LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Marco E. G. V. Cattaneo

Likelihood decision functions

Technical Report Number 128, 2012
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

LIKELIHOOD DECISION FUNCTIONS

By Marco E. G. V. Cattaneo

Department of Statistics, LMU Munich

In both classical and Bayesian approaches, statistical inference is unified and generalized by the corresponding decision theory. This is not the case for the likelihood approach to statistical inference, in spite of the manifest success of the likelihood methods in statistics. The goal of the present work is to fill this gap, by extending the likelihood approach in order to cover decision making as well. The resulting decision functions, called likelihood decision functions, generalize the usual likelihood methods (such as ML estimators and LR tests), in the sense that these methods appear as the likelihood decision functions in particular decision problems. In general, the likelihood decision functions maintain some key properties of the usual likelihood methods, such as equivariance and asymptotic optimality. By unifying and generalizing the likelihood approach to statistical inference, the present work offers a new perspective on statistical methodology and on the connections among likelihood methods.

1. Introduction. Wald (1950) tried to unify statistics in his theory of decision functions. However, many of the most appreciated statistical methods do not fit well in this setting. In particular, the likelihood methods (such as the maximum likelihood estimators and the likelihood ratio tests) are usually suboptimal in corresponding (finite-sample) decision problems. In fact, the post-data nature of likelihood methods is at variance with the pre-data evaluation of decision functions.

Since statistical methods based on the likelihood function are extremely successful as regards estimation and testing, it is natural to try extending the likelihood approach to more general decision problems. The topic of the present paper are criteria for basing decisions on the likelihood function alone. The resulting optimal decisions generalize the usual likelihood methods, in the sense that these methods are optimal in corresponding (finite-sample) decision problems. The approach of this paper offers a new perspective on statistical methodology and on the connections among likelihood methods.

Surprisingly, only very few authors have studied extensions of the like-

AMS 2000 subject classifications: Primary 62C05; secondary 62A01.

Keywords and phrases: Likelihood approach, decision theory, foundations, conditional inference, minimax, invariance, consistency, efficiency.

lihood approach to cover decision making. Besides the author (Cattaneo, 2005, 2007), only Lehmann and Romano (2005, Section 1.7, substantially unchanged since the first edition in 1959), Diehl and Sprott (1965), and Giang and Shenoy (2005) seem to have worked in this direction. However, the latter three approaches are not directly applicable to general statistical decision problems in the sense of Wald (1950), and their properties have not been investigated.

Many authors (such as Fisher, 1973; Barnard, 1967; Birnbaum, 1962; Hacking, 1964; Kalbfleisch, 1985; Sprott, 2000; Edwards, 1992; Lindsey, 1996; Azzalini, 1996; Royall, 1997; Reid, 2000; Pawitan, 2001; Hills, 2005) consider the likelihood function as a description of uncertain knowledge about the parameters of the statistical model. More precisely, the likelihood function describes the relative plausibility of the possible values of the parameters in the light of the observed data. The uncertainty in this description is non-probabilistic, and therefore the likelihood approach to decision making clearly differs from the Bayesian approach.

In particular, prior (uncertain) knowledge about the parameters is not needed in the likelihood approach to decision making: this is a fundamental advantage over the Bayesian approach. However, the fact that likelihood functions induced by independent data are combined by (pointwise) multiplication suggests the possibility of describing prior uncertain knowledge by a prior likelihood function (this idea is implicitly or explicitly considered for example by Barnard, 1949, 1972; Barnard, Jenkins and Winsten, 1962; Birnbaum, 1962; Edwards, 1969, 1970, 1992; Hudson, 1971; Leonard, 1978; Lindsey, 1996, 1999; Pawitan, 2001). In particular, complete ignorance about the values of the parameters is described by a constant (prior) likelihood function: the possibility of describing ignorance distinguish likelihood functions from probability measures (as descriptions of uncertain knowledge) and leads to the above fundamental advantage over the Bayesian approach.

Despite the different descriptions of uncertain knowledge about the parameters of the statistical model, the likelihood and Bayesian approaches to decision making share a basic property: they both satisfy the (strong) likelihood principle (see for example Birnbaum, 1962; Basu, 1975; Joshi, 1983; Berger and Wolpert, 1988; Evans, Fraser and Monette, 1986; Lindsey, 2005). This principle gives theoretical reasons for the likelihood approach to decision making, in addition to the pragmatic reasons mentioned above (that is, the successfulness of the likelihood methods). In particular, this approach can be applied post-data, avoiding the severe difficulties of pre-data evaluations (see for instance Kiefer, 1977; Robinson, 1979; Berger, 1985a; Goutis and Casella, 1995), and drastically reducing the complexity of the decision problems. However, in the tradition of Wald (1950), the pre-data properties of the resulting decision functions, called likelihood decision functions, will be studied in the present paper.

The paper is organized as follows. In the next section, basic definitions and notations are introduced. Section 3 presents criteria for basing decisions on the likelihood function alone. Pre-data properties of the resulting likelihood decision functions are the subject of Section 4 (the proofs of the theorems are in the appendix). The final section is devoted to conclusions and directions for further research.

2. Settings.

Let $(\Omega, \mathcal{F})$ be a measurable space, and for each $\theta \in \Theta$, let P_θ be a probability measure on $(\Omega, \mathcal{F})$. Random variables on Ω are denoted by X or X_n (with $n \in \mathbb{N}$), and their codomains by $\mathcal{X}$ and $\mathcal{X}_n$, respectively (it is assumed that all singleton subsets of $\mathcal{X}$ and $\mathcal{X}_n$ are measurable). The only assumption about Θ is that it is not empty. In particular, the statistical model $\{P_\theta : \theta \in \Theta\}$ can be parametric (in this case, θ describes the parameters of the statistical model) or nonparametric (in this case, θ simply indexes the probability measures).

2.1. *Likelihood function.*

Let Λ be the set of all functions $\lambda : \Theta \to [0, 1]$ such that $\sup_{\theta\in\Theta} \lambda(\theta) = 1$. If the event $A \in \mathcal{F}$ satisfies $P_\theta(A) > 0$ for some $\theta \in \Theta$, then the (relative) likelihood function given A is the unique function $\lambda \in \Lambda$ such that $\lambda(\theta) \propto P_\theta(A)$. When there is a unique $\theta \in \Theta$ such that $\lambda(\theta) = 1$, it is called maximum likelihood estimate of θ, and denoted by $\hat{\theta}$. For each subset $\mathcal{H}$ of Θ, with a slight abuse of notation, $\lambda(\mathcal{H})$ denotes the likelihood ratio test statistic for the null hypothesis $H_0 : \theta \in \mathcal{H}$ against the alternative $H_1 : \theta \in \Theta\setminus\mathcal{H}$. That is, $\lambda(\mathcal{H}) = \sup_{\theta\in\mathcal{H}} \lambda(\theta)$, with the convention that $\lambda(\varnothing) = 0$.

If $x \in \mathcal{X}$ satisfies $P_\theta(X = x) > 0$ for some $\theta \in \Theta$, then the likelihood function given $X = x$ is denoted by λ_x. This definition is not applicable when the random variable X is continuous for all $\theta \in \Theta$. In fact, it can be argued that the realization of a continuous random variable can never be observed with infinite precision: it is only possible to observe $X \in N$ for a suitable neighborhood N of x. The likelihood function λ_N given $X \in N$ is then usually well-defined. If for each $\theta \in \Theta$ the density f_θ of X with respect to a fixed σ-finite measure μ on $\mathcal{X}$ exists and satisfies $\sup_{\theta\in\Theta} f_\theta(x) \in \mathbb{R}_{>0}$, then λ_N is possibly well approximated by the unique function $f \in \Lambda$ such that $f(\theta) \propto f_\theta(x)$.

The likelihood function given $X = x$ is often simply defined as the function f, but the definition of likelihood in terms of probability (and the consequent interpretation of f as a mere approximation of λ_N) seems to be preferred by most authors who consider likelihood functions as descriptions of uncertain knowledge (see for example Edwards, 1992; Kalbfleisch, 1985; Barnard and Sprott, 1983; Lindsey, 1996, 1999; Sprott, 2000; Pawitan, 2001; Montoya, Díaz-Francés and Sprott, 2009). The reasons are that the post-data interpretation of the function f can be problematic, since the densities f_θ are not unique (but only unique μ-a.e.), and that f is not well-defined when $f_\theta(x)$ is an unbounded function of θ. However, for the pre-data properties studied in Section 4 the nonuniqueness of the densities is not a problem, and therefore, in order to simplify the results, f will then be called the likelihood function given $X = x$ and denoted by λ_x (when it is well-defined).

If the random variables X_1, X_2 are independent for all $\theta \in \Theta$, then the likelihood function given $(X_1, X_2) = (x_1, x_2)$ satisfies $\lambda_{(x_1,x_2)}(\theta) \propto \lambda_{x_1}(\theta)\,\lambda_{x_2}(\theta)$ (when these three functions are well-defined). As noted in Section 1, this suggests the possibility of describing prior uncertain knowledge by a prior likelihood function: when $X_2 = x_2$ is observed, the prior λ_{x_1} is updated to the posterior $\lambda_{(x_1,x_2)}$. The prior likelihood function is simply interpreted as the likelihood function given $X_1 = x_1$, regardless of whether the observation $X_1 = x_1$ is real or imagined. The choice of a prior likelihood function seems better supported by intuition than the choice of a prior probability measure: in particular, the likelihood function constant equal to 1 describes the complete ignorance about the value of $\theta \in \Theta$ (see also Cattaneo, 2007, Subsection 3.1.2). The penalty term in penalized likelihood methods can often be formally interpreted as a prior likelihood function (see for example Leonard, 1978).

2.2. *Decision problem.* A statistical decision problem is described by a loss (or weight) function $W : \Theta \times \mathcal{D} \to \mathbb{R}_{\geq 0}$, where $\mathcal{D}$ is the (nonempty) set of all possible decisions, one or more of which must be chosen. For each pair $(\theta, d) \in \Theta \times \mathcal{D}$, the value $W(\theta, d)$ represents the loss suffered by choosing the decision d when P_θ is the correct probability measure. It is assumed that the function W summarizes all important aspects of the decision problem. In particular, if randomized decisions are allowed, then they should already be contained in $\mathcal{D}$, and the corresponding loss described by W.

Let $\mathcal{W}$ be the set of all functions $w : \Theta \to \mathbb{R}_{\geq 0}$. To each decision $d \in \mathcal{D}$ can be associated the function $w_d \in \mathcal{W}$ such that $w_d(\theta) = W(\theta, d)$ for all $\theta \in \Theta$. The decision problem can be restated as the problem of choosing one or more functions w from the subset $\{w_d : d \in \mathcal{D}\}$ of $\mathcal{W}$, where the loss (as a function of θ) suffered by choosing w is represented by the function w itself. To each function w can correspond more than one decision $d \in \mathcal{D}$, but these decisions are equivalent from the standpoint of the decision problem.

When $X = x$ is observed, the likelihood function λ_x describes the relative plausibility of the possible values of θ, and can thus be useful for choosing a decision $d \in \mathcal{D}$. Possible criteria for this kind of post-data decision making are the subject of Section 3. Some pre-data properties of these decision criteria are then studied in Section 4. In order to do this, the chosen decision must be considered as a function of the observed realization of X. Such a function $\delta : \mathcal{X} \to \mathcal{D}$, describing a whole decision strategy, is called decision function.

3. Likelihood decision criteria.

Let $\lambda \in \Lambda$ be the likelihood function given the data (possibly including prior information), and let the loss function W on $\Theta \times \mathcal{D}$ describe a decision problem. The subject of this section are criteria for choosing, on the basis of λ and W, one or more decisions $d \in \mathcal{D}$, or equivalently, on the basis of λ, one or more functions $w \in \{w_d : d \in \mathcal{D}\}$. For instance, when the maximum likelihood estimate $\hat{\theta}$ is well-defined, a simple criterion for choosing d consists in minimizing $W(\hat{\theta}, d)$, or equivalently, for choosing w, in minimizing $w(\hat{\theta})$. That is, the criterion consists in minimizing the loss under the assumption that $P_{\hat{\theta}}$ is the correct probability measure. This simple criterion is often used in practical applications: for example when in the portfolio selection problem of Markowitz (1952) the parameters of the model are estimated by maximum likelihood (see for instance Levy and Sarnat, 1970; Board and Sutcliffe, 1994). The criterion was also formally, though hesitantly, considered by Diehl and Sprott (1965). However, besides being perhaps too optimistic about the quality of maximum likelihood estimates, this simple criterion is not always well-defined. Before considering alternative criteria, in the next subsection a general definition of likelihood decision criteria is introduced.

3.1. *General definition.*

A likelihood decision criterion for choosing one or more decisions $d \in \mathcal{D}$ consists in minimizing a certain evaluation $V(w_d, \lambda)$ of the corresponding loss w_d on the basis of the likelihood function λ, where the functional $V : \mathcal{W} \times \Lambda \to \overline{\mathbb{R}}$ must satisfy the following three properties:

(P1) If the functions $w, w' \in \mathcal{W}$ satisfy $w(\theta) \leq w'(\theta)$ for all $\theta \in \Theta$, then $V(w, \lambda) \leq V(w', \lambda)$ must hold for all functions $\lambda \in \Lambda$.

(P2) If the function $b : \Theta \to \Theta$ is bijective, then $V(w \circ b, \lambda \circ b) = V(w, \lambda)$ must hold for all pairs of functions $(w, \lambda) \in \mathcal{W} \times \Lambda$.

(P3) If the subset $\mathcal{H}$ of Θ and the sequence of functions $\lambda_n \in \Lambda$ (with $n \in \mathbb{N}$) satisfy $\lim_{n\to\infty} \lambda_n(\Theta \setminus \mathcal{H}) = 0$, then $\lim_{n\to\infty} V(c\, I_{\mathcal{H}} + c'\, I_{\Theta \setminus \mathcal{H}}, \lambda_n) = c$ must hold for all constants $c, c' \in \mathbb{R}_{\geq 0}$.

Before analyzing these properties, it is important to clarify what is meant by minimization of $V(w_d, \lambda)$. If there is a decision $d \in \mathcal{D}$ such that $V(w_d, \lambda) = \inf_{d' \in \mathcal{D}} V(w_{d'}, \lambda)$, then d is optimal according to the likelihood decision criterion described by the functional V. When there is no optimal decision, the criterion suggests the choice of a decision $d \in \mathcal{D}$ such that $V(w_d, \lambda) < \inf_{d' \in \mathcal{D}} V(w_{d'}, \lambda) + \varepsilon$, for a suitably small $\varepsilon \in \mathbb{R}_{>0}$.

(P1) can be interpreted as a property of monotonicity of the functional V, following directly from the assumption that the loss function W summarizes all important aspects of the decision problem. In fact, if the decisions $d, d' \in \mathcal{D}$ satisfy $W(\theta, d) \leq W(\theta, d')$ for all $\theta \in \Theta$, then it is unreasonable to prefer d' to d.

(P2) can be interpreted as a property of parametrization invariance, typical of the likelihood methods. This invariance is a consequence of the idea that everything important about θ is described by the loss function W and the likelihood function λ. In particular, (P2) excludes the Bayesian criteria when Θ is infinite. In fact, with some additional measurability restrictions, the Bayesian criterion with prior π is described by the functional $V_\pi : (w, \lambda) \mapsto \int w \, \lambda \, d\pi$. Hence, (P2) implies in particular the invariance $\pi \circ b^{-1} = \pi$ for all measurable bijections b, since $V_\pi(I_{\mathcal{H}}, I_\Theta) = \pi(\mathcal{H})$ for all measurable subsets $\mathcal{H}$ of Θ. This invariance can be satisfied only if Θ is finite (when π is the uniform prior) or if the measurability conditions are very restrictive.

(P3) can be interpreted as a minimal consistency property, implying that some information provided by the likelihood function is actually used by the likelihood decision criterion. In particular, it excludes the minimax criterion, described by the functional $(w, \lambda) \mapsto \sup_{\theta \in \Theta} w(\theta)$. Moreover, (P3) with $\mathcal{H} = \Theta$ implies the following calibration property: $V(c \, I_\Theta, \lambda) = c$ for all constants $c \in \mathbb{R}_{\geq 0}$ and all likelihood functions $\lambda \in \Lambda$. This property and (P1) imply in particular that $\inf_{\theta \in \Theta} w(\theta) \leq V(w, \lambda) \leq \sup_{\theta \in \Theta} w(\theta)$ holds for all pairs of functions $(w, \lambda) \in \mathcal{W} \times \Lambda$.

A simple example of likelihood decision criterion can be obtained by modifying the minimax criterion in order to satisfy (P3). It suffices to reduce Θ to the likelihood confidence region consisting of all θ whose likelihood exceeds a certain threshold $\beta \in \,]0, 1[$, before applying the minimax criterion. The resulting likelihood decision criterion is called Likelihood-based Region Minimax (LRM) criterion and is described by the functional $V_{LRM,\beta} : (w, \lambda) \mapsto \sup_{\theta \in \Theta \, : \, \lambda(\theta) > \beta} w(\theta)$. It has been applied for example in the problem of regression with imprecisely observed data (see for instance Cattaneo and Wiencierz, 2012).

If the maximum likelihood estimate $\hat{\theta} \in \Theta$ is well-defined and there is a topology on Θ such that $w \in \mathcal{W}$ is continuous at $\hat{\theta}$ and $\lambda(\Theta \setminus \mathcal{N}) < 1$ holds for all neighborhoods $\mathcal{N}$ of $\hat{\theta}$, then $\lim_{\beta \uparrow 1} V_{LRM,\beta}(w, \lambda) = w(\hat{\theta})$. Hence, the likelihood decision criterion described by the (pointwise) limit of $V_{LRM,\beta}$ when β tends to 1 is strictly related to the idea considered at the beginning of the present section, but has the advantage of being always well-defined. It is called Maximum Likelihood Decision (MLD) criterion and is described by the functional $V_{MLD} : (w, \lambda) \mapsto \lim_{\beta \uparrow 1} \sup_{\theta \in \Theta \,:\, \lambda(\theta) > \beta} w(\theta)$.

The MLD criterion clearly generalizes maximum likelihood estimation, while the LRM criterion can be seen as a generalization of likelihood ratio testing. In the next subsection, a likelihood decision criterion generalizing both these very successful components of the likelihood approach to statistics is considered in more detail.

3.2. *MPL criterion.* An alternative way of modifying the minimax criterion in order to satisfy (P3) consists in applying it after having weighted the loss associated to θ by means of the likelihood of θ (raised to a certain power $\alpha \in \mathbb{R}_{>0}$). The resulting likelihood decision criterion is called Minimax Plausibility-weighted Loss (MPL) criterion and is described by the functional $V_{MPL,\alpha} : (w, \lambda) \mapsto \sup_{\theta \in \Theta} w(\theta)\, \lambda(\theta)^{\alpha}$. It can be characterized among the likelihood decision criteria by few basic decision-theoretic properties, but this goes beyond the scope of the present paper (see Cattaneo, 2007, Subsection 4.1.2). The exponent $\alpha \in \mathbb{R}_{>0}$ plays a similar role for the MPL criterion as the threshold $\beta \in \,]0, 1[$ does for the LRM criterion. In fact, $\lim_{\alpha \downarrow 0} V_{MPL,\alpha}(w, \lambda) = \lim_{\beta \downarrow 0} V_{LRM,\beta}(w, \lambda)$ holds for all pairs of functions $(w, \lambda) \in \mathcal{W} \times \Lambda$, while $\lim_{\alpha \uparrow \infty} V_{MPL,\alpha}(w, \lambda) = \lim_{\beta \uparrow 1} V_{LRM,\beta}(w, \lambda) = V_{MLD}(w, \lambda)$ holds for all pairs of functions $(w, \lambda) \in \mathcal{W} \times \Lambda$ such that $V_{MPL,\alpha}(w, \lambda)$ is finite for some $\alpha \in \mathbb{R}_{>0}$.

The simple choice $\alpha = 1$ for the exponent of the likelihood function is supported by the analogy with the Bayesian criterion: the integral with respect to π in the functional V_π is replaced by the supremum with respect to θ in the functional $V_{MPL,1}$. The analogy of the Bayesian and MPL criteria (with $\alpha = 1$) emerges also when considering the likelihood ratio test statistic $\lambda(\mathcal{H})$ as a function of $\mathcal{H} \subseteq \Theta$. This set function is a completely maxitive measure in the terminology of Shilkret (1971), who introduced also the corresponding theory of integration: the integral of $w \in \mathcal{W}$ with respect to the completely maxitive measure λ is $V_{MPL,1}(w, \lambda)$. Hence, the MPL criterion with $\alpha = 1$ corresponds to minimizing the integral of the loss with respect to the likelihood ratio test statistic, interpreted as a completely maxitive measure describing the posterior information about θ.

Example 3.1 (maximum likelihood estimation). *The estimation of θ can be described as a decision problem with $\mathcal{D} = \Theta$. When Θ is finite, it makes sense to employ the simple loss function W such that $w_d = I_{\Theta \setminus \{d\}}$ for all $d \in \mathcal{D}$. In this case, if the maximum likelihood estimate $\hat{\theta}$ is well-defined, then it is the unique optimal decision according to the MLD and MPL criteria (independently of the exponent α), while for the LRM criterion this holds only if the threshold β is sufficiently large.*

These results can be generalized to the case with infinite Θ, for example when a metric on Θ is considered. For a suitably small $\varepsilon \in \mathbb{R}_{>0}$, it makes then sense to employ the simple loss function W such that $w_d = I_{\Theta \setminus B(d,\varepsilon)}$ for all $d \in \mathcal{D}$, where $B(d, \varepsilon)$ denotes the closed ball with center d and radius ε. It can be easily proved that in this case, if the maximum likelihood estimate $\hat{\theta}$ is well-defined, $B(\hat{\theta}, \varepsilon)$ is compact, and $\lambda(\Theta \setminus B(\hat{\theta}, \varepsilon)) < 1$, then for the MLD and MPL criteria (independently of the exponent α) optimal decisions exist and, even when they are not unique, they all lie in $B(\hat{\theta}, \varepsilon)$, while for the LRM criterion this holds only if the threshold β is sufficiently large.

Hence, the MPL and MLD criteria lead practically to maximum likelihood estimates in this simple decision-theoretic description of estimation, and therefore they can be interpreted as generalizations of maximum likelihood estimation (while this is not true for the LRM criterion).

Example 3.2 (likelihood ratio testing). *For each subset $\mathcal{H}$ of Θ, testing for the null hypothesis $H_0 : \theta \in \mathcal{H}$ against the alternative $H_1 : \theta \in \Theta \setminus \mathcal{H}$ can be described as a decision problem with $\mathcal{D} = \{1, 0\}$, where 1 and 0 represent the rejection and the acceptance (or non-rejection) of H_0, respectively. When constant losses $c_1, c_2 \in \mathbb{R}_{>0}$ (with $c_1 > c_2$) are assigned to errors of the first and of the second kind, respectively, the resulting loss function W satisfies $w_1 = c_1 \, I_{\mathcal{H}}$ and $w_0 = c_2 \, I_{\Theta \setminus \mathcal{H}}$. In this case, according to the MPL criterion with exponent α, rejection is the unique optimal decision if and only if $\lambda(\mathcal{H}) < ({c_2}/{c_1})^{1/\alpha}$, while acceptance is the unique optimal decision if and only if $\lambda(\mathcal{H}) > ({c_2}/{c_1})^{1/\alpha}$. Similarly, according to the LRM criterion with threshold β, rejection is the unique optimal decision if and only if $\lambda(\mathcal{H}) \leq \beta$, while acceptance is the unique optimal decision if and only if $\lambda(\mathcal{H}) > \beta$. Finally, according to the MLD criterion, rejection is the unique optimal decision if and only if $\lambda(\mathcal{H}) < 1$, while acceptance is the unique optimal decision if and only if $\lambda(\mathcal{H}) = 1$.*

Hence, the MPL and LRM criteria lead practically to likelihood ratio tests in this simple decision-theoretic description of hypothesis testing, and therefore they can be interpreted as generalizations of likelihood ratio testing (while this is not true for the MLD criterion).

4. Properties. Likelihood decision criteria were introduced in Section 3 as criteria for post-data decision making. The subject of the present section are pre-data properties of the resulting likelihood decision functions. Before considering some asymptotic results, in the next subsection finite-sample invariance properties are presented.

4.1. *Invariances.* Let X be a random variable such that the likelihood function $\lambda_x \in \Lambda$ is well-defined for all $x \in \mathcal{X}$, and let the loss function W on $\Theta \times \mathcal{D}$ describe a decision problem. A likelihood decision criterion described by the functional V can be applied for each possible realization x of X, by minimizing the evaluation $V(w_d, \lambda_x)$ over all decisions $d \in \mathcal{D}$. In this subsection, in order to simplify the results, it is assumed that for each possible realization x of X there is a unique optimal decision $\delta(x)$ according to the likelihood decision criterion. The resulting likelihood decision function $\delta : \mathcal{X} \rightarrow \mathcal{D}$ is then uniquely defined.

Some basic invariance properties follow directly from the fact that the likelihood approach to decision making satisfies the likelihood principle. In particular, if $s(X)$ is a sufficient statistic for θ, then $\delta(x) = \delta(x')$ holds for all $x, x' \in \mathcal{X}$ such that $s(x) = s(x')$, since in this case $\lambda_x = \lambda_{x'}$ (see for instance Schervish, 1995, Theorem 2.21 and Proposition 2.23). That is, the likelihood decision function δ is completely described by a function $\delta' : \mathcal{S} \rightarrow \mathcal{D}$ such that $\delta = \delta' \circ s$, where $\mathcal{S}$ is the codomain of s.

As noted in Subsection 3.1, a certain kind of parametrization invariance is implied by (P2). In fact, a bijection $b : \Theta \rightarrow \Theta$ can be interpreted as the description of a reparametrization of the statistical model, in which $\theta \in \Theta$ is replaced by $\vartheta \in \Theta$, with $b(\vartheta) = \theta$. For the reparametrized statistical model, the likelihood function given $X = x$ is $\lambda_x \circ b$, and the decision problem is described by the loss function $(\vartheta, d) \mapsto W(b(\vartheta), d)$. Hence, (P2) implies that the likelihood decision function δ is left invariant by this reparametrization of the statistical model.

Another direct consequence of (P2) is the following important invariance property. Given three bijections $g : \mathcal{X} \rightarrow \mathcal{X}$, $b : \Theta \rightarrow \Theta$, and $h : \mathcal{D} \rightarrow \mathcal{D}$, if for each $x \in \mathcal{X}$ the likelihood function given $X = g(x)$ is $\lambda_x \circ b$, and $w_{h(d)} = w_d \circ b$ holds for all $d \in \mathcal{D}$, then the likelihood decision function satisfies $\delta \circ g = h \circ \delta$. That is, if the decision problem is invariant, then δ is equivariant (see for example Berger, 1985b, Section 6.2; Schervish, 1995, Subsection 6.2.1). In particular, it is not even necessary to identify the symmetries of the decision problem: the likelihood decision functions are guaranteed to respect them anyway. Among the invariance properties considered in the present subsection, this is the only one that does not necessarily hold when a prior likelihood function is used. In fact, prior information can destroy the symmetries of the decision problem.

EXAMPLE 4.1 (variance components). *Let $X_1, \ldots, X_m$ be independent and n-variate normally distributed random variables (with $n \geq 2$) such that for all $i \in \{1, \ldots, m\}$, each component of X_i has expected value μ and variance $\tau^2 + \sigma^2$, and each pair of different components of X_i has covariance τ^2, where $\theta = (\mu, \tau, \sigma)$ and $\Theta = \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$. That is, each vector X_i represents the n observations in one of the m groups of a balanced one-way random effect model under normality assumptions (see for example Searle, Casella and McCulloch, 1992). In order to simplify the results, assume that the model is conditioned on the (a.s.) event that for no vector X_i all components are equal, and so $\mathcal{X}_1 = \cdots = \mathcal{X}_m = \mathbb{R}^n \setminus \{(y_1, \ldots, y_n) \in \mathbb{R}^n : y_1 = \cdots = y_n\}$.*

The problem of estimating the variance component τ^2 is particularly interesting, because the analysis of variance estimate can be negative. For this problem, Portnoy (1971) suggested the following location and scale invariant version of the squared error loss function (with $\mathcal{D} = \mathbb{R}$):

$$W : ((\mu, \tau, \sigma), d) \mapsto \frac{(\tau^2 - d)^2}{(\sigma^2 + n\,\tau^2)^2}.$$

For each $i \in \{1, \ldots, m\}$, let $\bar{X}_i$ and S_i be the mean and the sum of squared deviations from the mean, respectively, for the components of X_i. Furthermore, let $\bar{X}$ and S be the mean and the sum of squared deviations from the mean, respectively, for the sample $\bar{X}_1, \ldots, \bar{X}_m$. That is, $\bar{X}$ is the grand mean, while $n\,S$ and $\sum_{i=1}^m S_i$ are the sum of squares due to differences between groups and within groups, respectively. Finally, define the ratio

$$R = \frac{n\,S}{n\,S + \sum_{i=1}^m S_i}.$$

Since $(\bar{X},\, n\,S,\, \sum_{i=1}^m S_i)$ is a sufficient statistic for (μ, τ, σ), and the decision problem described by the loss function W is location and scale invariant, when a likelihood decision function $\delta : \mathcal{X}_1 \times \cdots \times \mathcal{X}_m \to \mathbb{R}$ is uniquely defined, it satisfies

$$\delta(X_1, \ldots, X_m) = (n\,S + \textstyle\sum_{i=1}^m S_i)\, \delta'(R)$$

for some function $\delta' : [0, 1[\to \mathbb{R}$. This holds in particular for the likelihood decision function resulting from the MPL criterion with exponent $\alpha = 1$: for each $r \in [0, 1[$, the value $\delta'(r)$ can be easily obtained numerically as the unique $d \in \mathbb{R}$ minimizing

$$\max_{(\tau,\sigma) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{>0}} \frac{(\tau^2 - d)^2}{(\sigma^2 + n\,\tau^2)^{\frac{m}{2}+2}\, \sigma^{(n-1)\,m}} \exp\left(-\frac{r}{2\,(\sigma^2 + n\,\tau^2)} - \frac{1-r}{2\,\sigma^2}\right).$$

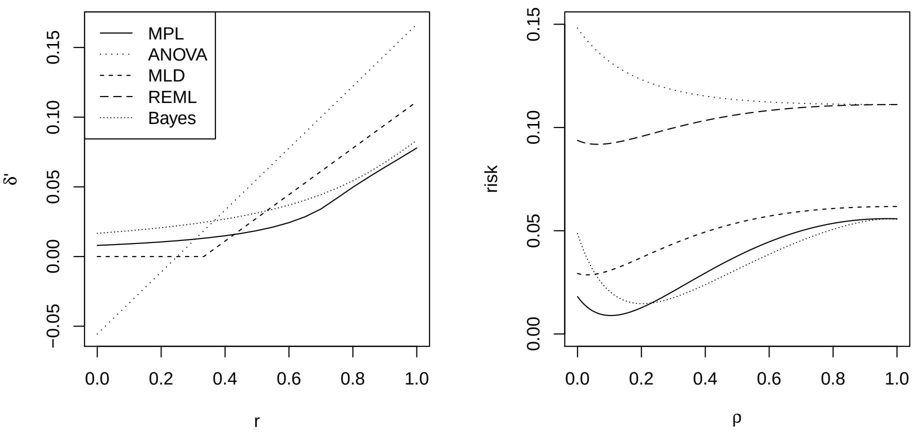

Fig 1. *Functions δ' and risk functions corresponding to the variance component estimators considered in Example 4.1.*

The resulting function δ' in the case $m = n = 3$ is plotted in the left panel of Figure 1, together with the functions δ' corresponding to some other decision criteria or estimation methods. The right panel of Figure 1 shows the expected loss (that is, the risk) of these estimators as a function of $\rho = \tau^2/(\tau^2+\sigma^2)$. Besides the MPL criterion, the methods considered are the analysis of variance, maximum likelihood (corresponding to the MLD criterion), restricted maximum likelihood (Thompson, 1962; the function δ' is the pointwise maximum of the ones corresponding to analysis of variance and maximum likelihood), and the Bayesian criterion with the Jeffreys' prior proposed by Tiao and Tan (1965). The results are qualitatively similar for other values of m and n.

Portnoy (1971) showed that the estimator resulting from the Bayesian criterion with the Jeffreys' prior proposed by Tiao and Tan (1965) is nearly minimax (from the standpoint of risk). Therefore, the MPL criterion leads to a nearly minimax estimator as well, and has the fundamental advantage of avoiding the difficult choice of a prior probability measure.

4.2. *Consistency.* Let the loss function W on $\Theta \times \mathcal{D}$ describe a decision problem, and consider a sequence of random variables X_n (with $n \in \mathbb{N}$). A sequence of decision functions $\delta_n : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathcal{D}$ (with $n \in \mathbb{N}$) is said to be (strongly) consistent at $\theta_0 \in \Theta$ if

$$\lim_{n \to \infty} W(\theta_0, \delta_n(X_1, \dots, X_n)) = \inf_{d \in \mathcal{D}} W(\theta_0, d)$$

holds P_{θ_0}-a.s. That is, consistency at θ_0 means that when P_{θ_0} is the correct probability measure, the sequence of decisions $\delta_n(X_1, \dots, X_n)$ tends to minimize the loss (almost surely). For example, if $\mathcal{D} = \Theta$, and W is a metric on Θ, then the decision problem describes the estimation of θ, and the sequence of estimators δ_n is (strongly) consistent in the usual sense if and only if it is consistent in the above sense at each $\theta \in \Theta$.

A sequence of decision functions $\delta_n : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathcal{D}$ (with $n \in \mathbb{N}$) is said to be optimal according to the likelihood decision criterion described by a functional V if

$$V(w_{\delta_n(x_1,\dots,x_n)}, \lambda_{(x_1,\dots,x_n)}) < \inf_{d \in \mathcal{D}} V(w_d, \lambda_{(x_1,\dots,x_n)}) + 2^{-n}$$

holds for all $n \in \mathbb{N}$ and all $(x_1, \dots, x_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ such that the likelihood function $\lambda_{(x_1,\dots,x_n)} \in \Lambda$ is well-defined. Hence, for each likelihood decision criterion an optimal sequence of decision functions δ_n always exists, though in general it is not unique and no single decision $\delta_n(x_1, \dots, x_n)$ needs to be optimal. However, this weak definition of optimality of a sequence of decision functions is strong enough to warrant important asymptotic results.

In general, a sequence of decision functions that is optimal according to a likelihood decision criterion can be consistent at $\theta_0 \in \Theta$ only if the likelihood tends to concentrate on θ_0, in the following sense. Given a topology on Θ, the likelihood is said to tend to concentrate on θ_0 if P_{θ_0}-a.s. the likelihood function $\lambda_{(X_1,\dots,X_n)} \in \Lambda$ is well-defined for sufficiently large n, and $\lim_{n\to\infty} \lambda_{(X_1,\dots,X_n)}(\Theta \setminus \mathcal{H}) = 0$ holds P_{θ_0}-a.s. for all neighborhoods $\mathcal{H}$ of θ_0. Sufficient conditions for the likelihood to tend to concentrate on θ_0 are well-known: see for example Wald (1949, Theorem 1), Kiefer and Wolfowitz (1956, (2.12)), or Bahadur (1967, (xxvii)). The tendency of the likelihood to concentrate on θ_0 is not affected by the use of a prior likelihood function bounded away from 0 in a neighborhood of θ_0.

As noted in Subsection 3.1, some kind of minimal consistency is implied by (P3). In fact, a simple consequence of (P3) and (P1) is that

$$\lim_{n \to \infty} V(w, \lambda_{(X_1,\dots,X_n)}) = w(\theta_0)$$

holds P_{θ_0}-a.s. when the function $w \in \mathcal{W}$ is bounded and there is a topology on Θ such that w is continuous at θ_0 and the likelihood tends to concentrate on θ_0. This implies in particular the consistency at θ_0 of all sequences of decision functions that are optimal according to some likelihood decision criterion, when $\mathcal{D}$ is finite and for each $d \in \mathcal{D}$ the loss w_d is bounded and there is a topology on Θ such that w_d is continuous at θ_0 and the likelihood tends to concentrate on θ_0. The following theorem shows that in the case of infinite $\mathcal{D}$ it suffices to replace the assumptions of continuity at θ_0 of the functions w_d with the stronger assumption of their equicontinuity at θ_0.

Theorem 4.1. *If the loss w_d is bounded for each decision $d \in \mathcal{D}$, the sequence of decision functions $\delta_n : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathcal{D}$ (with $n \in \mathbb{N}$) is optimal according to a likelihood decision criterion, and there are a $\theta_0 \in \Theta$ and a topology on Θ such that the likelihood tends to concentrate on θ_0 and the set of functions $\{w_d : d \in \mathcal{D}\}$ is equicontinuous at θ_0, then the sequence of decision functions δ_n is consistent at θ_0.*

Example 4.2 (hypothesis testing). *In the decision problem of Example 3.2, if there is a topology on Θ such that for each $\theta_0 \in \Theta$ the likelihood tends to concentrate on θ_0, then Theorem 4.1 implies the consistency at each $\theta \in \Theta \setminus \partial\mathcal{H}$ (where $\partial\mathcal{H}$ denotes the boundary of $\mathcal{H}$) of all sequences of decision functions that are optimal according to some likelihood decision criterion. That is, each likelihood decision criterion will P_θ-a.s. give the correct test result for sufficiently large n, for all $\theta \in \Theta \setminus \partial\mathcal{H}$.*

In Theorem 4.1, it is assumed that the functions w_d are bounded and equicontinuous at θ_0. As noted by Wald (1950, Subsection 3.1.2), such assumptions are not seriously restrictive from a practical point of view. However, they are not satisfied in many standard formulations of statistical decision problems, such as for example the estimation of θ when Θ is a Euclidean space and W represents squared error. In order to prove the consistency of sequences of likelihood decision functions in such standard decision problems as well, the assumptions of Theorem 4.1 can be replaced by the weaker, but more complex ones of the following theorem.

Theorem 4.2. *If the sequence of decision functions $\delta_n : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathcal{D}$ (with $n \in \mathbb{N}$) is optimal according to the likelihood decision criterion described by a functional V, and there are a $\theta_0 \in \Theta$, a topology on Θ, a constant $c \in \mathbb{R}_{>0}$ with $c > \inf_{d \in \mathcal{D}} W(\theta_0, d)$, and a neighborhood $\mathcal{H}$ of θ_0 such that the following three conditions are satisfied:*

(i) the likelihood tends to concentrate on θ_0,

(ii) the set of functions $\{w_d : d \in \mathcal{D}, \inf_{\theta \in \mathcal{H}} W(\theta, d) < c\}$ is equicontinuous at θ_0,

(iii) $\lim_{m\to\infty}\limsup_{n\to\infty}\big(V(w_d,\lambda_{(X_1,\dots,X_n)})-V(w_d\wedge m,\lambda_{(X_1,\dots,X_n)})\big)=0$ *(where $w_d\wedge m$ denotes the pointwise minimum of w_d and $m\in\mathbb{N}$) holds P_{θ_0}-a.s. for all $d\in\mathcal{D}$ such that $W(\theta_0,d)<c$,*

then the sequence of decision functions δ_n is consistent at θ_0.

Example 4.3 (uniform distribution). *Let the sequence of random variables X_n (with $n\in\mathbb{N}$ and $\mathcal{X}_n=\mathbb{R}_{>0}$) be independent and uniformly distributed on the interval $]0,\theta[$, where $\Theta=\mathbb{R}_{>0}$. Consider the problem of estimating θ with the scale invariant version $W:(\theta,d)\mapsto {|\theta-d|}/{\theta}$ of the absolute error loss function, where $\mathcal{D}=\mathbb{R}_{>0}$. For each $n\in\mathbb{N}$, since the maximum $X_{(n)}$ is a sufficient statistic of $X_1,\dots,X_n$ for θ, and the decision problem is scale invariant, when a likelihood decision function $\delta_n:\mathbb{R}^n_{>0}\to\mathbb{R}_{>0}$ is uniquely defined, it satisfies $\delta_n(X_1,\dots,X_n)=\kappa_n\,X_{(n)}$ for some constant $\kappa_n\in\mathbb{R}_{>0}$. More generally, for each likelihood decision criterion an optimal sequence of decision functions of the form $\delta_n(X_1,\dots,X_n)=\kappa_n\,X_{(n)}$ always exists.*

For each $\theta_0\in\mathbb{R}_{>0}$, the likelihood tends to concentrate on θ_0 with respect to the Euclidean topology, since $\lim_{n\to\infty}X_{(n)}=\theta_0$ holds P_{θ_0}-a.s., and $\lambda_{(X_1,\dots,X_n)}:\theta\mapsto({X_{(n)}}/{\theta})^n\,I_{]X_{(n)},\infty[}(\theta)$ for all $n\in\mathbb{N}$. Moreover, it can be easily checked that for any $c\in\mathbb{R}_{>0}$ and any bounded neighborhood $\mathcal{H}$ of θ_0, condition (ii) of Theorem 4.2 is satisfied, while condition (iii) holds for instance when the functional V satisfies $V(w,\lambda)=V(w\,I_{\lambda^{-1}(]0,1])},\lambda)$ for all pairs of functions $(w,\lambda)\in\mathcal{W}\times\Lambda$. That is, Theorem 4.2 implies the (strong) consistency of all sequences of estimators δ_n resulting from likelihood decision criteria with the property that each evaluation $V(w_d,\lambda)$ does not depend on the loss associated with values of θ with zero likelihood.

The three examples of likelihood decision criteria explicitly considered in Section 3 have this property, and for each $n\in\mathbb{N}$, they lead to uniquely defined likelihood decision functions $\delta_n:(x_1,\dots,x_n)\mapsto\kappa_n\,x_{(n)}$. Therefore, Theorem 4.2 implies $\lim_{n\to\infty}\kappa_n=1$. In fact, for the MLD criterion $\kappa_n=1$ holds for all $n\in\mathbb{N}$, while $\kappa_n=\kappa(\alpha\,n)$ and $\kappa_n=\kappa'(\sqrt[n]{\beta})$ hold for the MPL criterion with exponent α and the LRM criterion with threshold β, respectively, where $\kappa:\mathbb{R}_{>0}\to\,]1,2[$ and $\kappa':\,]0,1[\,\to\,]1,2[$ are decreasing bijections. More precisely, $\kappa':y\mapsto{2}/{(y+1)}$, while κ assigns to each $y\in\mathbb{R}_{>0}$ the unique solution $s>1$ of the equation $(s-1)\,s^y={y^y}/{(y+1)^{y+1}}$.

4.3. *Efficiency.* Stronger assumptions about the statistical model, the loss function, and the likelihood decision criterion allow asymptotic properties stronger than consistency for the sequences of likelihood decision functions. For example, in a parametric estimation problem, the following theorem gives simple sufficient conditions for a sequence of likelihood decision functions to be an asymptotically efficient sequence of estimators. Its proof uses the result (strictly related to the Bernstein–von Mises theorem) that, under some regularity conditions, the likelihood function tends to a normal density around the maximum likelihood estimate (see for example Brenner, Fraser and McDunnough, 1982; Fraser and McDunnough, 1984; Schervish, 1995, Subsection 7.4.2; van der Vaart, 1998, Section 10.2). When a continuous prior likelihood function taking only positive values is used, the theorem still holds. For simplicity, its statement is restricted to the estimation of the natural parameter of a minimal regular exponential family (see for example Brown, 1986) under a power loss function, and to the three examples of likelihood decision criteria explicitly considered in Section 3 (for a version of the theorem with weaker, but more complex assumptions see Cattaneo, 2007, Subsection 5.1.1). An example of a likelihood decision criterion for which the result does not hold is the minimin version of the LRM criterion, described by the functional $(w, \lambda) \mapsto \inf_{\theta \in \Theta \,:\, \lambda(\theta) > \beta} w(\theta)$ (for some threshold $\beta \in \,]0, 1[$).

Theorem 4.3. *Let the sequence of random variables X_n (with $n \in \mathbb{N}$ and $\mathcal{X}_n = \mathcal{X}$) be independent and identically distributed according to a minimal regular exponential family with natural parameter space $\Theta \subseteq \mathbb{R}^k$. Let W be the loss function $(\theta, d) \mapsto |\theta - d|^\gamma$, where $\mathcal{D} = \Theta$ and $\gamma \in \mathbb{R}_{>0}$. If the sequence of decision functions $\delta_n : \mathcal{X}^n \to \Theta$ (with $n \in \mathbb{N}$) is optimal according to the MPL criterion (for some exponent $\alpha \in \mathbb{R}_{>0}$), the LRM criterion (for some threshold $\beta \in \,]0, 1[$), or the MLD criterion, then it is asymptotically efficient.*

Example 4.4 (normal distribution). *Let the sequence of random variables X_n (with $n \in \mathbb{N}$ and $\mathcal{X}_n = \mathbb{R}$) be independent and normally distributed with expected value θ and variance 1, where $\Theta = \mathbb{R}$. Consider the problem of estimating θ with the power loss function $W : (\theta, d) \mapsto |\theta - d|^\gamma$, where $\mathcal{D} = \mathbb{R}$ and $\gamma \in \mathbb{R}_{>0}$. For each $n \in \mathbb{N}$, let $\bar{X}_n$ denote the mean of the sample $X_1, \ldots, X_n$. From (P2) with the reflection with respect to $\bar{X}_n$ as bijection $b : \mathbb{R} \to \mathbb{R}$ it follows that for each $n \in \mathbb{N}$, when a likelihood decision function $\delta_n : \mathbb{R}^n \to \mathbb{R}$ is uniquely defined, it satisfies $\delta_n(X_1, \ldots, X_n) = \bar{X}_n$. This holds in particular for the likelihood decision functions resulting from the MPL, LRM, and MLD criteria (independently of the exponent α and the threshold β), which is in accordance with Theorem 4.3, since the sequence of estimators $\bar{X}_n$ is asymptotically efficient.*

Asymptotic efficiency is not necessarily a desirable property when the loss function is asymmetric. Consider for instance the so-called pinball (or check) loss function $W : (\theta, d) \mapsto (\theta - d) \left(\tau - I_{]\theta, \infty[}(d)\right)$, where $\tau \in \,]0, 1[$. This loss *function penalizes the overestimation of θ more than its underestimation when $\tau < 1/2$, and vice versa when $\tau > 1/2$. For each $n \in \mathbb{N}$, since the mean $\bar{X}_n$ is a sufficient statistic of $X_1, \dots, X_n$ for θ, and the decision problem is location invariant, when a likelihood decision function $\delta_n : \mathbb{R}^n \to \mathbb{R}$ is uniquely defined, it satisfies $\delta_n(X_1, \dots, X_n) = \bar{X}_n + \kappa_n$ for some constant $\kappa_n \in \mathbb{R}$. More generally, for each likelihood decision criterion an optimal sequence of decision functions of the form $\delta_n(X_1, \dots, X_n) = \bar{X}_n + \kappa_n$ always exists. Such a sequence of estimators is asymptotically efficient if and only if $\lim_{n\to\infty} \sqrt{n}\,\kappa_n = 0$. However, when $\tau \neq 1/2$, a sequence of estimators with $\lim_{n\to\infty} \sqrt{n}\,\kappa_n \neq 0$ can have expected loss up to $\exp(z_\tau^2/2)$ times smaller than that of $\bar{X}_n$ (where z_τ denotes the τ-quantile of the standard normal distribution), independently of θ and n.*

In particular, if the likelihood decision function $\delta_1 : x_1 \mapsto x_1 + \kappa_1$ is uniquely defined, and the likelihood decision criterion is described by a functional V such that $V(c\,w, \lambda) = c\,V(w, \lambda)$ for all pairs of functions $(w, \lambda) \in \mathcal{W} \times \Lambda$ and all constants $c \in \mathbb{R}_{>0}$ (that is, the evaluation of the loss is scale equivariant), then $\delta_n : (x_1, \dots, x_n) \mapsto \bar{x}_n + \kappa_1/\sqrt{n}$ is the uniquely defined likelihood decision function for each $n \in \mathbb{N}$. This follows from (P2) with the scaling by $\sqrt{n}$ as bijection $b : \mathbb{R} \to \mathbb{R}$, and is true in particular for the likelihood decision functions resulting from the MPL, LRM, and MLD criteria. More precisely, $\kappa_1 = 0$ and $\kappa_1 = \sqrt{-2\ln\beta}\,(2\tau - 1)$ hold for the MLD criterion and LRM criterion with threshold β, respectively, while $\kappa_1 = s/\sqrt{\alpha}$ holds for the MPL criterion with exponent α, where s is the unique real solution of the equation $1 + s/2\,(s - \sqrt{s^2+4}) = (1/\tau - 1)\exp(s/2\,\sqrt{s^2+4})$. Therefore, the sequence of estimators $\bar{X}_n$ resulting from the MLD criterion is asymptotically efficient, but for example when $\tau = 1/10$, the estimators resulting from the MPL criterion with $\alpha = 1$ and the LRM criterion with $\beta = 1/2$ have expected losses approximately 2.21 and 2.13 times smaller than that of $\bar{X}_n$, respectively (independently of θ and n).

5. Conclusion. In the present paper, the likelihood approach to statistics is extended and unified by the concept of likelihood decision function. Such a decision function is obtained by a post-data evaluation of the possible decisions on the basis of the likelihood function (interpreted as a description of uncertain knowledge about the parameters of the statistical model, and possibly including prior information). Besides the conceptual and computational advantage of being based on post-data evaluations, likelihood decision functions have several invariance properties, and also (under regularity conditions) asymptotic properties such as consistency and efficiency. Moreover, in the likelihood approach to decision making, prior knowledge about the parameters is not needed, and some special cases of likelihood decision functions (such as maximum likelihood estimators and likelihood ratio tests) are among the most successful statistical methods.

Future work includes a detailed analysis of the decision-theoretic properties characterizing the MPL criterion (see Cattaneo, 2007, Sections 3.1 and 4.1), in connection with the theories of risk measurement (see for instance Föllmer and Schied, 2002; Artzner et al., 1999) and of nonadditive measures and integrals (see for example Cattaneo, 2007, Chapter 2; Denneberg, 1994; Shilkret, 1971). Applications of the likelihood approach to decision making will also be further developed, in particular in the fields of robust statistics (see Cattaneo and Wiencierz, 2012) and probabilistic graphical models (see Cattaneo, 2010; Antonucci, Cattaneo and Corani, 2012).

APPENDIX

Proofs of Theorems 4.1 and 4.2. Theorem 4.1 is a special case of Theorem 4.2, so it suffices to prove the latter. Define $i_0 = \inf_{d \in \mathcal{D}} W(\theta_0, d)$, and choose an $\varepsilon \in \,]0, {}^{(c-i_0)}/_{6}[$. Let $d' \in \mathcal{D}$ be a decision such that $w_{d'}(\theta_0) < i_0 + \varepsilon$. Condition (ii) implies that there is a neighborhood $\mathcal{H}'$ of θ_0 such that $\mathcal{H}' \subseteq \mathcal{H}$ and $|w_d(\theta') - w_d(\theta_0)| < \varepsilon$ for all $\theta' \in \mathcal{H}'$ and all $d \in \mathcal{D}$ with $\inf_{\theta \in \mathcal{H}} w_d(\theta) < c$. The assumptions of Theorem 4.2 ensure that there is an $m \in \mathbb{N}$ such that P_{θ_0}-a.s. the following five properties hold for sufficiently large n:

(a) the likelihood function $\lambda_{(X_1,\ldots,X_n)} \in \Lambda$ is well-defined (this is part of condition (i)),

(b) $V(w_{\delta_n(X_1,\ldots,X_n)}, \lambda_{(X_1,\ldots,X_n)}) < V(w_{d'}, \lambda_{(X_1,\ldots,X_n)}) + \varepsilon$ (as implied by the optimality of the sequence δ_n),

(c) $V(w_{d'}, \lambda_{(X_1,\ldots,X_n)}) < V(w_{d'} \wedge m, \lambda_{(X_1,\ldots,X_n)}) + \varepsilon$ (this is a consequence of condition (iii)),

(d) $V\big((w_{d'}(\theta_0) + \varepsilon)\, I_{\mathcal{H}'} + m\, I_{\Theta \setminus \mathcal{H}'}, \lambda_{(X_1,\ldots,X_n)}\big) < w_{d'}(\theta_0) + 2\,\varepsilon$ (as follows from (P3) and condition (i)),

(e) $V\big((i_0 + 6\,\varepsilon)\, I_{\mathcal{H}'}, \lambda_{(X_1,\ldots,X_n)}\big) > i_0 + 5\,\varepsilon$ (this is again a consequence of (P3) and condition (i)).

From the above choice of $\mathcal{H}'$ it follows that the (pointwise) inequality $w_{d'} \wedge m \leq (w_{d'}(\theta_0) + \varepsilon)\, I_{\mathcal{H}'} + m\, I_{\Theta \setminus \mathcal{H}'}$ is valid. Therefore, (P1) and the properties (a), (b), (c), and (d) imply that P_{θ_0}-a.s. the following result holds for sufficiently large n:

$$\begin{aligned}
&V(w_{\delta_n(X_1,\ldots,X_n)}, \lambda_{(X_1,\ldots,X_n)}) \\
&\quad < V(w_{d'}, \lambda_{(X_1,\ldots,X_n)}) + \varepsilon \\
&\quad < V(w_{d'} \wedge m, \lambda_{(X_1,\ldots,X_n)}) + 2\,\varepsilon \\
&\quad \leq V\big((w_{d'}(\theta_0) + \varepsilon)\, I_{\mathcal{H}'} + m\, I_{\Theta\setminus\mathcal{H}'}, \lambda_{(X_1,\ldots,X_n)}\big) + 2\,\varepsilon \\
&\quad < w_{d'}(\theta_0) + 4\,\varepsilon \\
&\quad < i_0 + 5\,\varepsilon
\end{aligned}$$

In order to complete the proof, it suffices to show that from this result and the properties (a) and (e) it follows that P_{θ_0}-a.s. the inequality $w_{\delta_n(X_1,\ldots,X_n)}(\theta_0) < i_0 + 7\,\varepsilon$ holds for sufficiently large n. In particular, it suffices to show that for any decision $d \in \mathcal{D}$ and any likelihood function $\lambda \in \Lambda$, the inequality $w_d(\theta_0) < i_0 + 7\,\varepsilon$ is implied by the inequalities $V(w_d, \lambda) < i_0 + 5\,\varepsilon$ and $V((i_0 + 6\,\varepsilon)\, I_{\mathcal{H}'}, \lambda) > i_0 + 5\,\varepsilon$.

This implication is a simple consequence of (P1) and the above choice of $\mathcal{H}'$, and can be proved as follows. First note that $\inf_{\theta\in\mathcal{H}'} w_d(\theta) < c$ holds, since otherwise (P1) would imply $V(w_d, \lambda) \geq V((i_0 + 6\,\varepsilon)\, I_{\mathcal{H}'}, \lambda) > i_0 + 5\,\varepsilon$. Now, from $\inf_{\theta\in\mathcal{H}'} w_d(\theta) < c$ and the above choice of $\mathcal{H}'$ it follows that the (pointwise) inequality $w_d > (w_d(\theta_0) - \varepsilon)\, I_{\mathcal{H}'}$ is valid. Therefore, the desired result $w_d(\theta_0) < i_0 + 7\,\varepsilon$ holds, because otherwise (P1) would imply $V(w_d, \lambda) \geq V((i_0 + 6\,\varepsilon)\, I_{\mathcal{H}'}, \lambda) > i_0 + 5\,\varepsilon$.

Proof of Theorem 4.3. For each $n \in \mathbb{N}$, let $\mathbf{X}_n$ denote the random variable $(X_1, \ldots, X_n)$, and define the function $\theta_n : \mathcal{X}^n \to \Theta$ as follows: $\theta_n(\mathbf{x}_n) = \hat{\theta}$ for all $\mathbf{x}_n \in \mathcal{X}^n$ such that the likelihood function $\lambda_{\mathbf{x}_n} \in \Lambda$ and the maximum likelihood estimate $\hat{\theta}$ are well-defined, and $\theta_n(\mathbf{x}_n) = \theta_0$ otherwise (with $\theta_0 \in \Theta$ arbitrary). Under each P_θ (with $\theta \in \Theta$), the probability that the likelihood function $\lambda_{\mathbf{X}_n} \in \Lambda$ and the maximum likelihood estimate $\hat{\theta}$ are well-defined tends to 1 as n tends to ∞, and

$$\sqrt{n}\,(\theta_n(\mathbf{X}_n) - \theta) \xrightarrow{d} \mathcal{N}_k\big(0, I(\theta)^{-1}\big)\,,$$

where $I(\theta)$ is the Fisher information matrix (see for example Schervish, 1995, Theorem 7.57). Hence, in order to prove the theorem it suffices to show that

$$\sqrt{n}\,(\delta_n(\mathbf{X}_n) - \theta_n(\mathbf{X}_n)) \xrightarrow{P_\theta} 0$$

for all $\theta \in \Theta$ (see for instance van der Vaart, 1998, Theorem 2.7 (iv)).

For each $n \in \mathbb{N}$, define the function $\lambda_n : \mathcal{X}^n \times \mathbb{R}^k \to [0, 1]$ as follows: $\lambda_n(\mathbf{x}_n, \tau) = \lambda_{\mathbf{x}_n}\big(\theta_n(\mathbf{x}_n) + (1/\sqrt{n})\, I(\theta_n(\mathbf{x}_n))^{-1/2}\, \tau\big)$ for all $(\mathbf{x}_n, \tau) \in \mathcal{X}^n \times \mathbb{R}^k$

such that the likelihood function $\lambda_{\mathbf{x}_n} \in \Lambda$ is well-defined and its argument $\theta_n(\mathbf{x}_n) + ({}^1\!/\!\sqrt{n})\, I(\theta_n(\mathbf{x}_n))^{-1/2}\, \tau$ lies in Θ, and $\lambda_n(\mathbf{x}_n, \tau) = 0$ otherwise. The following result (strictly related to the Bernstein–von Mises theorem) is implied by Theorem 7.89 of Schervish (1995), whose regularity conditions can be easily checked thanks to the analytic properties of exponential families (see for example Brown, 1986, Chapter 2):

$$\sup_{\tau \in \mathbb{R}^k \,:\, |\tau| < t} \left| \lambda_n(\mathbf{X}_n, \tau) - \exp\!\left(-\tfrac{1}{2}\, |\tau|^2\right) \right| \xrightarrow{P_\theta} 0$$

for all $t \in \mathbb{R}_{>0}$ and all $\theta \in \Theta$.

Assume that the sequence δ_n is optimal according to the MPL criterion with exponent $\alpha \in \mathbb{R}_{>0}$. For each $n \in \mathbb{N}$, define the function $v_n : \mathcal{X}^n \times \mathbb{R}^k \to \overline{\mathbb{R}}_{\geq 0}$ by $v_n(\mathbf{x}_n, \zeta) = n^{-\gamma/2} \sup_{\tau \in \mathbb{R}^k} \left| I(\theta_n(\mathbf{x}_n))^{-1/2}\, \tau - \zeta \right|^\gamma \lambda_n(\mathbf{x}_n, \tau)^\alpha$. Then

$$V_{MPL,\alpha}(w_d, \lambda_{\mathbf{x}_n}) = \sup_{\theta \in \Theta} |\theta - d|^\gamma\, \lambda_{\mathbf{x}_n}(\theta)^\alpha = v_n\!\left(\mathbf{x}_n, \sqrt{n}\, (d - \theta_n(\mathbf{x}_n))\right)$$

for all $n \in \mathbb{N}$, all $d \in \Theta$, and all $\mathbf{x}_n \in \mathcal{X}^n$ such that the likelihood function $\lambda_{\mathbf{x}_n} \in \Lambda$ is well-defined.

For each $\theta \in \Theta$, let $v_\theta : \mathbb{R}^k \to \mathbb{R}_{\geq 0}$ be the function defined by $v_\theta(\zeta) = \sup_{\tau \in \mathbb{R}^k} \left| I(\theta)^{-1/2}\, \tau - \zeta \right|^\gamma \exp\!\left(-\tfrac{\alpha}{2}\, |\tau|^2\right)$. Since the function $\tau \mapsto \lambda_n(\mathbf{x}_n, \tau)$ is logarithmically concave for all $\mathbf{x}_n \in \mathcal{X}^n$, and the Fisher information matrix is a continuous function of θ,

$$\sup_{\zeta \in \mathbb{R}^k \,:\, |\zeta| < z} \left| n^{\gamma/2}\, v_n(\mathbf{X}_n, \zeta) - v_\theta(\zeta) \right| \xrightarrow{P_\theta} 0$$

holds for all $z \in \mathbb{R}_{>0}$ and all $\theta \in \Theta$. For each $\mathbf{x}_n \in \mathcal{X}^n$, the function $\zeta \mapsto n^{\gamma/2}\, v_n(\mathbf{x}_n, \zeta)$ is quasiconvex, and for each $\theta \in \Theta$, the function v_θ is strictly quasiconvex with a unique minimum at $\zeta = 0$. Therefore, for each $\varepsilon \in \mathbb{R}_{>0}$ there is an $\eta \in \mathbb{R}_{>0}$ such that under each P_θ (with $\theta \in \Theta$), the probability that there is a $\zeta \in \mathbb{R}^k$ with $|\zeta| > \varepsilon$ and $n^{\gamma/2}\, (v_n(\mathbf{X}_n, \zeta) - v_n(\mathbf{X}_n, 0)) < \eta$ tends to 0 as n tends to ∞. Hence,

$$\sup_{\zeta \in \mathbb{R}^k \,:\, v_n(\mathbf{X}_n, \zeta) < v_n(\mathbf{X}_n, 0) + 2^{-n}} |\zeta| \xrightarrow{P_\theta} 0$$

for all $\theta \in \Theta$, and this proves the desired result for the MPL criterion. The proofs for the LRM and MLD criteria are analogous.

REFERENCES

Antonucci, A., Cattaneo, M. and Corani, G. (2012). Likelihood-based robust classification with Bayesian networks. In *Advances in Computational Intelligence, part 3* (S. Greco, B. Bouchon-Meunier, G. Coletti, M. Fedrizzi, B. Matarazzo and R. R. Yager, eds.) 491–500. Springer.

Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D. (1999). Coherent measures of risk. *Math. Finance* **9** 203–228.

Azzalini, A. (1996). *Statistical Inference: Based on the Likelihood.* Chapman & Hall.

Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38** 303–324.

Barnard, G. A. (1949). Statistical inference. *J. Roy. Statist. Soc. Ser. B* **11** 115–149.

Barnard, G. A. (1967). The use of the likelihood function in statistical practice. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. I: Statistics* (L. M. Le Cam and J. Neyman, eds.) 27–40. University of California Press.

Barnard, G. A. (1972). The logic of statistical inference. *British J. Philos. Sci.* **23** 123–132.

Barnard, G. A., Jenkins, G. M. and Winsten, C. B. (1962). Likelihood inference and time series. *J. Roy. Statist. Soc. Ser. A* **125** 321–372.

Barnard, G. A. and Sprott, D. A. (1983). Likelihood. In *Encyclopedia of Statistical Sciences, vol. 4* (S. Kotz, N. L. Johnson and C. B. Read, eds.) 639–644. Wiley.

Basu, D. (1975). Statistical information and likelihood. *Sankhyā Ser. A* **37** 1–71.

Berger, J. (1985a). The frequentist viewpoint and conditioning. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, vol. I* (L. M. Le Cam and R. A. Olshen, eds.) 15–44. Wadsworth.

Berger, J. O. (1985b). *Statistical Decision Theory and Bayesian Analysis*, second ed. Springer.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*, second ed. Institute of Mathematical Statistics.

Birnbaum, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.* **57** 269–326.

Board, J. L. G. and Sutcliffe, C. M. S. (1994). Estimation methods in portfolio selection and the effectiveness of short sales restrictions: UK evidence. *Management Sci.* **40** 516–534.

Brenner, D., Fraser, D. A. S. and McDunnough, P. (1982). On asymptotic normality of likelihood and conditional analysis. *Canad. J. Statist.* **10** 163–172.

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

Cattaneo, M. (2005). Likelihood-based statistical decisions. In *ISIPTA '05, Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications* (F. G. Cozman, R. Nau and T. Seidenfeld, eds.) 107–116. SIPTA.

Cattaneo, M. (2007). *Statistical Decisions Based Directly on the Likelihood Function.* PhD thesis, ETH Zurich, `doi:10.3929/ethz-a-005463829`.

Cattaneo, M. (2010). Likelihood-based inference for probabilistic graphical models: Some preliminary results. In *PGM 2010, Proceedings of the Fifth European Workshop on Probabilistic Graphical Models* (P. Myllymäki, T. Roos and T. Jaakkola, eds.) 57–64. HIIT Publications.

Cattaneo, M. and Wiencierz, A. (2012). Likelihood-based Imprecise Regression. *Internat. J. Approx. Reason.* in press, `doi:10.1016/j.ijar.2012.06.010`.

Denneberg, D. (1994). *Non-Additive Measure and Integral.* Kluwer.

Diehl, H. and Sprott, D. A. (1965). Die Likelihoodfunktion und ihre Verwendung beim statistischen Schluß [The likelihood function and its use in statistical inference (in German with English summary)]. *Statist. Hefte* **6** 112–134.

Edwards, A. W. F. (1969). Statistical methods in scientific inference. *Nature* **222** 1233–1237.

Edwards, A. W. F. (1970). Likelihood. *Nature* **227** 92–92.

Edwards, A. W. F. (1992). *Likelihood*, expanded ed. Johns Hopkins University Press.

Evans, M. J., Fraser, D. A. S. and Monette, G. (1986). On principles and arguments to likelihood. *Canad. J. Statist.* **14** 181–199.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*, third ed. Hafner Press.

Föllmer, H. and Schied, A. (2002). Convex measures of risk and trading constraints. *Finance Stoch.* **6** 429–447.

Fraser, D. A. S. and McDunnough, P. (1984). Further remarks on asymptotic normality of likelihood and conditional analyses. *Canad. J. Statist.* **12** 183–190.

Giang, P. H. and Shenoy, P. P. (2005). Decision making on the sole basis of statistical likelihood. *Artificial Intelligence* **165** 137–163.

Goutis, C. and Casella, G. (1995). Frequentist post-data inference. *Internat. Statist. Rev.* **63** 325–344.

Hacking, I. (1964). On the foundations of statistics. *British J. Philos. Sci.* **15** 1–26.

Hills, M. (2005). Likelihood. In *Encyclopedia of Biostatistics, vol. 4,* second ed. (P. Armitage and T. Colton, eds.) 2775–2779. Wiley.

Hudson, D. J. (1971). Interval estimation from the likelihood function. *J. Roy. Statist. Soc. Ser. B* **33** 256–262.

Joshi, V. M. (1983). Likelihood principle. In *Encyclopedia of Statistical Sciences, vol. 4* (S. Kotz, N. L. Johnson and C. B. Read, eds.) 644–647. Wiley.

Kalbfleisch, J. G. (1985). *Probability and Statistical Inference, vol. 2: Statistical Inference*, second ed. Springer.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *J. Amer. Statist. Assoc.* **72** 789–827.

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

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, third ed. Springer.

Leonard, T. (1978). Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc. Ser. B* **40** 113–146.

Levy, H. and Sarnat, M. (1970). International diversification of investment portfolios. *Amer. Econ. Rev.* **60** 668–675.

Lindsey, J. K. (1996). *Parametric Statistical Inference.* Oxford University Press.

Lindsey, J. K. (1999). Some statistical heresies. *The Statistician* **48** 1–40.

Lindsey, J. K. (2005). Likelihood principle. In *Encyclopedia of Biostatistics, vol. 4,* second ed. (P. Armitage and T. Colton, eds.) 2779–2782. Wiley.

Markowitz, H. (1952). Portfolio selection. *J. Finance* **7** 77–91.

Montoya, J. A., Díaz-Francés, E. and Sprott, D. A. (2009). On a criticism of the profile likelihood function. *Statist. Papers* **50** 195–202.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press.

Portnoy, S. (1971). Formal Bayes estimation with application to a random effects model. *Ann. Math. Statist.* **42** 1379–1402.

Reid, N. (2000). Likelihood. *J. Amer. Statist. Assoc.* **95** 1335–1340.

Robinson, G. K. (1979). Conditional properties of statistical procedures. *Ann. Statist.* **7** 742–755.

Royall, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm.* Chapman & Hall.

Schervish, M. J. (1995). *Theory of Statistics.* Springer.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components.*

Wiley.

Shilkret, N. (1971). Maxitive measure and integration. *Indag. Math.* **33** 109–116.

Sprott, D. A. (2000). *Statistical Inference in Science.* Springer.

Thompson, W. A. Jr. (1962). The problem of negative estimates of variance components. *Ann. Math. Statist.* **33** 273–289.

Tiao, G. C. and Tan, W. Y. (1965). Bayesian analysis of random-effect models in the analysis of variance. I. Posterior distribution of variance-components. *Biometrika* **52** 37–53.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics* **20** 595–601.

Wald, A. (1950). *Statistical Decision Functions.* Wiley.

Institut für Statistik
Ludwig-Maximilians-Universität München
Ludwigstrasse 33
80539 München, Germany
E-mail: cattaneo@stat.uni-muenchen.de