ on $\mathcal{P} \times \mathcal{D}$ consist in minimizing

$$\int S^d \text{lik} \, drP \quad \text{and} \quad \int S^d \text{lik} \, d\pi,$$

respectively (the second expression is a Lebesgue integral). If $\pi$ possesses a density (with respect to some measure on $\mathcal{P}$), then this is updated through multiplication with the likelihood functions induced by the observations of new data (the normalization of the resulting density is not necessary for decision making), exactly in the same way as the density function of $rP$ is updated. But there is a difference: the updating of the density function of $rP$ is symmetric, in the sense that two likelihood functions are multi-
plied, while the updating of the density of $\pi$ is asymmetric, in the sense that a likelihood function is multiplied with the density of a probability measure. A consequence of the symmetry is that when several prior relative plausibility measures on $\mathcal{P}$ are assumed to be induced by independent observations, they can be combined without problems by means of the operation $\odot$ (which is commutative and associative), while the combination of several prior probability measures is problematic.

However, the fundamental problem with prior probability measures is the impossibility of describing ignorance (in the sense of absence of information for discrimination between the models): the intuition that a (nearly) constant density function describes a situation of (near) ignorance is wrong for probability measures (see for example Shafer, 1982, and in particular the comment by DeGroot), while it is correct for relative plausibility measures. The choice of a prior likelihood function on $\mathcal{P}$ seems thus to be better supported by intuition than the choice of a prior probability measure on $\mathcal{P}$, also because of the possibility of analogies with likelihood functions encountered in past experience (we can only observe relative likelihoods of models, not probabilities of models), or with those induced by hypothetical data in the situation at hand (see also Dahl, 2005).

The relative plausibility measure $1^\uparrow$ on $\mathcal{P}$ describes complete ignorance (in the above sense), but partial ignorance can also be represented: for example, if $\mathcal{G}$ is a set, $g : \mathcal{P} \to \mathcal{G}$ is a mapping, and $f : \mathcal{G} \to \mathcal{P}$ describes the relative plausibility of the elements of $\mathcal{G}$, then $(f \circ g)^\uparrow$ contains the same information as $f$, in the sense that the relative plausibility of the elements of $\mathcal{P}$ is the one (described by $f$) of their images under $g$ (in this way we can in particular obtain a relative plausibility measure on $\mathcal{P}$ from a pseudo likelihood function $f$ on $\mathcal{G}$).

## 3.2 Hierarchical Model

Altogether, our mathematical representation of reality can be considered as a hierarchical model with two levels: the first one consists of a set $\mathcal{P}$ of statistical models, and the second one consists of a relative plausibility measure on $\mathcal{P}$. The two levels describe different kinds of uncertain knowledge: in the first level the uncertainty is stochastic, while in the second one it is about which of the statistical models in $\mathcal{P}$ is the best representation of the reality. When new data are observed, the hierarchical model can be
updated by conditioning each element of \( \mathcal{P} \), and by updating the relative plausibility measure on \( \mathcal{P} \) as considered in Subsection 3.1.2. The relative plausibility measure on \( \mathcal{P} \) is simply a convenient extension (from the elements of \( \mathcal{P} \) to the subsets of \( \mathcal{P} \) by means of the supremum, and thus in accordance with the likelihood-based inference methods) of the likelihood function on \( \mathcal{P} \) induced by the observed data, possibly after multiplication with a prior likelihood function. Therefore, under each model \( P \in \mathcal{P} \) the relative plausibility measure on \( \mathcal{P} \) tends to concentrate around (the conditioned version of) \( P \) when more and more data are observed, if some regularity conditions are satisfied.

When we face a decision situation depending on some aspects of the reality described by a hierarchical model consisting of a relative plausibility measure \( rp \) on a set \( \mathcal{P} \) of statistical models, if we can describe the decision problem by a loss function \( L \) on \( \mathcal{P} \times \mathcal{D} \) (under each model \( P \in \mathcal{P} \) the loss we would incur by making the decision \( d \in \mathcal{D} \) can be reduced to a representative value, such as the expected value), then we can apply a likelihood-based decision criterion, with respect to the (pseudo) likelihood function \( rp^\dagger \) on \( \mathcal{P} \). Statistical inferences can be obtained by applying the likelihood-based inference methods to \( rp^\dagger \) (in Subsection 1.3.1 we have seen that the MPL criterion leads to these methods, when applied to some standard form of the corresponding decision problems). In particular, if \( \mathcal{G} \) is a set, and \( g : \mathcal{P} \rightarrow \mathcal{G} \) is a mapping, then \( (rp \circ g^{-1})^\dagger \) is a pseudo likelihood function on \( \mathcal{G} \) with respect to \( g \), and we can obtain the maximum likelihood estimate and the likelihood-based confidence regions for \( g(P) \) by considering \( (rp \circ g^{-1})^\dagger \) as equivalent to the profile likelihood function on \( \mathcal{G} \) (they are in fact equivalent if \( rp \) is the relative plausibility measure on \( \mathcal{P} \) induced by the observed data).

Some authors have noted that if \( lik \) is a likelihood function on \( \mathcal{P} \), then \( lik^\dagger \) is a normalized possibility measure on \( \mathcal{P} \) assigning to each set \( \mathcal{H} \subseteq \mathcal{P} \) its likelihood ratio \( LR(\mathcal{H}) \), and they have thus used the apparatus of possibility theory to obtain inferences and decisions: see for example Smets (1982), Moral and de Campos (1991), Chen (1993), and Giang and Shenoy (2002). The connection with possibility theory is in fact tempting, but it can be misleading, since in this theory possibility measures are often used in ways incompatible with the meaning of likelihood functions.
3.2 Hierarchical Model

3.2.1 Classical and Bayesian Models

From an abstract point of view, the classical and Bayesian models can be considered as special cases of our hierarchical model, but the approaches to statistical decision problems are different. The classical model consists of a set $\mathcal{P}$ of statistical models, with complete ignorance about which of the elements of $\mathcal{P}$ is the best representation of the reality; in the classical approach to statistical decision problems the model is never updated, since only pre-data decision problems are considered. Hence the classical model corresponds to the hierarchical model with prior relative plausibility measure $1^\perp$ on the set $\mathcal{P}$, as long as only pre-data decision problems are considered. In particular, with this prior the MPL criterion applied to the pre-data decision problem described by the risk function corresponds to the minimax risk criterion. With a general prior relative plausibility measure on $\mathcal{P}$, the MPL criterion applied to this decision problem consists in using the prior likelihood of the models to weight the respective losses, before applying the minimax risk criterion. The use of a nondegenerate prior relative plausibility measure $\rho \mathcal{P}$ on $\mathcal{P}$ when applying the MPL criterion to the pre-data decision problem described by the risk function is thus a simple way to address the idea behind the restricted Bayesian criterion: we can use some prior information, but at the same time maintain the maximum expected loss under control. In fact, if $\lambda = \inf \rho \mathcal{P}^\perp > 0$, then the maximum expected loss of an optimal decision function according to the MPL criterion is at most $\frac{1}{\lambda}$ times the one of the optimal decision functions according to the minimax risk criterion.

In Subsection 1.1.2 we have seen that the (strict) Bayesian model consists of a single probability measure $P_\pi$ on $\mathcal{P} \times \Omega$, and it is updated by conditioning $P_\pi$ on the observed data. Hence the Bayesian model corresponds to the hierarchical model with as first level the single statistical model $P_\pi$, since there is a unique nondegenerate relative plausibility measure on the singleton $\{P_\pi\}$. Thus only stochastic uncertainty is considered in the Bayesian model, and this allows the coherence properties stated in Subsection 1.1.2. In fact, in the Bayesian approach to statistical decision problems the uncertainty about which of the statistical models in $\mathcal{P}$ is the best representation of the reality is eliminated by assuming that exactly one of them is “true” and mixing them by means of a prior probability measure $\pi$ on $\mathcal{P}$. In our approach a prior relative plausibility measure $\rho \mathcal{P}$ on $\mathcal{P}$ is used instead of $\pi$: the main differences between relative plausibility
and probability measures have been considered in the previous section. It is interesting to note that the Bayesian method of maximum a posteriori for the estimation of \( P \in \mathcal{P} \) corresponds to the method of maximum likelihood, when \( r_P \) is proportional to the density of \( \pi \) with respect to the considered dominating measure on \( \mathcal{P} \), but the method of maximum likelihood with prior \( r_P \) can be applied also when the choice of a dominating measure on \( \mathcal{P} \) is problematic (see for example Leonard, 1978). In fact, thanks to their simplicity, relative plausibility measures can be handled without problems also on sets \( \mathcal{P} \) that are so wide that using probability measures on them would be difficult, as in the following example.

**Example 3.2.** Consider the problem of nonparametric estimation of the mean of a probability distribution on \([0, 1]\), with squared error loss. Let \( \mathcal{P} \) be a family of probability measures such that under each of them the random variables \( X_1, \ldots, X_n \) are independent and identically distributed on \([0, 1]\), and such that each probability distribution on \([0, 1]\) for these random variables corresponds to an element of \( \mathcal{P} \). Let \( g : P \mapsto E_P(X_1) \) be the function on \( \mathcal{P} \) assigning to each model the mean of the corresponding probability distribution on \([0, 1]\), and let \( L : (P, d) \mapsto [g(P) - d]^2 \) be the loss function on \( \mathcal{P} \times [0, 1] \) describing the estimation problem. As noted at the beginning of Section 1.2, for a continuous distribution the observation \( X_i = x_i \) is impossible: we can only observe \( X_i \in N_i \), for a neighborhood \( N_i \) of \( x_i \). However, for the decision problem at hand, if the neighborhoods \( N_i \) are sufficiently small, then the results are practically equivalent to those obtained by considering only the discrete distributions on \([0, 1]\) and the exact observations \( X_i = x_i \).

For instance, in the case with \( n = 1 \) we obtain the following (approximate) profile likelihood function for the mean \( g(P) \) of the distribution of \( X_1 \): the function \( lik_g \) on \([0, 1]\) defined by

\[
lik_g(\gamma) = \begin{cases} 
\frac{\gamma}{x_1} & \text{if } 0 \leq \gamma < x_1, \\
1 & \text{if } \gamma = x_1, \\
\frac{1-\gamma}{1-x_1} & \text{if } x_1 < \gamma \leq 1.
\end{cases}
\]

If the function \( f : [0, 1] \to \mathbb{R} \) describes the prior relative plausibility of the possible values of the mean \( g(P) \), we can use this prior information simply by multiplying \( lik_g \) with \( f \): this amounts to using the prior relative plausibility measure \( (f \circ g)^\uparrow \) on \( \mathcal{P} \). If we do not use prior information, then the maximum likelihood estimate of \( g(P) \) is \( X_1 \), and the estimate obtained
by applying the MPL criterion is \( m(X_1) \), where \( m : [0, 1] \rightarrow [\frac{1}{4}, \frac{3}{4}] \) is an increasing bijection, plotted in the first diagram of Figure 3.1 (solid line) together with the bijection \( \text{id}_{[0,1]} \) corresponding to the maximum likelihood estimate (dashed line). It can be proved that the estimate \( m(X_1) \) obtained by applying the MPL criterion is optimal also with respect to the minimax risk criterion, since the least favorable distribution on \([0, 1]\) with mean \( \gamma \) is the binomial distribution with parameters \( n = 1 \) and \( p = \gamma \) (considered in Example 1.8); therefore the maximum expected losses (as functions of \( p = \gamma \in [0, 1] \)) for the estimates obtained by applying the MPL criterion and the method of maximum likelihood (MLD criterion) are plotted in the first diagram of Figure 1.3 (solid and dotted curved lines, respectively).

In the cases with \( n > 1 \) (without prior information), the maximum likelihood estimate is simply the arithmetic mean \( \bar{X} \) of the observations \( X_1, \ldots, X_n \), while the estimate obtained by applying the MPL criterion is not a function of \( \bar{X} \), because sets of observations with the same mean can induce rather different profile likelihood functions \( \text{lik}_X \) on \([0, 1]\) (but as \( n \) increases, the estimate obtained by applying the MPL criterion tends rapidly to \( \bar{X} \)). For example, in the case with \( n = 2 \) the pairs of observations \( \{X_1 = 0.4, \ X_2 = 0.4\} \) and \( \{X_1 = 0, \ X_2 = 0.8\} \) induce the profile likelihood functions plotted in the second diagram of Figure 3.1 (solid and dashed lines, respectively; the two functions have been scaled to have the same maximum): the estimates obtained by applying the MPL criterion are 0.449 and 0.401, respectively. The first estimate is larger than the second one because the corresponding profile likelihood function is more
skew; in fact, the estimates obtained by applying the LRM$\beta$ criterion with $\beta \approx 0.259$ (so that $-2 \log \beta$ is the 90%-quantile of the $\chi^2$ distribution with one degree of freedom) are 0.449 and 0.414, respectively. These estimates are the midpoints of the intervals corresponding to the likelihood-based confidence regions for $g(P)$ with cutoff point $\beta$: Owen (1988) proved that these regions have 90% asymptotic coverage probability.

\[\diamond\]

### 3.2.2 Imprecise Probability Models

As noted at the end of Subsection 1.1.2, the Bayesian model can be generalized by considering a whole set $\Gamma$ of prior probability measures $\pi$ on $\mathcal{P}$ (with complete ignorance about which of the elements of $\Gamma$ is the best prior). The resulting model consists of a set $\Pi$ of probability measures $P_\pi$ on $\mathcal{P} \times \Omega$, and it is updated by conditioning each of them on the observed data, but usually the induced likelihood function on $\Pi$ is not considered (some methods using the likelihood function have been cited in Subsections 1.2.1 and 1.3.2). Hence in general this model cannot be interpreted as a special case of the hierarchical one, because the prior relative plausibility measure $1^\dagger$ on $\Pi$ is not updated (that is, we remain in a state of complete ignorance about which of the elements of $\Pi$ is the best Bayesian model).

A set $\mathcal{P}$ of probability measures on a measurable space $(\Omega, \mathcal{A})$ can also be interpreted as an imprecise probability measure on $(\Omega, \mathcal{A})$: see for example Walley (1991) and Weichselberger (2001). From $\mathcal{P}$ we obtain in particular the upper probability measure $\sup \mathcal{P}$ (which is normalized, monotonic, and subadditive) and its dual measure $\inf \mathcal{P}$ (the lower probability measure). Gärdenfors and Sahlin (1982) advocated the necessity of considering the reliability of the probability measures in $\mathcal{P}$: they stated that this reliability should be described by a real-valued function $\rho$ on $\mathcal{P}$, without explaining precisely how $\rho$ should be updated, but their model seems to be in agreement with the hierarchical one (through identification of $\rho$ and $r\rho^\dagger$). The axiomatic approaches of Nau (1992), de Cooman and Walley (2002), and Giraud (2005) lead to models similar to the hierarchical one, at least as regards the static aspect. If we start with no information for discrimination between the probability measures in $\mathcal{P}$, and we observe an event $A \in \mathcal{A}$, then we obtain a likelihood function $lik$ on $\mathcal{P}$, and we can base inferences and decisions on $lik$. In particular, if we are interested in the probability of an event $B \in \mathcal{A}$, then we can consider the mapping
3.2 Hierarchical Model

$g : P \mapsto P(B|A)$ on $\mathcal{P}$, and base our inferences on the profile likelihood function $\text{lik}_g$ on $[0, 1]$, as we did in Example 1.7 for the profile likelihood functions plotted in Figure 1.1. In this regard it is important to note that $\text{lik}_g$ does not depend on the definition of $P(B|A)$ when $P(A) = 0$, since in this case $\text{lik}(P) = 0$. Of particular interest are the supremum and infimum of the likelihood-based confidence region for $g(P)$ with cutoff point $\beta \in (0, 1)$: these can be considered as improved upper and lower probabilities of $B$, conditional on the observation of $A$; they correspond to the upper and lower conditional probabilities of $B$ after the restriction of $\mathcal{P}$ to the likelihood-based confidence region $\mathcal{P}_\beta$ for $P$ with cutoff point $\beta$. It is important to note that $\mathcal{P}_\beta$ does not depend on the choice of the event $B \in \mathcal{A}$, and that the restriction of $\mathcal{P}$ to $\mathcal{P}_\beta$ is only temporary: no element of $\mathcal{P}$ should be really discarded, since the relative plausibility of the elements of $\mathcal{P}$ can change in the light of new data. The larger is $\beta$, the more precise are the upper and lower conditional probabilities based on $\mathcal{P}_\beta$ (in the sense that their distance is smaller), but the lower is the confidence in them: there is a sort of confidence-precision trade-off. The limits of the upper and lower conditional probabilities based on $\mathcal{P}_\beta$ as $\beta$ tends to 0 correspond to the “regular extension” considered by Walley (1991, Appendix J) and resulting from the “intuitive concept of conditional probability” studied by Weichselberger and Augustin (2003), while their limits as $\beta$ tends to 1 correspond to the result of a generalization of the rule for conditioning introduced by Dempster (1967) and having a central role in the theory of Shafer (1976): see also Moral and de Campos (1991) and Gilboa and Schmeidler (1993). Hence, for $\beta$ varying in $(0, 1)$, the restrictions of $\mathcal{P}$ to $\mathcal{P}_\beta$ build a continuum between the two extreme cases corresponding to “regular extension” and “generalized Dempster’s conditioning” (inferences other than upper and lower probabilities can also be based on $\mathcal{P}_\beta$: for instance upper and lower expected values of a random variable). As regards upper and lower conditional probabilities, other compromises between the two extreme cases have been proposed in the literature: for example, the use of $\int^S g\text{dlik}_g$ as the upper probability of $B$ conditional on the observation of $A$ (the lower probability can be obtained as dual measure) was considered by Chen (1993) for some particular cases of independent events $A$ and $B$, while Moral (1992) proposed to use the Choquet integral or the Sugeno integral instead of the Shilkret integral in the above expression.

Consider a statistical decision problem described by a loss function $L$ on $\mathcal{P} \times \mathcal{D}$: if we modify the Bayesian approach by using an imprecise prior probability measure $\Gamma$ on $(\mathcal{P}, \mathcal{C})$, instead of a (precise) prior probability
measure on \((\mathcal{P}, \mathcal{C})\), then we obtain the imprecise Bayesian model consisting of the set \(\Pi = \{P_\pi : \pi \in \Gamma\}\) of probability measures on \((\mathcal{P} \times \Omega, \mathcal{E})\), considered at the end of Subsection 1.1.2. When we observe an event \(A \in \mathcal{A}\), we can condition each element of the set \(\Pi\) on the event \(\mathcal{P} \times A\), and we obtain the likelihood function \(lik : P_\pi \mapsto P_\pi(\mathcal{P} \times A)\) on \(\Pi\). The decision criteria appearing in the literature on imprecise Bayesian models can usually be interpreted as criteria applied to the decision problem described by the loss function \(L'' : (P_\pi, d) \mapsto E_{P_\pi}[L(P, d)|\mathcal{P} \times A]\) on \(\Pi \times \mathcal{D}\) without using the information contained in the likelihood function \(lik\) (that is, a part of the information provided by the data is disregarded). A simple way of using \(lik\) for decision making is to (temporarily) restrict \(\Pi\) to the likelihood-based confidence region \(\Pi_\beta\) for \(P_\pi\) with cutoff point \(\beta \in (0, 1)\), when applying the decision criterion (that is, the criterion is applied to the decision problem described by \(L''|_{\Pi_\beta \times \mathcal{D}}\)). In particular, the \(\Gamma\)-minimax criterion applied after having restricted \(\Pi\) to \(\Pi_\beta\) (or equivalently, \(\Gamma\) to the likelihood-based confidence region \(\Gamma_\beta\) for \(\pi\) with cutoff point \(\beta\)) corresponds to the \(\text{LRM}_\beta\) criterion applied to the decision problem described by \(L''\). The information contained in the likelihood function \(lik\) can be used in a less limited way by applying other likelihood-based decision criteria (such as the MPL criterion) to the decision problem described by \(L''\). It is interesting to note that the application of the MPL criterion to the decision problem described by \(L''\) can be interpreted as a sort of \(\Gamma\)-minimax criterion with respect to the set of non-normalized conditional probability measures (Moral, 1992, argued that this set describes the conditional information): in fact, it consists in minimizing the supremum (over all \(\pi \in \Gamma\)) of

\[
P_\pi(\mathcal{P} \times A) E_{P_\pi}[L(P, d)|\mathcal{P} \times A] = E_{P_\pi}[L(P, d) 1_{\mathcal{P} \times A}].
\]

The state of complete ignorance about the models in \(\mathcal{P}\) can be described by the set \(\Gamma_0\) of all probability measures on \((\mathcal{P}, \mathcal{C})\). As noted at the end of Subsection 1.3.1, if \(\mathcal{C}\) contains all singletons of \(\mathcal{P}\), then using the imprecise prior probability measure \(\Gamma_0\) on \(\mathcal{P}\) and applying the MPL criterion to the decision problem described by \(L''\) is equivalent to using the prior relative plausibility measure \(1^\dagger\) on \(\mathcal{P}\) and applying the MPL criterion directly to the decision problem described by \(L\). This property is not possessed by other likelihood-based decision criteria, such as the \(\text{LRM}_\beta\) one, the MLD one, or the one obtained by substituting the Choquet integral for the Shilkret integral in the MPL criterion (that is, the “MPL* criterion” of Cattaneo, 2005), but in general the likelihood-based methods can be used without problems with the imprecise prior probability measure \(\Gamma_0\) on \(\mathcal{P}\).
while this is not the case for the methods that do not use the information contained in the likelihood function $lik$ on $\Pi$. In fact, when the imprecise Bayesian model is updated by means of “regular extension” (and thus $lik$ is used only to discard the $P_\pi \in \Pi$ that are impossible in the light of the observed data, in the sense that $P_\pi(\mathcal{P} \times A) = 0$), it is in general not possible to get out of the state of complete ignorance, because an important part of the information provided by the data is disregarded (see also the beginning of Subsection 4.2.2). More generally, a consequence of the disregard of the information contained in the likelihood function $lik$ is that the conclusions obtained on the basis of the “regular extension” can be useless if the imprecise prior probability measure on $\mathcal{P}$ has not been selected with extreme care (see for instance Kriegler, 2005, Chapter 4): this undermines the intuitive basis of the imprecise Bayesian model.

**Example 3.3.** In Examples 1.5 and 1.9 we considered (for the estimation problem of Example 1.1) the imprecise Bayesian model $\Pi$ with as imprecise prior probability measure on $\mathcal{P}$ the family $\Gamma$ of probability measures on $\mathcal{P}$ corresponding to the family of beta distributions for the parameter $p \in [0, 1]$. The observation of $X = x$ induces a likelihood function $lik$ on $\Pi$: a method using the information contained in $lik$ (such as the MPL criterion) can lead to reasonable conclusions, while a method disregarding this part of the information provided by the data (such as the $\Gamma$-minimax criterion) will lead to vacuous conclusions. However, non-vacuous conclusions can be obtained without using the information contained in the likelihood function $lik$, when instead of $\Gamma$ some narrower set of probability measures on $\mathcal{P}$ is used. Walley (1991, Section 5.3) proposed to use a “near-ignorance prior” on $\mathcal{P}$: that is, an imprecise prior probability measure on $\mathcal{P}$ such that for all possible observations the upper and lower probabilities are 1 and 0, respectively; in this way, an apparent initial state of complete ignorance is obtained (an example of near-ignorance prior is $\Gamma$ itself). In particular, Walley considered the near-ignorance priors $\Gamma_s \subset \Gamma$ corresponding to the sets of all the beta distributions with sum of the two parameters equal $s \in \mathbb{P}$: the resulting imprecise Bayesian models are special cases of the “imprecise Dirichlet models” studied in Walley (1996), where the particular choices $s = 2$ and $s = 1$ are advocated.

Let $g'$ be the function on $\Pi$ corresponding to the function $g$ on $\mathcal{P}$ considered in Example 1.6: that is, $g'$ assigns to each model $P_\pi$ the probability $E_{P_\pi}[g(P)|X = x]$ of observing at least 3 successes in 5 future independent binary experiments with the same success probability as the $n$
observed ones. Figure 3.2 shows the graphs of the profile likelihood functions $\text{lik}_g'$ on $[0, 1]$ induced by the observations $X = 8$ with $n = 10$ (first diagram), and $X = 33$ with $n = 50$ (second diagram), for the three imprecise Bayesian models with imprecise prior probability measures $\Gamma$ (dotted lines), $\Gamma_2$ (dashed lines), and $\Gamma_1$ (solid lines), respectively (the functions have been scaled to have the same maximum). The profile likelihood functions $\text{lik}_g'$ for the imprecise Bayesian model with prior $\Gamma$ correspond to the profile likelihood functions $\text{lik}_g$ for the model $\mathcal{P}$ (plotted in Figure 1.1: dotted and dashed lines, respectively), and also to the limits, as $s$ tends to infinity, of the profile likelihood functions $\text{lik}_g'$ for the imprecise Bayesian models with prior $\Gamma_s$; while as $s$ tends to 0 these functions tend to concentrate on $\gamma' = E_{\pi'}(g)$, where $\pi'$ is the probability measure on $\mathcal{P}$ corresponding to the (possibly degenerate) beta distribution with parameters $x$ and $n - x$ (for the two observations considered, $\gamma' \approx 0.905$ and $\gamma' \approx 0.771$, respectively; as $n$ increases, $\gamma'$ tends to the maximum likelihood estimate $\hat{\gamma}_{ML}$, considered in Example 1.7).

The profile likelihood function $\text{lik}_g'$ for the imprecise Bayesian model with prior $\Gamma$ always takes positive values on the whole interval $(0, 1)$, and therefore only vacuous conclusions can be obtained if the model is updated by means of "regular extension" (and the information contained in $\text{lik}_g'$ is thus disregarded). But with the priors $\Gamma_s$ non-vacuous conclusions can be obtained even if the model is updated by means of "regular extension", because in this case the range of the function $g'$ (corresponding to the support of $\text{lik}_g'$) is usually only a small subset of the interval $[0, 1]$: for the
two observations considered, with \( s = 2 \) we obtain the probability intervals \((0.758, 0.937)\) and \((0.733, 0.790)\), respectively, while with \( s = 1 \) we obtain the probability intervals \((0.833, 0.923)\) and \((0.752, 0.781)\), respectively (the endpoints of the probability intervals are the upper and lower probabilities of observing at least 3 successes in 5 future independent binary experiments with the same success probability as the \( n \) observed ones). The last column of Table 1.1 gives the probability intervals obtained by restricting the imprecise Bayesian model \( \Pi \) with prior \( \Gamma \) to the likelihood-based confidence region \( \Pi_\beta \) for \( P_\pi \) with cutoff point \( \beta \approx 0.036 \): this cutoff point is rather small, but for any reasonable choice of \( \beta \) the probability intervals based on \( \Pi_\beta \) would be much longer than the above ones (based on the “regular extension” of the imprecise Bayesian models with priors \( \Gamma_2 \) and \( \Gamma_1 \), respectively), as can be easily seen from the graphs of Figure 3.2. Anyway, the usefulness of the near-ignorance priors \( \Gamma_s \) is very limited: it suffices to slightly modify the model to obtain vacuous conclusions, when the modified model is updated by means of “regular extension” (see for example Piatti, Zaffalon, and Trojani, 2005).

It is interesting to consider the possibility of replacing the statistical models in \( P \) with imprecise statistical models: let \( \mathcal{IP} \) be a set of imprecise probability measures on a measurable space \((\Omega, \mathcal{A})\); that is, each \( IP \in \mathcal{IP} \) is a set of (precise) probability measures on \((\Omega, \mathcal{A})\). The simplest way to handle imprecise statistical models within our hierarchical model is to consider the set \( P = \bigcup \mathcal{IP} \) of (precise) statistical models, and the mapping \( g : P \to \mathcal{IP} \) defined by \( g^{-1}\{IP\} = IP \) (for all \( IP \in \mathcal{IP} \)). To be precise, \( g \) is well-defined only if the elements of \( \mathcal{IP} \) are disjoint, but this can always be obtained for example by expanding the measurable space \((\Omega, \mathcal{A})\) to its product with the measurable space \((\mathcal{IP}, 2^{\mathcal{IP}})\), and replacing each \( P \in IP \) with the product of \( P \) and the Dirac measure \( \delta_{IP} \) on \( \mathcal{IP} \) (for all \( IP \in \mathcal{IP} \)). If we observe an event \( A \in \mathcal{A} \) (or the corresponding event \( A \times \mathcal{IP} \), when \( \Omega \) has been expanded to \( \Omega \times \mathcal{IP} \)), then we obtain a likelihood function \( lik \) on \( P \), and we can base inferences and decisions on the profile likelihood function \( lik_g \) on \( \mathcal{IP} \).

\[
lik_g(IP) = \sup_{IP} lik = (\sup IP)(A) \quad \text{for all } IP \in \mathcal{IP},
\]

\( lik_g \) is simply the pseudo likelihood function on \( \mathcal{IP} \) based on the upper probability measures \( \sup IP \). For example, if we have a set \( P' \) of statistical models on \((\Omega, \mathcal{A})\), and for some \( \varepsilon \in (0, 1) \) we replace each \( P' \in P' \) with its \( \varepsilon \)-contamination class (that is, the imprecise model consisting of the
contaminated statistical models \((1 - \varepsilon) P' + \varepsilon P\), for all the probability measures \(P\) on \((\Omega, \mathcal{A})\), then instead of the likelihood function \(lik'\) on \(\mathcal{P}'\) we obtain the pseudo likelihood function \((1 - \varepsilon) lik' + \varepsilon\), which can be interpreted as an attenuate version of \(lik'\) (the information contained in \(lik'\) has been discounted, in the sense of Shafer, 1976).

In general, the profile likelihood function \(lik_g\) on \(\mathcal{I}\mathcal{P}\) favors the more imprecise models in \(\mathcal{I}\mathcal{P}\) (because it is based only on the upper probability measures \(\text{sup} IP\)): this behavior seems to be in agreement with the various interpretations of the imprecise probability measures, but if in a particular case it is disapproved, then another pseudo likelihood function on \(\mathcal{I}\mathcal{P}\), penalizing the more imprecise models (for example by explicitly taking into account also the lower probability measures \(\text{inf} IP\)), should be used instead of \(lik_g\). When we combine the imprecise statistical models in \(\mathcal{I}\mathcal{P}\) with a (precise or imprecise) prior probability measure on \(\mathcal{I}\mathcal{P}\), we obtain the imprecise Bayesian model consisting of the probability measures on \(\mathcal{I}\mathcal{P} \times \Omega\) that result from all the possible combinations of the averaging probability measures considered and the (precise) statistical models considered; but if we update the resulting model by means of “regular extension” (disregarding thus a part of the information provided by the data), then we can easily get unreasonable results (see for example Wilson, 2001).

**Example 3.4.** If in the estimation problem of Example 1.1 we replace the models for the results of each single binary experiment with their \(\varepsilon\)-contamination classes, but we still consider the \(n\) results as independent and identically distributed, then we obtain the problem of estimating \(p \in [0, 1]\) (with squared error loss) after having observed the number of successes in \(n\) independent binary experiments with common success probability \(q \in [(1 - \varepsilon)p, (1 - \varepsilon)p + \varepsilon]\). The first diagram of Figure 3.3 shows the graphs of the profile likelihood functions \(lik_g\) on \([0, 1]\) for the mapping \(g\) and the three observations considered in Example 1.6, under the contaminated models with \(\varepsilon = 0.1\) (the three functions have been scaled to have the same maximum); that is, it shows how the graphs of Figure 1.1 are modified by the \(\varepsilon\)-contaminations of the binary experiments. From these graphs we can see that there is an unavoidable uncertainty about \(p\): in fact, even if we knew the actual success probability \(q\), we could only conclude that \(p \in [\max\{\frac{q - \varepsilon}{1 - \varepsilon}, 0\}, \min\{\frac{q}{1 - \varepsilon}, 1\}]\).

The second diagram of Figure 3.3 shows (for the case with \(n = 100\)) the graphs of the maximum expected losses under the contaminated models with \(\varepsilon = 0.1\) (as functions of \(p \in [0, 1]\)) for the estimates obtained by
applying the MPL criterion to these models (upper solid line) and to the uncontaminated models (upper dashed line), and also the graphs of their expected losses under the uncontaminated models (lower solid and dashed lines, respectively). The estimate obtained by applying the MPL criterion to the uncontaminated models was considered in Example 1.8 (the lower dashed line of the second diagram of Figure 3.3 corresponds to the solid line of the third diagram of Figure 1.3); its maximum expected loss under the contaminated models is particularly high for values of $p$ near the endpoints of the interval $[0, 1]$: this is a consequence of the fact that the maximum $(\frac{1}{2} + |p - \frac{1}{2}|)\varepsilon$ of the possible difference between $p$ and the actual success probability $q$ is larger for values of $p$ near the endpoints than for values of $p$ near the center of the interval $[0, 1]$. The estimate obtained by applying the MPL criterion to the contaminated models corrects this behavior by favoring the values of $p$ near the endpoints of the interval $[0, 1]$. 

\diamond
Likelihood-Based Decision Criteria

This chapter begins with a general definition of likelihood-based decision criterion. Several additional decision theoretic properties are then considered and related to integral representations of the decision criteria. Finally, some statistical properties of the likelihood-based decision criteria are studied; in particular their asymptotic optimality.

4.1 Decision Theoretic Properties

Consider a statistical decision problem described by a loss function $L$ on $\mathcal{P} \times \mathcal{D}$, and let $rp$ be a relative plausibility measure on $\mathcal{P}$ describing the uncertain knowledge about the models in $\mathcal{P}$. By definition of $L$, all important aspects of the decisions $d \in \mathcal{D}$ are summarized by the respective functions $l_d \in \mathbb{P}\mathcal{P}$; we will consider only decision criteria that compare the possible decisions on the basis of some functional $V_{rp}: \mathbb{P}\mathcal{P} \to \mathbb{P}$ (depending on $rp$), in the sense that the criteria can be expressed as follows:

$$\text{minimize} \quad V_{rp}(l_d).$$

Each functional $V_{rp}$ represents a total preorder $\preceq_{rp}$ on $\mathbb{P}\mathcal{P}$, and $l_d \preceq_{rp} l_{d'}$ can be interpreted as “$d'$ is not preferred to $d$ on the basis of the decision criterion considered and of the relative plausibility measure $rp$”.

For the sake of generality, we replace $\mathcal{P}$ with a generic set of alternatives determining for each possible decision the loss incurred, and we consider thus decision criteria associating to each (nondegenerate) relative plausibility measure $rp$ (on some set $Q_{rp}$) a preference relation $\preceq_{rp}$ on $\mathbb{P}Q_{rp}$; when interpreting the definitions and the results, however, we regard $Q_{rp}$ as a set of statistical models, and each $l \in \mathbb{P}Q_{rp}$ as the uncertain
loss \( l_d \) incurred by making a particular decision \( d \). The following necessary condition of monotonicity for the preference relations \( \preceq_{rp} \) was stated at the beginning of Section 1.1 as a direct consequence of the interpretation of a loss function:

\[
    l \leq l' \quad \Rightarrow \quad l \preceq_{rp} l' \quad \text{for all } l, l' \in \overline{P}_{Q_{rp}}. \tag{4.1}
\]

Since the loss functions \( L \) and \( \alpha L \) are considered equivalent (for all \( \alpha \in \mathbb{P} \)), the following condition of scale invariance for the preference relations \( \preceq_{rp} \) is also necessary:

\[
    l \preceq_{rp} l' \quad \Rightarrow \quad \alpha l \preceq_{rp} \alpha l' \quad \text{for all } \alpha \in \mathbb{P} \text{ and all } l, l' \in \overline{P}_{Q_{rp}}. \tag{4.2}
\]

To avoid vacuous decision criteria, we can assume the following condition describing the ability of discrimination between the decisions with constant losses:

\[
    c \preceq_{rp} c' \quad \Rightarrow \quad c \leq c' \quad \text{for all } c, c' \in \overline{P}. \tag{4.3}
\]

The next theorem implies that, given the preceding three conditions, the following two are sufficient for a total preorder \( \preceq_{rp} \) to be represented by a functional \( V_{rp} : \overline{P}_{Q_{rp}} \rightarrow \overline{P} \); it can be easily proved that they are also necessary (except for the cases with \( \inf C = 0 \) and \( \sup C = \infty \), respectively).

\[
    l \preceq_{rp} c \text{ for all } c \in C \quad \Rightarrow \quad l \preceq_{rp} \inf C \quad \text{for all } C \subset \overline{P} \text{ and all } l \in \overline{P}_{Q_{rp}}. \tag{4.4}
\]

\[
    c \preceq_{rp} l \text{ for all } c \in C \quad \Rightarrow \quad \sup C \preceq_{rp} l \quad \text{for all } C \subset \overline{P} \text{ and all } l \in \overline{P}_{Q_{rp}}. \tag{4.5}
\]

**Theorem 4.1.** Let \( Q_{rp} \) be a nonempty set. A total preorder \( \preceq_{rp} \) on \( \overline{P}_{Q_{rp}} \) fulfills the conditions (4.1), (4.2), (4.3), (4.4), and (4.5) if and only if it is represented by a homogeneous, monotone functional \( V_{rp} : \overline{P}_{Q_{rp}} \rightarrow \overline{P} \) such that \( V_{rp}(c) = c \) for all \( c \in \overline{P} \); the functional \( V_{rp} \) is unique.

**Proof.** The “if” part is simple: the first two conditions follow from the monotonicity and the homogeneity of \( V_{rp} \), respectively, and the last three conditions are implied by the property that \( V_{rp}(c) = c \) for all \( c \in \overline{P} \).

For the “only if” part, define \( V_{rp}(l) = \sup\{c \in \overline{P} : c \preceq_{rp} l\} \) for all \( l \in \overline{P}_{Q_{rp}} \). Condition (4.5) implies that \( V_{rp}(l) \preceq_{rp} l \), and condition (4.4)
implies that \( l \preceq_{rp} V_{rp}(l) \), because \( l \preceq_{rp} c \) for all \( c > V_{rp}(l) \); therefore \( l \) is equivalent to the constant loss \( V_{rp}(l) \). The desired results can be easily proved using this equivalence and the correspondence of the relation \( \preceq_{rp} \) for the constant losses with the order \( \preceq \) on \( \overline{\mathbb{P}} \), implied by conditions (4.1) and (4.3).

As regards the connection between the relative plausibility measures \( rp \) and the respective preference relations \( \preceq_{rp} \), we can assume that if a relative plausibility measure is derived from \( rp \) by means of a function on \( Q_{rp} \), then the respective preference relation can be derived from \( \preceq_{rp} \) in the corresponding way:

\[
l \preceq_{rp \circ T^{-1}} l' \iff l \circ t \preceq_{rp} l' \circ t \quad \text{for all sets } T, \text{ all } t \in T_{Q_{rp}}^{\circ}, \quad \text{and all } l, l' \in \overline{\mathbb{P}}^T.
\]

(4.6)

A likelihood-based decision criterion is a function associating to each nondegenerate relative plausibility measure \( rp \) (on some set \( Q_{rp} \)) a total preorder \( \preceq_{rp} \) on \( \overline{\mathbb{P}}^{Q_{rp}} \), in such a way that the conditions (4.1), (4.2), (4.3), (4.4), (4.5), and (4.6) are fulfilled (for all nondegenerate relative plausibility measures \( rp \)).

Consider a statistical decision problem described by a loss function \( L \) on \( \mathcal{P} \times \mathcal{D} \), and let \( lik \) be a likelihood function on \( \mathcal{P} \). If \( \mathcal{G} \) is a set, \( g : \mathcal{P} \to \mathcal{G} \) is a mapping, and the loss function \( L \) depends on the models \( P \) only through \( g(P) \), in the sense that there is a function \( L' \) on \( \mathcal{G} \times \mathcal{D} \) such that \( L(P, d) = L'(g(P), d) \) for all \( P \in \mathcal{P} \) and all \( d \in \mathcal{D} \), then condition (4.6) states that applying a likelihood-based decision criterion to \( L \) with respect to \( lik \) is equivalent to applying it to \( L' \) with respect to the profile likelihood function \( lik_g \). If another pseudo likelihood function \( lik' \) on \( \mathcal{G} \) (with respect to \( g \)) is expected to give better results than \( lik_g \), then the criterion can be applied to \( L' \) with respect to \( lik' \); in this sense, condition (4.6) states that the profile likelihood function is the default choice of a pseudo likelihood function, in accordance with the likelihood-based inference methods.

There is a close connection between the conditions appearing in the definition of a likelihood-based decision criterion and those appearing in Theorem 3.1. In the present section we assume that the consequences of the decisions are in \( \overline{\mathbb{P}} \), and thus no quantitative evaluation of the consequences is necessary; in particular, \( c_0 \) corresponds to 0, and so condition (3.1) is clearly fulfilled. The monotonicity (4.1) is a special case of the worst-case evaluation (3.5), and the scale invariance (4.2) corresponds to
(3.2). The only effect of assuming condition (4.3) is the exclusion of the trivial relations $\lesssim_{rp}$ for which all the elements of $\overline{P}^{Q_{rp}}$ are equivalent: this nontriviality is not assumed in Theorem 3.1, but it is necessary and sufficient for the uniqueness of $rp$. Conditions (4.4) and (4.5) can be rewritten as a continuity condition corresponding to (3.3) and two conditions expressing the exclusion of “infinitesimal evaluations” and of “transfinite evaluations” (the cases with $\inf C = 0$ and $\sup C = \infty$, respectively): the first of these last two conditions corresponds to (3.4), and the second one is a special case of (3.5), with the difference that in Subsection 3.1.1 the boundedness of the consequences is assumed, and thus no consequence corresponds to $\infty$. Condition (4.6) cannot be expressed in the framework of Subsection 3.1.1 (because in that framework the existence of the relative plausibility measure $rp$ is not assumed), but it is related to the limited version of the worst-case evaluation stated at the end of Subsection 3.1.1. Thus, apart from the assumptions that the consequences are in $\overline{P}$, that $rp$ describes the uncertain knowledge about the elements of $Q_{rp}$, and that the preference relation on $\overline{P}^{Q_{rp}}$ is nontrivial, the difference between the conditions appearing in the definition of a likelihood-based decision criterion and those appearing in Theorem 3.1 is that the worst-case evaluation (3.5) is weakened by replacing it with two special cases (the monotonicity and the exclusion of “transfinite evaluations”) and condition (4.6). The next theorem states that a likelihood-based decision criterion corresponds to a functional on the set

$$S_{\overline{P}} = \{ \varphi \in [0, 1]^{\overline{P}} : \sup \varphi = 1 \};$$

the MPL criterion of Theorem 3.1 corresponds to the particular functional $\varphi \mapsto \sup \text{id}_{\overline{P}} \varphi$ on $S_{\overline{P}}$.

**Theorem 4.2.** A function associating to each nondegenerate relative plausibility measure $rp$ (on some set $Q_{rp}$) a binary relation $\lesssim_{rp}$ on $\overline{P}^{Q_{rp}}$ is a likelihood-based decision criterion if and only if there is a functional $S : S_{\overline{P}} \rightarrow \overline{P}$ such that each relation $\lesssim_{rp}$ is represented by the functional $V_{rp} : l \mapsto S[(rp \circ l^{-1}) \downarrow]$ and the following three conditions are fulfilled:

$$S(I_{\{1\}}) = 1,$$

$$S[\varphi(\cdot)] = c S(\varphi) \quad \text{for all } \varphi \in S_{\overline{P}} \text{ and all } c \in \overline{P},$$

$$\sup_{[x, \infty]} \varphi \leq \sup_{[x, \infty]} \psi \quad \text{and} \quad \sup_{[0, x]} \varphi \geq \sup_{[0, x]} \psi \quad \Rightarrow \quad S(\varphi) \leq S(\psi) \quad \text{for all } \varphi, \psi \in S_{\overline{P}}.$$
Proof. Theorem 4.1 implies that we have a likelihood-based decision criterion if and only if condition (4.6) is fulfilled and each relation $\chi_{rp}$ is represented by a homogeneous, monotonic functional $V_{rp} : \mathbb{R}^{Q_{rp}} \rightarrow \mathbb{R}$ such that $V_{rp}(1) = 1$ (since then $V_{rp}(c) = c$ for all $c \in \mathbb{R}$ follows from homogeneity and monotonicity). Condition (4.6) corresponds to the existence of a functional $S : S_{\mathbb{R}} \rightarrow \mathbb{R}$ such that $V_{rp}(l) = S[(rp \circ l^{-1})]$. In fact, the existence of such a functional implies condition (4.6), because $rp \circ (l \circ t)^{-1} = (rp \circ t^{-1}) \circ l^{-1}$; conversely, when condition (4.6) holds we can define $S(\phi) = V_{\phi}(id_\mathbb{R})$, since

$$V_{rp}(l) = V_{rp}(id_\mathbb{R} \circ l) = V_{rp}[V_{rp}(l) \circ l] = V_{rp}(l) = V_{rp}(l) = V_{rp}(l),$$

where the last equality is a consequence of the second one. Hence, to prove the theorem it suffices to show that the property $V_{rp}(1) = 1$, the homogeneity, and the monotonicity of the functionals $V_{rp}$ correspond to the conditions (4.7), (4.8), and (4.9), respectively. The first two correspondences follow easily from the equalities $(rp \circ l^{-1}) = I_{[1]}$, and $(rp \circ (c_1 l)^{-1})(x) = (rp \circ l^{-1})(\frac{x}{c})$, respectively. The monotonicity of the functionals $V_{rp}$ follows from condition (4.9), since by defining $\phi = (rp \circ l)^{-1}$ and $\psi = [rp \circ (l')^{-1}]$, the inequality $l \leq l'$ implies the premise of condition (4.9): in fact, $\sup_{[x, \infty]} \phi = \sup_{[x, \infty]} \psi = \sup_{[0, x]} \phi = \sup_{[0, x]} \psi$, and analogously for $\psi$ and $l'$. To prove the converse it suffices to show that for all $\phi, \psi \in S_{\mathbb{R}}$ satisfying the premise of condition (4.9) there are a nondegenerate relative plausibility measure $rp$ (on some set $Q_{rp}$) and two functions $l, l' \in \mathbb{R}^{Q_{rp}}$ satisfying $l \leq l'$ and (4.10). Define the set $Q_{rp} \subseteq \mathbb{R} \times [0, 1] \times \{0, 1\}$ as follows:

$$Q_{rp} = \{(x, y) \in \mathbb{R}^2 : y < \phi(x) \} \cup \{(x, y) \in \mathbb{R}^2 : y < \psi(x) \} \times \{0, 1\},$$

and let $rp$ be the nondegenerate relative plausibility measure on $Q_{rp}$ defined by $rp_{\downarrow}(x, y, z) = y$. Since $\sup_{[x, \infty]} \phi \leq \sup_{[x, \infty]} \psi$, for each $(x, y) \in \mathbb{R}^2$ such that $y < \phi(x)$ there is an $x' \in [x, \infty]$ such that $y < \psi(x')$: define $l(x, y, 0) = x$ and $l'(x, y, 0) = x'$. Analogously, since $\sup_{[0, x]} \phi \geq \sup_{[0, x]} \psi$, for each $(x, y) \in \mathbb{R}^2$ such that $y < \psi(x)$ there is an $x' \in [0, x]$ such that $y < \phi(x')$: define $l(x, y, 1) = x'$ and $l'(x, y, 1) = x$. We obtain two functions $l, l' \in \mathbb{R}^{Q_{rp}}$ such that $l \leq l'$; the equalities of (4.10) are also satisfied, since $l(x, y, z) = x'$ implies $y < \phi(x')$, and conversely $y < \phi(x')$ implies $l(x', y, 0) = x'$: therefore $(rp \circ l^{-1})_{\downarrow}(x') = \phi(x')$, and analogously for $\psi$ and $l'$.

$\square$
Theorem 4.2 implies that a likelihood-based decision criterion associates to each nondegenerate relative plausibility measure \( r_p \) (on some set \( Q_{r_p} \)) a functional \( V_{r_p} \) on \( \mathbb{P}^{Q_{r_p}} \) such that \( V_{r_p}(l) \) depends only on \( (r_p \circ l^{-1})^\downarrow \). Thus the values taken by \( l \) on \( \{r_p^\downarrow = 0\} \) have no influence on \( V_{r_p}(l) \), and in particular we have
\[
\inf_{\{r_p^\downarrow > 0\}} l \leq V_{r_p}(l) \leq \sup_{\{r_p^\downarrow > 0\}} l \quad \text{for all } l \in \mathbb{P}^{Q_{r_p}}.
\]

It can be easily proved that the functionals \( V_{r_p} = \sup_{\{r_p^\downarrow > 0\}} l \) correspond to a likelihood-based decision criterion, which can be called **essential minimax criterion**, since \( \sup_{\{r_p^\downarrow > 0\}} l = \text{ess}_{r_p} \sup l \). If \( Q_{r_p} = \mathcal{P} \) is a set of statistical models, and \( r_p \) is updated as described in Subsection 3.1.2 when new data are observed, then the density functions of the updated relative plausibility measures will still take the constant value 0 on \( \{r_p^\downarrow = 0\} \). Therefore the models in \( \{r_p^\downarrow = 0\} \) can be definitively discarded, since they will never have any influence on the likelihood-based decisions and inferences. A likelihood-based decision criterion discards thus the models \( P \) such that \( r_p\{P\} = 0 \), but it can be useful only if the models \( P \) such that \( r_p\{P\} \) is small have only a weak influence on the evaluation of a decision (the smaller \( r_p\{P\} \), the weaker the influence of \( P \)). That is, to be useful a likelihood-based decision criterion must satisfy some sort of continuity at 0 of the influence of a model, with respect to its relative plausibility: we shall consider this sort of continuity more precisely in Subsection 4.2.2, but it is clear that the essential minimax criterion does not satisfy it, since it uses \( r_p \) only to discard the models in \( \{r_p^\downarrow = 0\} \).

Theorem 4.2 states that a likelihood-based decision criterion corresponds to a functional \( S \) on \( \mathcal{S}_P \); in particular the two functions \( f_1, f_0 \) on \([0, 1]\) can be defined as follows on the basis of \( S \):
\[
f_1(x) = S(I_{\{0\}} + x I_{\{1\}}) \quad \text{and} \quad f_0(x) = S[(1 - x) I_{\{0\}} + I_{\{1\}}] - f_1(1),
\]
for all \( x \in [0, 1] \). The properties of \( S \) imply that \( f_1 \) and \( f_0 \) are nondecreasing, \( f_1(0) = f_0(0) = 0 \), and \( f_1(1) + f_0(1) = 1 \). The “continuity at 0 of the influence of a model, with respect to its relative plausibility” would imply the continuity of \( f_1 \) and \( f_0 \) at 0 and 1, respectively; in fact, for the essential minimax criterion, \( f_1 = I_{\{0,1\}} \) and \( f_0 = 0 \). If \( r_p \) is a nondegenerate relative plausibility measure on a set \( Q \), then the functional \( V_{r_p} \) on \( \mathbb{P}^Q \) can be defined as in Theorem 4.2: for all \( A \subseteq Q \) we have
\[
V_{r_p}(I_A) = S[r_p(Q \setminus A) I_{\{0\}} + r_p(A) I_{\{1\}}] = f_1[r_p(A)] + f_0[r_p(A)],
\]
where $\overline{rp}$ is the dual of the normalized representative of $rp$. In particular, if $rp = 1^\top$, then $V_{11}(I_A) = f_1(1)$ for all nonempty $A \subseteq Q$; this implies that two decisions can be considered equivalent by all likelihood-based decision criteria, even when one decision dominates the other, their uncertain losses are bounded, and $rp^\perp > 0$. Moreover, since $V_{11}(1) = 1$, no likelihood-based decision criterion can correspond to an additive functional $V_{11}$ on $\mathbb{P}^Q$ when $Q$ has at least three elements.

Consider a statistical decision problem described by a loss function $L$ on $\mathcal{P} \times \mathcal{D}$, let $rp$ be a nondegenerate relative plausibility measure on $\mathcal{P}$ describing the uncertain knowledge about the models in $\mathcal{P}$, and let $d, d' \in \mathcal{D}$ be two decisions such that $d$ dominates $d'$. The monotonicity (4.1) of the preference relation $\lesssim_{rp}$ assures that $d'$ is not preferred to $d$, but as shown above, it is possible that $d$ and $d'$ are equivalent according to all likelihood-based decision criteria (this is in particular the case when $rp \circ l_d^{-1} = rp \circ l_{d'}^{-1}$). However, as noted at the beginning of Section 1.1, if we have to choose between $d$ and $d'$, then we must choose $d$; hence, if necessary, a likelihood-based decision criterion can be refined by discarding from the set of optimal decisions those that are dominated (by other optimal decisions). Anyway, if $d'$ is strictly dominated by $d$, in the sense that $\inf(l_{d'} - l_d) > 0$, then a necessary and sufficient condition for $d$ to be preferred to $d'$ according to all likelihood-based decision criteria is that $l_d$ is essentially bounded with respect to $rp$, in the sense that $\text{ess}_{rp} \sup l_d < \infty$. This condition is necessary, because the functionals $V_{rp} = \text{ess}_{rp} \sup$ correspond to the essential minimax criterion, and it is sufficient, because if $\varepsilon = \inf(l_{d'} - l_d)$ and $s = \text{ess}_{rp} \sup l_d$, then $\frac{s + \varepsilon}{s} l_d \leq l_{d'}$ on $\{rp^\perp > 0\}$, and thus $V_{rp}(l_d) < V_{rp}(l_{d'})$ for all likelihood-based decision criteria, since $V_{rp}$ is monotonic and homogeneous, $V_{rp}(l_d) \leq s$, and $V_{rp}(l_{d'}) \geq \varepsilon$.

In Subsection 1.3.1 we have seen that the MPL criterion leads to the usual likelihood-based inference methods, when applied to some standard form of the corresponding decision problems; this is true also for a general likelihood-based decision criterion, if the corresponding functions $f_1$ and $f_0$ fulfill some conditions. For the estimation problems described by the loss functions $I_W$ and $I_{W_0}$ considered in Subsection 1.3.1, it can be easily proved that when a maximum likelihood estimate exists and is (really) unique, a likelihood-based decision criterion leads (practically) to it if and only if $f_1(x) < f_1(1) + f_0(1 - x)$ for all $x \in [0, 1)$. When applied to the hypothesis testing problem considered in Subsection 1.3.1 (with $c_1 > c_2$), a likelihood-based decision criterion leads practically to a likelihood ratio
test with some critical value $\beta$ depending on $\frac{c_2}{c_1}$ (where “practically” means that we do not bother about the case with $LR(H) = \beta$) if and only if the function $f : x \mapsto \frac{f_1(x)}{f_1(1) + f_0(1-x)}$ on $[0,1]$ is well-defined, it is strictly increasing on $f^{-1}(0,1)$, and it satisfies $\sup_{(0,1)} f = 1$ and $\inf_{(0,1)} f = 0$. The critical value $\beta$ is a strictly increasing function of $\frac{c_2}{c_1}$ if and only if $f$ is continuous: in this case, $\beta = f^{-1}(\frac{c_2}{c_1})$ (the inverse function $f^{-1}$ is well-defined on $(0,1)$). For the problem of selecting a confidence region considered in Subsection 1.3.1, a likelihood-based decision criterion is able to discriminate between regions $d \in \mathcal{D}$ with different values of $LR\{g \notin d\}$ if and only if $f_1$ is strictly increasing. In conclusion, a likelihood-based decision criterion leads to the usual likelihood-based inference methods if and only if $f_1$ is continuous and strictly increasing, and $f_0$ is continuous on $[0,1)$; in fact, for the MPL criterion, $f_1 = id_{[0,1]}$ and $f_0 = 0$.

4.1.1 Attitudes Toward Ignorance

Let $\mathcal{P}$ be a set of statistical models, and let $lik : \mathcal{P} \to [0,1]$ be the likelihood function induced by the observed data. If $lik = I_{\{P\}}$ for some model $P \in \mathcal{P}$, then all likelihood-based decisions and inferences would be the same as if we were certain of $P$ (that is, as if $P$ were the only representation of reality that we consider). If $lik \geq I_{\{P\}}$, then we would be more ignorant about which of the models in $\mathcal{P}$ is the best representation of the reality: in fact, the data would contain less information for discrimination between $P$ and the other models in $\mathcal{P}$ (in the sense of Kullback and Leibler, 1951); the larger $lik$, the more ignorant we would be. In the extreme case with $lik = I_{\{P\}}$, the data contain infinite information for discrimination between $P$ and the other models in $\mathcal{P}$; while at the other extreme we have the case of complete ignorance, in which $lik = 1$ and the data contain no information for discrimination between the models. In this sense, if $r_p$ and $r_p'$ are two nondegenerate relative plausibility measures on $\mathcal{P}$ such that $r_p \leq r_p'$, then we are more ignorant about which of the models in $\mathcal{P}$ is the best representation of the reality when $r_p'$ describes our uncertain knowledge about the models than when $r_p$ describes it.

A likelihood-based decision criterion reveals ignorance aversion if the corresponding functionals $V_{r_p}$ satisfy

$$r_p \leq r_p' \Rightarrow V_{r_p} \leq V_{r_p'}$$

for all pairs of nondegenerate relative plausibility measures $r_p, r_p'$ on the same set $\mathcal{Q}$. This property can be expressed as follows in terms of preference
relations:
\[ c \preceq_{rp} l \text{ and } rp \leq rp' \implies c \preceq_{rp'} l \text{ for all } c \in \overline{P} \text{ and all } l \in \overline{P}^\cap. \]

Since the evaluation of the decisions with constant losses \( c \in \overline{P} \) does not depend on our uncertain knowledge about the models, we can use these decisions to compare our preferences in different states of knowledge: ignorance aversion means that if we do not prefer a particular decision to the constant loss \( c \), then neither would we when we were more ignorant.

**Theorem 4.3.** Consider a likelihood-based decision criterion and the corresponding functional \( S : S_{\overline{P}} \rightarrow \overline{P} \); the following four assertions are equivalent:

1. The decision criterion reveals ignorance aversion,
2. \( S \) is monotone,
3. \( S \) is \( \cap \)-independent,
4. each relation \( \preceq_{rp} \) is represented by the functional \( V_{rp} : l \mapsto S(rp\{l \geq \cdot\}) \).

**Proof.** Let \( \varphi, \psi \in S_{\overline{P}} \) such that \( \varphi \leq \psi \). If the decision criterion reveals ignorance aversion, then \( S(\varphi) = V_{\varphi\{id_{\overline{P}}\}} \leq V_{\psi\{id_{\overline{P}}\}} = S(\psi) \), and thus \( S \) is monotone. If \( \varphi_{[0,\infty]} = \psi_{[0,\infty]} \), then \( S(\psi) \leq S(\varphi) \) follows from condition (4.9), and therefore \( S \) is \( \cap \)-independent when it is monotone. If \( S \) is \( \cap \)-independent, and \( \chi = (rp \circ l^{-1})^{+} \), then

\[ V_{rp}(l) = S(\chi) = S(I_{\{0\}} + \chi I_{(0,\infty)}) = S(rp\{l \geq \cdot\}), \]

where the last equality is implied by condition (4.9) with \( \varphi = I_{\{0\}} + \chi I_{(0,\infty)} \) and \( \psi : x \mapsto \sup_{[x,\infty]} \chi = rp\{l \geq x\} \). Finally, if the fourth assertion holds, then the ignorance aversion is a consequence of condition (4.9), since \( \sup_{[x,\infty]} rp\{l \geq \cdot\} = rp\{l \geq x\} \) and \( \sup_{[0,x]} rp\{l \geq \cdot\} = 1. \)

Let \( \mathcal{H} \) be a nonempty subset of \( \mathcal{P} \), and consider the (nondegenerate) relative plausibility measure \( (I_{\mathcal{H}})^\dagger \) on \( \mathcal{P} \). Theorem 4.3 implies that the functional on \( \overline{P}^\mathcal{P} \) corresponding to a likelihood-based decision criterion revealing ignorance aversion is simply \( V_{(I_{\mathcal{H}})^\dagger} = \sup_{\mathcal{H}} \). In particular, in the case of complete ignorance about which of the models in \( \mathcal{P} \) is the best representation of the reality we have \( V_{1^{\dagger}} = \sup \), and therefore the likelihood-based decision criteria revealing ignorance aversion can be considered as generalizations of the minimax criterion. Moreover, for these cri-
teria \( f_1(1) = S(I_{[0,1]}), S(I_{[1]}) = 1 \), and thus \( f_0 = 0 \); hence a likelihood-based decision criterion revealing ignorance aversion leads to the usual likelihood-based inference methods if and only if \( f_1 : [0, 1] \rightarrow [0, 1] \) is an increasing bijection. The fourth assertion of Theorem 4.3 suggests the idea that the functionals \( V_{r_p} \) corresponding to a decision criterion revealing ignorance aversion can be expressed by means of an integral respecting distributional dominance; the next theorem states that the injectivity of \( f_1 \) is a sufficient condition for this integral representation.

**Theorem 4.4.** Consider a likelihood-based decision criterion such that the corresponding function \( f_1 : [0, 1] \rightarrow [0, 1] \) is strictly increasing, and let \( \mathcal{M} \) be the class of all monotone measures taking values in the range of \( f_1 \). The decision criterion reveals ignorance aversion if and only if there is a support-based, homogeneous integral respecting distributional dominance on \( \mathcal{M} \) such that each relation \( \preceq_{r_p} \) is represented by the functional \( V_{r_p} : l \mapsto \int l d(f_1 \circ r_p) \); the integral is unique.

**Proof.** The “if” part is simple: the decision criterion reveals ignorance aversion because \( f_1 \) is nondecreasing and an integral respecting distributional dominance is monotone with respect to measures (Theorem 2.12).

For the “only if” part, we have \( V_{r_p}(I_A) = f_1[r_p(A)] \) for all nondegenerate relative plausibility measures \( r_p \) and all \( A \subseteq Q_{r_p} \), and so the equality \( \int l d(f_1 \circ r_p) = V_{r_p}(l) \) defines an integral on the class \( \mathcal{M}' \) of all measures of the form \( f_1 \circ \mu \), for all normalized, completely maxitive measures \( \mu \), since for each \( \nu \in \mathcal{M}' \) there is a unique nondegenerate relative plausibility measure \( r_p \) such that \( \nu = f_1 \circ r_p \). Theorem 4.3 implies that the integral respects distributional dominance, and thus it corresponds to a functional on \( \mathcal{D}_F(\mathcal{M}') \) that can be extended to \( \mathcal{D}_F(\mathcal{M}) \) by means of \( 0 \)-independence. The integral on \( \mathcal{M} \) corresponding to the extended functional satisfies the desired properties, and its uniqueness can be easily proved by exploiting the fact that \( \mathcal{D}_F(\mathcal{M}) = \mathcal{D}_F(\mathcal{M}'') \) for the class \( \mathcal{M}'' \) of all measures of the form \( f_1 \circ \mu \), for all possibility measures \( \mu \). \( \square \)

The integral obtained in Theorem 4.4 can be easily (but not uniquely) extended to a support-based, homogeneous integral respecting distributional dominance on the class of all monotone measures, when \( f_1 \) is left-continuous; the next example shows that such an extension is in general not possible when \( f_1 \) is not left-continuous.
Example 4.5. Let $\delta$ be the function on $\overline{P}$ defined in Example 2.2, and consider the likelihood-based decision criterion associating to each nondegenerate relative plausibility measure $rp$ (on some set $Q_{rp}$) the relation $\prec_{rp}$ on $\overline{P} \otimes_{rp}$ represented by the functional $V_{rp} : l \mapsto \int l \, d(\delta \circ rp)$. This decision criterion reveals ignorance aversion and $f_1 = \delta|_{[0,1]}$ is strictly increasing, but the results of Example 2.11 imply that no integral respecting distributional dominance on the class of all monotonic measures can correspond to the rectangular integral on the class of all monotonic measures taking values in the range of $f_1$.

In Theorem 4.4 the injectivity of $f_1$ is needed to define the integral on the basis of the functionals $V_{rp}$, but in general for all nondecreasing functions $f_1 : [0, 1] \to [0, 1]$ such that $f_1(0) = 0$ and $f_1(1) = 1$, and for all homogeneous integrals respecting distributional dominance on the class of all normalized, monotonic measures taking values in the range of $f_1$, the functionals $V_{rp} : l \mapsto \int l \, d(f_1 \circ rp)$ correspond to a likelihood-based decision criterion revealing ignorance aversion. In particular, if the range of $f_1$ is $\{0, 1\}$, then the choice of the integral has no influence: $V_{rp} = \text{ess} f_{1 \circ rp} \sup$. If $f_1 = I_{(\beta,1]}$ for some $\beta \in (0,1)$, then the functionals $V_{rp} = \text{ess} I_{(\beta,1]} \circ rp \sup$ correspond to the LRM$\beta$ criterion; while if $f_1 = I_{(1)}$, then the functionals $V_{rp} = \text{ess} I_{(1)} \circ rp \sup$ correspond to the MLD criterion. These two likelihood-based decision criteria reveal thus ignorance aversion, and the corresponding functionals $V_{rp}$ can be expressed as integrals with respect to a distortion of $rp$.

Theorem 4.3 implies that a likelihood-based decision criterion revealing ignorance aversion corresponds to a $0$-independent functional $S$ on $S_{\overline{P}}$, but since the functionals $V_{rp} : l \mapsto S(rp\{l \geq \cdot\})$ on $\overline{P} \otimes_{rp}$ depend on $rp$ through its normalized representative only, in general they are not “support-based”, in the sense that in general $rp\{l > 0\} > 0$ does not imply $V_{rp}(l) = V_{rp|_{\{l > 0\}}}(l|_{\{l > 0\}})$. In fact, the essential minimax criterion is the only likelihood-based decision criterion such that the corresponding functionals $V_{rp}$ are “support-based” in this sense. However, for decision criteria revealing ignorance aversion (and thus generalizing the minimax criterion) it is reasonable to assume that the preference between two decisions does not depend on the models for which both decisions are correct; that is, the preference between $l$ and $l'$ is left unchanged when we restrict attention to the subset $\{l > 0\} \cup \{l' > 0\} = \{l + l' > 0\}$ of $Q_{rp}$. A likelihood-based decision criterion is said to be $0$-independent if for each nondegenerate
relative plausibility measure \( rp \) (on some set \( Q_{rp} \))

\[
l \preceq_{rp} l' \iff l|_{\{l + l' > 0\}} \preceq_{rp|_{\{l + l' > 0\}}} l'|_{\{l + l' > 0\}} \quad \text{for all } l, l' \in \mathcal{P}_{Q_{rp}} \text{ such that } rp\{l + l' > 0\} > 0.
\]

The next theorem implies in particular that all \( 0 \)-independent likelihood-based decision criteria reveal ignorance aversion.

**Theorem 4.6.** A likelihood-based decision criterion is \( 0 \)-independent if and only if either it is the essential minimax criterion, or there are a support-based, regular integral on the class of all finite, monotone measures, and a positive real number \( \alpha \) such that each relation \( \preceq_{rp} \) is represented by the functional \( l \mapsto \int l \, d(rp^\alpha) \); in this case, the integral is unique.

**Proof.** For the “if” part, let \( A \subseteq Q_{rp} \) be a set such that \( rp(A) > 0 \) and \( l|_{Q_{rp} \setminus A} = 0 \). The functionals \( V_{rp} \) corresponding to the likelihood-based decision criterion satisfy \( V_{rp}(l) = V_{rp|_{A}}(l|_{A}) \) in the case of the essential minimax criterion, while in the other case they satisfy

\[
V_{rp}(l) = \int l \, d(rp^\alpha) = \int l|_{A} \, d(rp^\alpha)|_{A} = [rp(A)]^\alpha V_{rp|_{A}} (l|_{A}),
\]

because the integral is support-based and bihomogeneous. To obtain the \( 0 \)-independence it suffices to choose \( A = \{l + l' > 0\} \).

For the “only if” part, first note that the function \( f_1 \) on \([0, 1]\) corresponding to the \( 0 \)-independent likelihood-based decision criterion considered is positive on \((0, 1]\). To prove this, let \( rp \) be the (nondegenerate) relative plausibility measure on \( \{0, 1\} \) defined by \( rp^\perp(1) = c \in (0, 1] \) and \( rp^\perp(0) = 1 \). Then \( f_1(c) = V_{rp}(I_{\{1\}}) > V_{rp}(0) = 0 \), because \( rp\{1\} > 0 \) and \( V_{rp|_{\{1\}}}(I_{\{1\}}|_{\{1\}}) = 1 > 0 = V_{rp|_{\{1\}}}(0|_{\{1\}}) \). We can now prove that the functional \( S \) on \( \mathcal{S}_F \) corresponding to the decision criterion satisfies the following equality for all \( c \in (0, 1] \) and all \( \varphi \in \mathcal{S}_F^* \):

\[
S(I_{\{0\}} + c \varphi I_{(0, \infty)}) = f_1(c) S(\varphi).
\] (4.11)

Define \( Q = \mathcal{P} \times \{0, 1\} \times \{0, 1\} \) and \( A = \{(x, y, z) \in Q : z = 1\} \), let \( rp' \) be the (nondegenerate) relative plausibility measure on \( Q \) defined by \( (rp')^\perp(x, y, z) = c z \varphi(x) + (1 - z) \), and let \( l, l' \) be the two functions on \( Q \) defined by \( l(x, y, z) = x z \) and \( l'(x, y, z) = \frac{S(\varphi)}{f_1(1)} y z \), respectively. If \( S(\varphi) = 0 \), then \( \varphi = I_{\{0\}} \), because \( f_1 \) is positive on \((0, 1]\), and \( S \) fulfills
4.1 Decision Theoretic Properties

conditions (4.8) and (4.9). Hence if \( S(\varphi) = 0 \), then the equality (4.11) is obvious; while if \( S(\varphi) > 0 \), then \( rp'l[l + l' > 0] > 0 \), and (4.11) can be obtained as follows (the choice \( \varphi = I_{\{1\}} \) implies \( f_1(1) = 1 \):

\[
S(I_{\{0\}} + c\varphi I_{\{0,\infty\}}) = V_{rp'}(l) = V_{rp'}(l') = S\left(I_{\{0\}} + c I_{\left\{ \frac{g(\varphi)}{f_1(1)} \right\}} \right) = f_1(c) \frac{S(\varphi)}{f_1(1)},
\]

where the second equality is implied by 0-independence and

\[
V_{rp'|A}(l|A) = S(\varphi) = S\left(I_{\left\{ \frac{g(\varphi)}{f_1(1)} \right\}} \right) = V_{rp'|A}(l'|A),
\]

because \( A \cap \{ l|A + l'|A > 0 \} = \{ l + l' > 0 \} \) follows from \( (l + l')|\emptyset \setminus A = 0. \) If \( n \) is a positive integer, and \( \varphi = I_{\{0\}} + c^n I_{\{1\}} \), then equality (4.11) implies \( f_1(c^{n+1}) = f_1(c) f_1(c^n) \). By induction we obtain that \( f_1(c^y) = [f_1(c)]^y \) holds for all positive integers \( y \), and from this we obtain that it holds also for all positive rational numbers \( y \). Hence the two functions \( y \mapsto f_1(c^y) \) and \( y \mapsto [f_1(c)]^y \) on \( \mathbb{P} \) are nonincreasing and they correspond on the set of all positive rational numbers: since the second one is continuous, they are equal. Let \( c = e^{-1} \) and \( \alpha = - \log[f_1(e^{-1})] \); for all \( x \in (0, 1) \) we obtain \( f_1(x) = [f_1(e^{-1})]^{- \log x} = x^\alpha \). If \( \alpha = 0 \), then \( f_1 = I_{\{0,1\}} \), and the decision criterion is the essential minimax criterion, because for all nondegenerate relative plausibility measures \( rp \) and all \( q \in \{ rp \downarrow > 0 \} \) we have

\[
V_{rp}(l) \geq V_{rp}[l(q) I_{\{q\}] \geq l(q) f_1(rp\{q\}) = l(q).
\]

If \( \alpha \in \mathbb{P} \), then let \( F \) be the functional on \( \Psi = \{ \varphi \in \mathcal{NI}_\mathbb{P} : \varphi(0) \in \mathbb{P} \} \) defined by

\[
F(\varphi) = \varphi(0) S\left[ \left( \frac{\varphi}{\varphi(0)} \right)^{\frac{1}{\alpha}} \right].
\]

Equality (4.11) implies that \( F \) is 0-independent, since for all \( \varphi, \psi \in \Psi \) such that \( \varphi|_{\{0,\infty\}} = \psi|_{\{0,\infty\}} \) and \( \varphi(0) > \psi(0) \) we have

\[
F(\varphi) = \varphi(0) S\left[ I_{\{0\}} + \left( \varphi(0) \right)^{\frac{1}{\alpha}} \left( \frac{\psi}{\varphi(0)} \right)^{\frac{1}{\alpha}} I_{\{0,\infty\}} \right] =
\]

\[
= \varphi(0) f_1 \left[ \left( \frac{\psi}{\varphi(0)} \right)^{\frac{1}{\alpha}} \right] S\left[ \left( \frac{\psi}{\varphi(0)} \right)^{\frac{1}{\alpha}} \right] = F(\psi).
\]

Therefore \( F \) is also monotonic, because for all \( \varphi, \psi \in \Psi \) such that \( \varphi \leq \psi \) we have

\[
F(\varphi) = F[\psi(0) I_{\{0\}} + \varphi I_{\{0,\infty\}}] \leq F(\psi),
\]

where the inequality follows from the monotonicity of \( S \), which is implied by Theorem 4.3, since equality (4.11) with \( c = 1 \) expresses the
0-independence of $S$. Moreover, $F$ is clearly bihomogeneous, and for all $y \in [0, \infty)$ we have

$$F(I_{[0,1]} + y I_{(0,)} = (y + 1) f_1 \left( \left( \frac{1}{y+1} \right)^k \right) = 1.$$ 

Hence $F$ is the restriction to $\Psi$ of the calibrated, 0-independent, bihomogeneous, monotonic functional $\varphi \mapsto \sup_{\{\psi \in \Psi, \psi \leq \varphi\}} F$ on $\mathcal{N}I_{\mathbb{P}}$ corresponding to a support-based, regular integral on the class of all monotonic measures (Theorem 2.13). The desired integral representation follows from Theorem 4.3:

$$V_{rp}(l) = S(rp\{l \geq \cdot\}) = F[(rp^\alpha)\{l \geq \cdot\}] = \int l \, d(rp^\alpha),$$

and the uniqueness of the integral on the class of all finite, monotonic measures follows from the uniqueness of the functionals $V_{rp}$ (implied by Theorem 4.1).

Theorem 4.6 implies in particular that a 0-independent likelihood-based decision criterion leads to the usual likelihood-based inference methods if and only if it is not the essential minimax criterion, since for this criterion $f_1 = I_{(0,1]}$, while otherwise there is an $\alpha \in \mathbb{P}$ such that $f_1 : x \mapsto x^\alpha$. In this case, if $L$ is a loss function on $\mathcal{P} \times \mathcal{D}$, and $lik$ is the likelihood function on $\mathcal{P}$ induced by the observed data (while before observing the data we had no information for discrimination between the models in $\mathcal{P}$), then the 0-independent decision criterion applied to the decision problem described by $L$ consists in minimizing $\int l_d \, d(lik^\alpha)$, where the integral is regular and support-based on the class of all monotonic measures. In fact, each support-based, regular integral on the class of all finite, monotonic measures can be extended to a support-based, regular integral on the class of all monotonic measures; but the extension is not unique. However, the extension used in the proof of Theorem 4.6 is the most natural one, since it is the one characterized by the following continuity property:

$$\int f \, d\mu = \lim_{c \uparrow \infty} \int f \, d\min\{\mu, c\} \quad (4.12)$$

for all sets $Q$, all monotonic measures $\mu$ on $Q$, and all functions $f : Q \to \mathbb{P}$. Thanks to the extension of the integral, the criterion can be applied also when $lik$ is an unbounded pseudo likelihood function (obtained for example when using the densities of a continuous random object to approximate
the likelihood function). The decision criterion can be interpreted as evaluating the loss we would incur by making each decision \( d \in D \) by means of the integral of \( l_d \) with respect to a completely maxitive measure whose density function is proportional to the distorted likelihood function \( lik^\alpha \).

If \( \alpha \in (0,1) \), then the distortion represents an attenuation of the likelihood function \( lik \), while if \( \alpha \in (1,\infty) \), then the distortion represents an intensification of \( lik \); in fact, \( lik^\alpha \) contains \( \alpha \) times the information in \( lik \) for discrimination between the models in \( \mathcal{P} \) (in the sense of Kullback and Leibler, 1951). Hence, \( \alpha \in \mathbb{P} \) can be interpreted as a parameter expressing the confidence in the information provided by the likelihood function \( lik \), or more generally, by the (nondegenerate) relative plausibility measure \( rp \). In fact, the limits as \( \alpha \) tends to 0 of the functionals \( V_{rp} : l \mapsto \int l \, d(rp^\alpha) \) corresponding to the decision criterion considered are the functionals \( V_{rp} : l \mapsto \int l \, d(I_{[0,1]} \circ rp) \) corresponding to the essential minimax criterion (which usually amounts to disregarding the information provided by \( rp \)); while the limit of \( rp^\alpha \) as \( \alpha \) tends to infinity is \( I_{[1]} \circ rp \), and the functionals \( V_{rp} : l \mapsto \int l \, d(I_{[1]} \circ rp) \) correspond to the MLD criterion (which usually amounts to considering the model maximizing \( rp \) as certain). The limit of \( \int l \, d(rp^\alpha) \) as \( \alpha \) tends to infinity is not necessarily \( \int l \, d(I_{[1]} \circ rp) \), but if \( \text{ess}_{rp} \sup l < \infty \), and the integral is quasi-subadditive, then \( \lim_{\alpha \uparrow \infty} \int l \, d(rp^\alpha) = \int l \, d(I_{[1]} \circ rp) \) follows easily from Theorem 2.23.

The attitude toward ignorance opposed to aversion can be called attraction. A likelihood-based decision criterion reveals ignorance attraction if the corresponding functionals \( V_{rp} \) satisfy

\[
(rp \leq rp') \implies V_{rp} \geq V_{rp'}
\]

for all pairs of nondegenerate relative plausibility measures \( rp, rp' \) on the same set \( Q \). This property can be expressed as follows in terms of preference relations:

\[
l \lesssim_{rp} c \quad \text{and} \quad rp \leq rp' \quad \Rightarrow \quad l \lesssim_{rp'} c \quad \text{for all} \quad c \in \mathbb{P} \quad \text{and all} \quad l \in \mathbb{P}^2;
\]

ignorance attraction means that if we do not prefer the constant loss \( c \) to a particular decision, then neither would we when we were more ignorant.

The next theorem parallels Theorem 4.3, and implies in particular that the function \( f_1 \) on \([0,1]\) corresponding to a likelihood-based decision criterion revealing ignorance attraction satisfies \( f_1(1) = S(I_{[0,1]}) = S(I_{[0,1]}) = 0 \). Therefore, a decision criterion cannot reveal both ignorance aversion and attraction, and a decision criterion such that \( f_1(1) \in (0,1) \) reveals nei-
ther ignorance aversion nor attraction (we shall consider examples of such decision criteria in a moment).

**Theorem 4.7.** Consider a likelihood-based decision criterion and the corresponding functional \( S : S_p \rightarrow P \); the following four assertions are equivalent:

1. the decision criterion reveals ignorance attraction,
2. \( \varphi \preceq \psi \Rightarrow S(\varphi) \geq S(\psi) \) for all \( \varphi, \psi \in S_p \),
3. \( S \) is \( \infty \)-independent,
4. each relation \( \preceq_{rp} \) is represented by the functional \( V_{rp} : l \mapsto S(rp\{l \leq \cdot\}) \), where \( rp\{l \leq \cdot\} \) denotes the function \( x \mapsto rp\{l \leq x\} \) on \( P \).

**Proof.** Let \( \varphi, \psi \in S_p \) such that \( \varphi \preceq \psi \). If the decision criterion reveals ignorance attraction, then \( S(\varphi) = V_{\varphi} (id_P) \geq V_{\psi} (id_P) = S(\psi) \), and thus the second assertion holds. If \( \varphi|_{(0,\infty)} = \psi|_{(0,\infty)} \), then \( S(\varphi) \leq S(\psi) \) follows from condition (4.9), and therefore \( S \) is \( \infty \)-independent when the second assertion holds. If \( S \) is \( \infty \)-independent, and \( \chi = (rp \circ l^{-1})^† \), then

\[
V_{rp}(l) = S(\chi) = S(\chi I_{[0,\infty)} + I_{\{\infty\}}) = S(rp\{l \leq \cdot\}),
\]

where the last equality follows from condition (4.9) with \( \varphi = \chi I_{[0,\infty)} + I_{\{\infty\}} \) and \( \psi : x \mapsto \sup_{[0,x]} \chi = rp\{l \leq l \} \). Finally, if the fourth assertion holds, then the ignorance attraction is a consequence of condition (4.9), since \( \sup_{[x,\infty]} rp\{l \leq \cdot\} = 1 \) and \( \sup_{[0,x]} rp\{l \leq \cdot\} = rp\{l \leq x\} \). \( \square \)

Let \( H \) be a nonempty subset of \( P \), and consider the (nondegenerate) relative plausibility measure \( (I_H)^† \) on \( P \). Theorem 4.7 implies that the functional on \( P^F \) corresponding to a likelihood-based decision criterion revealing ignorance attraction is simply \( V(I_H)^† = \inf_H \). In particular, in the case of complete ignorance about which of the models in \( P \) is the best representation of the reality we have \( V_{1,1} = \inf \), and therefore the likelihood-based decision criteria revealing ignorance attraction can be considered as generalizations of the minimin criterion. The counterpart of Theorem 4.4 for a decision criterion revealing ignorance attraction can be easily obtained: it states in particular that if the corresponding function \( f_0 \) on \([0,1]\) is strictly increasing, then each relation \( \preceq_{rp} \) is represented by the functional \( V_{rp} : l \mapsto \int l d(f_0 \circ rp) \), for a (unique) support-based, homogeneous integral on the class of all monotonic measures taking values in
the range of \( f_0 \), but the integral would “respect distributional dominance” only if this property were defined in terms of the “distribution function” \( \mu\{f > \cdot\} \) instead of \( \mu\{f \geq \cdot\} \). However, in general for all nondecreasing functions \( f_0 : [0, 1] \rightarrow [0, 1] \) such that \( f_0(0) = 0 \) and \( f_0(1) = 1 \), and for all homogeneous integrals respecting distributional dominance on the class of all normalized, monotonic measures taking values in the range of \( f_0 \), the functionals \( V_{rp} : l \mapsto \int l d(f_0 \circ \overline{\mathcal{P}}) \) correspond to a likelihood-based decision criterion revealing ignorance attraction. In particular, if the range of \( f_0 \) is \( \{0, 1\} \), then the choice of the integral has no influence: \( V_{rp} = \mathrm{ess}_{f_0 \circ \overline{\mathcal{P}}} \sup \). If \( f_0 = I_{[\beta, 1]} \) for some \( \beta \in (0, 1) \), then the functionals \( V_{rp} = \mathrm{ess}_{I_{[\beta, 1]} \circ \overline{\mathcal{P}}} \sup = \inf_{\{\tau_{\mathcal{P}} \mid \beta\}} \) correspond to what can be called likelihood-based region minimin criterion (in analogy with the LRM\( _\beta \) criterion): it consists in reducing \( \mathcal{P} \) to the likelihood-based confidence region for \( P \) with cutoff point \( \beta \), before applying the minimin criterion. If \( f_0 = I_{(0,1]} \), then the functionals \( V_{rp} = \mathrm{ess}_{I_{(0,1]} \circ \overline{\mathcal{P}}} \sup = \lim_{\beta \uparrow 1} \inf_{\{\tau_{\mathcal{P}} \mid \beta\}} \) correspond to the ignorance attracted analogue of the MLD criterion.

Consider a statistical decision problem described by a loss function \( L \) on \( \mathcal{P} \times \mathcal{D} \), and let \( \text{lik} \) be the likelihood function on \( \mathcal{P} \) induced by the observed data (while before observing the data we had no information for discrimination between the models in \( \mathcal{P} \)). If a maximum likelihood estimate \( \hat{P}_{ML} \) of \( P \) exists, then each decision \( d \in \mathcal{D} \) that is correct for \( \hat{P}_{ML} \) is optimal according to all likelihood-based decision criteria revealing ignorance attraction, since \( (\text{lik}^{-1} \circ l_d^{-1})^{-1}(0) = 1 \) implies \( V_{lik}(l_d) = S(1) = 0 \). Moreover, if \( l' \in \mathcal{D} \) is not correct for \( \hat{P}_{ML} \), and there is a topology on \( \mathcal{P} \) such that \( l_{d'} \) is continuous, and \( LR(P \setminus N) < 1 \) for all neighborhoods \( N \) of \( \hat{P}_{ML} \), then \( V_{lik}(l_{d'}) > 0 \) for all likelihood-based decision criteria such that \( f_0 \) is positive on \( (0, 1] \). That is, the decision criteria revealing ignorance attraction and such that \( f_0 \) is positive on \( (0, 1] \) are almost equivalent to the MLD criterion, at least when a maximum likelihood estimate \( \hat{P}_{ML} \) of \( P \) exists and there is a correct decision for \( \hat{P}_{ML} \). So the likelihood-based decision criteria revealing ignorance attraction are not so useless for statistical decision problems as the minimin criterion (of which they are generalizations), but since \( f_1 = 0 \), they do not lead to the usual likelihood-based inference methods (except for the maximum likelihood method, if \( f_0 \) is positive on \( (0, 1] \)).

Let \( rp \) be a nondegenerate relative plausibility measure on \( \mathcal{P} \), and let \( l : \mathcal{P} \rightarrow \overline{\mathcal{P}} \) be a function. It can be easily proved that if \( V_{rp}' \) is the functional on \( \overline{\mathcal{P}}^\mathcal{P} \) corresponding to a likelihood-based decision criterion revealing ig-
orance aversion, and \( V''_{rp} \) is the functional on \( P^P \) corresponding to a decision criterion revealing ignorance attraction, then the following inequalities hold:

\[
\inf_{\{rp^1 > 0\}} l \leq V''_{rp}(l) \leq \lim_{\beta \uparrow 1} \inf_{\{rp^1 > \beta\}} l \leq \lim_{\beta \uparrow 1} \sup_{\{rp^1 > \beta\}} l \leq V'_{rp}(l) \leq \sup_{\{rp^1 > 0\}} l.
\]

That is, a likelihood-based decision criterion revealing ignorance aversion can be interpreted as comparing the decisions \( d \in D \) by means of an upper evaluation \( V'_{rp}(l_d) \) of \( l_d \), while a decision criterion revealing ignorance attraction compares them by means of a lower evaluation \( V''_{rp}(l_d) \) of \( l_d \). The sixth value in the inequalities (4.13) is the upper evaluation of \( l \) considered by the essential minimax criterion, and analogously the ignorance attracted decision criterion using the first value in (4.13) as the lower evaluation of \( l \) can be called essential minimin criterion. The likelihood-based decision criteria using the fourth and third values in (4.13) as evaluations of \( l \) are the MLD criterion and its ignorance attracted analogue, respectively.

If \( \delta : [0,1] \rightarrow [0,1] \) is a nondecreasing function such that \( \delta(0) = 0 \), then the dual of \( \delta \) is the function \( \bar{\delta} \) on \([0,1]\) defined by

\[
\bar{\delta}(x) = \delta(1) - \delta(1 - x) \quad \text{for all } x \in [0,1].
\]

Note that \( \bar{\delta} \) is nondecreasing, with \( \bar{\delta}(0) = 0 \) and \( \bar{\delta}(1) = \delta(1) \); the use of the term “dual” is justified by the equality \( \bar{\delta} = \delta \) and by the property that if \( \mu \) is a normalized, monotone measure, then \( \bar{\delta} \circ \mu = \delta \circ \bar{\mu} \). If there are two nondecreasing functions \( f_1, f_0 : [0,1] \rightarrow [0,1] \) such that \( f_1(0) = f_0(0) = 0 \) and \( f_1(1) = f_0(1) = 1 \), and a homogeneous integral respecting distributional dominance on the class of all normalized, monotone measures, such that \( V''_{rp} \) and \( V'_{rp} \) can be written as integrals with respect to \( f_1 \circ rp \) and \( f_0 \circ rp \), respectively, then the functionals \( V'_{rp} \) and \( V''_{rp} \) can be considered as giving conjugate upper and lower evaluations if and only if \( f_1 \circ rp \) and \( f_0 \circ rp \) are dual to each other (and this holds for all nondegenerate relative plausibility measures \( rp \) if and only if \( f_1 \) and \( f_0 \) are dual to each other).

In this case, the inequalities (4.13) can be rewritten as follows:

\[
\int l \ d(I_{[0,1]} \circ rp) \leq \int l \ d(f_1 \circ rp) \leq \int l \ d(I_{[1]} \circ rp) \leq \int l \ d(I_{[0,1]} \circ rp),
\]

(4.14)
because the functions $I_{[0,1]}$ and $I_{[1]}$ on $[0, 1]$ are dual to each other. Thus in particular the essential minimax criterion and its ignorance attracted analogue (the essential minimin criterion) correspond to functionals giving conjugate upper and lower evaluations, and the same holds for the MLD criterion and its ignorance attracted analogue. The interpretation of the fifth and second values in (4.14) as conjugate upper and lower evaluations of $l$ is particularly clear when $f_1$ is left-continuous and the integral is regular and symmetric on the class of all monotone measures: in this case, Theorem 2.18 can be applied, and in particular, if $f_1|_{(0,1)}$ has range in $(0,1)$, then the inequalities (4.14) can be rewritten as follows, using the notation of Theorem 2.18:

\[
\liminf_{x \uparrow 1} \{ f_1 \circ \mathbb{F}^\uparrow \geq 1 - x \} \ l \leq F(I_{[0,1]} \ \inf_{f_1 \circ \mathbb{F}^\uparrow \geq 1 - x} \ l) \leq \liminf_{x \rightarrow 0} \{ f_1 \circ \mathbb{F}^\uparrow \geq 1 - x \} \ l \\
\leq \limsup_{x \downarrow 1} \{ f_1 \circ \mathbb{F}^\uparrow \geq x \} \ l \leq F(I_{[0,1]} \ \sup_{f_1 \circ \mathbb{F}^\uparrow \geq x} \ l) \leq \limsup_{x \rightarrow 0} \{ f_1 \circ \mathbb{F}^\uparrow \geq x \} \ l.
\]

The functionals $V'_{rp}$ and $V''_{rp}$ corresponding to two likelihood-based decision criteria revealing respectively ignorance aversion and attraction can thus be interpreted as giving respectively upper and lower evaluations: these can be combined to obtain intermediate evaluations, by means of an averaging function $a : H(id_{\mathbb{P}}) \rightarrow \mathbb{P}$, where $H(id_{\mathbb{P}}) = \{(x, y) \in \mathbb{P}^2 : x \geq y\}$ is the hypograph of $id_{\mathbb{P}}$. Using Theorem 4.2, it can be easily proved that the resulting functionals $V_{rp} = a(V'_{rp}, V''_{rp})$ correspond to a likelihood-based decision criterion if and only if $a(1, 1) = 1$, the function $a$ is positively homogeneous (that is, $a(cx, cy) = ca(x, y)$ for all $(x, y) \in H(id_{\mathbb{P}})$ and all $c \in \mathbb{P}$), and it is nondecreasing in both arguments. In the case of complete ignorance about which of the models in $\mathcal{P}$ is the best representation of the reality, the resulting decision criterion corresponds to the functional $V_{11} = a(\sup, \inf)$ on $\mathbb{P}^2$; therefore the resulting decision criterion reveals ignorance aversion or attraction if and only if $a(x, y) = x$ or $a(x, y) = y$ for all $(x, y) \in H(id_{\mathbb{P}})$, while in all other cases it reveals neither ignorance aversion nor attraction. In particular, if $\lambda \in (0, 1)$, and $a_{\lambda}$ is the function on $H(id_{\mathbb{P}})$ defined by $a_{\lambda}(x, y) = \lambda y + (1 - \lambda) x$ (for all $(x, y) \in H(id_{\mathbb{P}})$), then the likelihood-based decision criteria obtained by means of the averaging function $a_{\lambda}$ can be considered as generalizations of the Hurwicz criterion with optimism parameter $\lambda$.

Consider a regular integral on the class of all monotone measures, and let $f_1, f_0 : [0,1] \rightarrow [0,1]$ be two nondecreasing functions such that $f_1(0) = f_0(0) = 0$ and $f_1(1) + f_0(1) = 1$. A likelihood-based decision
criterion with corresponding functions $f_1, f_0$ can be defined for example through the functionals

$$V_{rp} : l \mapsto \int l \, d(f_1 \circ r_p) + \int l \, d(f_0 \circ r_p),$$

or through the functionals

$$V_{rp} : l \mapsto \int l \, d(f_1 \circ r_p + f_0 \circ r_p).$$

If $f_1(1) = 1$ or $f_1(1) = 0$, then the two definitions correspond, and the decision criterion reveals respectively ignorance aversion or attraction; but if $f_1(1) \in (0, 1)$, then in general the two definitions are different, although both decision criteria generalize the Hurwicz criterion with optimism parameter $\lambda = f_0(1)$. In this case, the first definition corresponds to the application of the averaging function $\alpha_\lambda$ to the integrals with respect to $\frac{1}{1-\lambda} f_1 \circ r_p$ and $\frac{1}{\lambda} f_0 \circ r_p$, respectively; while in general the second definition cannot be obtained as $a(V'_{rp}, V''_{rp})$ for an averaging function $a$ on $H(id_\mathbb{R})$. However, for the Choquet integral the two definitions correspond (thanks to the additivity with respect to measures), and in particular, if $\frac{1}{1-\lambda} f_1$ and $\frac{1}{\lambda} f_0$ are dual to each other, and the function $\delta = \frac{1}{1-\lambda} f_1 = \frac{1}{\lambda} f_0$ is left-continuous, then using Theorem 2.18 we can clarify the connection with the Hurwicz criterion by expressing the functionals $V_{rp}$ as follows by means of the Lebesgue integral (with respect to the Lebesgue measure on $[0, 1]$):

$$V_{rp} : l \mapsto \int_0^1 \left[ \lambda \inf_{\{\delta \circ r_p \leq x\}} l + (1 - \lambda) \sup_{\{\delta \circ r_p \geq x\}} l \right] dx.$$

Since the Choquet integral is translation equivariant, it can be used to evaluate the decisions also when the decision problem is described by a bounded utility function $U : \mathcal{P} \times \mathcal{D} \to \mathbb{R}$. In particular, Theorem 2.28 implies that if we apply the likelihood-based decision criterion corresponding to the functionals $V_{rp} : l \mapsto \int C l \, d(f_1 \circ r_p + f_0 \circ r_p)$ to the decision problem described by the loss function $c - U$, where $c$ is an upper bound of $U$ (we have considered this kind of loss function at the end of Subsection 3.1.1), then we obtain the following decision criterion:

$$\max \int C u_d \, d(f_0 \circ r_p + f_1 \circ r_p),$$
where the integral is the extension of the Choquet integral by means of equality (2.10), and for each \( d \in \mathcal{D} \) the function \( u_d \) on \( \mathcal{P} \) is defined by \( u_d(P) = U(P, d) \) (for all \( P \in \mathcal{P} \)). In particular, the ignorance averse decision criterion corresponding to the functionals \( V_{\mathcal{R} \mathcal{P}} : l \mapsto \int_{\mathcal{C}} l \, d((f_1 \circ \mathcal{R} \mathcal{P}) \circ l) \) compares the decisions \( d \in \mathcal{D} \) by means of the lower evaluation \( \int_{\mathcal{C}} u_d \, d((f_1 \circ \mathcal{R} \mathcal{P}) \circ l) \) of \( u_d \); while the ignorance attracted decision criterion corresponding to the functionals \( V_{\mathcal{R} \mathcal{P}} : l \mapsto \int_{\mathcal{C}} l \, d(f_0 \circ \mathcal{R} \mathcal{P}) \) compares them by means of the upper evaluation \( \int_{\mathcal{C}} u_d \, d((f_0 \circ \mathcal{R} \mathcal{P}) \circ l) \) of \( u_d \).

By contrast, as noted at the end of Subsection 3.1.1, the Shilkret integral cannot be used directly to evaluate the decisions when the decision problem is described by a utility function \( U : \mathcal{P} \times \mathcal{D} \to \mathbb{R} \), because the utility functions \( U \) and \( \alpha U + \beta \) are considered equivalent (for all \( \alpha \in \mathbb{P} \) and all \( \beta \in \mathbb{R} \)). But if the positive values of \( U \) are interpreted as gains, and the negative ones as losses, then the functions \( U \) and \( \alpha U + \beta \) are equivalent only if \( \beta = 0 \) (because of the peculiar meaning of the value 0), and the Shilkret integral can be used to evaluate separately gains and losses. If we evaluate gains and losses by means of the Shilkret integral with respect to \( f_1 \circ \mathcal{R} \mathcal{P} \), then the uncertain gain resulting from a decision \( d \in \mathcal{D} \) can be indifferently defined as \( \max\{u_d, 0\} \), as \( u_d\{u_d > 0\} \), or as \( u_d\{u_d > 0\} \) (and analogously for the uncertain loss), since the Shilkret integral is support-based:

\[
\int_{\mathcal{S}} \max\{u_d, 0\} \, d(f_1 \circ \mathcal{R} \mathcal{P}) = \int_{\mathcal{S}} u_d\{u_d > 0\} \, d(f_1 \circ \mathcal{R} \mathcal{P}\{u_d > 0\}).
\]

By contrast, the use of \( \mathcal{R} \mathcal{P}\) in the integrals would pose some problems, because in general if \( \mathcal{H} \) is a subset of \( \mathcal{P} \), then \( \mathcal{R} \mathcal{P}|_{\mathcal{H}} \) is different from \( \mathcal{R} \mathcal{P}\{\mathcal{H}\} \), since the latter depends also on \( \mathcal{R} \mathcal{P}(\mathcal{P} \setminus \mathcal{H}) \). The decisions \( d \in \mathcal{D} \) can thus be compared on the basis of the two evaluations \( \int_{\mathcal{S}} \max\{u_d, 0\} \, d(f_1 \circ \mathcal{R} \mathcal{P}) \) and \( \int_{\mathcal{S}} \max\{-u_d, 0\} \, d(f_1 \circ \mathcal{R} \mathcal{P}) \); for instance, maximizing their difference corresponds to maximizing \( \int_{\mathcal{S}} u_d \, d(f_1 \circ \mathcal{R} \mathcal{P}) \), where the integral is the extension of the Shilkret integral by means of equality (2.11). It is important to note that both evaluations of gains and losses are “upper evaluations”, and therefore the resulting decision criteria are ignorance attracted as regards gains, and ignorance averse as regards losses. When \( f_1 \) is continuous, the comparison of decisions on the basis of separate evaluations of gains and losses by means of the Shilkret integral seems to be in agreement with the theory of Shackle (1949).
4.1.2 Sure-Thing Principle

A likelihood-based decision criterion satisfies the **sure-thing principle** if for all disjoint sets \( A, B \), and all nondegenerate relative plausibility measures \( rp \) on \( Q = A \cup B \) such that \( rp(A) > 0 \) and \( rp(B) > 0 \), the following condition holds:

\[
 l|_A \preceq_{rp|_A} l'|_A \text{ and } l|_B \preceq_{rp|_B} l'|_B \implies l \preceq_{rp} l' \text{ for all } l, l' \in \mathcal{F}^Q.
\]

Note that the conditions on \( rp(A) \) and \( rp(B) \) are necessary only because \( \preceq_{rp} \) is undefined if \( rp' \) is a relative plausibility measure with constant value 0, but when for instance \( rp(B) = 0 \), for all likelihood-based decision criteria we have \( l|_A \preceq_{rp|_A} l'|_A \) if and only if \( l \preceq_{rp} l' \). The above sure-thing principle corresponds to the first part of the “sure-thing principle” of Savage (1954), while the second part states that if one of the two preferences on the left-hand side of the implication is strict, then so is the preference on the right-hand side. This would imply that if \( l \leq l' \), and \( rp(l < l') > 0 \), then \( l \) must be preferred to \( l' \); therefore, no likelihood-based decision criterion can satisfy the second part of the “sure-thing principle” of Savage: the next example shows that this is true also if the dominated decisions are discarded.

**Example 4.8.** Let \( Q = \{a, b, c\} \) be a set with three elements, and consider the relative plausibility measure \( 1^\uparrow \) on \( Q \). The two functions \( I_{\{a\}} \) and \( I_{\{b,c\}} \) on \( Q \) are considered equivalent by all likelihood-based decision criteria, since \( V_1^\uparrow(I_{\{a\}}) = f_1(1) = V_1^\uparrow(I_{\{b,c\}}) \). But \( I_{\{a\}}|_{\{c\}} = 0 \) is certainly preferred to \( I_{\{b,c\}}|_{\{c\}} = 1 \), while the equivalence of \( I_{\{a\}}|_{\{a,b\}} \) and \( I_{\{b,c\}}|_{\{a,b\}} = I_{\{b\}}|_{\{a,b\}} \) is implied by symmetry. \( \diamond \)

However, the second part of the “sure-thing principle” of Savage does not really deserve its name, since we are not “sure” of the strict preference: we do not assume that the strict preference holds in both alternative cases \( A \) and \( B \). If we were assuming this, then we would obtain the “strict version” of the above sure-thing principle, in which each one of the three preferences is strict: it can be easily proved that if a likelihood-based decision criterion satisfies the sure-thing principle, then it satisfies also its strict version. Moreover, since relative plausibility measures are (equivalence classes of) finitely maxitive measures, the content of the sure-thing
principle is left unchanged when the assumption that $A$ and $B$ are disjoint is dropped (and this is valid also for the strict version).

Consider a statistical decision problem described by a loss function $L$ on $\mathcal{P} \times \mathcal{D}$, and let $rp$ be a nondegenerate relative plausibility measure on $\mathcal{P}$ describing the uncertain knowledge about the models in $\mathcal{P}$. The sure-thing principle states that for all coverings $\mathcal{H}, \mathcal{H'}$ of $\mathcal{P}$ (that is, $\mathcal{H}, \mathcal{H'} \subseteq \mathcal{P}$ and $\mathcal{P} = \mathcal{H} \cup \mathcal{H'}$) such that $rp(\mathcal{H}) > 0$ and $rp(\mathcal{H'}) > 0$, if the preference between two decisions in $\mathcal{D}$ is the same when we restrict attention to the models in $\mathcal{H}$ or to the models in $\mathcal{H'}$, then this preference holds also when we consider all the models in $\mathcal{P}$. In particular, if $d \in \mathcal{D}$ is optimal, according to a likelihood-based decision criterion satisfying the sure-thing principle, for both the decision problems described by $L|_{\mathcal{H} \times \mathcal{D}}$ and by $L|_{\mathcal{H}' \times \mathcal{D}}$, then $d$ is optimal also for the decision problem described by $L$. Moreover, if for both the restricted decision problems $d$ is the unique optimal decision in $\mathcal{D}$, then this holds also for the unrestricted decision problem; but this conclusion is in general not valid when $d$ is the unique optimal decision only for one of the two restricted decision problems (since the second part of the “sure-thing principle” of Savage does not hold).

The “if” part of the condition of 0-independence is a special case of the sure-thing principle, while the “only if” part is a special case of the second part of Savage’s “sure-thing principle”. Theorem 4.6 implies that a likelihood-based decision criterion can be 0-independent only if the corresponding function $f_1$ is positive on $(0, 1]$: the next theorem states that in this case the 0-independence is implied by the sure-thing principle.

**Theorem 4.9.** A likelihood-based decision criterion satisfying the sure-thing principle is 0-independent if and only if the corresponding function $f_1 : [0, 1] \rightarrow [0, 1]$ is positive on $(0, 1]$.

**Proof.** The “only if” part follows from Theorem 4.6, and to prove the “if” part it suffices to show that if the function $f_1$ is positive on $(0, 1]$, and $A$ is a subset of $\mathcal{Q}$ such that $l|_{\mathcal{Q}\setminus A} = 0$, and both $rp(A)$ and $rp(\mathcal{Q}\setminus A)$ are positive, then $V_{rp}(l) = V_{rp}(l|A)V_{rp/A}(l|A)$, because $V_{rp}(l|A) \geq f_1(rp(A)) > 0$. The above equality follows from the equivalence of $l$ and $V_{rp/A}(l|A)I_A$, which is implied by the sure-thing principle and the equivalence of $l|_A$ and $V_{rp/A}(l|A)$. □

The sure-thing principle does not follow from the 0-independence: for instance, the likelihood-based decision criterion such that each relation
$\gtrsim_{rp}$ is represented by the functional $l \mapsto \int^C l \, drp$ (that is, the "MPL* criterion" of Cattaneo, 2005) is 0-independent (Theorem 4.6), but the next example shows that it does not satisfy the sure-thing principle. In fact, the subsequent theorem implies that a likelihood-based decision criterion such that each relation $\gtrsim_{rp}$ is represented by the functional $l \mapsto \int l \, drp$ can satisfy the sure-thing principle only if the integral is maxitive.

Example 4.10. Consider the likelihood-based decision criterion corresponding to the functionals $V_{rp} : l \mapsto \int^C l \, drp$, and let $rp$ be the (nondegenerate) relative plausibility measure on $Q = \{0, 1, 2\}$ defined by $rp(0) = rp(1) = 1$ and $rp(2) = \frac{1}{2}$. We have $V_{rp|\{0,2\}}(id_Q|\{0,2\}) = V_{rp|\{1\}}(id_Q|\{1\}) = 1$, but $V_{rp}(id_Q) = \frac{3}{2}$; that is, $id_Q|\{0,2\} \gtrsim_{rp|\{0,2\}} 1|\{0,2\}$ and $id_Q|\{1\} \gtrsim_{rp|\{1\}} 1|\{1\}$, but not $id_Q \gtrsim_{rp} 1$.

Theorem 4.11. Consider a likelihood-based decision criterion satisfying the sure-thing principle. The corresponding functions $f_1, f_0 : [0, 1] \to [0, 1]$ fall into one of the following five alternative cases:

- $f_0 = 0$ and $f_1 = I_{\{0,1\}}$,
- $f_1 = 0$ and $f_0 = I_{\{1\}}$,
- $f_0 = 0$ and $f_1 = I_{\{1\}}$,
- $f_1 = 0$ and $f_0 = I_{\{0,1\}}$,
- $f_0 = 0$ and $f_1 : x \mapsto x^\alpha$ for an $\alpha \in \mathbb{P}$.

If $f_0 = 0$, then the decision criterion reveals ignorance aversion, and for each nondegenerate relative plausibility measure $rp$ (on some set $Q_{rp}$) the corresponding functional $V_{rp}$ on $\mathbb{P}^Q_{\mathbb{Q}}$ fulfills the following condition:

$$V_{rp}(\max\{l, l'\}) = \max\{V_{rp}(l), V_{rp}(l')\} \quad \text{for all } l, l' \in \mathbb{P}^Q_{\mathbb{Q}}.$$ 

Proof. Consider first the case with $f_1 = 0$; let $rp$ be the (nondegenerate) relative plausibility measure on $Q = \{0, 1, 2\}$ defined by $rp\{0\} = 1$ and $rp\{1\} = rp\{2\} = 1 - x$ for an $x \in [0, 1]$, and let $l = I_{\{0\}}$ and $l' = f_0(x) I_{\{0,1\}}$ be two functions on $Q$. We have

$$f_0(x) = V_{rp|\{0,1\}}(l|\{0,1\}) = V_{rp|\{0,1\}}(l'|\{0,1\}) = V_{rp}(l) = V_{rp}(l') = [f_0(x)]^2,$$

where the fourth equality is implied by the sure-thing principle, since $l|\{2\} = 0 = l'|\{2\}$. Hence the range of $f_0$ is $\{0, 1\}$; to obtain that $f_0 = I_{\{1\}}$.
or \( f_0 = I_{\{0,1\}} \), it suffices to show that \( \sup\{f_0 = 0\} \in (0,1) \) leads to a contradiction: assume thus that it is true and define \( y = (1 - \sup\{f_0 = 0\})^\frac{1}{3} \). Let \( r'p' \) be the (nondegenerate) relative plausibility measure on \( Q = \{0,1,2\} \) defined by \((r'p')^k : k \mapsto y^k\), and consider the two functions \( I_{\{0\}} \) and \( I_{\{0,1\}} \) on \( Q \). We have

\[
V_{r'p'}(I_{\{0\}}) = f_0(1-y) = 0 < 1 = f_0(1-y^2) = V_{r'p'}(I_{\{0,1\}});
\]

this contradicts the sure-thing principle, since

\[
V_{r'p'}|_{\{1,2\}}(I_{\{0,1\}}|_{\{1,2\}}) = f_0(1-y) = 0 = V_{r'p'}|_{\{1,2\}}(I_{\{0\}}|_{\{1,2\}})
\]

and \( I_{\{0,1\}}|_{\{0\}} = 1 = I_{\{0\}}|_{\{0\}} \).

Consider now the case with \( f_1(1) > 0 \); we first prove that the decision criterion reveals ignorance aversion (and thus in particular \( f_0 = 0 \)). By Theorem 4.3 it suffices to show that the corresponding functional \( S \) on \( S_\mathbb{P} \) is 0-independent; that is, it suffices to show that for all \( \varphi \in S_\mathbb{P} \) we have equality (4.11) with \( c = 1 \) and \( f_1(1) = 1 \). Define \( Q, A, r'p', l, \) and \( l' \) as in the proof of Theorem 4.6, but with \( c = 1 \); since \( l|_{Q \setminus A} = 0 = l'|_{Q \setminus A} \) and \( V_{r'p'}|_A(l|_A) = S(\varphi) = V_{r'p'}|_A(l'|_A) \), the sure-thing principle implies

\[
S(I_{\{0\}} + \varphi I_{\{0,\infty\}}) = V_{r'p'}(l) = V_{r'p'}(l') = V_{r'p'}|_A(l'|_A) = S(\varphi).
\]

We now prove that \( \sup\{f_1 = 0\} \in \{0,1\} \): analogously to the case with \( f_1 = 0 \), assume that \( \sup\{f_1 = 0\} \in (0,1) \) and define \( y = (\sup\{f_1 = 0\})^\frac{1}{3} \). Let \( r'p \) be the (nondegenerate) relative plausibility measure on \( Q = \{0,1,2\} \) defined by \((r'p)^k : k \mapsto y^k\), and consider the two functions \( l = I_{\{2\}} \) and \( l' = f_1(y) I_{\{1,2\}} \) on \( Q \). We have \( V_{r'p}(l) = f_1(y^2) = 0 < [f_1(y)]^2 = V_{r'p}(l') \): this contradicts the sure-thing principle, since \( l|_{\{0\}} = 0 = l'|_{\{0\}} \); and \( V_{r'p}|_{\{1,2\}}(l|_{\{1,2\}}) = f_1(y) = V_{r'p}|_{\{1,2\}}(l'|_{\{1,2\}}) \). Therefore, either we have \( \sup\{f_1 = 0\} = 1 \) and thus \( f_1 = I_{\{1\}} \), or we have \( \sup\{f_1 = 0\} = 0 \) and thus \( f_1 = I_{\{0,1\}} \) or \( f_1 : x \mapsto x^\alpha \) for an \( \alpha \in \mathbb{P} \) (Theorems 4.9 and 4.6). Since the decision criterion reveals ignorance aversion, Theorem 4.3 implies

\[
V_{r'p}(\max\{l, l'\}) = S(\max\{l, l'\}) = S(\max\{r'p(l \geq \cdot), r'p(l' \geq \cdot)\}) \geq \max\{V_{r'p}(l), V_{r'p}(l')\},
\]

where the inequality follows from the monotonicity of \( S \). To prove the last statement of the theorem, it suffices thus to show that this inequality is an equality; that is, it suffices to show that for all nonincreasing \( \varphi, \psi \in S_\mathbb{P} \) we have \( S(\max\{\varphi, \psi\}) \leq \max\{S(\varphi), S(\psi)\} \). Define \( A = H(\varphi) \times \{0\} \) and
\[ B = H(\varphi) \times \{1\}, \] where \( H(\varphi) \) and \( H(\psi) \) are the hypographs of \( \varphi \) and \( \psi \), respectively, and thus \( A \cup B \subseteq \mathbb{F} \times [0, 1] \times \{0, 1\} \); let \( l \) and \( r_p \) be respectively the function and the (nondegenerate) relative plausibility measure on \( A \cup B \) defined by \( l(x, y, z) = x \) and \( r_p(x, y, z) = y \). Since \( r_p|_A(x|_A \geq x) = \varphi(x) \) and \( r_p(A) = 1 \), we have \( V_{r_p}|_A(l|_A) = S(\varphi) \); and analogously \( V_{r_p}|_B(l|_B) = S(\psi) \). The sure-thing principle implies thus that \( V_{r_p}(l) \leq \max\{S(\varphi), S(\psi)\} \), but \( V_{r_p}(l) = S(\max\{\varphi, \psi\}) \), because \( r_p(l \geq x) = \max\{\varphi(x), \psi(x)\} \). \( \square \)

Theorem 4.11 implies in particular that neither the \( \text{LRM}_\beta \) criterion (for any \( \beta \in (0, 1) \)), nor any likelihood-based decision criterion with \( f_1(1) \in (0, 1) \) (such as one generalizing the Hurwicz criterion) can satisfy the sure-thing principle. If a likelihood-based decision criterion satisfies it, and \( f_0 = 0 \), then the decision criterion reveals ignorance aversion, and thus in particular it generalizes the minimax criterion; but the next example shows that a likelihood-based decision criterion satisfying the sure-thing principle does not necessarily reveal ignorance attraction when \( f_1 = 0 \).

**Example 4.12.** Let \( S \) be the functional on \( S_\mathbb{F} \) defined by
\[
S(\varphi) = \begin{cases} 
0 & \text{if } \varphi(0) > 0, \\
\sup\{\varphi > 0\} & \text{if } \varphi(0) = 0,
\end{cases}
\]
for all \( \varphi \in S_\mathbb{F} \).

It can be easily proved that \( S \) corresponds to a likelihood-based decision criterion satisfying the sure-thing principle with \( f_1 = 0 \) and \( f_0 = I_{(1)} \) (the second case of Theorem 4.11), but revealing neither ignorance aversion nor attraction. \( \Diamond \)

Five other examples of likelihood-based decision criteria satisfying the sure-thing principle and falling in the five alternative cases listed in Theorem 4.11 are the essential minimax criterion and its ignorance attracted analogue (the essential minimin criterion), the MLD criterion and its ignorance attracted analogue, and the MPL criterion, respectively. Let \( S \) be the functional on \( S_\mathbb{F} \) corresponding to a likelihood-based decision criterion, and consider the functional \( S' \) on \( S_\mathbb{F} \) defined by
\[
S'(\varphi) = \begin{cases} 
S(\varphi) & \text{if } \sup\{\varphi > 0\} < \infty, \\
\infty & \text{if } \sup\{\varphi > 0\} = \infty,
\end{cases}
\]
for all \( \varphi \in S_\mathbb{F} \).

It can be easily proved that \( S' \) corresponds to a likelihood-based decision criterion with the same functions \( f_1, f_0 \) as the original one (that is,
the one corresponding to $S$). If the original decision criterion satisfies the sure-thing principle or reveals ignorance aversion, then so does the one corresponding to $S'$; but this one cannot reveal ignorance attraction, even when the original one does. The functional on $\mathcal{S}_P$ corresponding to the essential minimax criterion is $S : \varphi \mapsto \sup\{\varphi > 0\}$, and in this case $S' = S$; in fact, using Theorems 4.9 and 4.6 we obtain that the essential minimax criterion is the only likelihood-based decision criterion satisfying the sure-thing principle with $f_0 = 0$ and $f_1 = I_{(0,1)}$. But if $S$ is the functional on $\mathcal{S}_P$ corresponding to one of the other four likelihood-based decision criteria listed above, then $S'$ is different from $S$, and thus only the first one of the five alternative cases of Theorem 4.11 uniquely determines a decision criterion satisfying the sure-thing principle.

In particular, Theorems 4.9, 4.6, and 4.11 imply that if a likelihood-based decision criterion satisfies the sure-thing principle with $f_0 = 0$ and $f_1 : x \mapsto x^\alpha$ for an $\alpha \in \mathbb{P}$, then there is a unique support-based, regular integral on the class of all finite, monotonic measures, such that each relation $\preceq_{\mu,\rho}$ is represented by the functional $l \mapsto \int l \, d(\rho^\alpha)$, and the integral is maxitive (on the class of all finitely maxitive measures). Since the integral is maxitive, homogeneous, and transformation invariant, we have $\int f \, d\mu = \int S f \, d\mu$ when $f$ is essentially bounded with respect to $\mu$ (that is, $\text{ess}_\mu \sup f < \infty$). But to obtain the Shilkret integral also when $\text{ess}_\mu \sup f = \infty$, we need something more: for instance countable maxitivity, or the following continuity property, which parallels condition (4.12):

$$\int f \, d\mu = \lim_{c \uparrow \infty} \int \min\{f, c\} \, d\mu,$$

for all sets $Q$, all finite, monotonic measures $\mu$ on $Q$, and all functions $f : Q \to \mathbb{P}$. This continuity property corresponds to the following condition for preference relations:

$$\min\{l, c\} \preceq_{\mu,\rho} l' \text{ for all } c \in \mathbb{P} \implies l \preceq_{\mu,\rho} l' \text{ for all } l, l' \in \mathbb{P}^Q_{\mathbb{P},\rho}. \quad (4.15)$$

The next theorem implies that this condition suffices to determine a unique likelihood-based decision criterion satisfying the sure-thing principle in each one of the three cases with $f_0 = 0$ listed in Theorem 4.11; but this is not true in the two cases with $f_1 = 0$: for instance, the condition is fulfilled by both the essential minimin criterion and the decision criterion of Example 4.12.
Theorem 4.13. Consider a likelihood-based decision criterion such that the corresponding function \( f_1 : [0, 1] \to [0, 1] \) satisfies \( f_1(1) > 0 \), and for each nondegenerate relative plausibility measure \( r_p \) (on some set \( Q_{rp} \)) condition (4.15) is fulfilled. The decision criterion satisfies the sure-thing principle if and only if either it is the essential minimax criterion or the MLD criterion, or there is a positive real number \( \alpha \) such that each relation \( \mathcal{Z}_{rp} \) is represented by the functional \( l \mapsto \int S l \, d(r_p^\alpha) \).

Proof. The “if” part is simple: the essential minimax criterion satisfies the sure-thing principle, because

\[
V_{rp}(l) = \text{ess}_{rp} \sup l = \max\{V_{rp|A}(l|A), V_{rp|B}(l|B)\};
\]

the MLD criterion satisfies it because if \( r_p(A) = r_p(B) = 1 \), then

\[
V_{rp}(l) = \text{ess}_{I_{\{1\} \circ r_p}} \sup l = \max\{V_{rp|A}(l|A), V_{rp|B}(l|B)\},
\]

while if \( r_p(B) < 1 \), then \( V_{rp}(l) = V_{rp|A}(l|A) \). Finally, in the third case, using Theorem 2.6 we obtain

\[
V_{rp}(l) = \sup l \ (r_p l) = \max\{[r_p(A)]^\alpha V_{rp|A}(l|A), [r_p(B)]^\alpha V_{rp|B}(l|B)\}.
\]

For the “only if” part, Theorem 4.11 implies that \( f_0 = 0 \) and we have three alternative cases for \( f_1 \). If \( f_1 = I_{\{0,1\}} \), then Theorems 4.9 and 4.6 imply that the decision criterion is the essential minimax criterion. For \( f_1 = I_{\{1\}} \), we have to prove that if \( b \in (\text{ess}_{I_{\{1\} \circ r_p}} \sup l, \infty) \), then \( V_{rp}(l) \leq b \); in fact, in this case the decision criterion is the MLD criterion, because \( V_{rp}(l) \geq \text{ess}_{I_{\{1\} \circ r_p}} \sup l \) follows from ignorance aversion (implied by Theorem 4.11). Thanks to condition (4.15), it suffices to show that \( V_{rp}(\min\{l, c\}) \leq b \) for all \( c \in (b, \infty) \): this follows from the last statement of Theorem 4.11, since \( \min\{l, c\} \leq \max\{b, cI_{\{l>b\}}\} \) and \( V_{rp}(I_{\{l>b\}}) = f_1(r_p(l > b)) = 0 \). Finally, if \( f_1 : x \mapsto x^\alpha \) for an \( \alpha \in \mathbb{P} \), then Theorems 4.9, 4.6, and 4.11 imply that there is a maxitive, regular integral on the class of all finite, completely maxitive measures, such that each relation \( \mathcal{Z}_{rp} \) is represented by the functional \( V_{rp} : l \mapsto \int l \, d(r_p^\alpha) \). We have to prove that \( V_{rp}(l) = \int l \, d(r_p^\alpha) \), and thanks to Theorem 2.15 and condition (4.15), it suffices to show that \( \int \min\{l, c\} \, d(r_p^\alpha) \leq \int l \, d(r_p^\alpha) \) for all \( c \in \mathbb{P} \). Now, for a bounded function \( f \) with finite range, the value of \( \int f \, d(r_p^\alpha) \) is determined by maxitivity and homogeneity, since \( f \) can be written as the maximum of a finite number of functions of
the form \( a I_A \) with \( a \in \mathbb{P} \). Therefore \( \int f \, d(\mathbb{P}^\alpha) = \int S f \, d(\mathbb{P}^\alpha) \), but then \( \int \min\{l, c\} \, d(\mathbb{P}^\alpha) = \int S \min\{l, c\} \, d(\mathbb{P}^\alpha) \) follows for instance from Theorem 2.23.  

Theorem 4.13 implies that the MPL criterion (possibly applied after having distorted the likelihood function by raising its values to the power \( \alpha \in \mathbb{P} \)) is the only likelihood-based decision criterion that satisfies the sure-thing principle and the continuity condition (4.15), and that leads to the usual likelihood-based inference methods, when applied to some standard form of the corresponding decision problems. Moreover, the MPL criterion possesses also the following property. A likelihood-based decision criterion is said to be **decomposable** if the corresponding functionals \( V_{rp} \) satisfy

\[
V_{rp}(l) = V_{rp^{\otimes t^{-1}}} \left[ V_{rp|\{t=\}}(l|\{t=\}) \right]
\]  

(4.16)

for all sets \( Q \) and \( T \), all nondegenerate relative plausibility measures \( rp \) on \( Q \), all functions \( l : Q \to \mathbb{P} \), and all functions \( t : Q \to T \) such that \( rp\{t = \tau\} > 0 \) for all \( \tau \in T \); where \( V_{rp|\{t=\}}(l|\{t=\}) \) denotes the function \( \tau \mapsto V_{rp|\{t=\}}(l|\{t=\}) \) on \( T \). Note that the condition on \( rp\{t = \tau\} \) is necessary only because \( V_{rp'} \) is undefined if \( rp' \) is a relative plausibility measure with constant value 0; alternatively, we could drop this condition and assume simply that \( V_{rp|\{t=\}}(l|\{t=\}) \) denotes a function \( l' : T \to \mathbb{P} \) such that \( l'(\tau) = V_{rp|\{t=\}}(l|\{t=\}) \) for all \( \tau \in \{(rp \circ t^{-1}) / > 0\} \), since the values taken by \( l' \) outside this set have no influence on \( V_{rp^{\otimes t^{-1}}} \).

The equality \( V_{rp^{\otimes t^{-1}}} \left[ V_{rp|\{t=\}}(l|\{t=\}) \right] = V_{rp} \left[ V_{rp|\{t=\}}(l|\{t=\}) \circ t \right] \) holds for all likelihood-based decision criteria; if a decision criterion is decomposable, then the function \( V_{rp|\{t=\}}(l|\{t=\}) \circ t \) can be interpreted as the “conditional evaluation” of \( l \) given \( t \), in analogy with the conditional expectation \( E(X|Y) \) of a random variable \( X \) given a random object \( Y \). The “conditional evaluation” corresponds to the conditional expectation in many respects; in particular, equality (4.16) corresponds to the property \( E(X) = E[E(X|Y)] \). That is, a likelihood-based decision criterion is decomposable if it allows the introduction of a concept of “conditional evaluation” presenting many analogies with the concept of conditional expectation.

It can be easily proved that a decomposable decision criterion satisfies the sure-thing principle: it suffices to choose \( T = \{0, 1\} \) and \( t = I_A \). In fact, the property of decomposability can also be expressed as follows: for all sets \( A \) of disjoint sets, and all nondegenerate relative plausibility measures
rp on $Q = \bigcup A$ such that $rp(A) > 0$ for all $A \in A,$

$$l|_A \preceq_{rp|_A} l'|_A \text{ for all } A \in A \Rightarrow l \preceq_{rp} l' \text{ for all } l, l' \in \mathbb{P}_Q.$$  

The sure-thing principle corresponds to the case with finite $A,$ and it is thus equivalent to the finite decomposability (that is, decomposability with the additional condition that $T$ is finite). By exploiting the fact that there is a finite or countable subset $Q' \subset Q$ such that $rp|_{Q'} \circ (l|_{Q'})^{-1} = rp \circ l^{-1},$ it can be proved that decomposability is equivalent to countable decomposability (for which $T$ and $A$ are at most countable). As in the case with finite $A,$ also in the general case the assumption that the elements of $A$ are disjoint can be dropped without consequences. But contrary to the case with finite $A,$ the above formulation of decomposability does not imply its strict version (in which both preferences are strict); in fact, no likelihood-based decision criterion can satisfy the strict version even for countable sets $A,$ since this would imply that if $l$ and $l'$ have countable range, and $l \prec l',$ then $l$ is preferred to $l',$ while it is possible that $rp \circ l^{-1} = rp \circ (l')^{-1}.$

**Theorem 4.14.** Consider a likelihood-based decision criterion such that the corresponding function $f_1 : [0, 1] \to [0, 1]$ is continuous at 0, and the corresponding function $f_0 : [0, 1] \to [0, 1]$ is continuous at 1. The decision criterion is decomposable if and only if there is a positive real number $\alpha$ such that each relation $\preceq_{rp}$ is represented by the functional $l \mapsto \int_0^1 l \, d(rp^\alpha).$

**Proof.** The “if” part is simple: using Theorem 2.6 we obtain

$$V_{rp \circ t^{-1}} [V_{rp|_{\{t=\tau\}}} (l|_{\{t=\tau\}})] = \sup_{\tau \in T} V_{rp|_{\{t=\tau\}}} (l|_{\{t=\tau\}}) \left( rp\{t = \tau\}\right)^\alpha = \sup_{\tau \in T} \sup_{l|_{\{t=\tau\}}} \left( rp^\tau\right)^\alpha = V_{rp}(l).$$

For the “only if” part, let $l, t$ and $rp$ be respectively the functions and the (nondegenerate) relative plausibility measure on $Q = (0, 1) \times \{0, 1\}$ defined by $l(x,y) = y$ and $t(x,y) = x,$ and $rp^\tau(x,y) = xy + (1 - y).$ We have $rp\{t = \tau\} = 1$ and $V_{rp|_{\{t=\tau\}}} (l|_{\{t=\tau\}}) = f_1(\tau)$ for all $\tau \in T = (0, 1);$ and therefore

$$f_1(1) = V_{rp}(l) = V_{rp}(f_1 \circ t) \leq \sup(f_1 \circ t) = \sup_{(0,1)} f_1.$$  

That is, $f_1$ is continuous at 1; the continuity of $f_0$ at 0 can be proved analogously. Theorem 4.11 implies thus that the decision criterion reveals
ignorance aversion, and \( f_1 : x \mapsto x^\alpha \) for an \( \alpha \in \mathbb{P} \); we have to show that \( V_{rp}(l) = \sup l(\{rp \downarrow\}^\alpha) \) for all nondegenerate relative plausibility measures \( rp \) (on some set \( Q_{rp} \)), and all functions \( l : Q_{rp} \to \overline{\mathbb{P}} \). Let \( l', t' \) and \( rp' \) be respectively the functions and the (nondegenerate) relative plausibility measure on \( Q' = Q_{rp} \times \{0, 1\} \) defined by \( l'(q, y) = l(q) y \) and \( t'(q, y) = q \), and \( (rp')^{-1}(q, y) = rp^\downarrow(q) y + (1 - y) \). We have \( V_{rp'}(l') = V_{rp}(l) \), because \( rp'[l' \geq x] = rp'([l \geq x] \times \{1\}) = rp([l \geq x]) \) for all \( x \in \overline{\mathbb{P}} \) (Theorem 4.3); and since \( rp'[t' = \tau] = 1 \) and \( V_{rp'}[t' = \tau]([l'|_{t' = \tau}] \downarrow) = l([rp^\downarrow(\tau)]^\alpha) \) for all \( \tau \in T' = Q_{rp} \), we obtain

\[
V_{rp}(l) = V_{rp'}(l') = V_{rp'}[t']^{-1}[l(\{rp \downarrow\}^\alpha)] = \sup l(\{rp \downarrow\}^\alpha),
\]

because \( rp' \circ (t')^{-1} = 1 \) on \( Q_{rp} \).

At the beginning of the present section we noted that a likelihood-based decision criterion can be useful only if it satisfies some sort of continuity at 0 of the influence of a model, with respect to its relative plausibility; and thus in particular only if the corresponding functions \( f_1 \) and \( f_0 \) on \([0, 1]\) are continuous at 0 and 1, respectively. In this sense, Theorem 4.14 implies that the MPL criterion (possibly applied after having distorted the likelihood function by raising its values to the power \( \alpha \in \mathbb{P} \)) is the only decomposable likelihood-based decision criterion that can be useful; and it is also the only decomposable decision criterion that leads to the usual likelihood-based inference methods, when applied to some standard form of the corresponding decision problems. The only consequence of the assumption about \( f_1 \) in Theorem 4.14 is the exclusion of the essential minimax criterion, while the assumption about \( f_0 \) does not exclude only the essential minimin criterion, but for instance also the decision criterion of Example 4.12.

Theorem 4.2 states that a likelihood-based decision criterion compares the decisions on the basis of a functional \( S : \mathcal{S}_\overline{\mathbb{P}} \to \overline{\mathbb{P}} \) applied to the functions \( (rp \circ l^{-1})^\downarrow \) describing the relative plausibility of the possible values of the loss incurred. That is, a decision criterion corresponds to a preference relation (represented by \( S \)) on the set \( \mathcal{S}_\overline{\mathbb{P}} \), which can be identified with the set of all nondegenerate relative plausibility measures on the set \( \overline{\mathbb{P}} \) of the possible values of the loss. Moreover, the property of decomposability can be interpreted also as a rule for “combining” the nondegenerate relative plausibility measures \( rp \circ (l|_{t=\tau})^{-1} \) on \( \overline{\mathbb{P}} \) (for each \( \tau \in T \)) on the basis of the “marginal” nondegenerate relative plausibility measure \( rp \circ t^{-1} \) on
Hence, if we assume the sure-thing principle (that is, the finite decomposability), then we can consider a framework similar to the one of the representation theorem of von Neumann and Morgenstern (1944): a decision criterion corresponds to a preference relation on an abstract set of consequences (the nondegenerate relative plausibility measures on $\mathcal{P}$), on which for each $r \in \mathcal{P}$ we have a binary operation (the combination of two consequences with plausibility ratio $r$).

The axiomatic approach to qualitative decision making with possibility theory of Dubois and Prade (1995) uses a similar framework, and results in particular in two decision criteria differing from one another only in their attitudes toward ignorance: one is based on an assumption of ignorance aversion, while the other is based on an assumption of ignorance attraction. By replacing these assumptions with an axiom of "qualitative monotonicity", and adapting other axioms to the new attitude toward ignorance and to the use of profile likelihood functions, we obtain the decision criterion of Giang and Shenoy (2002), briefly introduced at the end of Subsection 1.3.2. If we identify the best value $(1, 0)$ of the binary utility with the loss 0, and the worst value $(0, 1)$ with the loss 1, then the assumption of "qualitative monotonicity" corresponds to the assumption that both functions $f_1$ and $f_0$ on $[0, 1]$ are strictly increasing. Theorem 4.11 implies that this assumption is incompatible with the sure-thing principle, on which the framework of Giang and Shenoy is based; but there is no contradiction, because their criterion is not a "likelihood-based decision criterion" in the sense introduced at the beginning of the present section.

In our framework, the justification of the axiom of "qualitative monotonicity" by Giang and Shenoy can be expressed as follows: if one of the functions $f_1$, $f_0$ is not strictly increasing, then there is a decision problem in which a decision is considered equivalent to another decision dominating it. But this is unavoidable when the decisions are evaluated solely on the basis of the profile likelihood function on the set of the possible consequences; moreover, two decisions can be considered equivalent by the criterion of Giang and Shenoy even when one strictly dominates the other and they have only a finite number of possible consequences (that is, this decision criterion does not satisfy the strict version of the sure-thing principle). In particular, when we are in the state of complete ignorance about the models considered, and a decision has only a finite number of possible consequences, if all these consequences are better than the intermediate value $(1, 1)$, then the criterion of Giang and Shenoy evaluates the decision by its worse consequence, while if all the possible consequences are worse
4.2 Statistical Properties

Consider a statistical decision problem described by a loss function $L$ on $\mathcal{P} \times \mathcal{D}$, where $\mathcal{P}$ is a set of probability measures on a measurable space $(\Omega, \mathcal{A})$. If we regard the observed data as a particular realization of a random object $X : \Omega \rightarrow \mathcal{X}$, then we can consider a loss function $L'$ on $\mathcal{P} \times \Delta$ (where $\Delta$ is a set of decision functions $\delta : \mathcal{X} \rightarrow \mathcal{D}$), as done in Subsection 1.1.1. A likelihood-based decision criterion can be applied either to the pre-data decision problem described by $L'$, or to the post-data decision problem described by $L$ (conditional on the observed realization of $X$). When applied to the pre-data decision problem described by the risk function, all likelihood-based decision criteria revealing ignorance aversion reduce to the minimax risk criterion (if no prior likelihood function is used). In general the application of a decision criterion to a pre-data decision problem is rather difficult; therefore it can be interesting to consider the decision function $\delta$ obtained by conditional application of a likelihood-based decision criterion (to the post-data decision problem): that is, $\delta(x)$ is the decision that we would select after having observed $X = x$. If $X$ is continuous (under each model in $\mathcal{P}$), then it is usually possible to base the decisions $\delta(x)$ on the pseudo likelihood functions on $\mathcal{P}$ obtained from the densities of $X$ and considered at the beginning of Section 1.2 as approximate likelihood functions. If a statistic $s : \mathcal{X} \rightarrow \mathcal{S}$ is sufficient for $\mathcal{X}$, then $\delta$ depends on $x$ only through $s(x)$; that is, the decision function obtained by conditional application of a likelihood-based decision criterion depends only on sufficient statistics. In general the decision function $\delta$ is not optimal from the pre-data point of view (it is not even necessarily in...
but it can nevertheless be useful. The pre-data evaluation of $\delta$ in terms of $L'$ can be of interest even when we are primarily concerned with the post-data decision problem described by $L$; in particular, it can be interesting to consider the expected loss of the decision function $\delta$ under each model in $\mathcal{P}$.

The pre-data performances of the decision functions obtained by conditional application of a likelihood-based decision criterion can be arbitrarily bad in particular decision problems. But the same is true for the pre-data performances of the likelihood-based inference methods, and yet these are the most appreciated general methods for the respective inference problems, because they are intuitive and their pre-data performances are usually good in actual applications. A reasonable likelihood-based decision criterion can be intuitive too, and if it leads to the likelihood-based inference methods when applied to some standard form of the corresponding decision problems, then we can expect that in many actual problems the decision functions obtained by conditional application of this criterion will perform well from the pre-data point of view. When this is not the case, the pre-data performance of these decision functions can often be improved by replacing the profile likelihood function (which is automatically employed by the decision criterion) with another pseudo likelihood function (this technique is used also for the likelihood-based inference methods).

4.2.1 Invariance Properties

The likelihood-based decision criteria are obviously “parametrization invariant”, since they do not depend on a parametrization of the set $\mathcal{P}$ of statistical models considered; they possess many other invariance properties: in particular the following one.

**Theorem 4.15.** Consider a statistical decision problem described by a loss function $L$ on $\mathcal{P} \times \mathcal{D}$ such that there are two functions $f : \mathcal{P} \to \mathcal{P}$ and $g : \mathcal{D} \to \mathcal{D}$, and a positive real number $c$ satisfying $L(P, d) = c L[f(P), g(d)]$ for all $P \in \mathcal{P}$ and all $d \in \mathcal{D}$. All likelihood-based decision criteria have the following property: the decision $d \in \mathcal{D}$ is optimal with respect to the nondegenerate relative plausibility measure $\rho_P$ on $\mathcal{P}$, if $g(d)$ is optimal with respect to $\rho_P \circ f^{-1}$.

In particular, if $\rho_P \circ f^{-1} = \rho_P$, then for all likelihood-based decision criteria the set of all optimal decisions (with respect to $\rho_P$) is invariant
under \(g^{-1}\), and thus if there is a unique optimal decision, then it is a fixed point of \(g\).

Proof. Since for all \(x \in \mathcal{P}\) we have
\[
\{l_d = x\} = \{P \in \mathcal{P} : L(P, d) = x\} = f^{-1}\{l_g(d) = \frac{x}{c}\},
\]
Theorem 4.2 implies
\[
V_{rp}(l_d) = S(r_{P} \circ l^{-1}_d) = S[(r_{P} \circ f^{-1} \circ l^{-1}_g(d))\{\frac{1}{c}\}] = c V_{r_{P} \circ f^{-1}}(l_g(d)),
\]
and therefore \(d\) minimizes \(V_{rp}(l_d)\) if \(d' = g(d)\) minimizes \(V_{r_{P} \circ f^{-1}}(l_{d'})\). □

The invariance property stated by Theorem 4.15 can be very useful when the decision problem presents some symmetries. In particular, the next theorem states that the decision function \(\delta\) obtained by conditional application of a likelihood-based decision criterion is equivariant, if it is well-defined and the decision problem is invariant. In fact, the premise of the theorem is the definition of an invariant decision problem: only the assumption about the prior relative plausibility measure \(r_P\) and the restriction to discrete random objects \(X\) are not usual. But the assumption about \(r_P\) is clearly satisfied when we use no prior information (that is, when \(r_P = 1\)); and the theorem is valid also for continuous \(X\), if we base \(\delta\) on the densities of \(X\) (and we consider the usual definition of invariant decision problem). Moreover, the theorem is valid also if condition (4.6) is restricted to the functions \(t\) that are bijective, and this restricted condition can be satisfied also when the profile likelihood function is replaced by another pseudo likelihood function.

**Theorem 4.16.** Let \(\mathcal{P}\) be a set of probability measures on a measurable space \((\Omega, \mathcal{A})\), and let \(X : \Omega \to X\) be a random object that is discrete under each model in \(\mathcal{P}\). Consider a statistical decision problem described by a loss function \(L \colon \mathcal{P} \times \mathcal{D}\), with prior (nondegenerate) relative plausibility measure \(r_P\) on \(\mathcal{P}\). Let \(G, G', G''\) be three permutation groups on \(X, \mathcal{P}, \mathcal{D}\), respectively, and let \(g \mapsto g'\) and \(g \mapsto g''\) be two homomorphisms of \(G\) to \(G'\) and \(G''\), respectively, such that for all \(g \in G\), all \(x \in X\), all \(P \in \mathcal{P}\), and all \(d \in \mathcal{D}\) we have \(r_P \circ (g')^{-1} = r_P\),
\[
P\{X = g(x)\} = [g'(P)]\{X = x\}, \quad \text{and} \quad L[g'(P), d] = L[P, g''(d)].
\]
All likelihood-based decision criteria have the following property (for all \(g \in G\), all \(d \in \mathcal{D}\), and all \(x \in X\) such that there is a \(P \in \{r_P^{-1} > 0\}\) with
$P\{X = x\} > 0$: the decision $d$ is optimal given the observation $X = x$ if and only if $g''(d)$ is optimal given $X = g(x)$.

In particular, if the decision function $\delta : X \rightarrow D$ obtained by conditional application of a likelihood-based decision criterion is well-defined, then it is equivariant: that is, $\delta \circ g = g'' \circ \delta$ for all $g \in G$.

**Proof.** Since $rp^\dagger(P) P\{X = g(x)\} = rp^\dagger[g'(P)] [g'(P)]\{X = x\}$, the desired result can be easily obtained by applying Theorem 4.15 twice. \(\square\)

In the classical approach, when the loss function $L$ on $P \times D$ describes an invariant decision problem (that is, when the premise of Theorem 4.16 is satisfied), the equivariance can be imposed on the decision functions $\delta \in \Delta$, assuring that they present symmetries corresponding to those of the decision problem; this is particularly useful when $G''$ acts transitively on $P$, because then $L'(P, \delta)$ (the representative value of the random loss of $\delta$) does not depend on $P$. For a decision function $\delta$ obtained by conditional application of a likelihood-based decision criterion, the equivariance is assured by Theorem 4.16, without need of considering the symmetries of the decision problem (and only if these are not invalidated by asymmetric prior information). However, the recognition of these symmetries simplifies the construction of $\delta$ (as in the next example); in particular, if $G$ acts transitively on $X$, then $\delta$ is uniquely determined by any of its values $\delta(x)$.

**Example 4.17.** The estimation problem of Example 1.1 is invariant with respect to the permutation group $G = \{id_X, n-id_X\}$ on $X = \{0, \ldots, n\}$, the permutation group $G'' = \{id_D, 1-id_D\}$ on $D = [0, 1]$, and the permutation group $G'$ on $P$ that corresponds to $G''$ by means of the parametrization $P_p \mapsto p$ of $P$ with parameter space $D$ (considered in Example 1.6); the two homomorphisms are obvious. Let $\delta : \mathcal{X} \rightarrow \mathcal{D}$ be the decision function obtained by conditional application of a likelihood-based decision criterion (without using prior information). Theorem 4.16 implies that if $\delta$ is well-defined (this is for instance the case when we use the MPL, LRM$\beta$, or MLD criteria considered in Examples 1.8 and 1.10), then $\delta(n - x) = 1 - \delta(x)$ for all $x \in \mathcal{X}$, and thus $\delta$ is uniquely determined by its values $\delta(x)$ for $x < \frac{n}{2}$.

A very important aspect of a statistical procedure is its robustness under deviations from the assumed models (see for example Hampel,
Ronchetti, Rousseeuw, and Stahel, 1986, Chapter 1). In general the decisions obtained by conditional application of a likelihood-based decision criterion are not robust, but they can often be robustified by enlarging the set $\mathcal{P}$ of statistical models considered. In this regard it is important to consider the possibility of using a prior relative plausibility measure on the enlarged set $\mathcal{P}'$; for instance the one with density function proportional to $I_P + \beta I_{P\cap \mathcal{P}}$ for a $\beta \in [0,1]$: the prior relative plausibility of the models in $\mathcal{P}' \setminus \mathcal{P}$ is $\beta$ times the one of the models in $\mathcal{P}$ (the extreme cases with $\beta = 0$ and $\beta = 1$ correspond to using the sets $\mathcal{P}$ and $\mathcal{P}'$, respectively, without prior information). When $\mathcal{P}$ is enlarged to $\mathcal{P}'$, the loss function $L$ on $\mathcal{P} \times \mathcal{D}$ must be extended to a loss function $L'$ on $\mathcal{P}' \times \mathcal{D}$; this extension is simple if each model $P \in \mathcal{P}$ is replaced by a set $\mathcal{H}_P \subseteq \mathcal{P}'$ of models involving the same losses as $P$ (with $P \in \mathcal{H}_P$ and $\mathcal{P}' = \bigcup_{P \in \mathcal{P}} \mathcal{H}_P$); we have $L'(P', d) = L[g(P'), d]$ for all $P' \in \mathcal{P}'$ and all $d \in \mathcal{D}$, where $g : \mathcal{P}' \to \mathcal{P}$ is the mapping defined by $g^{-1}\{P\} = \mathcal{H}_P$ (for all $P \in \mathcal{P}$). Of course, $g$ and $L'$ are well-defined only if all the sets $\mathcal{H}_P$ are disjoint, but when this is not the case we can “disjoin” them as suggested at the end of Subsection 3.2.2 for the imprecise statistical models; in fact, if the density function of the prior relative plausibility measure is constant on each set $\mathcal{H}_P$, then enlarging $\mathcal{P}$ to $\mathcal{P}'$ corresponds to replacing each model $P \in \mathcal{P}$ with the imprecise statistical model $\mathcal{H}_P$.

As regards the enlargement of $\mathcal{P}$ to $\mathcal{P}'$ by replacing each model $P \in \mathcal{P}$ with a set $\mathcal{H}_P$ of models involving the same losses as $P$, it is interesting to consider the stability of the decisions obtained by conditional application of a likelihood-based decision criterion. Let $r_P$ be the nondegenerate relative plausibility measure on $\mathcal{P}'$ describing our uncertain knowledge about the models after having observed the data, and assume that $r_P|_{\mathcal{P}'}$ is nondegenerate too. For each decision $d \in \mathcal{D}$ the uncertain loss with respect to $\mathcal{P}'$ is given by $l_d \circ g : \mathcal{P}' \to \mathbb{R}$; hence for all likelihood-based decision criteria the preferences between the possible decisions remain unchanged (when $\mathcal{P}$ is enlarged to $\mathcal{P}'$) if $r_P \circ g^{-1} = r_P|_{\mathcal{P}'}$. This can happen in particular when each $P \in \mathcal{P}$ is the most plausible model in $\mathcal{H}_P$ in the light of the observed data, or when the models in $\mathcal{H}_P$ that are more plausible than $P$ in the light of the data have been sufficiently penalized by the prior relative plausibility measure on $\mathcal{P}'$. More generally, for each likelihood-based decision criterion we can consider the change $V_{r_P \circ g^{-1}}(l) - V_{r_P|_{\mathcal{P}'}}(l)$ in the evaluation of the uncertain loss $l \in \mathbb{P}^{\mathcal{P}'}$ when $\mathcal{P}$ is enlarged to $\mathcal{P}'$, and in particular the continuity at $r_P' = (r_P \circ g)^+$ of the mapping $r_P' \mapsto V_{r_P' \circ g^{-1}}(l)$ (defined on the set of all nondegenerate relative plausibility measures $r_P'$ on $\mathcal{P}'$), with
respect to \(\|rp' - rp''\|\) (note that \(rp'' \circ g^{-1} = rp|_P\)). If a likelihood-based decision criterion always satisfies this continuity when \(l\) is bounded, then the corresponding functions \(f_1, f_0\) on \([0,1]\) are continuous; but if \(f_1\) is positive on \((0,1]\), then the above continuity cannot always be satisfied when \(l\) is unbounded. Theorem 2.23 implies that the likelihood-based decision criterion corresponding to the functionals \(V_{rp} : l \mapsto \int l d(f_1 \circ rp)\) satisfies the above continuity for all bounded functions \(l\), if \(f_1 : [0,1] \rightarrow [0,1]\) is a nondecreasing surjection, and the integral is regular and quasi-subadditive on the class of all completely maxitive measures. The Shilkret integral satisfies these properties, and its values are particularly stable with respect to changes in the measure, since the Shilkret integral depends on the distribution function only through a supremum. Moreover, if \(f_1 : x \mapsto x^\alpha\) for an \(\alpha \in \mathbb{P}\), then the likelihood-based decision criterion corresponding to the functionals \(V_{rp} : l \mapsto \int \gamma l d(f_1 \circ rp)\) (that is, the MPL criterion applied after having distorted \(rp\) by raising its values to the power \(\alpha \in \mathbb{P}\)) satisfies the sure-thing principle (and its infinite version: the decomposability), which can also be interpreted as expressing some sort of stability of the preferences between the possible decisions: in particular, if a preference between two decisions holds with respect to both \(V\) and \(V_1\), then this preference remains unchanged when \(V\) is enlarged to \(V'\).

It is also interesting to consider the stability of the decisions obtained by conditional application of a likelihood-based decision criterion, when the values of the loss function \(L : \mathcal{P} \times \mathcal{D} \rightarrow \mathbb{P}\) are changed. If some important symmetry of the decision problem is maintained, then the optimal decisions can remain unchanged; but in general when the change in the values of \(L\) is substantial, the resulting decision problem is very different from the original one, and so are the optimal decisions. As with the enlargement of \(\mathcal{P}\), we can consider the continuity at \(l \in \mathbb{P}^{\mathcal{P}}\) of the evaluation \(l' \mapsto V_{rp}(l')\) (that is, of the functional \(V_{rp}\) on \(\mathbb{P}^{\mathcal{P}}\)), with respect to \(\|l' - l\|\); we obtain in particular that the likelihood-based decision criterion corresponding to the functionals \(V_{rp} : l \mapsto \int l d(f_1 \circ rp)\) satisfies this continuity for all functions \(l\), if the integral is regular and quasi-subadditive on the class of all finitely maxitive measures, and \(f_1 : [0,1] \rightarrow [0,1]\) is nondecreasing, \(f_1(0) = 0\), and \(f_1(1) = 1\). The decisions obtained by conditional application of a likelihood-based decision criterion revealing ignorance aversion can sometimes be “robustified” with respect to the choice of the loss function \(L\), by replacing it with a whole set \(\mathcal{L}\) of loss functions on \(\mathcal{P} \times \mathcal{D}\), and using as evaluation of each decision \(d \in \mathcal{D}\) the supremum of \(V_{rp}(l_d)\) as \(L\) ranges over \(\mathcal{L}\) (it is important to note that all the loss functions in \(\mathcal{L}\)
must be expressed in the same scale). Theorem 2.6 implies that if \( f_1 \) is continuous, and \( V_{rp} : l \mapsto \int S l d(f_1 \circ rp) \), then the same decisions can be obtained by applying the likelihood-based decision criterion corresponding to the functionals \( V_{rp} \) (that is, by applying the MPL criterion after having distorted \( rp \) by means of \( f_1 \)) to the decision problem described by the loss function \( \sup L \).

**Example 4.18.** Consider the problem of estimating the mean of a normal distribution with known variance, without prior information about the relative plausibility of the possible values of the mean (in order to simplify the results, in the present example the variance is assumed to be known, but analogous results would be obtained without this assumption). Let \( \mathcal{P} = \{ P_\mu : \mu \in \mathbb{R} \} \) be a family of probability measures such that under each model \( P_\mu \) the random variables \( X_1, \ldots, X_n \) are independent and normally distributed with mean \( \mu \) and variance 1. These random variables are continuous: assume that each realization \( X_i = x_i \) is observed with precision \( \delta \in \mathbb{P} \), in the sense that in fact we observe \( X_i \in [x_i - \frac{\delta}{2}, x_i + \frac{\delta}{2}] \); for instance, if the observations are rounded off to \( m \) digits after the decimal point, then \( \delta = 10^{-m} \). Suppose that \( \delta \) is sufficiently small to allow the use of the function \( P_\mu \mapsto \delta \varphi(x_i - \mu) \) on \( \mathcal{P} \) (where \( \varphi \) is the continuous density of the standard normal distribution, with respect to the Lebesgue measure on \( \mathbb{R} \)) as approximation of the likelihood function on \( \mathcal{P} \) induced by the (imprecise) observation \( X_i = x_i \). The arithmetic mean \( \bar{X} \) is sufficient for \( X_1, \ldots, X_n \), and it is normally distributed with mean \( \mu \) and variance \( \frac{1}{n} \). Hence, for a likelihood-based method the problem reduces to the estimation of \( \mu \) on the basis of a single observation from a normal distribution with mean \( \mu \) and variance \( \frac{1}{n} \), and in this reduced problem the only reasonable point estimate is the observation itself (that is, the arithmetic mean \( \bar{X} \)), at least if we consider overestimation and underestimation as equally harmful.

In fact, using Theorem 4.2 it can be easily proved that \( \bar{X} \) is an optimal decision, according to all likelihood-based decision criteria, for all symmetric, nondecreasing estimation errors; that is, for all the decision problems described by loss functions \( L : (P_\mu, d) \mapsto f(|\mu - d|) \) on \( \mathcal{P} \times \mathbb{R} \), where \( f : [0, \infty) \rightarrow [0, \infty) \) is nondecreasing. In general this optimal decision is not unique: for instance, according to the essential minimax criterion all the decisions in \( \mathbb{R} \) are equivalent (independently of the choice of \( f \)); but if \( f \) is strictly increasing, and \( \lim_{x \uparrow \infty} f(x) [\varphi(x)]^n = 0 \), then the arithmetic mean is the unique optimal decision according to the MPL, LRM_\beta, and MLD.
Figure 4.1. Profile likelihood functions and maximum likelihood estimates of the number of degrees of freedom from Example 4.18.

criteria. Anyway, for the decision problem described by $L$, if the decision obtained by conditional application of a likelihood-based decision criterion is well-defined, then it is the arithmetic mean, which is not robust with respect to deviations from the normal models (in particular it is vulnerable to outliers). Consider for instance the six observations $X_1 \approx 1.176$, $X_2 \approx -0.563$, $X_3 \approx 0.235$, $X_4 \approx -1.443$, $X_5 \approx -1.079$, and $X_6 = 5$; the first five have been generated using the model $P_0$ (their arithmetic mean is $\bar{X} \approx -0.335$), while the sixth one is an artificial outlier, strongly influencing the arithmetic mean (which is $\bar{X} \approx 0.554$ when all six observations are considered). The likelihood functions on $\mathbb{R}$ (with respect to the parameterization $P_\mu \mapsto \mu$ of $P$) induced by the first five observations and by all six observations are plotted in the first diagram of Figure 4.1 for $\mu \in [-2, 2]$ (solid and dashed lines, respectively; the two functions have been scaled to have the same maximum). These functions give no information about how well the normal models fit the data, and thus no method based only on the likelihood function can recognize $X_6$ as an outlier, if $P$ does not contain some other model to which the normal ones can be compared.

In order to robustify the decisions obtained by conditional application of a likelihood-based decision criterion, we can for example enlarge $\mathcal{P}$ to the set $\mathcal{P}'$ consisting of all the probability measures such that the random variables $X_1, \ldots, X_n$ are independent and identically distributed with variance 1 and either normal or $t$ distribution (with any number $\nu \in (2, \infty)$ of degrees of freedom; the normal distribution corresponds to
the limit $\nu \to \infty$). Let $m : P \mapsto \mu = E_P[X_1]$ be the function on $\mathcal{P}'$ assigning to each model the expected value of the random variables $X_i$, and let $L' : (P, d) \mapsto f[|m(P) - d|]$ be the obvious extension of $L$ to $\mathcal{P}' \times \mathcal{D}$. As above, assume that the likelihood function on $\mathcal{P}'$ induced by the (imprecise) observation $X_i = x_i$ can be approximated using the continuous densities of the random variable $X_i$ (with respect to the Lebesgue measure on $\mathbb{R}$). The profile likelihood functions $\text{lik}_m$ on $\mathbb{R}$ induced by the first five observations and by all six observations are plotted in the first diagram of Figure 4.1 for $\mu \in [-2, 2]$ (solid and dotted lines, respectively; the two functions have been scaled to have the same maximum). When considering only the first five observations, the enlargement of $\mathcal{P}$ to $\mathcal{P}'$ has no real consequence (the $t$ models do not fit these observations considerably better than the normal models): the profile likelihood functions $\text{lik}_m$ and $\text{lik}_{m|P}$ are almost equal on $[-2, 2]$ (in fact, their graphs are indistinguishable), and tend very rapidly to 0 outside this interval; hence, each reasonable likelihood-based decision criterion will give (almost) the same results when applied to the decision problem described by $L'$ as when applied to the one described by $L$, if the function $f$ is not too extreme. But when we consider all six observations, the enlargement of $\mathcal{P}$ to $\mathcal{P}'$ has important consequences: the profile likelihood functions $\text{lik}_m$ and $\text{lik}_{m|P}$ are completely different; that is, the $t$ models fit these observations much better than the normal models. In fact, the second diagram of Figure 4.1 shows the graph of the maximum likelihood estimates of the number $\nu$ of degrees of freedom for the $t$ models in $m^{-1}\{\mu\}$ (as a function of $\mu \in [-2, 2]$): the estimated values of $\nu$ are quite small (far away from the limit $\nu \to \infty$ corresponding to the normal models). The profile likelihood functions $\text{lik}_m$ induced by the first five observations and by all six observations are rather similar, and thus each reasonable likelihood-based decision criterion applied to the decision problem described by $L'$ will give similar results when the outlier $X_6$ is considered and when it is discarded, if the function $f$ is not too extreme. For instance, when applied to the decision problem described by $L'$ with $f : x \mapsto x^2$ (that is, with squared error loss), the MPL and MLD criteria lead respectively to the estimates $-0.286$ and $-0.323$ when $X_6$ is considered, and to the estimate $-0.335$ when $X_6$ is discarded.

The $t$ models with unknown number of degrees of freedom have been successfully used instead of the normal models to obtain maximum likelihood estimates: see for instance Lange, Little, and Taylor (1989) and Pinheiro, Liu, and Wu (2001); since the resulting likelihood functions typically have several local maxima, even better estimates can perhaps be
obtained by using other likelihood-based decision criteria. However, the \( t \) models do not cover all the possible deviations from the normal models; in particular, they have problems with asymmetric distributions and extreme outliers. In order to cope with asymmetric distributions, several generalizations of the \( t \) models have been proposed in recent years: see for instance Azzalini and Capitanio (2003). To allow the possibility of extreme outliers, we could enlarge \( \mathcal{P} \) by replacing the normal distributions of the random variables \( X_1, \ldots, X_n \) with their \( \varepsilon \)-contamination classes for some \( \varepsilon \in (0, 1) \), while still considering these random variables as independent and identically distributed, as done by Huber (1964). By using the MLD criterion we could then obtain reasonable results, but since the pseudo likelihood function \( \text{lik}' \) on \( \mathcal{P} \) would satisfy \( \inf \text{lik}' > 0 \) (independently of the number \( n \) of observations), each likelihood-based decision criterion revealing ignorance aversion and such that the corresponding function \( f_1 : [0, 1] \rightarrow [0, 1] \) is positive on \( (0, 1] \) would consider all the decisions in \( \mathbb{R} \) as equivalent, when applied (with respect to \( \text{lik}' \)) to the decision problem described by the loss function \( L \) corresponding to an unbounded function \( f \). The problem is that under each contaminated model the random variables \( X_1, \ldots, X_n \) are independent and identically distributed, and so for each normal model the event that all \( n \) observations will be arbitrarily extreme outliers has a probability larger than \( \left( \frac{\varepsilon}{n} \right)^n \) under a suitable contaminated model; this is one of the reasons why Huber evaluated the estimators by means of the maximum asymptotic variance (and not the maximum variance) under the contaminated models.

To avoid this problem, we can for example enlarge \( \mathcal{P} \) to the set \( \mathcal{P}'' \) consisting of all the probability measures such that \( n - k \) of the random variables \( X_1, \ldots, X_n \) are independent and normally distributed with the same mean and variance 1 (while the other \( k \) random variables can have any distribution), where \( k \) is any nonnegative integer not larger than the fixed, positive integer \( K < \frac{n}{2} \). Let \( g \) be the function on \( \mathcal{P}'' \) assigning to each model \( P \) the mean \( \mu \) of the majority of the random variables \( X_1, \ldots, X_n \), and let \( L'' : (P, d) \mapsto f[[g(P) - d]] \) be the obvious extension of \( L \) to \( \mathcal{P}'' \times \mathcal{D} \). By using \( \mathcal{P}'' \), we allow at most \( K \) observations to be considered as outliers: the profile likelihood function \( \text{lik}_g \) on \( \mathbb{R} \) induced by the \( n \) observations is the maximum of the likelihood functions \( \text{lik}_g|_\mathcal{P} \) induced by all possible subsets of \( n - K \) observations; in particular, we certainly have \( \inf \text{lik}_g = 0 \), and thus the above problems with unbounded functions \( f \) are resolved. Each value \( \text{lik}_g(\mu) \) of the profile likelihood function induced by the \( n \) observations corresponds to the value \( \text{lik}(P_\mu) \) of the likelihood function induced by the
n—K observations nearest to \( \mu \); that is, if we use no prior information, then the number of observations discarded as outliers is always the maximum possible, because there is no penalty for discarding observations.

This can be changed by using a prior relative plausibility measure on \( \mathcal{P}'' \) with density function proportional to \( r \circ \kappa \), where \( \kappa \) is the function on \( \mathcal{P}'' \) assigning to each model \( P \) the smallest number \( k \) such that \( n - k \) of the random variables \( X_1, \ldots, X_n \) are independent and normally distributed with mean \( g(P) \) and variance 1, while the function \( r \) on \( \{0, \ldots, K\} \) gives the prior relative plausibility of the number of outliers. Each plausibility ratio \( \frac{r_k(0)}{r_k(k)} \) can also be interpreted as the penalty factor for discarding \( k \) observations: if we apply the same penalty factor \( c \in (0,1) \) for each discarded observation, then we obtain the function \( r : k \mapsto c^k \). In particular, if \( c = \delta \varphi(x) \) for an \( x \in \mathbb{P} \), and \( r_P \) is the relative plausibility measure after having observed the data, then for each \( \mu \in \mathbb{R} \) the value \( r_P\{g = \mu\} \) corresponds (up to a positive multiplicative constant independent of \( \mu \)) to the value \( \text{lik}(P_\mu) \) of the likelihood function induced by the \( n \) “huberized” observations: that is, the \( n \) observations after having replaced each value \( x_i \) such that \( |x_i - \mu| > x \) with the value \( \mu + x \text{sign}(x_i - \mu) \) (when there are more than \( K \) such values \( x_i \), only the \( K \) values farthest from \( \mu \) are replaced). For example, when using \( x = 4 \) and observing the first five of the six realizations \( X_i = x_i \) considered above, the density function of the resulting relative plausibility measure \( r_P \circ g^{-1} \) and the likelihood function \( \text{lik}_g \) induced by these observations are proportional on the interval \([-2.824, 2.557]\); on the interval \([-2.824, 1]\) these two functions are also proportional to the density function of the relative plausibility measure \( r_P \circ g^{-1} \) obtained after having observed all six realizations \( X_i = x_i \). Outside these intervals the three functions tend very rapidly to 0, and thus each reasonable likelihood-based decision criterion will give (almost) the same results when applied to the decision problem described by \( L'' \) (with or without considering \( X_6 \)) as when applied to the one described by \( L \) without considering \( X_6 \), if the function \( f \) is not too extreme. For instance, if \( f : x \mapsto x^2 \), then the arithmetic mean of the first five observations is the estimate obtained by applying the MPL and MLD criteria to the decision problem described by \( L'' \) with or without considering \( X_6 \); that is, the outlier \( X_6 \) has no influence on the estimate (this holds also for the LRM_\beta criterion, if \( \beta \geq 0.012 \)).
4.2.2 Asymptotic Optimality

Consider a statistical decision problem described by a loss function $L$ on $\mathcal{P} \times \mathcal{D}$. If the data contain only little information for discrimination between the models in $\mathcal{P}$, then different (conditional) decision criteria can lead to completely different conclusions, when applied to the decision problem described by $L$; but if the data contain a lot of information, then we can expect that the decision criteria lead to similar conclusions. Consider in particular the simplest nontrivial case, in which $L$ is bounded, $\mathcal{P} = \{P_1, P_2\}$, and $\mathcal{D} = \{d_1, d_2\}$, and let $d_1$ be the best decision according to $P_1$; a conditional decision criterion can be reasonable only if it tends to select $d_1$, when in the light of the observed data $P_1$ tends to be infinitely more plausible than $P_2$ (and $P_1$ was not discarded by the prior information). For example, the Bayesian criterion is reasonable in this sense, while in general a decision criterion based on an imprecise Bayesian model updated by means of “regular extension” will tend to select $d_1$ only if the lower prior probability of $P_1$ is positive (that is, some prior information is needed). It can be easily proved that the following property is necessary for a likelihood-based decision criterion to be reasonable in the above sense; we shall see that it is also sufficient. A likelihood-based decision criterion is said to be **consistent** if the corresponding functional $S : S_\mathcal{P} \to \mathbb{R}$ satisfies

$$\lim_{y \to 0} S(I\{1\} + y I\{x\}) = 1 \quad \text{for all } x \in \mathbb{R} \setminus \{1, \infty\}.$$

When compared with condition (4.7), the property of consistency appears as a minimal requirement of “continuity at 0 of the influence of a model, with respect to its relative plausibility”. At the beginning of the previous section we noted that this sort of continuity is necessary for a likelihood-based decision criterion to be useful, and in fact a decision criterion that does not tend to select $d_1$ in the above simple situation cannot be really useful. The consistency of a likelihood-based decision criterion implies that the corresponding functions $f_1$ and $f_0$ on $[0, 1]$ are continuous at 0 and 1, respectively (the continuity of $f_0$ at 1 corresponds to the case with $x = 0$, and the continuity of $f_1$ at 0 follows easily from the next theorem), and thus in particular the essential minimax and essential minimin criteria are not consistent. It can be easily proved that the likelihood-based decision criterion corresponding to the functionals $V_{r_p} : l \mapsto \int l d(f_1 \circ r_p)$ is consistent, if $f_1 : [0, 1] \to [0, 1]$ is nondecreasing and continuous at 0.
with \( f_1(0) = 0 \) and \( f_1(1) = 1 \), and the integral is homogeneous, quasi-subadditive, and respects distributional dominance on the class of all normalized, finitely maxitive measures taking values in the range of \( f_1 \). Hence in particular the MPL, LRM\(_{\beta} \), and MLD criteria are consistent; and Theorem 4.14 implies that the MPL criterion (possibly applied after having distorted the likelihood function by raising its values to the power \( \alpha \in \mathbb{P} \)) is the only likelihood-based decision criterion that is decomposable and consistent.

**Theorem 4.19.** A likelihood-based decision criterion is consistent if and only if the corresponding functionals \( V_{rp} \) have the following property: if \( A \) and \( B \) are two disjoint sets, the sequence \( rp_1, rp_2, \ldots \) of nondegenerate relative plausibility measures on \( Q = A \cup B \) satisfies \( \lim_{n \to \infty} rp_n(A) = 0 \), and \( l : Q \to \mathbb{P} \) is a function such that \( l|_A \) is bounded and \( l|_B \) has constant value \( c \in \mathbb{P} \), then \( \lim_{n \to \infty} V_{rp_n}(l) = c \).

**Proof.** To prove the “if” part it suffices to consider the functions \( l \) such that \( l|_A \) has constant value \( x \in \mathbb{P} \setminus \{1, \infty\} \), and \( l|_B \) has constant value 1. For the “only if” part, if \( c' \in (c, \infty) \), then \( l' = \frac{\sup A}{c'} I_A + I_B \) satisfies \( c' l' \geq l \) and \( \lim_{n \to \infty} V_{rp_n}(l') = 1 \), and thus \( \limsup_{n \to \infty} V_{rp_n}(l) \leq c' \). Analogously, if \( c'' \in (0, c) \), then \( l'' = \frac{\inf A}{c''} I_A + I_B \) satisfies \( c'' l'' \leq l \) and \( \liminf_{n \to \infty} V_{rp_n}(l'') = 1 \), and thus \( \liminf_{n \to \infty} V_{rp_n}(l) \geq c'' \). Hence \( \lim_{n \to \infty} V_{rp_n}(l) = c \). \( \square \)

The continuity property appearing in Theorem 4.19 corresponds to a special case of the continuity of the evaluations \( V_{rp}(l) \) of the uncertain losses \( l \in \mathbb{P} \) when \( \mathcal{P} \) is enlarged to \( \mathcal{P}' \), considered in the previous subsection. In particular, no likelihood-based decision criterion such that the corresponding function \( f_1 : [0, 1] \to [0, 1] \) is positive on \( (0, 1] \) could satisfy it, when the assumption that \( l|_A \) is bounded were dropped. If \( Q = \mathcal{P} \), and the sequence \( rp_1, rp_2, \ldots \) describes the evolution of the uncertain knowledge about the models in \( \mathcal{P} \) obtained by observing more and more data, then Theorem 4.19 implies that a consistent likelihood-based decision criterion tends to evaluate a bounded uncertain loss \( l \) correctly, when \( l \) is constant on \( B \) and the models in \( A \) tend to be discarded by the information obtained. In particular, a consistent likelihood-based decision criterion would tend to select \( d_1 \) in the situation considered at the beginning of the present subsection. Consider a likelihood-based decision criterion and the corresponding functionals \( V_{rp} \); a sequence \( d_1, d_2, \ldots \in \mathcal{D} \) of decisions is said
to be obtained by applying the decision criterion to \( L \) with respect to the sequence \( r_p_1, r_p_2, \ldots \) of nondegenerate relative plausibility measures on \( \mathcal{P} \), if
\[
\lim_{n \to \infty} \| V_{r_p_n}(l_{d_n}) - \inf_{d \in \mathcal{D}} V_{r_p_n}(l_d) \| = 0.
\]
That is, the single decisions \( d_n \) do not need to be strictly optimal according to the decision criterion: only the limiting behavior of the evaluations \( V_{r_p_n}(l_{d_n}) \) is well-defined, but this suffices in the next theorems.

**Theorem 4.20.** Consider a statistical decision problem described by a loss function \( L \) on \( \mathcal{P} \times \mathcal{D} \), and a consistent likelihood-based decision criterion with corresponding functionals \( V_{r_p} \). Let \( d_1, d_2, \ldots \in \mathcal{D} \) be a sequence of decisions obtained by applying the decision criterion to \( L \) with respect to a sequence \( r_p_1, r_p_2, \ldots \) of nondegenerate relative plausibility measures on \( \mathcal{P} \). If
\[
\lim_{n \to \infty} \liminf_{c \to \infty} \left[ V_{r_p_n}(l_d) - V_{r_p_n}(\min\{l_d, c\}) \right] = 0 \quad \text{for all } d \in \mathcal{D}, \tag{4.17}
\]
and there are a model \( P \in \mathcal{P} \) and a topology on \( \mathcal{P} \) such that \( \{l_d : d \in \mathcal{D}\} \) is equicontinuous at \( P \), and \( \lim_{n \to \infty} r_{p_n}(\mathcal{P} \setminus \mathcal{H}) = 0 \) for all neighborhoods \( \mathcal{H} \) of \( P \), then
\[
\lim_{n \to \infty} L(P, d_n) = \inf_{d \in \mathcal{D}} L(P, d).
\]

**Proof.** Let \( \varepsilon \in (0, 1) \), and let \( d' \in \mathcal{D} \) such that \( L(P, d') \leq \inf_{d \in \mathcal{D}} L(P, d) + \varepsilon \). Since \( \{l_d : d \in \mathcal{D}\} \) is equicontinuous at \( P \), there is a neighborhood \( \mathcal{H} \) of \( P \) such that \( \| l_d \|_{\mathcal{H}} - L(P, d) \| \leq \varepsilon \) for all \( d \in \mathcal{D} \). Condition (4.17) implies that there is a \( c' \in \mathcal{P} \) such that \( \lim \sup_{n \to \infty} \left[ V_{r_p_n}(l_{d'}) - V_{r_p_n}(\min\{l_{d'}, c'\}) \right] \leq \varepsilon \).

Let \( l = [L(P, d') + \varepsilon] I_{\mathcal{H}} + \min\{l_{d'}, c'\} I_{\mathcal{P} \setminus \mathcal{H}} \). For sufficiently large \( n \) we have \( V_{r_p_n}(I_{\mathcal{H}}) \geq 1 - \varepsilon \) (since \( f_0 \) is continuous at 1), \( V_{r_p_n}(l) \leq L(P, d') + 2 \varepsilon \) (by Theorem 4.19), \( V_{r_p_n}(l_{d'}) - V_{r_p_n}(\min\{l_{d'}, c'\}) \leq 2 \varepsilon \) (thanks to the above choice of \( c' \)), and \( V_{r_p_n}(l_{d_n}) \leq \inf_{d \in \mathcal{D}} V_{r_p_n}(l_d) + \varepsilon \) (by definition of the sequence \( d_1, d_2, \ldots \)). Therefore we obtain
\[
\inf_{d \in \mathcal{D}} L(P, d) \geq L(P, d') - \varepsilon \geq V_{r_p_n}(l) - 3 \varepsilon \geq V_{r_p_n}(\min\{l_{d'}, c'\}) - 3 \varepsilon \geq V_{r_p_n}(l_{d'}) - 5 \varepsilon \geq \inf_{d \in \mathcal{D}} V_{r_p_n}(l_d) - 5 \varepsilon \geq V_{r_p_n}(l_{d_n}) - 6 \varepsilon \geq V_{r_p_n}(\max\{L(P, d_n) - \varepsilon, 0\} I_{\mathcal{H}}) - 6 \varepsilon \geq \max\{L(P, d_n) - \varepsilon, 0\} V_{r_p_n}(I_{\mathcal{H}}) - 6 \varepsilon \geq [L(P, d_n) - \varepsilon](1 - \varepsilon) - 6 \varepsilon.
\]
This implies the desired result, since \( \varepsilon \in (0, 1) \) was chosen arbitrarily. \( \Box \)
Note that the boundedness of the uncertain losses $l_d$ is sufficient for condition (4.17) to be satisfied, and their continuity at $P$ is necessary for $\{l_d : d \in D\}$ to be equicontinuous at $P$. Theorem 4.20 states that the decisions obtained by applying a consistent likelihood-based decision criterion to $L$ with respect to $r_{P_1}, r_{P_2}, \ldots$ tend to be optimal according to the model $P$, if condition (4.17) is satisfied (that is, unbounded functions $l_d$ pose no problems), and there is a topology on $\mathcal{P}$ that is sufficiently strong to allow the equicontinuity of $\{l_d : d \in D\}$ at $P$ and sufficiently weak to allow $\lim_{n \to \infty} r_{P_n}(\mathcal{P} \setminus \mathcal{H}) = 0$ for all neighborhoods $\mathcal{H}$ of $P$. A sequence $d_1, d_2, \ldots$ of decisions depending on the realizations of a sequence of random objects is said to be (strongly) asymptotically optimal under the model $P \in \mathcal{P}$ (for the decision problem described by $L$) if the decisions $d_n \in D$ are well-defined for sufficiently large $n$ a.e. $[P]$, and

$$\lim_{n \to \infty} L(P, d_n) = \inf_{d \in D} L(P, d) \quad \text{a.e. } [P].$$

When a sequence $r_{P_1}, r_{P_2}, \ldots$ of nondegenerate relative plausibility measures on $\mathcal{P}$ is obtained by observing the realizations of a sequence of random objects, the asymptotic optimality (under the model $P$) of the decisions obtained by applying a consistent likelihood-based decision criterion to $L$ with respect to $r_{P_1}, r_{P_2}, \ldots$ follows easily from the proof of Theorem 4.20, if condition (4.17) holds a.e. $[P]$, and there is a topology on $\mathcal{P}$ such that $\{l_d : d \in D\}$ is equicontinuous at $P$, and $\lim_{n \to \infty} r_{P_n}(\mathcal{P} \setminus \mathcal{H}) = 0$ a.e. $[P]$ for all neighborhoods $\mathcal{H}$ of $P$. In particular, if under each model in $\mathcal{P}$ the random objects are independent and identically distributed, then the last condition can often be proved by means of the next theorem, which is essentially based on the version of Schervish (1995, Subsection 7.3.2) of the theorem of Wald (1949) on the (strong) consistency of the maximum likelihood estimates; in order to state the theorem, we need some definitions.

Let $\mathcal{P}$ be a set of probability measures on a measurable space $(\Omega, \mathcal{A})$, let $X : \Omega \to \mathcal{X}$ be a random object that is discrete under each model in $\mathcal{P}$, and for each $x \in \mathcal{X}$ let $r_{P_x}$ be the relative plausibility measure on $\mathcal{P}$ induced by the observation $X = x$. For each $P \in \mathcal{P}$ and each $\mathcal{H} \subseteq \mathcal{P}$ define

$$\mathcal{I}_X(P, \mathcal{H}) = E_P \left( \log \frac{r_{P_x}\{P\}}{r_{P_x}(\mathcal{H})} \right),$$

where $\frac{\theta}{0}$ is defined as $1$ (or any other value in $\overline{\mathbb{P}}$; this choice has no influence on $\mathcal{I}_X(P, \mathcal{H})$). If $\mathcal{H} = \{P'\}$, then $\mathcal{I}_X(P, \mathcal{H})$ is the expected value (under $P$)
of the information in $X$ for discrimination in favor of $P$ against $P'$, introduced by Kullback and Leibler (1951); they proved that $I_X(P, \{P'\}) \in \mathbb{P}$, and that $I_X(P, \{P'\}) = 0$ if and only if $X$ has the same distribution under $P$ as under $P'$. In general, $I_X(P, \mathcal{H})$ can be interpreted as the expected value (under $P$) of the information in $X$ for discrimination in favor of $P$ against $\mathcal{H}$; note that $I_X(P, \mathcal{H})$ is not necessarily in $\mathbb{P}$ (it is not even necessarily well-defined). For each $x \in X$ let $\text{lik}_x$ be the likelihood function on $\mathcal{P}$ induced by the observation $X = x$; the value

$$H_X(P) = E_P(-\log[\text{lik}_X(P)])$$

is the entropy of the distribution of $X$ under the model $P \in \mathcal{P}$, studied by Shannon (1948). Since $\text{lik}_x \leq 1$ for all $x \in X$, we have $H_X(P) \in \mathbb{P}$, and if $H_X(P)$ is finite, then $I_X(P, \mathcal{H}) \geq -H_X(P)$ for all $\mathcal{H} \subseteq \mathcal{P}$ (in particular, $I_X(P, \mathcal{H})$ is well-defined for all $\mathcal{H} \subseteq \mathcal{P}$). In the next theorems only discrete random objects are considered, but the results are valid also for continuous ones, if the approximate likelihood functions based on the densities are used, and are bounded a.e. $[P]$ (when this is not the case, the approximation is unreasonable); the approximation must be explicitly considered when determining the entropy of the distributions.

**Theorem 4.21.** Let $\mathcal{P} \cup \{P\}$ be a set of probability measures on a measurable space $(\Omega, \mathcal{A})$, and let $X_1, X_2, \ldots : \Omega \to X$ be a sequence of random objects such that under each model in $\mathcal{P} \cup \{P\}$ the random objects are discrete, independent, and identically distributed. Let $\tau_P$ be the prior (nondegenerate) relative plausibility measure on $\mathcal{P}$, and let $\tau_{P_1}, \tau_{P_2}, \ldots$ be the sequence of relative plausibility measures on $\mathcal{P}$ obtained by observing the realizations of the random objects $X_1, X_2, \ldots$. If $P' \in \{\tau_P > 0\}$ is the unique $P'' \in \mathcal{P}$ minimizing $I_{X_1}(P, \{P''\})$, and $\mathcal{P}$ is equipped with a topology such that there are a compact subset $\mathcal{P}'$ of $\mathcal{P}$ and a set $\mathcal{A}$ of relatively open subsets of $\mathcal{P}'$ such that $\bigcup \mathcal{A} = \mathcal{P}' \setminus \{P'\}$ and $I_{X_1}(P, A) > I_{X_1}(P, \{P'\})$ for all $A \in \mathcal{A}$ and also for $A = \mathcal{P} \setminus \mathcal{P}'$, then the relative plausibility measures $\tau_{P_1}, \tau_{P_2}, \ldots$ are nondegenerate a.e. $[P]$, and $\lim_{n \to \infty} \tau_{P_n}(\mathcal{P} \setminus \mathcal{H}) = 0$ a.e. $[P]$ for all neighborhoods $\mathcal{H}$ of $P'$.

In particular, the set $\mathcal{A}$ certainly exists if the relative topology on $\mathcal{P}'$ is metrizable, the likelihood function on $\mathcal{P}'$ induced by observing the realization of $X_1$ is upper semicontinuous a.e. $[P]$, and $H_{X_1}(P)$ is finite.

**Proof.** Let $\text{lik}_k$ be the likelihood function on $\mathcal{P}$ induced by observing the realization of $X_k$, and let $Y_k = \inf_A \log \frac{\text{lik}_k(P')}{\text{lik}_k}$, where $0^0$ is defined as 1,
and $A$ is a subset of $\mathcal{P}$ such that $\mathcal{I}_{X_1}(P, A) > \mathcal{I}_{X_1}(P, \{P'\})$. Since such a set $A$ certainly exists (for example $A = \mathcal{P} \setminus \mathcal{P}'$), we obtain in particular that $\mathcal{I}_{X_1}(P, \{P'\})$ is finite, and thus $\text{lik}_k(P') > 0$ a.e. $[P]$. Therefore $r_{P_1}, r_{P_2}, \ldots$ are nondegenerate a.e. $[P]$, and

$$\log \frac{r_{P_n}(\{P'\})}{r_{P_n}(A)} = \inf_A \left( \log \frac{r_{P_n}(P')}{r_{P_n}(P)} + \sum_{k=1}^{n} \log \frac{\text{lik}_k(P')}{\text{lik}_k(P)} \right) \geq \log \frac{r_{P_n}(P')}{r_{P_n}(A)} + \sum_{k=1}^{n} Y_k$$

holds a.e. $[P]$. The strong law of large numbers implies that the right-hand side of the inequality tends to infinity a.e. $[P]$ as $n \to \infty$, since the $Y_k$ are independent and identically distributed (under $P$), and

$$E_{P}(Y_1) = E_{P}\left( \log \frac{\text{lik}_1(P)}{\sup_A \text{lik}_1} \right) - E_{P}\left( \log \frac{\text{lik}_1(P)}{\text{lik}_1(P')} \right) > 0.$$  

That is, $\lim_{n \to \infty} r_{P_n}(A) = 0$ a.e. $[P]$ for all the subsets $A$ of $\mathcal{P}$ such that $\mathcal{I}_{X_1}(P, A) > \mathcal{I}_{X_1}(P, \{P'\})$. Let $\mathcal{H}$ be a neighborhood of $P'$. There is an open neighborhood $\mathcal{H}'$ of $P'$ such that $\mathcal{H}' \subseteq \mathcal{H}$; and since $\mathcal{P}' \setminus \mathcal{H}'$ is a compact subset of $\bigcup A$, there are $A_1, \ldots, A_m \in \mathcal{A}$ such that $\mathcal{P}' \setminus \mathcal{H}'$ is a subset of $\bigcup_{j=1}^{m} A_j$. Hence

$$r_{P_n}(\mathcal{P} \setminus \mathcal{H}) \leq \max\{r_{P_n}(\mathcal{P} \setminus \mathcal{P}'), r_{P_n}(A_1), \ldots, r_{P_n}(A_m)\},$$

and the right-hand side tends to 0 a.e. $[P]$ as $n \to \infty$.

To prove the last statement of the theorem, we restrict attention to the set $\mathcal{P}'$ equipped with the relative topology. Since this is metrizable, for each $P'' \in \mathcal{P}\setminus\{P'\}$ there is a sequence $\mathcal{H}_1 \supseteq \mathcal{H}_2 \supseteq \ldots$ of closed neighborhoods of $P''$ such that $\bigcap_{m=1}^{\infty} \mathcal{H}_m = \{P''\}$. Since $\mathcal{P}'$ is compact, if $\text{lik}_1 | P'$ is upper semicontinuous, then each $\mathcal{H}_m$ contains a model $P_m$ such that $\text{lik}_1(P_m) = \sup_{\mathcal{H}_m} \text{lik}_1$, and $\lim_{m \to \infty} \sup_{\mathcal{H}_m} \text{lik}_1(P_m) < \text{lik}_1(P'')$. That is, $\lim_{m \to \infty} \sup_{\mathcal{H}_m} \text{lik}_1 = \text{lik}_1(P'')$ a.e. $[P]$. Since $\mathcal{I}_{X_1}(P)$ is finite, we have $\mathcal{I}_{X_1}(P, \mathcal{A}) > -\infty$ for all $A \subseteq \mathcal{P}'$. If $\mathcal{I}_{X_1}(P, \mathcal{H}_1) = \infty$, then

$$\mathcal{I}_{X_1}(P, A) \geq \mathcal{I}_{X_1}(P, \mathcal{P} \setminus \mathcal{P}') > \mathcal{I}_{X_1}(P, \{P'\}).$$

If $\mathcal{I}_{X_1}(P, \mathcal{H}_1) < \infty$, then $\sup_{\mathcal{H}_1} \text{lik}_1 > 0$ a.e. $[P]$, and

$$\liminf_{m \to \infty} \mathcal{I}_{X_1}(P, \mathcal{H}_m) = \liminf_{m \to \infty} E_{P}\left( \log \frac{\sup_{\mathcal{H}_m} \text{lik}_1}{\sup_{\mathcal{H}_m} \text{lik}_1} \right) + \mathcal{I}_{X_1}(P, \mathcal{H}_1) \geq$$

$$\geq E_{P}\left( \log \frac{\sup_{\mathcal{H}_1} \text{lik}_1}{\text{lik}_1(P'')} \right) + \mathcal{I}_{X_1}(P, \mathcal{H}_1) =$$

$$= \mathcal{I}_{X_1}(P, \{P''\}) > \mathcal{I}_{X_1}(P, \{P'\}),$$

and
where the first inequality follows from Fatou's lemma. Thus in any case we have $\mathcal{I}_{X_1}(P, \mathcal{H}_m) > \mathcal{I}_{X_1}(P, \{P''\})$ for a sufficiently large $m$; let $A \subseteq \mathcal{H}_m$ be an open neighborhood of $P''$: since $\mathcal{I}_{X_1}(P, A) \geq \mathcal{I}_{X_1}(P, \mathcal{H}_m)$, the collection $\mathcal{A}$ of these sets $A$ (for all $P'' \in \mathcal{P}' \setminus \{P''\}$) has the desired properties. \qed

In Theorem 4.21, the model $P$ generating the data is not assumed to be in $\mathcal{P}$. If the premise of Theorem 4.21 is satisfied, the loss function $L$ on $\mathcal{P} \times \mathcal{D}$ is bounded, and $\{l_d : d \in \mathcal{D}\}$ is equicontinuous, then the decisions obtained by applying a consistent likelihood-based decision criterion to $L$ with respect to $r_{P_1}, r_{P_2}, \ldots$ tend to be optimal (a.e. $[P]$) according to the unique model $P' \in \mathcal{P}$ that is expected to be the most compatible with the data, in the sense that $P'$ minimizes the expected information for discrimination in favor of $P$ contained in each observation. The situation is more complicated when this expected information is minimized by several models in $\mathcal{P}'$, because then the resulting relative plausibility measures can be asymptotically unstable on the set of these models (see for example Berk, 1966). If $P \in \mathcal{P}$, and the distribution of $X_1$ under $P$ is different from the distribution of $X_1$ under any other model in $\mathcal{P}$, then $P$ is the unique $P'' \in \mathcal{P}$ minimizing $\mathcal{I}_{X_1}(P, \{P''\})$, and thus Theorems 4.20 and 4.21 can be joined to prove the asymptotic optimality of the decisions obtained by applying a consistent likelihood-based decision criterion.

**Theorem 4.22.** Let $\mathcal{P}$ be a finite set of probability measures on a measurable space $(\Omega, \mathcal{A})$, and let $X_1, X_2, \ldots : \Omega \to X$ be a sequence of random objects such that under each model in $\mathcal{P}$ the random objects are discrete, independent, and identically distributed, and their distributions are different under different models in $\mathcal{P}$. Consider a statistical decision problem described by a finite loss function $L$ on $\mathcal{P} \times \mathcal{D}$, with prior (nondegenerate) relative plausibility measure $r_P$ on $\mathcal{P}$. All sequences of decisions obtained by applying a consistent likelihood-based decision criterion to $L$ (with respect to the sequence of relative plausibility measures on $\mathcal{P}$ obtained by observing the realizations of the random objects $X_1, X_2, \ldots$) are asymptotically optimal under all models in $\{r_P > 0\}$.

**Proof.** Let $P \in \{r_P > 0\}$, and consider the discrete topology on $\mathcal{P}$. If $\mathcal{P}' = \mathcal{P}$, and $A$ is the set of all singletons of $\mathcal{P} \setminus \{P\}$, then the premise of Theorem 4.21 is satisfied, with $P' = P$. Since $l_d$ is bounded for all $d \in \mathcal{D}$, and $\{l_d : d \in \mathcal{D}\}$ is trivially equicontinuous with respect to the discrete topology on $\mathcal{P}$, the premise of Theorem 4.20 holds a.e. $[P]$, and the desired result follows. \qed
When the set $\mathcal{P}$ of statistical models considered is infinite, the uncertain losses $l_d$ can be finite but unbounded, and usually the premise of Theorem 4.21 cannot be satisfied when using the discrete topology on $\mathcal{P}$; hence in general the condition (4.17) and the equicontinuity of $\{l_d : d \in \mathcal{D}\}$ are not trivially satisfied. In some cases they pose no problem, as in the following example, while sometimes the subsequent generalization of Theorem 4.20 can be useful.

Example 4.23. Consider the estimation problem of Example 1.1, and let the random variables $X_1, X_2, \ldots : \Omega \rightarrow \{0, 1\}$ describe the results of each single binary experiment (that is, $X = \sum_{k=1}^n X_k$). With respect to the topology on $\mathcal{P}$ induced by the metric $\rho : (\mathcal{P}, \mathcal{M}) \rightarrow |p - p'|$, both likelihood functions $P_p \mapsto 1 - p$ and $P_p \mapsto p$ on $\mathcal{P}$ induced by observing the two possible realizations 0 and 1 of $X_1$ are continuous, $\mathcal{P}$ is compact, and $|l_d(P) - l_d(P')| \leq 2\rho(P, P')$ for all $d \in [0,1]$ and all $P, P' \in \mathcal{P}$. Since $L$ is bounded and $H(X_1) = \log 2$ for all $P \in \mathcal{P}$, Theorems 4.20 and 4.21 imply that as $n$ tends to infinity, all sequences of decisions obtained by applying a consistent likelihood-based decision criterion to the decision problem described by $L$ are asymptotically optimal under all models in $\mathcal{P}$. To be precise, in order to obtain this result we must consider that in the proof of Theorem 4.20, for each $\varepsilon \in (0,1)$ we use the assumption $\liminf_{d \to P} r_{P_n}((d, \mathcal{H})) = 0$ only for a particular neighborhood $\mathcal{H}$ of $P$.

Theorem 4.24. Consider a statistical decision problem described by a loss function $L$ on $\mathcal{P} \times \mathcal{D}$, and a consistent likelihood-based decision criterion with corresponding functionals $V_{r_p}$. Let $r_p, r_{p_2}, \ldots$ be a sequence of relative plausibility measures on $\mathcal{P}$, depending on the realizations of a sequence of random objects. If there are a model $P \in \mathcal{P}$ and a set $\mathcal{M}$ of nondegenerate relative plausibility measures on $\mathcal{P}$ such that $r_{P_n} \in \mathcal{M}$ for sufficiently large $n$ a.e. $[P]$, and there is a subset $\mathcal{D}'$ of $\mathcal{D}$ such that

$$\lim_{c \to \infty} \sup_{r_p \in \mathcal{M}} [V_{r_p}(l_d) - V_{r_p}(\min\{l_d, c\})] = 0 \quad \text{for all } d \in \mathcal{D'},$$

$$\inf_{d \in \mathcal{D}'} L(P, d) = \inf_{d \in \mathcal{D}} L(P, d),$$

there is a positive real number $\varepsilon$ such that

all $d' \in \mathcal{D}$ satisfying $V_{r_p}(l_{d'}) \leq \inf_{d \in \mathcal{D}} V_{r_p}(l_d) + \varepsilon$ for some $rp \in \mathcal{M}$ are in $\mathcal{D}'$, and there is a topology on $\mathcal{P}$ such that $\{l_d : d \in \mathcal{D}'\}$ is equicontinuous at $P$, and $\lim_{n \to \infty} r_{P_n}(\mathcal{P} \setminus \mathcal{H}) = 0$ a.e. $[P]$ for all neighborhoods $\mathcal{H}$ of $P$, then all sequences of decisions obtained by applying the decision criterion to $L$ with respect to $r_p, r_{p_2}, \ldots$ are asymptotically optimal under $P$.

Proof. It suffices to adapt the proof of Theorem 4.20. $\square$
If the premise of Theorem 4.21 is satisfied, with $P' = P$, then its conclusion can be useful for selecting a set $\mathcal{M}$ of nondegenerate relative plausibility measures on $\mathcal{P}$ such that $r_{p_n} \in \mathcal{M}$ for sufficiently large $n$ a.e. $[P]$. Even when we consider a loss function $L$ and a consistent likelihood-based decision criterion that do not fulfill the condition (4.17) and the equicontinuity of $\{l_d : d \in \mathcal{D}\}$ at $P$, it is sometimes possible to select a subset $\mathcal{D}'$ of $\mathcal{D}$ such that the premise of Theorem 4.24 is satisfied (for some $\varepsilon \in \mathbb{P}$), as in the next example.

**Example 4.25.** Consider the estimation problem of Example 4.18, with $P_\mu \rightarrow \delta \varphi(x_i - \mu)$ as approximation of the likelihood function on $\mathcal{P}$ induced by the (imprecise) observation $X_i = x_i$. With respect to the topology on $\mathcal{P}$ induced by the metric $\rho : (P_\mu, P_{\mu'}) \mapsto |\mu - \mu'|$, the set $\mathcal{P}' = \{P_{\mu'} : \mu' \in [\mu - 2, \mu + 2]\}$ is compact for all $\mu \in \mathbb{R}$, and all the likelihood functions on $\mathcal{V}$ induced by observing the possible realizations of $X_1$ are continuous. Since $\mathcal{I}_{X_1}(P_\mu, \mathcal{P} \setminus \mathcal{P}') \approx 0.398$ for all $\mu \in \mathbb{R}$, and $H_{X_1}(P) = \frac{1}{2} + \log \frac{\sqrt{2 \pi}}{\delta}$, Theorem 4.21 (more precisely, its version for continuous random objects) implies that $\lim_{n \rightarrow \infty} r_{p_n}(\mathcal{P} \setminus \mathcal{H}) = 0$ a.e. $[P_{\mu}]$ for all $\mu \in \mathbb{R}$ and all neighborhoods $\mathcal{H}$ of $P_{\mu}$, where $r_{p_1}, r_{p_2}, \ldots$ is the sequence of (nondegenerate) relative plausibility measures on $\mathcal{P}$ obtained by observing the realizations of the random variables $X_1, X_2, \ldots$.

Consider for instance the MPL criterion and the decision problem described by the loss function $L : (P_\mu, d) \mapsto (\mu - d)^2$ (that is, the estimation of $\mu$ with squared error loss); since $\{l_d : d \in \mathcal{D}\}$ is nowhere equicontinuous on $\mathcal{P}$, Theorem 4.20 cannot be directly applied. But for each $\mu \in \mathbb{R}$ the above result implies that $r_{p_n} \in \mathcal{M}_{\mu}$ for sufficiently large $n$ a.e. $[P_\mu]$, when $\mathcal{M}_{\mu}$ is defined for instance as the set of all (nondegenerate) relative plausibility measures on $\mathcal{P}$ with density function proportional to a function of the form $P_{\mu'} \mapsto [\varphi(x - \mu')]^m$, for all $x \in [\mu - 1, \mu + 1]$ and all positive integers $m$. For each $\mu \in \mathbb{R}$ let $\mathcal{D}_\mu' = [\mu - 2, \mu + 2]$; Theorems 2.21 and 2.6 imply that for all $\mu \in \mathbb{R}$ and all $d \in \mathcal{D}_\mu'$

\[
\lim_{c \uparrow \infty} \sup_{r_{p_n} \in \mathcal{M}_{\mu}} \left( \int_0^S l_d \, dr_{p} - \int_0^S \min\{l_d, c\} \, dr_{p} \right) \leq 0
\]

\[
\leq \lim_{c \uparrow \infty} \sup_{r_{p_n} \in \mathcal{M}_{\mu}} \int_0^S \max\{l_d - c, 0\} \, dr_{p} \leq 0
\]

\[
\leq \lim_{c \uparrow \infty} \sup_{y \in (\sqrt{c}, \infty)} (y^2 - c) \exp\left(-\frac{(y-3)^2}{2}\right) \leq \lim_{c \uparrow \infty} c \exp\left(-\frac{(\sqrt{c}-3)^2}{2}\right) = 0.
\]
Moreover, for each $\mu \in \mathbb{R}$ we have $\inf_{d \in D'} L(P_\mu; d) = 0 = \inf_{d \in D} L(P_\mu; d)$, and since the Shilkret integral respects distributional dominance (on the class of all monotone measures), for $\varepsilon = 1 - 2e^{-1} \approx 0.264$, all $d' \in D \setminus D'$, and all $\rho \in \mathcal{M}_\mu$ we also have

$$\int_{S} l_{d'} \, d\rho \geq \int_{S} l_{d'} \, d(I_{(1)} \circ \rho) > 1 + 2e^{-1} + \varepsilon \geq \inf_{d \in D} \int_{S} l_{d} \, d\rho + \varepsilon.$$ 

Since $\sup_{d \in D'} |l_{d} - l_{d}(P_\mu)| = \max\{l_{\mu-2}, l_{\mu+2}\} - 4$ is continuous at $P_\mu$ (with respect to the topology on $\mathcal{P}$ induced by the metric $\rho$) for all $\mu \in \mathbb{R}$, Theorem 4.24 implies that all sequences of decisions obtained by applying the MPL criterion to $L$ with respect to $\rho_1, \rho_2, \ldots$ are asymptotically optimal under all models in $\mathcal{P}$. This simply means that as $n$ tends to infinity, the arithmetic mean of $X_1, \ldots, X_n$ is strongly consistent for $\mu$, as implied also by the strong law of large numbers.

Of course, sometimes the premise of Theorem 4.24 cannot be satisfied, simply because the decisions obtained by applying the likelihood-based decision criterion considered are not asymptotically optimal, as in the next example.

**Example 4.26.** In the situation of Example 4.25, consider the loss function $L$ on $\mathcal{P} \times \{d_1, d_2\}$ defined by $l_{d_1} = c_1 I_{\mathcal{H}}$ and $l_{d_2} = c_2 I_{\mathcal{P} \setminus \mathcal{H}}$, where $c_1, c_2 \in \mathbb{R}$ and $\mathcal{H} = \{P_\mu : \mu \in \mathbb{R} \setminus \mathbb{P}\}$. That is, consider the problem of testing if the mean of a normal distribution with known variance is positive or not, without prior information about the relative plausibility of the possible values of the mean. Since $L$ is bounded, and $\{l_{d_1}, l_{d_2}\}$ is equicontinuous at $P_\mu$ for all $\mu \in \mathbb{R} \setminus \{0\}$, with respect to the topology on $\mathcal{P}$ considered in Example 4.25, Theorem 4.24 implies that all sequences of decisions obtained by applying a consistent likelihood-based decision criterion to $L$ with respect to $\rho_1, \rho_2, \ldots$ are asymptotically optimal under all models in $\mathcal{P} \setminus \{P_0\}$ (it suffices to choose $D' = \{d_1, d_2\}$, and define $\mathcal{M}$ as the set of all nondegenerate relative plausibility measures on $\mathcal{P}$). But $l_{d_1}$ and $l_{d_2}$ are continuous at $P_0$ only with respect to a topology on $\mathcal{P}$ such that $\mathcal{H}$ is a neighborhood of $P_0$, while $\lim_{n \to \infty} \rho_{2n} (\mathcal{P} \setminus \mathcal{H}) = 0$ a.e. $[P_0]$ does not hold, since $P_0[\rho_{2n} (\mathcal{P} \setminus \mathcal{H}) = 1] = \frac{1}{2}$ for all $n$. That is, the premise of Theorem 4.24 cannot be satisfied when $P = P_0$, and in fact it can be easily proved that no sequence of decisions obtained by applying a consistent likelihood-based decision criterion to $L$ with respect to $\rho_1, \rho_2, \ldots$ can be asymptotically optimal under $P_0$. \[\diamond\]
5

Point Estimation

The present chapter focuses on the estimation of points in the Euclidean space \( \mathbb{R}^k \): many statistical decision problems can be formulated as problems of point estimation in \( \mathbb{R}^k \) by expressing the estimation error in some specific way (see also Wald, 1939). We shall consider the application of likelihood-based decision criteria to the problems of point estimation in \( \mathbb{R}^k \), with particular attention to the asymptotic efficiency of the resulting sequences of estimators, and to some important properties of those likelihood-based decision criteria that can be represented by means of the Shilkret integral.

5.1 Likelihood-Based Estimates

Let \( \mathcal{P} \) be a set of statistical models, and let \( g : \mathcal{P} \rightarrow \mathbb{R}^k \) be a mapping, where \( k \) is a positive integer. Consider the problem of estimating \( g(P) \) on the basis of a nondegenerate relative plausibility measure \( r_P \) on \( \mathbb{R}^k \) describing the uncertain knowledge about the elements of \( \mathbb{R}^k \); that is, consider the problem of estimating \( g(P) \) on the basis of the pseudo likelihood function \( r_P^\dagger \). When some data are observed, \( r_P^\dagger \) can be defined as equivalent to the induced profile likelihood function on \( \mathbb{R}^k \); but other pseudo likelihood functions with respect to \( g \) can also be used, and prior information can be included. In particular, the complexity of the models in \( \mathcal{P} \) can be taken into account: for example by penalizing it in accordance with the criteria introduced by Akaike (1973) or Rissanen (1978). Moreover, as seen in Example 4.18, in general an estimate of \( g(P) \) based on the pseudo likelihood function \( r_P^\dagger \) can be robust only if robustness aspects were considered when defining \( r_P^\dagger \).
When the estimation problem is stated without reference to a particular loss function, the usual likelihood-based inference methods can be applied to the pseudo likelihood function \( r_p \downarrow \), by considering it as equivalent to the profile likelihood function on \( R^k \) with respect to \( g \). If the maximum likelihood estimate \( \hat{g}_{ML} \) of \( g(P) \) exists and is really unique, in the sense that \( r_p[t^{-1}_{\hat{g}_{ML}}(B_\varepsilon)] < r_p\{\hat{g}_{ML}\} \) for all \( \varepsilon \in P \), then it is certainly a most natural point estimate of \( g(P) \). But the method of maximum likelihood uses \( r_p \downarrow \) in a very limited way: the additional consideration of a likelihood-based confidence region for \( g(P) \) with cutoff point \( \beta \in (0, 1) \) can be very informative. In the case with \( k = 1 \), such a confidence region is often an interval, and its endpoints can be generalized by the following conjugate upper and lower evaluations of \( g(P) \):

\[
\int C id_{\mathbb{R}} d(\delta \circ r_p) \quad \text{and} \quad \int C id_{\mathbb{R}} d(\delta \circ r_p),
\]

where \( \delta : [0, 1] \to [0, 1] \) is a nondecreasing function such that \( \delta(0) = 0 \) and \( \delta(1) = 1 \), and the integrals are defined as the limits of the integrals with \( id_{\mathbb{R}} \) and \( r_p \) restricted to the interval \([a, b]\), when \(-a\) and \( b\) tend to infinity and the Choquet integral is extended by means of equality (2.10). In the case with \( \delta = I_{[\beta, 1]} \) for some \( \beta \in (0, 1) \), the above conjugate upper and lower evaluations of \( g(P) \) are well-defined and correspond respectively to the supremum and to the infimum of the likelihood-based confidence region for \( g(P) \) with cutoff point \( \beta \).

It can be easily proved that if the maximum likelihood estimate \( \hat{g}_{ML} \) of \( g(P) \) exists, and \( k = 1 \), then the above conjugate upper and lower evaluations of \( g(P) \) are well-defined and can be expressed as

\[
\hat{g}_{ML} + \int C id_{\mathbb{R}} d(\delta \circ r_p \circ t^{-1}_{\hat{g}_{ML}}|\mathbb{P}) \quad \text{and} \quad \hat{g}_{ML} - \int C id_{\mathbb{R}} d[\delta \circ r_p \circ (t_{\hat{g}_{ML}})^{-1}|\mathbb{P}],
\]

respectively. Any regular integral (on the class of all monotonic measures) could be substituted for the Choquet integral in these two expressions, and the resulting upper and lower evaluations of \( g(P) \) could be combined to obtain a point estimate of \( g(P) \), but the meaning of this estimate would not be very clear. In order to obtain a point estimate of \( g(P) \) depending on the pseudo likelihood function \( r_p \downarrow \) in a less limited way than the maximum likelihood estimate \( \hat{g}_{ML} \) does, it is better to explicitly describe the estimation problem by means of a loss function and then apply a likelihood-based decision criterion.
A function \( f : \mathbb{R}^k \to \overline{P} \) is said to be quasiconvex if
\[
\lambda x + (1 - \lambda) y \leq \max\{f(x), f(y)\} \quad \text{for all } x, y \in \mathbb{R}^k \text{ and all } \lambda \in (0, 1);
\]
\( f \) is said to be strictly quasiconvex if the inequality is strict when \( x \neq y \).

An (estimation) error function for the problem of estimating \( \gamma \in \mathbb{R}^k \) is a function \( L : \mathbb{R}^k \times \mathbb{R}^k \to \overline{P} \) such that
\[
L(\gamma, d) = w(\gamma) f(\gamma - d) \quad \text{for all } \gamma, d \in \mathbb{R}^k,
\]
where \( w : \mathbb{R}^k \to P \) is a function, and \( f : \mathbb{R}^k \to \overline{P} \) is a quasiconvex function with \( f(0) = 0 \). The statistical decision problem of estimating \( g(P) \) with error function \( L \) is described by a loss function \( L' \) on \( P \times \mathcal{D} \), where \( \mathcal{D} \) is a subset of \( \mathbb{R}^k \), and \( L'(P, d) = L[g(P), d] \) for all \( P \in \mathcal{P} \) and all \( d \in \mathcal{D} \); obviously, the definition of \( L \) outside the set \( \{(g(P), d) : P \in \mathcal{P}, \ d \in \mathcal{D}\} \) is irrelevant for the decision problem described by \( L' \). Condition (4.6) implies that applying a likelihood-based decision criterion to \( L' \) with respect to a likelihood function \( \text{lik} \) on \( \mathcal{P} \) is equivalent to applying it to \( L|_{\mathbb{R}^k \times \mathcal{D}} \) with respect to the profile likelihood function \( \text{lik}_\theta \) on \( \mathbb{R}^k \). Therefore, more generally, we can consider a nondegenerate relative plausibility measure \( \mathcal{R} \) on \( \mathbb{R}^k \) and compare the estimates \( d \in \mathcal{D} \subseteq \mathbb{R}^k \) by means of the evaluation \( V_{\mathcal{R}}(l_d) \) of \( l_d \), where \( V_{\mathcal{R}} \) is the functional on \( \overline{P} \mathbb{R}^k \) corresponding to a likelihood-based decision criterion, and for each \( d \in \mathbb{R}^k \) the function \( l_d \) on \( \mathbb{R}^k \) is defined by \( l_d(\gamma) = L(\gamma, d) \) (for all \( \gamma \in \mathbb{R}^k \)).

**Theorem 5.1.** Consider the problem of estimating \( \gamma \in \mathbb{R}^k \) on the basis of a nondegenerate relative plausibility measure \( \mathcal{R} \) on \( \mathbb{R}^k \), with an error function \( L \) such that \( f : x \mapsto f'(|Mx|) \), where \( f' \) is a function on \( \overline{P} \), and \( M \in \mathbb{R}^{k \times k} \) is an invertible matrix. Let \( \mathcal{G} \) be the closed convex hull of \( \{\mathcal{R} \mathcal{P}^\perp > 0\} \). For each \( d \in \mathbb{R}^k \) there is a \( d' \in \mathcal{G} \) such that according to no likelihood-based decision criterion \( d \) is preferred to \( d' \).

Moreover, if \( \mathcal{G} \) is bounded, and \( L \) is such that \( f' \) is strictly increasing, \( \inf\{\mathcal{R} \mathcal{P}^\perp > 0\} w > 0 \), and \( \sup\{\mathcal{R} \mathcal{P}^\perp > 0\} w < \infty \), then for each \( d \in \mathbb{R}^k \setminus \mathcal{G} \) there is a \( d' \in \mathcal{G} \) that is preferred to \( d \) according to all likelihood-based decision criteria.

**Proof.** Let \( d \not\in \mathcal{G} \) (otherwise we can choose \( d' = d \)), and let \( d' \) be the element of \( \mathcal{G} \) such that \( Md' \) is the projection of \( Md \) on the closed, convex set \( \{M\gamma : \gamma \in \mathcal{G}\} \). Since \( |M(\gamma - d')| < |M(\gamma - d)| \) for all \( \gamma \in \mathcal{G} \), we have \( l_{d'}(\mathcal{R} \mathcal{P}^\perp > 0) \leq l_d(\mathcal{R} \mathcal{P}^\perp > 0) \), and the first statement of the theorem follows from Theorem 4.2.
At the beginning of Section 4.1 we noted that $d'$ is preferred to $d$ according to all likelihood-based decision criteria, if $\text{ess}_{\mathcal{R}} \sup_{d'} l_{d'} < \infty$ and $\inf(l_d - l_{d'}) > 0$. Theorem 4.2 implies that the second condition can be weakened to $\inf_{\{r_{p} > 0\}} (l_{d} - l_{d'}) > 0$, while the first one can be expressed as $\sup_{\{r_{p} > 0\}} l_{d'} < \infty$; these two conditions follow easily from the premise of the second statement of the theorem. \hfill \Box

Theorem 5.1 implies that for an important class of error functions, when applying a likelihood-based decision criterion to the problem of estimating $\gamma \in \mathbb{R}^k$ on the basis of a nondegenerate relative plausibility measure $r_{\mathcal{R}}$ on $\mathbb{R}^k$, we can restrict attention to the estimates in the closed convex hull of $\{r_{p} > 0\}$. Theorem 5.1 can be easily generalized; in particular, in the case with $k = 1$, the assumption about $f$ is not necessary for the first statement, while for the second one it can be replaced by the assumption that $f$ is strictly quasiconvex. It is important to note that for the problem of simultaneous estimation of the components of $\gamma \in \mathbb{R}^k$ (when $k > 1$), it is generally difficult to obtain a reasonable error function by combining in some way the estimation errors of the single components. In fact, the scale in which the estimation error is expressed is irrelevant (in the sense that the error functions $L$ and $\alpha L$ are considered equivalent, for all $\alpha \in \mathbb{P}$), but the relative scale of the estimation errors of the single components can be very important. That is, either $\gamma \in \mathbb{R}^k$ is a genuine vector (in the sense that we are really interested in $\gamma$, and not only in its components with respect to the standard basis of $\mathbb{R}^k$), in which case reasonable error functions are usually of the form considered in Theorem 5.1, or else it is probably better to estimate the $k$ components of $\gamma$ separately; therefore, the case with $k = 1$ is by far the most important one.

5.1.1 Asymptotic Efficiency

Let the loss function $L'$ on $\mathcal{P} \times \mathcal{D}$ describe the statistical decision problem of estimating $g(P)$ with an error function $L$ on $\mathbb{R}^k \times \mathbb{R}^k$, where $k$ is a positive integer, $g : \mathcal{P} \to \mathbb{R}^k$ is a mapping, and $\mathcal{D}$ is a subset of $\mathbb{R}^k$ containing the image of $\mathcal{P}$ under $g$. In Subsection 4.2.2 we have seen that under suitable regularity conditions, all sequences of decisions obtained by applying a consistent likelihood-based decision criterion to $L'$ (with respect to the sequence of relative plausibility measures on $\mathcal{P}$ obtained by observing the realizations of a sequence of random objects) are asymptotically optimal.
under all models in $\mathcal{P}$. This implies that the corresponding sequences of estimators of $g(P)$ are (strongly) consistent, if the error function $L$ has the following property: $\inf_{B_{\delta}} f > 0$ for all $\varepsilon \in \mathcal{P}$. The condition $f^{-1}\{0\} = \{0\}$ is necessary for this property to hold; in the case with $k = 1$ it is also sufficient, while in the case with $k > 1$ it is sufficient when $f$ is continuous.

Under suitable regularity conditions, the method of maximum likelihood leads to an asymptotically efficient sequence of estimators; and under very similar regularity conditions, the likelihood function tends to be equivalent to the density of a normal distribution with as mean the maximum likelihood estimate (see for example Fraser and McDunnough, 1984). In this case, the next theorem can be used to show that the estimates obtained by applying a certain kind of likelihood-based decision criterion (such as the MPL criterion) tend to the maximum likelihood estimates sufficiently fast to ensure that the corresponding sequences of estimators are asymptotically efficient as well.

**Theorem 5.2.** Consider the problem of estimating $\gamma \in \mathbb{R}^k$ with an error function $L$ such that $w$ is continuous at $x' \in \mathbb{R}^k$, and $f$ satisfies

$$\limsup_{c \to 0} x \in B_\delta \left| c^{-\alpha} f(c x) - |M x|^\alpha \right| = 0 \quad \text{for all } \delta \in \mathcal{P},$$

where $M \in \mathbb{R}^{k \times k}$ is an invertible matrix, and $\alpha \in \mathcal{P}$. Let $r_1, r_2, \ldots$ be a sequence of nondegenerate relative plausibility measures on $\mathbb{R}^k$ such that

$$\limsup_{n \to \infty} x \in B_\delta \left| r_{n+1}(x_n + \Sigma_n x) - \exp \left( -\frac{|x|^2}{2} \right) \right| = 0 \quad \text{for all } \delta \in \mathcal{P},$$

where the limit of the sequence $x_1, x_2, \ldots \in \mathbb{R}^k$ is $x'$, and $\Sigma_1, \Sigma_2, \ldots \in \mathbb{R}^{k \times k}$ is a sequence of matrices such that $\lim_{n \to \infty} \sqrt{n} \Sigma_n$ exists and is invertible. Consider a likelihood-based decision criterion corresponding to the functionals $V_{r_\psi} : l \mapsto \int l d(f_1 \circ r_\psi)$ for a nondecreasing surjection $f_1 : [0, 1] \to [0, 1]$ and a regular integral (on the class of all monotone measures) such that the corresponding functional $F$ on $\mathcal{N}I_{\mathcal{P}}$ satisfies

$$\psi > \varphi \text{ on } \mathcal{P} \Rightarrow \liminf_{c \to \infty} [F(\psi I_{[0,c]}) - F(\varphi I_{[0,c]})] > 0 \quad \text{for all } \psi \in \mathcal{N}I_{\mathcal{P}},$$

where $\varphi \in \mathcal{N}I_{\mathcal{P}}$ is defined by $\varphi(x) = f_1 \left[ \exp \left( -x^2/2 \right) \right]$ (for all $x \in \overline{\mathcal{P}}$). If

$$\lim_{\delta \to 0} \limsup_{n \to \infty} n^{2/3} \left\| V_{r_{n+1}}(x_n) - V_{r_{n-1}}(x_n + \Sigma_n x : x \in B_\delta) \right\| = 0,$$  \hspace{1cm} (5.1)
then all sequences \(d_1, d_2, \ldots \in \mathbb{R}^k\) satisfy
\[
\lim_{n \to \infty} n^{\frac{\alpha}{2}} \left\| V_{r \sigma_n}(l_{x_n}) - \inf_{d \in \mathbb{R}^k} V_{r \sigma_n}(l_d) \right\| = 0 \quad \Rightarrow \quad \lim_{n \to \infty} \sqrt{n} (d_n - x_n) = 0.
\]

Proof. Let \(\Sigma = \lim_{n \to \infty} \sqrt{n} \Sigma_n\); since \(\Sigma\) is invertible, for sufficiently large \(n\) the matrices \(\Sigma_n\) are invertible too. Conditions (5.1) and (4.6) imply that for all \(\varepsilon \in \mathbb{P}\) there is a \(\delta_\varepsilon \in \mathbb{P}\) such that for all \(\delta \in [\delta_\varepsilon, \infty)\) and sufficiently large \(n\) we have
\[
n^{\frac{\alpha}{2}} V_{r \sigma_n}(l_{x_n}) \leq n^{\frac{\alpha}{2}} V_{r \sigma_n \circ (\Sigma_n^{-1} \circ t_{x_n})^{-1}} [(w \circ t_{x_n}^{-1} \circ \Sigma_n) (f \circ \Sigma_n) I_{B_\delta}] + \varepsilon.
\]
For each \(\gamma \in \mathbb{R}^k\) and each \(\delta \in \mathbb{P}\), let \(f'_\gamma\) and \(\mu_\delta\) be respectively the functions and the completely maxitive measures on \(\mathbb{R}^k\) defined by
\[
f'_\gamma(x) = |M \Sigma (x - \gamma)|^\alpha \quad \text{and} \quad \mu_\delta(x) = \begin{cases} f_1 \left[ \exp \left( -\frac{|x|^2}{2} \right) \right] & \text{if } x \in B_\delta, \\ 1 & \text{if } x \in B_\delta^c. \end{cases}
\]
As \(n\) tends to infinity, the completely maxitive measures on \(\mathbb{R}^k\) with density functions \((f_1 \circ r_{p_n}^{-1} \circ t_{x_n}^{-1} \circ \Sigma_n) I_{B_\delta} + I_{B_\delta^c}\) converge uniformly to \(\mu_\delta\), and the functions \((w \circ t_{x_n}^{-1} \circ \Sigma_n) n^{\frac{\alpha}{2}} (f \circ \Sigma_n) I_{B_\delta}\) converge uniformly to \(w(x') f_0 I_{B_\delta}\). Hence, using Theorem 2.23 we obtain that for all \(\delta \in [\delta_\varepsilon, \infty)\) and sufficiently large \(n\)
\[
n^{\frac{\alpha}{2}} V_{r \sigma_n}(l_{x_n}) \leq w(x') \int f_0 I_{B_\delta} d\mu_\delta + 2 \varepsilon,
\]

since the integral is regular and condition (2.6) is fulfilled. Let \(a, b \in \mathbb{P}\) be respectively the infimum and the supremum of \(|M \Sigma x| : x \in \mathbb{R}^k, |x| = 1\), and let \(c = w(x')(b \delta_1)^\alpha + 2\); we have \(n^{\frac{\alpha}{2}} V_{r \sigma_n}(l_{x_n}) \leq c\) for sufficiently large \(n\). Since for all \(d \in \mathbb{R}^k\) and all \(n\) we have
\[
n^{\frac{\alpha}{2}} V_{r \sigma_n}(l_d) = n^{\frac{\alpha}{2}} \int l_d \, d(f_1 \circ r_{p_n}) \geq f_1 [r_{p_n}^{-1}(x_n)] w(x_n) n^{\frac{\alpha}{2}} f(x_n - d),
\]
for a sufficiently large \(\delta' \in \mathbb{P}\) and sufficiently large \(n\) we obtain
\[
\inf_{d \notin \{x_n + \Sigma_n x : x \in B_{\delta'}\}} n^{\frac{\alpha}{2}} V_{r \sigma_n}(l_d) \geq w(x') (a \delta')^\alpha - 1 \geq c + 1,
\]
and thus \(d_n \in \{x_n + \Sigma_n x : x \in B_{\delta'}\}\) for sufficiently large \(n\). Assume that the conclusion of the theorem does not hold; then there is an \(\varepsilon' \in \mathbb{P}\) such
that \( d_n \not\in \{x_n + \Sigma_n x : x \in B_{\varepsilon'}\} \) for infinitely many \( n \), and thus there is a subsequence \( d_{n_1}, d_{n_2}, \ldots \) such that \( \lim_{m \to \infty} \Sigma_n^{-1}(d_{n_m} - x_{n_m}) = \gamma' \in \mathbb{R}^k \), with \( \varepsilon' \leq |\gamma'| \leq \delta' \). Conditions (4.1) and (4.6) imply that for all \( \delta \in \mathbb{P} \) and sufficiently large \( n \) we have

\[
V_{r\pi_n}(l_{d_n}) \geq V_{r\pi_n \circ (\Sigma_n^{-1} \circ t_{x_n})^{-1}} \left( (\varphi \circ t^{-1}_{x_n} \circ \Sigma_n)(f \circ \Sigma_n \circ t_{\Sigma_n^{-1}(d_n - x_{n})}) I_{B_{\delta}} \right),
\]

and as above, using Theorem 2.23 we obtain that for all \( \delta \in \mathbb{P} \) and sufficiently large \( m \)

\[
n_{\delta} V_{r\pi_m}(l_{d_{n,m}}) \geq w(x') \int f'_{\gamma'} I_{B_{\delta}} \ d\mu_{\delta} - \varepsilon.
\]

Now, for all \( \delta \in \mathbb{P} \) we have

\[
\int f'_{0} I_{B_{\delta}} \ d\mu_{\delta} = (\sqrt{2}b)^{\alpha} F\left( \psi I_{\left[0, \left(\frac{\delta}{\sqrt{2}}\right)^{\alpha}\right]} \right),
\]

since \( F \) is bihomogeneous (Theorem 2.13) and

\[
\mu_{\delta}\{f'_{0} I_{B_{\delta}} \geq x\} = \begin{cases} f_{1}\left[\exp\left(-\frac{x^{2}}{2(b\delta)^{2}}\right)\right] & \text{if } 0 \leq x \leq (b\delta)^{\alpha}, \\ 0 & \text{if } (b\delta)^{\alpha} < x \leq \infty. \end{cases}
\]

Analogously, we obtain that there is a \( \psi \in \mathcal{N}I_{\mathbb{P}} \) such that \( \psi > \varphi \) on \( \mathbb{P} \), and for all \( \delta \in \mathbb{P} \)

\[
\int f'_{\gamma'} I_{B_{\delta}} \ d\mu_{\delta} \geq (\sqrt{2}b)^{\alpha} F\left( \psi I_{\left[0, \left(\frac{\delta}{\sqrt{2}}\right)^{\alpha}\right]} \right).
\]

Hence, there are \( \varepsilon, \delta'' \in \mathbb{P} \) such that for all \( \delta \in [\delta'', \infty) \) we have

\[
w(x') \int f'_{\gamma'} I_{B_{\delta}} \ d\mu_{\delta} - w(x') \int f'_{0} I_{B_{\delta}} \ d\mu_{\delta} \geq 4\varepsilon,
\]

and thus with \( \delta = \max\{\delta_{\varepsilon}, \delta''\} \) we obtain that for infinitely many \( n \)

\[
n_{\delta} \left[ V_{r\pi_n}(l_{d_n}) - V_{r\pi_n}(l_{x_n}) \right] \geq \varepsilon > 0,
\]

in contradiction with the assumption about the sequence \( d_1, d_2, \ldots \) \( \square \)

Theorem 5.2 can be generalized to a wider class of likelihood-based decision criteria: in particular, it holds also for the LRM\(_{\beta}\) and MLD criteria. In Theorem 5.2, the assumptions about \( L \) and about the sequence
$rp_1, rp_2, \ldots$ ensure that locally near $x_n$ the estimation problem tends to a simple form; and the optimality of $x_n$ in this simple problem is ensured by the assumptions about $f_1$ and $F$. The condition on $F$ is rather weak, when $f_1$ is not too extreme; in particular, it is satisfied for both the Shilkret integral and the Choquet integral, if there are $a, b \in \mathbb{R}$ such that $f_1(x) \leq ax^b$ for all $x \in [0,1]$. Condition (5.1) implies that it suffices to consider the estimation problem locally near $x_n$; the validity of this condition depends on the particular problem at hand. For example, if $f_1$ satisfies the above requirement, $w$ is bounded, $f : x \mapsto |Mx|, a$, and $rp_n^+$ is logarithmically concave for sufficiently large $n$, then condition (5.1) is fulfilled for both the Shilkret integral and the Choquet integral.

In Theorem 5.2 it is assumed that locally near 0 the function $f$ tends to be centrally symmetric about 0; the next example shows that when this condition is not fulfilled, the conclusion of the theorem does not necessarily hold, even when all other assumptions are satisfied. In fact, all asymptotically efficient sequences of estimators of $g(P)$ are asymptotically normal with mean $g(P)$ (under each model $P \in \mathcal{P}$), while a biased asymptotic distribution can be advantageous when the assumption about $f$ of Theorem 5.2 does not hold.

Example 5.3. In the situation of Examples 4.18 and 4.25, consider the problem of estimating $\mu \in \mathbb{R}$ with an error function $L$ such that $w$ is constant and $f$ is defined by

$$f(x) = \begin{cases} -2x & \text{if } -\infty < x \leq 0, \\ x & \text{if } 0 \leq x < \infty. \end{cases}$$

This estimation error penalizes the overestimation of $\mu$ more than its underestimation; and in fact, if the estimate obtained by applying a particular likelihood-based decision criterion is well-defined, then there is a nonnegative real number $c$ such that for each $n$ this estimate is $\bar{X} - \frac{c}{\sqrt{n}}$ (where as usual $\bar{X}$ denotes the arithmetic mean of the observations $X_1, \ldots, X_n$). The corresponding sequence of estimators of $\mu$ is asymptotically efficient if and only if $c = 0$; that is, if and only if the resulting estimates are the maximum likelihood estimates $\bar{X}$. For the MPL criterion $c \approx 0.345$, and thus the corresponding sequence of estimators is not asymptotically efficient; but the expected loss is 0.915 times the expected loss for the maximum likelihood estimates (independently of $n$ and $\mu$).
5.2 Shilkret Integral and Quasiconvexity

Consider the problem of estimating \( \gamma \in \mathbb{R}^k \) with an error function \( L \), on the basis of a nondegenerate relative plausibility measure \( \rho \) on \( \mathbb{R}^k \). Let \( S \) be the functional on \( \mathbb{F}_S \) corresponding to a likelihood-based decision criterion revealing ignorance aversion. Consider first the case with \( k = 1 \), and for each \( d \in \mathbb{R} \) define \( \varphi_d = \rho \{ l_d I_{(0,\infty)} \geq \cdot \} \) and \( \psi_d = \rho \{ l_d I_{(-\infty,0)} \geq \cdot \} \). Theorem 4.3 implies that \( S \) is monotonic, and \( V_{\rho \rho}(l_d) = S(\max \{ \varphi_d, \psi_d \}) \) for all \( d \in \mathbb{R} \), where \( V_{\rho \rho} \) is the functional on \( \mathbb{F}_S \) corresponding to the likelihood-based decision criterion considered. If \( d, d' \in \mathbb{R} \) with \( d \leq d' \), then \( \varphi_{d'} \leq \varphi_d \) and \( \psi_{d'} \geq \psi_d \); that is, if we raise the estimate \( d \), then the values of \( \varphi_d \) decrease and those of \( \psi_d \) increase, while if we reduce the estimate \( d \), then the values of \( \varphi_d \) increase and those of \( \psi_d \) decrease; an optimal estimate corresponds thus to some sort of equilibrium point. In the case with \( k > 1 \), the situation is more complicated, because there are infinitely many directions in which the estimate \( d \) can be moved; but when \( L \) is of the form considered in Theorem 5.1, for each of these directions we can obtain two functions with the properties of \( \varphi_d \) and \( \psi_d \) by dividing \( \mathbb{R}^k \) into two half-spaces. Theorems 5.1 and 5.2 exploit two special cases: in the first one, for a particular direction we have \( \varphi_{d'} \leq \varphi_d \) and \( \psi_{d'} = \psi_d = 0 \), and thus \( d \) cannot be preferred to \( d' \); in the second one, for each direction we have \( \varphi_d = \psi_d \), and thus \( d \) is optimal. The general case can be simplified when the likelihood-based decision criterion corresponds to the functional \( V_{\rho \rho} : l \mapsto \int S l d(f_1 \circ \rho) \) for a nondecreasing function \( f_1 : [0, 1] \to [0, 1] \) such that \( f_1(0) = 0 \) and \( f_1(1) = 1 \); in fact, the maximality of the Shilkret integral implies that \( V_{\rho \rho}(l_d) = \max \{ S(\varphi_d), S(\psi_d) \} \). In the case with \( k = 1 \), the function \( d \mapsto V_{\rho \rho}(l_d) \) on \( \mathbb{R} \) is then quasiconvex (since it is the maximum of a nonincreasing function and a nondecreasing function); the next theorem generalizes this result.

**Theorem 5.4.** Consider the problem of estimating \( \gamma \in \mathbb{R}^k \) with an error function \( L \), on the basis of a nondegenerate relative plausibility measure \( \rho \) on \( \mathbb{R}^k \). Let \( f_1 : [0, 1] \to [0, 1] \) be a nondecreasing function such that \( f_1(0) = 0 \) and \( f_1(1) = 1 \), and let \( V_{\rho \rho} \) be the functional \( l \mapsto \int S l d(f_1 \circ \rho) \) on \( \mathbb{F}_S \). The function \( d \mapsto V_{\rho \rho}(l_d) \) on \( \mathbb{R}^k \) is quasiconvex.

Moreover, if \( \{ \rho \rho^l > 0 \} \) is bounded, and \( L \) is such that \( f \) is continuous and strictly quasiconvex, \( \inf_{\{ \rho \rho^l > 0 \}} w > 0 \), and \( \sup_{\{ \rho \rho^l > 0 \}} w < \infty \), then the function \( d \mapsto V_{\rho \rho}(l_d) \) on \( \mathbb{R}^k \) is continuous and strictly quasiconvex.
Proof. For each $\gamma \in \mathbb{R}^k$ the function $d \mapsto l_d(\gamma)$ is quasiconvex, and

\[
V_{rp}(l_d) = \int_S l_d \, d(f_1 \circ rp) = \sup_{x \in P} x \, f_1(r_p\{l_d \geq x\}) = \sup_{x \in P} \sup_{\gamma_1, \gamma_2, \ldots \in \mathbb{R}^k} x \, f_1\left[\liminf_{n \to \infty} rp^\perp(\gamma_n)\right] = \sup_{\gamma_1, \gamma_2, \ldots \in \mathbb{R}^k} \left[\limsup_{n \to \infty} l_d(\gamma_n)\right] \, f_1\left[\liminf_{n \to \infty} rp^\perp(\gamma_n)\right].
\]

Hence the function $d \mapsto V_{rp}(l_d)$ on $\mathbb{R}^k$ is quasiconvex, since the limit superior of a sequence of quasiconvex functions is quasiconvex, and so is the supremum of a set of quasiconvex functions.

To prove the second statement of the theorem, let $d'$ and $d''$ be two distinct elements of $\mathbb{R}^k$, and let $d = \lambda d' + (1 - \lambda) d''$ for some $\lambda \in (0, 1)$. The assumptions about $rp$ and $L$ imply that there are $a, b \in \mathbb{P}$ such that on $\{rp^\perp > 0\}$ we have $\max\{l_{d'}, l_{d''}\} \geq l_d + a$ and $l_d \leq b$; hence $V_{rp}(l_d) \leq b$, and $\max\{l_{d'}, l_{d''}\} \geq \max\{(1 + \frac{a}{b}) l_d, a\}$ on $\{rp^\perp > 0\}$. Since the Shilkret integral is monotonic, homogeneous, maxitive, and transformation invariant on the class of all finitely maxitive measures, and $f_1 \circ rp$ is finitely maxitive, using Theorem 2.8 we obtain

\[
\max\{V_{rp}(l_{d'}), V_{rp}(l_{d''})\} = \int_S \max\{l_{d'}, l_{d''}\} \, d(f_1 \circ rp) \geq \int_S \max\{(1 + \frac{a}{b}) l_d, a\} \, d(f_1 \circ rp) = \max\{(1 + \frac{a}{b}) V_{rp}(l_d), a\} > V_{rp}(l_d);
\]

and this proves the strict quasiconvexity of the function $d \mapsto V_{rp}(l_d)$ on $\mathbb{R}^k$. Since the Shilkret integral is also subadditive on the class of all finitely maxitive measures, using Theorem 2.8 again we obtain

\[
|V_{rp}(l_d) - V_{rp}(l_{d'})| \leq \sup_{\{rp^\perp > 0\}} |l_d - l_{d'}| \quad \text{for all } d, d' \in \mathbb{R}^k.
\]

Hence, the continuity at $d' \in \mathbb{R}^k$ of the function $d \mapsto V_{rp}(l_d)$ on $\mathbb{R}^k$ follows from the equicontinuity at $d'$ of the set of all functions $d \mapsto l_d(\gamma)$ on $\mathbb{R}^k$ with $\gamma \in \{rp^\perp > 0\}$; and the equicontinuity at each $d' \in \mathbb{R}^k$ of this set of functions is implied by the assumptions about $rp$ and $L$. 

\[\square\]
Consider a likelihood-based decision criterion corresponding to the functional $V_{rp} : l \mapsto \int S d(f_1 \circ rp)$, where $f_1 : [0, 1] \rightarrow [0, 1]$ is a nondecreasing function such that $f_1(0) = 0$ and $f_1(1) = 1$ (for example the MPL, LRM$_\beta$, or MLD criteria). The decision criterion applied to the estimation problem described by $L$ consists in minimizing $V_{rp}(l_d)$, and Theorem 5.4 states that the function $d \mapsto V_{rp}(l_d)$ on $\mathbb{R}^k$ is quasiconvex. Therefore, in particular, the set of all optimal estimates in $\mathbb{R}^k$ is convex, and if the function $d \mapsto V_{rp}(l_d)$ has a strict local minimum at $d' \in \mathbb{R}^k$, then $d'$ is the unique optimal estimate in $\mathbb{R}^k$. If the premise of the second statement of the theorem is satisfied, then the function $d \mapsto V_{rp}(l_d)$ on $\mathbb{R}^k$ is also continuous and strictly quasiconvex, and so there is at most one optimal estimate in $\mathbb{R}^k$. It can be easily proved that if the premise of the second statement of Theorem 5.4 is satisfied, and either $k = 1$, or $L$ is of the form considered in Theorem 5.1, then there is a unique optimal estimate in $\mathbb{R}^k$, and it lies in the closed convex hull of $\{rp^{\perp} > 0\}$.

Consider now the case when $f_1$ is left-continuous; then $f_1 \circ rp$ is completely maxitive, and using Theorem 2.6 we obtain

$$V_{rp}(l_d) = \sup_{\gamma \in \mathbb{R}^k} f(\gamma - d) w(\gamma) f_1[rp^{\perp}](\gamma) = \int S (f \circ t_d) d\mu \quad \text{for all } d \in \mathbb{R}^k,$$

where $\mu$ is the completely maxitive measure on $\mathbb{R}^k$ with density function $\mu^{\perp} = w(f_1 \circ rp^{\perp})$. When $\mu$ is finite, applying the considered likelihood-based decision criterion to $L$ with respect to $rp$ corresponds thus to applying the MPL criterion to the error function $L' : (\gamma, d) \mapsto f(\gamma - d)$ with respect to the (nondegenerate) relative plausibility measure on $\mathbb{R}^k$ proportional to $\mu$. In particular, the assumptions about $w$ of Theorem 5.4 can be replaced by the assumptions that $f_1$ is left-continuous and $w(f_1 \circ rp^{\perp})$ is bounded. An algorithm for approximating the optimal estimates of $\gamma \in \mathbb{R}^k$ (that is, for minimizing the function $d \mapsto V_{rp}(l_d)$ on $\mathbb{R}^k$) is suggested by the following two expressions:

$$V_{rp}(l_d) = \sup_{\mu^{\perp} > 0} (f \circ t_d) \mu^{\perp} \quad \text{for all } d \in \mathbb{R}^k,$$

$$\{d \in \mathbb{R}^k : V_{rp}(l_d) \leq c\} = \bigcap_{\gamma \in \{\mu^{\perp} > 0\}} \left\{ f \circ (-t_\gamma) \leq \frac{c}{\mu^{\perp}(\gamma)} \right\} \quad \text{for all } c \in \mathbb{P}.$$

Let $d \in \mathbb{R}^k$ be an initial estimate, and let $\gamma_1, \ldots, \gamma_n \in \{\mu^{\perp} > 0\}$ be suitable approximations of the local maxima of $(f \circ t_d) \mu^{\perp}$. Then the value

$$c = \max\{f(\gamma_1 - d) \mu^{\perp}(\gamma_1), \ldots, f(\gamma_n - d) \mu^{\perp}(\gamma_n)\}$$

approximates $V_{rp}(l_d)$, and if actually \( \inf_{d' \in \mathbb{R}^k} V_{rp}(l_{d'}) \leq c \), then the optimal estimates lie in the set

\[
A = \{ d' \in \mathbb{R}^k : f(\gamma_1 - d') \leq \frac{c}{\mu_i(\gamma_1)}, \ldots, f(\gamma_n - d') \leq \frac{c}{\mu_i(\gamma_n)} \}.
\]

The set \( A \) is convex (since \( f \) is quasiconvex), and the estimate \( d \) usually lies on the border of \( A \); other estimates \( d' \in A \) lying in the interior of \( A \) can be expected to be better than \( d \), in the sense that \( V_{rp}(l_{d'}) < V_{rp}(l_d) \). The rule for selecting a new estimate \( d' \in A \) can exploit the properties of the particular problem at hand, such as the differentiability of \( f \) or \( \mu_i \). Once the estimate \( d' \in A \) has been selected, a new step of the algorithm can be carried out by using \( d' \) as the initial estimate; and \( \gamma_1, \ldots, \gamma_n \) can be used as starting points for approximating the local maxima of \( (f \circ t_{d'}) \mu_i \). That is, the algorithm can be interpreted as searching for the local maxima of a function that depends on the obtained approximations of the local maxima: some steps of the maximization algorithms are alternated with some steps of an algorithm that redefines the function to be maximized (by selecting a new estimate \( d' \)). The algorithm is particularly simple and effective in the case with \( k = 1 \), but it has proved successful also in multidimensional problems with \( k \) up to 10; however, the speed of convergence depends strongly on which properties of the estimation problem can be assumed and exploited when implementing the algorithm, and in particular when defining the rule for selecting the new estimate \( d' \in A \).

### 5.2.1 Impossible Estimates

Consider the problem of estimating \( \gamma \in \mathbb{R}^k \) with an error function \( L \), on the basis of a nondegenerate relative plausibility measure \( rp \) on \( \mathbb{R}^k \). The elements of \( \{ rp^\perp = 0 \} \) can be considered as impossible, since they are infinitely less plausible than any element of \( \{ rp^\perp > 0 \} \). If \( P \) is a set of statistical models, \( g : P \rightarrow \mathbb{R}^k \) is a mapping, and \( rp^\perp \) is proportional to the profile likelihood function \( lik_g \) induced by the observation of some data, then \( rp^\perp \) vanishes at \( \gamma \in \mathbb{R}^k \) if and only if either \( \gamma \) is physically impossible (in the sense that it does not lie in the image of \( P \) under \( g \)), or it is impossible in the light of the observed data (in the sense that the likelihood function on \( P \) induced by the observed data vanishes on \( \{ g = \gamma \} \)). Impossible optimal estimates can be problematic, and even a unique optimal estimate lying on the border of \( \{ rp^\perp > 0 \} \) can be disturbing.
Consider a likelihood-based decision criterion corresponding to the functional \( V_{rp} : l \mapsto \int \mathcal{S} l \, d(f_1 \circ \tau_p) \), where \( f_1 : [0, 1] \to [0, 1] \) is a non-decreasing function such that \( f_1(0) = 0 \) and \( f_1(1) = 1 \). We have seen that under some conditions, if either \( k = 1 \), or \( L \) is of the form considered in Theorem 5.1, then there is a unique optimal estimate \( \hat{\gamma} \in \mathbb{R}^k \), and \( \hat{\gamma} \) lies in the closed convex hull of \( \{ rp^{+} > 0 \} \). Of course, \( \hat{\gamma} \) can be impossible (in the sense that \( rp^{+}(\hat{\gamma}) = 0 \)) when \( \{ rp^{+} > 0 \} \) is not convex; for instance, if \( k = 2 \), the error function is \( L : (\gamma, d) \mapsto |\gamma - d| \), and \( rp^{+} \) is the indicator function of the unit circle \( C = \{ x \in \mathbb{R}^2 : |x| = 1 \} \), then \( \hat{\gamma} = 0 \) is the unique optimal estimate according to all likelihood-based decision criteria revealing ignorance aversion. If we try to minimize the function \( d \mapsto V_{rp}(l_d) \) on \( D = \{ rp^{+} > 0 \} \), then in general the minimum does not exist when \( D \) is not closed. In fact, \( \hat{\gamma} \) can be impossible also when \( \{ rp^{+} > 0 \} \) is convex; for example, with \( k \) and \( L \) as above, if \( rp^{+} \) is the indicator function of an open unit half-disc \( H = \{ x \in \mathbb{R}^2 : |x| < 1, \ x \cdot y > 0 \} \), where \( y \in \mathbb{R}^2 \setminus \{0\} \), then \( \hat{\gamma} = 0 \) is the unique optimal estimate according to all likelihood-based decision criteria revealing ignorance aversion. The next theorem implies that under some conditions \( \hat{\gamma} \) cannot be an extreme point of the closed convex hull of \( \{ rp^{+} > 0 \} \); the above example does not contradict this theorem, since \( 0 \) is not an extreme point of the closure of \( H \) (the extreme points of the closure of \( H \) are the \( x \in C \) such that \( x \cdot y \geq 0 \)). The next theorem is particularly interesting in the case with \( k = 1 \), because then the closed convex hull of \( \{ rp^{+} > 0 \} \) is an interval \( I \subset \mathbb{R} \), and if \( \hat{\gamma} \) is not an extreme point of \( I \), then it lies in the interior of \( I \), and thus also in the interior of \( \{ rp^{+} > 0 \} \), when this is convex.

**Theorem 5.5.** Consider the problem of estimating \( \gamma \in \mathbb{R}^k \) on the basis of a nondegenerate relative plausibility measure \( rp \) on \( \mathbb{R}^k \), with an error function \( L \) such that \( \sup_{\{ rp^{+} > 0 \}} w < \infty \) and \( f : x \mapsto f'(|Mx|) \), where \( f' : \mathbb{P} \to \mathbb{P} \) is a function continuous at 0, and \( M \in \mathbb{R}^{k \times k} \) is an invertible matrix. Let \( f_1 : [0, 1] \to [0, 1] \) be a nondecreasing function such that \( f_1(0) = 0 \) and \( f_1(1) = 1 \), and consider a likelihood-based decision criterion corresponding to the functional \( V_{rp} : l \mapsto \int \mathcal{S} l \, d(f_1 \circ \tau_p) \). If \( \hat{\gamma} \in \mathbb{R}^k \) is the unique optimal estimate, and \( V_{rp}(l_{\hat{\gamma}}) > 0 \), then \( \hat{\gamma} \) is not an extreme point of the closed convex hull of \( \{ rp^{+} > 0 \} \).

**Proof.** Let \( G \) be the closed convex hull of \( \{ rp^{+} > 0 \} \), and assume that \( \hat{\gamma} \) is an extreme point of \( G \). Then \( M\hat{\gamma} \) is an extreme point of the closed, convex set \( G' = \{ M\gamma : \gamma \in G \} \), and thus for each \( \varepsilon \in \mathbb{P} \) there is an open half-space...
162  5 Point Estimation

$H_\varepsilon \subset \mathbb{R}^k$ such that $M\hat{\gamma} \in H_\varepsilon$ and $G' \cap H_\varepsilon \subset t_{M\hat{\gamma}}^{-1}(B_\varepsilon)$. For each $\varepsilon \in \mathbb{P}$ let $\gamma_\varepsilon$ be the element of $\mathbb{R}^k$ such that $M\gamma_\varepsilon$ is the projection of $M\hat{\gamma}$ on the border of $H_\varepsilon$: we have $|M(\gamma - \gamma_\varepsilon)| < |M(\gamma - \hat{\gamma})|$ for all $\gamma \in M^{-1}(\mathbb{R}^k \setminus H_\varepsilon)$. Since $G'$ is not a singleton (because otherwise $V_{rp}(l_{\hat{\gamma}}) = 0$), for sufficiently small $\varepsilon \in \mathbb{P}$ we have $|M(\gamma_\varepsilon - \hat{\gamma})| \leq \varepsilon$; and the assumptions about $L$ thus imply that there is an $\varepsilon' \in \mathbb{P}$ such that $l_{\gamma_\varepsilon} \leq V_{rp}(l_{\hat{\gamma}})$ on $G \cap M^{-1}(H_\varepsilon')$. Since the Shilkret integral is monotone, homogeneous, maxitive, and transformation invariant on the class of all finitely maxitive measures, and $f_1 \circ rp$ is finitely maxitive, using Theorem 2.8 we obtain

$$V_{rp}(l_{\gamma_\varepsilon}) = \int S \max\{l_{\gamma_\varepsilon} \cap M^{-1}(H_\varepsilon'), l_{\gamma_\varepsilon} \cap M^{-1}(\mathbb{R}^k \setminus H_\varepsilon')\} d(f_1 \circ rp) \leq \int S \max\{V_{rp}(l_{\hat{\gamma}}), l_{\hat{\gamma}}\} d(f_1 \circ rp) = V_{rp}(l_{\hat{\gamma}}),$$

in contradiction with the assumption that $\hat{\gamma}$ is uniquely optimal.  

Theorem 5.5 can be easily generalized; in particular, in the case with $k = 1$, the assumption about $f$ can be replaced by the assumption that $f$ is continuous at 0. Moreover, as in Theorem 5.4, the assumption about $w$ of Theorem 5.5 can be replaced by the assumptions that $f_1$ is left-continuous and $w(f_1 \circ rp^\perp)$ is bounded. When $f^{-1}\{0\} = \{0\}$, a sufficient condition for $V_{rp}(l_{\hat{\gamma}})$ to be positive is that $f_1 \circ rp^\perp > 0$ is not a singleton; in particular, if (as in the following example) $rp^\perp$ is continuous on $\{rp^\perp > 0\}$, and $\{rp^\perp > 0\}$ is convex and is not a singleton, then $V_{rp}(l_{\hat{\gamma}}) = 0$ is possible only if $f_1 = I_{\{1\}}$ (that is, only for the MLD criterion).

**Example 5.6.** The problem of estimating (with squared error loss) the mean of a normal distribution with known variance is made much more complex by the additional assumption that the mean lies in a particular bounded interval. Casella and Strawderman (1981) proved that if $\mathcal{P'} = \{P_\mu : \mu \in (-m, m)\}$ is a family of probability measures such that under each model $P_\mu$ the random variable $X$ is normally distributed with mean $\mu$ and variance 1, and $m \in (0, 1.056)$, then the minimax risk criterion applied to the decision problem described by the loss function $L' : (P_\mu, d) \mapsto (\mu - d)^2$ on $\mathcal{P'} \times \mathbb{R}$ leads to the estimate $m \tanh(m X)$. For larger values of $m$, the estimate obtained by applying the minimax risk criterion is in general unknown, but Donoho, Liu, and MacGibbon (1990) showed that the maximum expected loss for this estimate is at least 0.8
times the one for the estimate $\frac{m^2}{m^2 + 1} X$. When $|X| \geq m + \frac{1}{m}$, this estimate is impossible; an estimate that is never impossible and has smaller expected loss (under all models in $\mathcal{P}'$) can be obtained by using the results of Moors (1981):

$$\text{sign}(X) \min \left\{ \frac{m^2}{m^2 + 1} |X|, m \tanh(m |X|) \right\}. \quad (5.2)$$

The maximum expected loss for this estimate is thus at most 1.25 times the one for the estimate obtained by applying the minimax risk criterion.

For the hierarchical model described in Section 3.2, the assumption that the mean $\mu$ lies in the interval $(-m, m)$ simply corresponds to a particular prior relative plausibility measure on the set $\mathcal{P} = \{P_\mu : \mu \in \mathbb{R}\} \supseteq \mathcal{P}'$ considered in Example 4.18 (with $n = 1$ and $X = X_1$): the prior relative plausibility measure with density function proportional to the function $P_\mu \mapsto I_{(-m, m)}(\mu)$ on $\mathcal{P}$. No probability measure on $\mathcal{P}$ can describe this kind of prior information, but a reasonable estimate can be obtained by applying to the decision problem described by the loss function $L : (P_\mu, d) \mapsto (\mu - d)^2$ on $\mathcal{P} \times \mathbb{R}$ the Bayesian criterion with as initial averaging probability measure on $\mathcal{P}$ the one corresponding to the uniform distribution on $(-m, m)$ for the parameter $\mu$ (see Gatsonis, MacGibbon, and Strawderman, 1987). When the realization $X = x$ is observed, the above prior relative plausibility measure can be updated by combining it (as described in Subsection 3.1.2) with the relative plausibility measure on $\mathcal{P}$ whose density function is proportional to the approximate likelihood function $P_\mu \mapsto \delta \varphi(x - \mu)$ on $\mathcal{P}$ considered in Example 4.18. A likelihood-based decision criterion can then be applied to the decision problem described by $L$: consider the decision criteria corresponding to the functionals $V_{r,p} : l \mapsto \int \|l d(f_1 \circ r_p)\)$ for some nondecreasing function $f_1 : [0, 1] \to [0, 1]$ such that $f_1(0) = 0$ and $f_1(1) = 1$. Theorems 5.1 and 5.4 imply that there is a unique optimal estimate in $\mathbb{R}$, and it lies in the closed interval $[-m, m]$; and Theorem 5.5 implies that if $f_1 \neq I_{(1)}$, then the estimate lies in the open interval $(-m, m)$. That is, the MLD criterion is the only one of the considered likelihood-based decision criteria that can lead to an impossible estimate, and in fact it leads to the estimate $\text{sign}(X) \min \{|X|, m\}$, which is impossible when $|X| \geq m$. When $|X| < m$, this estimate corresponds to the maximum likelihood estimate $X$; more generally, if $f_1$ is not too extreme (it suffices that there are $a, b \in \mathbb{P}$ such that $f_1(x) \leq ax^b$ for all $x \in [0, 1]$), then there is a $c \in \mathbb{P}$ such that the resulting estimate is $X$ when $|X| \leq m - c$ (for all $m \geq c$): for instance, $c = \sqrt{2}$ when $f_1 = id_{[0,1]}$ (that is, for the MPL
This can be considered as a stability property: when the observation \( x \) lies in the interval \((-m, m)\) and is sufficiently far away from its endpoints, the additional assumption that \( \mu \in (-m, m) \) has no influence on the estimate of \( \mu \in \mathbb{R} \).

The first diagram of Figure 5.1 shows the graphs of some estimates of \( \mu \), as functions of the observation \( x \in [0, 6] \) (all the estimators considered are odd functions), in the case with \( m = 3 \): the estimate (5.2) (dotted, almost polygonal line), and the estimates obtained by applying the MPL criterion (dashed line), the MLD criterion (upper solid line), the \( \text{LRM}_\beta \) criterion with \( \beta \approx 0.259 \) (so that \(-2 \log \beta \) is the 90%-quantile of the \( \chi^2 \) distribution with one degree of freedom; lower solid line), and the Bayesian criterion with initial averaging probability distribution uniform on \((-3, 3)\) (dotted curved line). The second diagram of Figure 5.1 shows the graphs of the corresponding expected losses (as functions of \( \mu \in (-3, 3) \)): for the estimate (5.2) (dotted line with maximum 0.809 occurring near \( \mu = 0 \)), and for the estimates obtained by applying the MPL criterion (dashed line with maximum 0.947), the MLD criterion (solid line with maximum 0.995 at \( \mu = 0 \)), the \( \text{LRM}_\beta \) criterion with \( \beta \approx 0.259 \) (solid line with supremum 1.001 occurring at the endpoints of the interval), and the Bayesian criterion with initial averaging probability distribution uniform on \((-3, 3)\) (dotted line with supremum 1.000 occurring at the endpoints of the interval). A consequence of the stability property considered above is that the three estimates obtained by applying a likelihood-based decision criterion have a relatively high expected loss under the model \( P_0 \).\[\diamond\]
Using the previous results, it can be easily proved that when the MPL criterion is applied to the problem of estimating \( \gamma \in \mathbb{R} \) with an error function \( L \) on the basis of a nondegenerate relative plausibility measure \( r_p \) on \( \mathbb{R} \), if \( L \) is such that \( f \) is continuous and strictly quasiconvex, \( w r_p \) is bounded, and

\[
\lim_{x \to -\infty} f(x - d) w(x) r_p(x) = \lim_{x \to -\infty} f(x - d) w(x) r_p(x) = 0 \quad \text{for all } d \in \mathbb{R},
\]

then there is a unique optimal estimate in \( \mathbb{R} \), and it lies in the interior of the convex hull of \( \{ r_p > 0 \} \), when this is not a singleton. Theorem 2.6 implies that applying the MPL criterion to \( L \) with respect to \( r_p \) corresponds to applying it to the error function \( L' : (\gamma, d) \mapsto f(\gamma - d) \) with respect to \( w' \circ r_p \); that is, the weight function \( w \) plays the role of a prior likelihood function on \( \mathbb{R} \), to be combined with the pseudo likelihood function \( r_p \).

More generally, when \( \mathcal{P} \) is a set of statistical models, \( g : \mathcal{P} \to \mathbb{R} \) is a mapping, and we estimate \( \gamma = g(P) \), a weight function \( w' \) can be defined directly on \( \mathcal{P} \): if \( r_p' \) is a relative plausibility measure on \( \mathcal{P} \), then applying the MPL criterion to \( L' \) with respect to \( [(w') - r_p'] \circ g^{-1} \) corresponds to applying it (with respect to \( r_p' \)) to the decision problem described by the loss function \( L'' : (P, d) \mapsto w'(P) f[g(P) - d] \) on \( \mathcal{P} \times \mathbb{R} \). When \( w' \) depends on \( P \) only through \( g(P) \), we recover the above situation, with \( r_p = r_p' \circ g^{-1} \) and \( w' = w \circ g \). The weight function \( w' \) must be chosen carefully: in general, the resulting estimates have better pre-data performances when \( w' \) is chosen so that estimates with bounded expected losses are possible.

**Example 5.7.** Famous examples of impossible estimates are the negative estimates of variance components in random effects models (see for instance Searle, Casella, and MacCulloch, 1992). Consider in particular the balanced one-way layout under normality assumptions; that is, consider a family \( \mathcal{P}' = \{ P_{\mu, v_a, v_e} : (\mu, v_a, v_e) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R} \} \) of probability measures such that under each model \( P_{\mu, v_a, v_e} \) the random variables \( X_{i,j} \) are normally distributed with mean \( \mu \) and variance \( v_a + v_e \) (for all \( i \in \{1, \ldots, n_a\} \) and all \( j \in \{1, \ldots, n_e\} \)), and the correlation coefficient of two different random variables \( X_{i,j} \) and \( X_{i',j'} \) is \( \rho = \frac{v_a}{v_a + v_e} \) when \( i = i' \), and 0 otherwise. Then, under each model \( P_{\mu, v_a, v_e} \) the random variables \( \frac{S_a}{n_a v_a + v_e} \) and \( \frac{S_e}{v_e} \) are \( \chi^2 \) distributed with \( n_a - 1 \) and \( n_a (n_e - 1) \) degrees of freedom, respectively, where

\[
S_a = n_a \sum_{i=1}^{n_a} (\bar{X}_{i,j} - \bar{X}_{i,:})^2 \quad \text{and} \quad S_e = \sum_{i=1}^{n_a} \sum_{j=1}^{n_e} (X_{i,j} - \bar{X}_{i,:})^2;
\]
with \( \bar{X}_{i,j} = \frac{1}{n_e} \sum_{j=1}^{n_e} X_{i,j} \) and \( \bar{X}_{i,.} = \frac{1}{n_a} \sum_{i=1}^{n_a} X_{i,.} \). The estimates of the variance components \( \nu_a \) and \( \nu_e \) obtained by applying the method of analysis of variance are

\[
\hat{\nu}_a = \frac{1}{n_e} \left( \frac{S_a}{n_a - 1} - \frac{S_e}{n_a (n_e - 1)} \right) \quad \text{and} \quad \hat{\nu}_e = \frac{S_e}{n_a (n_e - 1)},
\]

respectively; these estimates are unbiased and have minimum variance among all unbiased estimates, but \( \hat{\nu}_a \) can be negative. The problems of estimating \( \nu_a \) and \( \nu_e \) can be described as statistical decision problems by means of two loss functions \( L_a \) and \( L_e \) on \( \mathcal{P} \times \mathbb{R} \); consider in particular the loss functions

\[
L_a : (\mu, \nu_a, \nu_e, d_a) \mapsto \frac{(\nu_a - d_a)^2}{n_e \nu_a + \nu_e} \quad \text{and} \quad L_e : (\mu, \nu_a, \nu_e, d_e) \mapsto \frac{(\nu_e - d_e)^2}{\nu_e^2}.
\]

The squared error loss is not particularly well suited for scale parameters, but it accords with the estimates \( \hat{\nu}_a \) and \( \hat{\nu}_e \) (which have minimum variance among all unbiased estimates). The weights are chosen so that the decision problems are location and scale invariant, and estimates with bounded expected losses are possible (the naive choice \( \frac{1}{\nu_a^2} \) for \( L_a \) would in fact impose the estimate \( d_a = 0 \)).

Figure 5.2 shows the graphs of some equivariant estimators of \( \nu_a \) and \( \nu_e \), and of the corresponding expected losses, in the case with \( n_a = n_e = 3 \); the other cases are qualitatively similar, although the differences are less marked when \( n_a \) and \( n_e \) are larger. All the estimates considered depend on the observations \( X_{i,j} \) only through \( S_a \) and \( S_e \); the estimates of \( \nu_a \) and \( \nu_e \) divided by \( S_a + S_e \) are plotted (as functions of \( r = \frac{S_a}{S_a + S_e} \in (0,1) \)) in the first and second upper diagrams, respectively, while the corresponding expected losses are plotted (as functions of \( \rho \in (0,1) \)) in the first and second lower diagrams, respectively. The estimates \( \hat{\nu}_a \) and \( \hat{\nu}_e \) obtained by applying the method of analysis of variance (solid straight lines) have relatively large expected losses (upper solid lines); in particular, the expected loss for \( \hat{\nu}_a \) is large when \( \rho \) is small, because \( \hat{\nu}_a \) is negative when \( r \) is small: in fact, the truncated estimate \( \max\{\hat{\nu}_a, 0\} \) has smaller expected loss (upper dashed line). This truncated estimate is biased; when the requirement of unbiasedness is dropped, many generalizations of the estimate \( \hat{\nu}_a \) are possible. For instance, Hartung (1981) proposed an estimate of \( \nu_a \) (dotted straight line) that minimizes the bias among the nonnegative, invariant, quadratic estimates, but the corresponding expected loss (upper dotted line) is even larger than the one for \( \hat{\nu}_a \) when \( \rho \) is small.
Besides \( \hat{\vartheta}_a \) and \( \hat{\vartheta}_e \), the most appreciated estimates of \( \vartheta_a \) and \( \vartheta_e \) seem to be those based on the method of maximum likelihood; consider the density-based approximation of the likelihood function on \( P \) induced by the observations \( X_{i,j} \). The maximum likelihood estimate of \( \vartheta_a \) (dashed line) has a relatively small expected loss (lower dashed line), while for the maximum likelihood estimate of \( \vartheta_e \) (lower dashed line) the expected loss (lower dashed line) is small only when \( \rho \) is small; these are the estimates obtained by applying the MLD criterion to the decision problems described by \( L_a \) and \( L_e \), respectively, without using prior information. The restricted maximum likelihood estimates considered by Thompson (1962) are obtained by applying the method of maximum likelihood to a particular pseudo likelihood function on \( P \): they are appreciated because they are nonnegative and correspond to the estimates obtained by applying the

\[ \text{Figure 5.2. Estimators and corresponding expected losses from Example 5.7.} \]
method of analysis of variance when no one of these estimates is negative. In fact, the restricted maximum likelihood estimate of \( v_a \) is \( \max \{ \hat{v}_a, 0 \} \), and the one of \( v_e \) (upper dashed line) is \( \hat{v}_e \) when \( \hat{v}_a \geq 0 \); but the corresponding expected losses (upper dashed lines) are larger than the ones for the maximum likelihood estimates (lower dashed lines). Anyway, these two estimates of \( v_a \) based on the method of maximum likelihood can be impossible (in the sense that they can be 0); and even when the set \( \mathcal{P} \) is extended to include the models with \( v_a = 0 \), it can be disturbing to have estimates lying on the border of the set of possible values for \( v_a \).

To obtain positive estimates, the likelihood function must be used in a less limited way; for instance, the Bayesian criterion can be applied to the decision problems described by \( L_a \) and \( L_e \), but the choice of the initial averaging probability measure on \( \mathcal{P} \) is not simple. To simplify this choice, the Bayesian criterion can be extended by allowing infinite measures on \( \mathcal{P} \) as initial “averaging measures”; but unfortunately the usual choices lead to measures on \( \mathcal{P} \) that are still infinite when conditioned on the observed data. Tiao and Tan (1965) proposed the infinite measure on \( \mathcal{P} \) corresponding to the one with density \( \frac{1}{v_e (v_e v_a + v_a)} \) on \( \mathbb{R} \times \mathbb{P} \times \mathbb{P} \) (with respect to the Lebesgue measure) for the parameter (\( \mu, v_a, v_e \)); when conditioned on the observations \( X_{i,j} \), this measure is finite and thus proportional to a probability measure that can be used in the Bayesian criterion. This initial “averaging measure” has been strongly criticized: see for instance Hill (1965) and Stone and Springer (1965); in particular, it cannot be interpreted as describing prior information about the models in \( \mathcal{P} \), since it depends on \( n_e \). Anyway, Klotz, Milton, and Zacks (1969) showed that when using the (unweighted) squared error loss, the estimates of \( v_a \) and \( v_e \) obtained by applying the Bayesian criterion with this initial “averaging measure” have poor pre-data performances. But Portnoy (1971) showed that the estimates obtained by applying the Bayesian criterion (with this initial “averaging measure”) to the decision problems described by \( L_a \) and \( L_e \) (dotted curved lines) have very good pre-data performances: the maximum of the expected loss for the estimate of \( v_a \) (lower dotted line) seems to be nearly the minimum possible, and the expected loss for the estimate of \( v_e \) (dotted line) is also particularly small.

The estimates obtained by applying the MPL criterion to the decision problems described by \( L_a \) and \( L_e \) without using prior information (solid curved lines) are positive too, and they have very good pre-data performances, since the corresponding expected losses (lower solid lines)
are similar to the ones for the estimates studied by Portnoy. Compared to the Bayesian criterion, the MPL criterion has the important advantage of avoiding the difficult choice of the initial “averaging measure”, when no prior information is available, or when we do not want to use the available prior information.
References


Index

\[ \text{Index} \]

\[ \begin{align*}
&\text{Index} \\
&r_p & 64 \\
&\mu^l & 27 \\
&f^l & 63 \\
&f^S & 31 \\
&f^C & 47 \\
&f^r & 36 \\
&f^{rr} & 45 \\
&\mu|_A & 29 \\
&\bar{\mu} & 28, 112 \\
&\zeta_{rp} & 95 \\
&\odot & 79 \\
&\mu\{f \geq \cdot\} & 39 \\
&\varphi_+ & 42 \\
&\varphi_- & 50 \\
&\varphi_- & 42 \\
&\hat{c} & 70 \\
&a.e. & 26 \\
&\text{approach to statistical decision problems} \\
&\text{Bayesian} & 5–7, 17–19, 64–78, 80–81, 83–84, 138 \\
&\text{classical} & 4–5, 18, 83, 127–131, 141–147, 152–156 \\
&\text{imprecise Bayesian} & 7–8, 12–13, 19–20, 22, 64–65, 70–78, 86–92, 138 \\
&\text{MPL} & 13–20, 63–93, 116–169 \\
&\text{asymptotic efficiency} & 152–156 \\
&\text{asymptotic optimality} & 138–147 \\
&\text{Bayesian criterion} & 5 \\
&\text{class of measures} \\
&\text{closed under restrictions} & 29 \\
&\text{closed under transformations} & 28 \\
&\text{exhaustive} & 42 \\
&\text{regular} & 29 \\
&\text{coherence} \\
&\text{temporal} & 6, 18 \\
&\text{with respect to additive loss} & 6 \\
&\text{with respect to maxitive loss} & 19 \\
&\text{comonotonic functions} & 58 \\
&\text{complete ignorance} & 3, 79–81, 88–91, 102–103, 110 \\
&\text{conditional method} & 5 \\
&\text{consequence} & 66 \\
&\text{randomized} & 70–71 \\
&\text{uncertain} & 66 \\
&\mathcal{D} & 1–2 \\
&\text{decision} & 2, 66 \\
&\text{correct} & 2, 70, 76 \\
&\text{domination} & 2 \\
&\text{strict} & 101 \\
&\text{randomized} & 71 \\
&\text{decision criterion} & 2 \\
&\text{Bayesian} & 5 \\
&\Gamma\text{-minimax} & 7 \\
&\text{Hurwicz} & 3 \\
&\text{likelihood-based} & 97 \\
\end{align*} \]
0-independent 105
consistent 138
decomposable 123
revealing ignorance attraction 109
revealing ignorance aversion 102
satisfying the sure-thing principle 116
LRMβ 20, 105
minimax 3
essential 100
minimax risk 4
minimin 3
essential 112
MLD 21, 105
ignorance attracted analogue of 111
MPL 13, 64
approximation algorithm 159–160
restricted Bayesian 7
decision function 4
density function 27, 47
DF 41
distribution function 38, 45
distributional dominance 39
dual function 112
ε-contamination class 91
equivariance 129–130
error function 151–152, 165
asymmetric 156
ess 36
essential minimax criterion 100
essential minimin criterion 112
essential supremum 36
estimation problem 2–3, 149–152
evaluation
conditional 123
conjugate 112–113, 150
in terms of loss 67, 76–77
in terms of utility 76–77
lower 112–113, 150
post-data 6
pre-data 4–6
upper 112–113, 150
worst-case 73, 75–76, 78
f0 100–102
f1 100–102
functional
bihomogeneous 42
calibrated 42
homogeneous 42
monotonic 42
subadditive 52
x-independent 42
Γ-minimax criterion 7
hierarchical model 81–93
Hurwicz criterion 3
hypograph 48
impossible estimate 160–169
indicator property 30
integral 30
bihomogeneous 41
bimonotonic 40
Choquet 47, 61
comonotonic additive 58
generalized Imaoka 49
generalized Sugeno 32
homogeneous 30
maxitive 33
completely 33
monotonic 31
quasi-subadditive 52
rectangular 36
regularized 45
regular 41
respecting distributional dominance 39
Shilkret 31
Šipoš 61
subadditive 53
support-based 38
symmetric 50
transformation invariant 35
translation equivariant 60

$L$ 1–3
$l_d$ 2, 151
$lik$ 8–10
$lik_g$ 9
likelihood function 8–10, 67
prior 80–81
profile 9
pseudo 9
updating 78–79
likelihood ratio 11
likelihood ratio test 11
likelihood-based confidence region 11
likelihood-based inference 10–13, 15–17, 101–102
calibration problem of 11
likelihood ratio test 11
likelihood-based confidence region 11
maximum likelihood estimate 10
likelihood-based region minimax 20
likelihood-based region minimin 111
loss function 1–3
$LR$ 11
$LRM_\beta$ criterion 20, 105

maximum likelihood decision 21
maximum likelihood estimate 10
measure 26
2-alternating 26
continuous from below 28
dual 28
maxitive
completely 27
finitely 27
monotonic 26
nonzero 26
normalized 26

possibility 28
probability 64–65, 80–81
imprecise 7, 86–93
lower 86–93
prior 80–81, 87–91
upper 86–93
relative plausibility 63–65, 78–81
nondegenerate 63
prior 80–81
updating of 79
subadditive 26
minimax criterion 3
minimax plausibility-weighted loss 13
minimax risk criterion 4
minimin criterion 3
MLD criterion 21, 105
ignorance attracted analogue of 111
MPL criterion 13, 64
approximation algorithm 159–160
$\mathcal{N}\mathcal{L}$ 42
$\mathbb{P}$ vii
$\mathcal{P}$ 1
parametrization 9
parametrization invariance 128
preference relation 66–79
monotonic 68
representation 66
quasiconvex function 151, 159
strictly 151, 159
regular extension 87–92, 138
representation theorem 66–79
restricted Bayesian criterion 7
risk function 4
robustness 92–93, 130–137
$rp$ 63–65, 78–81
$\mathcal{S}_\varphi$ 98
sequence of decisions
  asymptotically optimal 141
  obtained by applying a likelihood-based decision criterion 140
statistical decision problem 1–3
statistical model 1, 64–65
  imprecise 91–93, 131
sure-thing principle 116–127, 132

Savage’s 68–69, 116–117
  strict version 116–117

t_y viii

uncertainty aversion 69–72, 76–77

V_{r_p} 95–102

total preorder 66
Curriculum Vitae

Personal Data:

Name : Marco E. G. V. Cattaneo
Born : September 19, 1976 in Kaiserslautern, West Germany
Nationality : Swiss, citizen of Capriasca, Ticino

Education:

1995–2000 : ETH Zurich, study of mathematics
2002–2007 : ETH Zurich, doctoral studies in mathematics

Teaching Experience:

1997–1998 : ETH Zurich, undergraduate teaching assistant
2000–2002 : ETH Zurich, teaching assistant in mathematics
2002–2006 : ETH Zurich, teaching assistant in statistics