Likelihood-Based Statistical Decisions

Marco Cattaneo
Seminar for Statistics
ETH Zürich, Switzerland

December 2, 2005

Statistical Decision Problem

Let $\mathcal{P}$ be a set of **statistical models**:

- $\mathcal{P}$ is a set of probability measures on a measurable space $(\Omega, \mathcal{A})$;
- no structure is imposed on $\mathcal{P}$.

Statistical Decision Problem

Let $\mathcal{P}$ be a set of **statistical models**:

- $\mathcal{P}$ is a set of probability measures on a measurable space $(\Omega, \mathcal{A})$;
- no structure is imposed on $\mathcal{P}$.

The observation of an event $A \in \mathcal{A}$ gives us some information about the relative plausibility of the models in $\mathcal{P}$. We want to use this information to select a decision $d \in \mathcal{D}$ (the set of possible decisions).

The **loss function** $L : \mathcal{P} \times \mathcal{D} \to [0, \infty)$ is assumed to summarize all aspects of the possible decisions that should be considered in their evaluation: $L(P, d)$ is the loss we would incur, according to the model P, by making the decision d.

Likelihood Function

The **likelihood function** $lik : \mathcal{P} \to [0, 1]$ based on the observation $A \in \mathcal{A}$:

- is defined by $lik(P) = P(A)$;
- measures the relative plausibility of the models $P \in \mathcal{P}$, on the basis of the observation A alone;
- is not calibrated: proportional likelihood functions contain the same statistical information.

Likelihood Function

The **likelihood function** $lik : \mathcal{P} \rightarrow [0,1]$ based on the observation $A \in \mathcal{A}$:

- is defined by $lik(P) = P(A)$;
- measures the relative plausibility of the models $P \in \mathcal{P}$, on the basis of the observation A alone;
- is not calibrated: proportional likelihood functions contain the same statistical information.

If we observe the realization $A = \{X = x\}$ of a continuous random object X, we have $P(A) = 0$ for all $P \in \mathcal{P}$. But in reality, because of the finite precision of any observation, we only know that X lies in a neighborhood N of x; thus $lik(P) = P\{X \in N\}$.

If f_P is a density of X under the model P, it can be useful to consider the approximation $P\{X \in N\} \approx \delta\, f_P(x)$. If this holds for all $P \in \mathcal{P}$, we obtain an approximate likelihood function, which is proportional to the function $P \mapsto f_P(x)$.

Pre-Data Evaluation (Classical Decision Theory)

If the observation A is considered as a particular realization $A = \{X = x_A\}$ of the random object $X : \Omega \to \mathcal{X}$, a **decision function** $\delta : \mathcal{X} \to \mathcal{D}$ can be evaluated on the basis of the **pre-data expected loss**

$$R_\delta : P \mapsto E_P[L(P, \delta(X))] \; .$$

Pre-Data Evaluation (Classical Decision Theory)

If the observation A is considered as a particular realization $A = \{X = x_A\}$ of the random object $X : \Omega \to \mathcal{X}$, a **decision function** $\delta : \mathcal{X} \to \mathcal{D}$ can be evaluated on the basis of the **pre-data expected loss**

$$\boxed{R_\delta : P \mapsto E_P[L(P, \delta(X))]} \, .$$

To select a decision $d \in \mathcal{D}$, we must select a decision function δ and apply it to the observed realization x_A of X, obtaining the decision $d = \delta(x_A)$.

To select a decision function δ, we must adopt a decision criterion, for example the **minimax criterion**: $\boxed{\text{minimize } \sup R_\delta}$.

Post-Data (Conditional) Evaluation

A **conditional decision criterion** (based on post-data evaluation):

- depends only on the observation $A = \{X = x_A\}$, not on the other possible realizations of X;
- selects a decision $d \in \mathcal{D}$, not a decision function $\delta : \mathcal{X} \to \mathcal{D}$;

Post-Data (Conditional) Evaluation

A **conditional decision criterion** (based on post-data evaluation):

- depends only on the observation $A = \{X = x_A\}$, not on the other possible realizations of X;
- selects a decision $d \in \mathcal{D}$, not a decision function $\delta : \mathcal{X} \to \mathcal{D}$;
- can only be based on the likelihood function lik, the loss function L and the prior information about the models in $\mathcal{P}$ (if available).

Bayesian Decision Theory

The (subjective) Bayesian conditional decision criterion consists in minimizing the **posterior expected loss** $E_{\pi|A}[L(P,d)]$.

Bayesian Decision Theory

The (subjective) Bayesian conditional decision criterion consists in minimizing the **posterior expected loss** $E_{\pi|A}[L(P,d)]$.

The Bayesian theory has many important properties, whereas its main problem is the need of a probability measure π on $\mathcal{P}$ describing the prior information (to be combined with the likelihood function lik to obtain the posterior probability measure $\pi|A$).

In fact, **complete ignorance** can not be described by a probability measure. In other words, we need prior information even when it is not available.

Bayesian Decision Theory

The (subjective) Bayesian conditional decision criterion consists in minimizing the **posterior expected loss** $\boxed{E_{\pi|A}[L(P,d)]}$.

The Bayesian theory has many important properties, whereas its main problem is the need of a probability measure π on $\mathcal{P}$ describing the prior information (to be combined with the likelihood function lik to obtain the posterior probability measure $\pi|A$).

In fact, **complete ignorance** can not be described by a probability measure. In other words, we need prior information even when it is not available.

Many generalizations of the Bayesian theory allowing an **imprecise prior** have been proposed: most of them are formally equivalent with the choice of some set of prior probability measures on $\mathcal{P}$.

Although these approaches allow some imprecision in the prior, they have problems with the representation of ignorance. In every situation, some compromise between the ignorance in the prior and the power of the conclusions must be reached.

Likelihood-Based Statistical Inference

Some of the most appreciated inference methods are based directly on the likelihood function (without a prior probability measure): in particular, the **maximum likelihood estimator** and the tests and confidence regions based on the **likelihood ratio statistic** $LR(\mathcal{H}) = \sup_{\mathcal{H}} \overline{lik}$ (for $\mathcal{H} \subseteq \mathcal{P}$), where $\overline{lik}$ is the **normed likelihood function** defined by $\overline{lik} \propto lik$ and $\sup \overline{lik} = 1$.

Likelihood-Based Statistical Inference

Some of the most appreciated inference methods are based directly on the likelihood function (without a prior probability measure): in particular, the **maximum likelihood estimator** and the tests and confidence regions based on the **likelihood ratio statistic** $LR(\mathcal{H}) = \sup_{\mathcal{H}} \overline{lik}$ (for $\mathcal{H} \subseteq \mathcal{P}$), where $\overline{lik}$ is the **normed likelihood function** defined by $\overline{lik} \propto lik$ and $\sup \overline{lik} = 1$.

Since inference problems can be seen as particular cases of decision problems, we can try to generalize the likelihood-based inference methods to a likelihood-based (conditional) decision criterion.

Likelihood-Based Statistical Inference

Some of the most appreciated inference methods are based directly on the likelihood function (without a prior probability measure): in particular, the **maximum likelihood estimator** and the tests and confidence regions based on the **likelihood ratio statistic** $LR(\mathcal{H}) = \sup_{\mathcal{H}} \overline{lik}$ (for $\mathcal{H} \subseteq \mathcal{P}$), where $\overline{lik}$ is the **normed likelihood function** defined by $\overline{lik} \propto lik$ and $\sup \overline{lik} = 1$.

Since inference problems can be seen as particular cases of decision problems, we can try to generalize the likelihood-based inference methods to a likelihood-based (conditional) decision criterion.

A generalization of the maximum likelihood estimator often used in practice is the following: select the decision $d \in \mathcal{D}$ which is optimal under the model $\widehat{P}_{ML} = \arg\max lik$ (if it exists). That is: minimize $L(\widehat{P}_{ML}, d)$.

Likelihood-Based Decision Criteria

The function $l_d = L(\cdot, d) : P \mapsto L(P, d)$ summarizes all aspects of d that should be considered in its evaluation.

A straightforward way to obtain a likelihood-based decision criterion is to associate to every decision $d \in \mathcal{D}$ a quantity $F(l_d, \overline{lik}) \in [0, \infty]$ (an evaluation of l_d on the basis of lik) and to select d by minimizing $F(l_d, \overline{lik})$.

Likelihood-Based Decision Criteria

The function $l_d = L(\cdot, d) : P \mapsto L(P, d)$ summarizes all aspects of d that should be considered in its evaluation.

A straightforward way to obtain a likelihood-based decision criterion is to associate to every decision $d \in \mathcal{D}$ a quantity $F(l_d, \overline{lik}) \in [0, \infty]$ (an evaluation of l_d on the basis of lik) and to select d by minimizing $F(l_d, \overline{lik})$.

For instance, the **maximum likelihood (ML) decision criterion** is based on

$$F_{ML}(l_d, \overline{lik}) = L(\widehat{P}_{ML}, d) = l_d(\arg\max \overline{lik}) \quad \text{(if } \widehat{P}_{ML} \text{ exists).}$$

Likelihood-Based Decision Criteria

The function $l_d = L(\cdot, d) : P \mapsto L(P, d)$ summarizes all aspects of d that should be considered in its evaluation.

A straightforward way to obtain a likelihood-based decision criterion is to associate to every decision $d \in \mathcal{D}$ a quantity $F(l_d, \overline{lik}) \in [0, \infty]$ (an evaluation of l_d on the basis of lik) and to select d by minimizing $F(l_d, \overline{lik})$.

For instance, the **maximum likelihood (ML) decision criterion** is based on

$$F_{ML}(l_d, \overline{lik}) = L(\widehat{P}_{ML}, d) = l_d(\arg\max \overline{lik}) \quad \text{(if } \widehat{P}_{ML} \text{ exists).}$$

A decision criterion obtained in this way (for a particular F) is obviously **parametrization invariant**.

Prior Information

Let lik_A and lik_B be the likelihood functions based on the observation of A and of B, respectively. If A and B are independent (under all models $P \in \mathcal{P}$), the likelihood function based on the observation of $A \cap B$ is $lik_A\, lik_B$.

Prior Information

Let lik_A and lik_B be the likelihood functions based on the observation of A and of B, respectively. If A and B are independent (under all models $P \in \mathcal{P}$), the likelihood function based on the observation of $A \cap B$ is $lik_A \, lik_B$.

It is thus natural to allow the expression of the prior information about the models in $\mathcal{P}$ in terms of a **prior likelihood function** lik_{prior}, which can be multiplied with lik_A when A is observed.

A constant likelihood function describes **complete ignorance**: in fact, using it as a prior is equivalent to using no prior.

Prior Information

Let lik_A and lik_B be the likelihood functions based on the observation of A and of B, respectively. If A and B are independent (under all models $P \in \mathcal{P}$), the likelihood function based on the observation of $A \cap B$ is $lik_A \, lik_B$.

It is thus natural to allow the expression of the prior information about the models in $\mathcal{P}$ in terms of a **prior likelihood function** lik_{prior}, which can be multiplied with lik_A when A is observed.

A constant likelihood function describes **complete ignorance**: in fact, using it as a prior is equivalent to using no prior.

We can therefore assume that the prior information (if available) is contained in lik.

Necessary Properties of F

Since $F(l, \overline{lik})$ is a likelihood-based evaluation of the loss $l : \mathcal{P} \rightarrow [0, \infty)$ of a decision, the following properties can be considered necessary ($c \geq 0$ is a constant).

Necessary Properties of F

Since $F(l, \overline{lik})$ is a likelihood-based evaluation of the loss $l : \mathcal{P} \to [0, \infty)$ of a decision, the following properties can be considered necessary ($c \geq 0$ is a constant).

- **Conditionality**: $b : \mathcal{P} \xrightarrow{\text{bij}} \mathcal{P}' \;\Rightarrow\; F(l \circ b^{-1}, \overline{lik} \circ b^{-1}) = F(l, \overline{lik})$

From Conditionality follows in particular the **equivariance** of the decision functions obtained by the conditional application (to an invariant problem) of the decision criterion defined by F.

Necessary Properties of F

Since $F(l, \overline{lik})$ is a likelihood-based evaluation of the loss $l : \mathcal{P} \to [0, \infty)$ of a decision, the following properties can be considered necessary ($c \geq 0$ is a constant).

- **Conditionality**: $b : \mathcal{P} \overset{\text{bij}}{\to} \mathcal{P}' \;\Rightarrow\; F(l \circ b^{-1}, \overline{lik} \circ b^{-1}) = F(l, \overline{lik})$

From Conditionality follows in particular the **equivariance** of the decision functions obtained by the conditional application (to an invariant problem) of the decision criterion defined by F.

- **Monotonicity**: $l \leq l' \;\Rightarrow\; F(l, \overline{lik}) \leq F(l', \overline{lik})$

Necessary Properties of F

Since $F(l, \overline{lik})$ is a likelihood-based evaluation of the loss $l : \mathcal{P} \to [0, \infty)$ of a decision, the following properties can be considered necessary ($c \geq 0$ is a constant).

- **Conditionality**: $b : \mathcal{P} \overset{\text{bij}}{\to} \mathcal{P}' \;\Rightarrow\; F(l \circ b^{-1}, \overline{lik} \circ b^{-1}) = F(l, \overline{lik})$

From Conditionality follows in particular the **equivariance** of the decision functions obtained by the conditional application (to an invariant problem) of the decision criterion defined by F.

- **Monotonicity**: $l \leq l' \;\Rightarrow\; F(l, \overline{lik}) \leq F(l', \overline{lik})$
- **Calibration**: $F(c, \overline{lik}) = c$

Necessary Properties of F

Since $F(l, \overline{lik})$ is a likelihood-based evaluation of the loss $l : \mathcal{P} \rightarrow [0, \infty)$ of a decision, the following properties can be considered necessary ($c \geq 0$ is a constant).

- **Conditionality**: $b : \mathcal{P} \stackrel{\text{bij}}{\rightarrow} \mathcal{P}' \;\Rightarrow\; F(l \circ b^{-1}, \overline{lik} \circ b^{-1}) = F(l, \overline{lik})$

From Conditionality follows in particular the **equivariance** of the decision functions obtained by the conditional application (to an invariant problem) of the decision criterion defined by F.

- **Monotonicity**: $l \leq l' \;\Rightarrow\; F(l, \overline{lik}) \leq F(l', \overline{lik})$
- **Calibration**: $F(c, \overline{lik}) = c$

Usually in statistics no meaning is attached to the unit of the loss, hence we can assume that $F(c\,l, \overline{lik}) \leq F(c\,l', \overline{lik})$ when $F(l, \overline{lik}) \leq F(l', \overline{lik})$. This implies the following property.

- **Homogeneity**: $F(c\,l, \overline{lik}) = c\,F(l, \overline{lik})$

Nonadditive Measures and Integrals

A **normed monotonic measure** μ on a set $\mathcal{Q}$ is a function $\mu : 2^{\mathcal{Q}} \rightarrow [0, 1]$, with $\mu(\emptyset) = 0$, $\mu(\mathcal{Q}) = 1$ and $\mu(\mathcal{S}_1) \leq \mu(\mathcal{S}_2)$ if $\mathcal{S}_1 \subseteq \mathcal{S}_2$.

Nonadditive Measures and Integrals

A **normed monotonic measure** μ on a set $\mathcal{Q}$ is a function $\mu : 2^{\mathcal{Q}} \to [0, 1]$, with $\mu(\emptyset) = 0$, $\mu(\mathcal{Q}) = 1$ and $\mu(\mathcal{S}_1) \leq \mu(\mathcal{S}_2)$ if $\mathcal{S}_1 \subseteq \mathcal{S}_2$.

A **homogeneous integral** $\int l \, \mathrm{d}\mu$ of nonnegative functions $l : \mathcal{Q} \to [0, \infty)$ with respect to normed monotonic measures μ on $\mathcal{Q}$ must satisfy the following properties ($c \geq 0$ is a constant).

- **Invariance**: $b : \mathcal{Q} \overset{\text{bij}}{\to} \mathcal{Q}' \;\Rightarrow\; \int l \circ b^{-1} \, \mathrm{d}(\mu \circ b^{-1}) = \int l \, \mathrm{d}\mu$
- **Monotonicity**: $l \leq l' \;\Rightarrow\; \int l \, \mathrm{d}\mu \leq \int l' \, \mathrm{d}\mu$
- **Homogeneity**: $\int c \, l \, \mathrm{d}\mu = c \int l \, \mathrm{d}\mu$
- **Extension**: $\int I_{\mathcal{S}} \, \mathrm{d}\mu = \mu(\mathcal{S})$

Two Homogeneous Integrals

Choquet integral: $$\int l \, \mathrm{d}\mu = \int_0^\infty \mu\{l > x\} \, \mathrm{d}x$$

Shilkret integral: $$\int l \, \mathrm{d}\mu = \sup_{x>0} x \, \mu\{l > x\}$$

Two Homogeneous Integrals

Choquet integral: $$\int l \, \mathrm{d}\mu = \int_0^\infty \mu\{l > x\} \, \mathrm{d}x$$

Shilkret integral: $$\int l \, \mathrm{d}\mu = \sup_{x>0} x \, \mu\{l > x\}$$

The Shilkret integral is the homogeneous special case of the generalized Sugeno integral (fuzzy integral). These two integral satisfy the following additional properties.

- **Subadditivity**: $\int (l + l') \, \mathrm{d}\mu \leq \int l \, \mathrm{d}\mu + \int l' \, \mathrm{d}\mu$
- **Cumulativity**: $\int l \, \mathrm{d}\mu = G(\mu\{l > \cdot\})$

Likelihood Ratio Statistic

A **possibility measure** μ on $\mathcal{Q}$ is a normed monotonic measure which is completely maxitive: $\mu(\bigcup_{j \in J} \mathcal{S}_j) = \sup_{j \in J} \mu(\mathcal{S}_j)$. A possibility measure is uniquely determined by its **distribution function** $Q \mapsto \mu\{Q\}$ on $\mathcal{Q}$.

The likelihood ratio statistic LR (defined by $LR(\mathcal{H}) = \sup_{\mathcal{H}} \overline{lik}$) is the possibility measure on $\mathcal{P}$ with distribution function $\overline{lik}$.

Likelihood Ratio Statistic

A **possibility measure** μ on $\mathcal{Q}$ is a normed monotonic measure which is completely maxitive: $\mu(\bigcup_{j \in J} \mathcal{S}_j) = \sup_{j \in J} \mu(\mathcal{S}_j)$. A possibility measure is uniquely determined by its **distribution function** $Q \mapsto \mu\{Q\}$ on $\mathcal{Q}$.

The likelihood ratio statistic LR (defined by $LR(\mathcal{H}) = \sup_{\mathcal{H}} \overline{lik}$) is the possibility measure on $\mathcal{P}$ with distribution function $\overline{lik}$.

If F satisfies $F(I_{\mathcal{H}}, \overline{lik}) = f[LR(\mathcal{H})]$ for some function f on $[0,1]$, it can be written as a homogeneous integral: $F(l, \overline{lik}) = \int l \, \mathrm{d}(f \circ LR)$.

In fact, from the necessary properties it follows that f is nondecreasing with $f(0) = 0$ and $f(1) = 1$. Thus $f \circ LR$ is a normed monotonic measure on $\mathcal{P}$; it is a possibility measure if f is continuous.

Likelihood Ratio Statistic

A **possibility measure** μ on $\mathcal{Q}$ is a normed monotonic measure which is completely maxitive: $\mu(\bigcup_{j \in J} \mathcal{S}_j) = \sup_{j \in J} \mu(\mathcal{S}_j)$. A possibility measure is uniquely determined by its **distribution function** $Q \mapsto \mu\{Q\}$ on $\mathcal{Q}$.

The likelihood ratio statistic LR (defined by $LR(\mathcal{H}) = \sup_{\mathcal{H}} \overline{lik}$) is the possibility measure on $\mathcal{P}$ with distribution function $\overline{lik}$.

If F satisfies $F(I_{\mathcal{H}}, \overline{lik}) = f[LR(\mathcal{H})]$ for some function f on $[0,1]$, it can be written as a homogeneous integral: $F(l, \overline{lik}) = \int l \, \mathrm{d}(f \circ LR)$.

In fact, from the necessary properties it follows that f is nondecreasing with $f(0) = 0$ and $f(1) = 1$. Thus $f \circ LR$ is a normed monotonic measure on $\mathcal{P}$; it is a possibility measure if f is continuous.

For instance, a possible general definition of the ML decision criterion is $F_{ML}(l, \overline{lik}) = \int l \, \mathrm{d}(I_{\{1\}} \circ LR)$, for a cumulative homogeneous integral.

Generalization of Inference Methods

Consider the conditional decision criterion defined by the homogeneous integral $\int l \,\mathrm{d}(f \circ LR)$, where $f : [0,1] \to [0,1]$ is nondecreasing with $f(0) = 0$ and $f(1) = 1$.

Generalization of Inference Methods

Consider the conditional decision criterion defined by the homogeneous integral $\int l \,\mathrm{d}(f \circ LR)$, where $f : [0,1] \to [0,1]$ is nondecreasing with $f(0) = 0$ and $f(1) = 1$.

- In order to generalize the **maximum likelihood estimator**, f must be strictly increasing in 1, in the sense that $f(x) < 1$ if $x < 1$.

  In this case we have for example that for a finite $\mathcal{P} = \mathcal{D}$ and the simple loss function $L(P, d) = \begin{cases} 0 & \text{if } P = d \\ 1 & \text{if } P \neq d \end{cases}$, if $\widehat{P}_{ML}$ exists, $d = \widehat{P}_{ML}$ minimizes $\int l_d \,\mathrm{d}(f \circ LR)$ and will therefore be selected.

Generalization of Inference Methods

Consider the conditional decision criterion defined by the homogeneous integral $\int l \,\mathrm{d}(f \circ LR)$, where $f : [0,1] \to [0,1]$ is nondecreasing with $f(0) = 0$ and $f(1) = 1$.

- In order to generalize the **maximum likelihood estimator**, f must be strictly increasing in 1, in the sense that $f(x) < 1$ if $x < 1$.

  In this case we have for example that for a finite $\mathcal{P} = \mathcal{D}$ and the simple loss function $L(P,d) = \begin{cases} 0 & \text{if } P = d \\ 1 & \text{if } P \neq d \end{cases}$, if $\widehat{P}_{ML}$ exists, $d = \widehat{P}_{ML}$ minimizes $\int l_d \,\mathrm{d}(f \circ LR)$ and will therefore be selected.

- In order to generalize the tests and confidence regions based on the **likelihood ratio statistic**, f must be bijective (that is, continuous and strictly increasing).

  An example with non-bijective f is the ML decision criterion: it has $f = I_{\{1\}}$ and will reject a null hypothesis $\mathcal{H}_0$ as soon as $LR(\mathcal{H}_0) < 1$.

Asymptotic Optimality

- $X_1, X_2, \ldots$ iid (discrete) random objects

Asymptotic Optimality

- $X_1, X_2, \ldots$ iid (discrete) random objects
- $\mathcal{P}$ and L satisfy some regularity conditions

Asymptotic Optimality

- $X_1, X_2, \ldots$ iid (discrete) random objects
- $\mathcal{P}$ and L satisfy some regularity conditions
- after having observed $X_1, \ldots, X_n$, using the conditional decision criterion defined by a subadditive homogeneous integral $\int l \, \mathrm{d}(f \circ LR)$ with f continuous in 0, we obtain the decision d_n

Asymptotic Optimality

- $X_1, X_2, \ldots$ iid (discrete) random objects
- $\mathcal{P}$ and L satisfy some regularity conditions
- after having observed $X_1, \ldots, X_n$, using the conditional decision criterion defined by a subadditive homogeneous integral $\int l \, \mathrm{d}(f \circ LR)$ with f continuous in 0, we obtain the decision d_n

$$\Rightarrow \boxed{L(P, d_n) \to \inf L(P, \cdot) \;\; P\text{-a.s}}, \text{ for all } P \in \mathcal{P}$$

Asymptotic Optimality

- $X_1, X_2, \ldots$ iid (discrete) random objects
- $\mathcal{P}$ and L satisfy some regularity conditions
- after having observed $X_1, \ldots, X_n$, using the conditional decision criterion defined by a subadditive homogeneous integral $\int l \,\mathrm{d}(f \circ LR)$ with f continuous in 0, we obtain the decision d_n

$$\Rightarrow \boxed{L(P, d_n) \to \inf L(P, \cdot) \ \ P\text{-a.s}}, \text{ for all } P \in \mathcal{P}$$

In estimation problems with symmetric loss functions, asymptotic efficiency can be obtained, but the integral needs to satisfy stronger properties. If the loss function is asymmetric, the asymptotic efficiency can be a drawback for a decision criterion.

Cumulativity

The cumulativity of the integral, $\int l \, \mathrm{d}(f \circ LR) = G[f(LR\{l > \cdot\})]$, has in particular the following consequences.

Cumulativity

The cumulativity of the integral, $\int l \, \mathrm{d}(f \circ LR) = G[f(LR\{l > \cdot\})]$, has in particular the following consequences.

- Nuisance parameters are automatically eliminated by means of the **profile likelihood function** (in order to use other pseudo-likelihood functions, we have to apply them before the conditional decision criterion).

Cumulativity

The cumulativity of the integral, $\int l \, \mathrm{d}(f \circ LR) = G[f(LR\{l > \cdot\})]$, has in particular the following consequences.

- Nuisance parameters are automatically eliminated by means of the **profile likelihood function** (in order to use other pseudo-likelihood functions, we have to apply them before the conditional decision criterion).
- The evaluation $\int l \, \mathrm{d}(f \circ LR)$ of the loss l is not influenced by models $P \in \mathcal{H}$ which are dominated by another model P' in the sense that $l|_{\mathcal{H}} \leq l(P')$ and $LR(\mathcal{H}) \leq \overline{lik}(P)$. Moreover, the evaluation of l is not influenced by impossible models (that is, models $P \in \mathcal{H}$ with $LR(\mathcal{H}) = 0$).

Cumulativity

The cumulativity of the integral, $\boxed{\int l \, \mathrm{d}(f \circ LR) = G[f(LR\{l > \cdot\})]}$, has in particular the following consequences.

- Nuisance parameters are automatically eliminated by means of the **profile likelihood function** (in order to use other pseudo-likelihood functions, we have to apply them before the conditional decision criterion).
- The evaluation $\int l \, \mathrm{d}(f \circ LR)$ of the loss l is not influenced by models $P \in \mathcal{H}$ which are dominated by another model P' in the sense that $l|_{\mathcal{H}} \leq l(P')$ and $LR(\mathcal{H}) \leq \overline{lik}(P)$. Moreover, the evaluation of l is not influenced by impossible models (that is, models $P \in \mathcal{H}$ with $LR(\mathcal{H}) = 0$).
- **Conditional Minimax**: $F(l, I_{\mathcal{H}}) = \sup_{\mathcal{H}} l$

If F generalizes the likelihood-based inference methods and satisfies Conditional Minimax, it can be written as a homogeneous integral: $F(l, \overline{lik}) = \int l \, \mathrm{d}(f \circ LR)$.

Choquet Integral

The Choquet integral $\int l \, \mathrm{d}(f \circ LR) = \int_0^\infty f[LR(\{l > x\})] \, \mathrm{d}x$ can be easily extended to bounded functions $l : \mathcal{P} \to (-\infty, \infty)$ and it is **translation invariant**: $\int (l + c) \, \mathrm{d}(f \circ LR) = \int l \, \mathrm{d}(f \circ LR) + c$ (c is a constant). Hence it can be used in **general decision problems**, where the loss function L can take negative values and is usually defined only up to positive affine transformations (that is, the loss functions L and $L\prime = b\, L + c$ are equivalent, for all constants $b > 0$ and c).

Choquet Integral

The Choquet integral $\int l \,\mathrm{d}(f \circ LR) = \int_0^\infty f[LR(\{l > x\})] \,\mathrm{d}x$ can be easily extended to bounded functions $l : \mathcal{P} \to (-\infty, \infty)$ and it is **translation invariant**: $\int (l + c) \,\mathrm{d}(f \circ LR) = \int l \,\mathrm{d}(f \circ LR) + c$ (c is a constant). Hence it can be used in **general decision problems**, where the loss function L can take negative values and is usually defined only up to positive affine transformations (that is, the loss functions L and $L\prime = b\,L + c$ are equivalent, for all constants $b > 0$ and c).

The **dual measure** μ^c of a normed monotonic measure μ on a set $\mathcal{Q}$ is the normed monotonic measure on $\mathcal{Q}$ defined by $\mu^c(\mathcal{S}) = 1 - \mu(\mathcal{Q} \setminus \mathcal{S})$.

$F(l, \overline{lik}) = \int l \,\mathrm{d}[(f \circ LR)^c]$ satisfies the 4 necessary properties and **Conditional Minimin**: $F(l, I_{\mathcal{H}}) = \inf_{\mathcal{H}} l$. Furthermore we have:

$$\int l \,\mathrm{d}[(f \circ LR)^c] \leq F_{ML}(l, \overline{lik}) \leq \int l \,\mathrm{d}(f \circ LR) .$$

The two integrals can be interpreted as lower and upper conditional evaluation of the loss l.

Shilkret Integral

The Shilkret integral $\int l \, \mathrm{d}(f \circ LR) = \sup_{x>0} x \, f[LR(\{l > x\})]$ is not translation invariant, but in **statistical decision problems** the loss function L is usually not considered equivalent to $L' = L + c$ (for a constant $c > 0$).

Shilkret Integral

The Shilkret integral $\int l \,\mathrm{d}(f \circ LR) = \sup_{x>0} x\, f[LR(\{l > x\})]$ is not translation invariant, but in **statistical decision problems** the loss function L is usually not considered equivalent to $L' = L + c$ (for a constant $c > 0$).

If f is continuous, the Shilkret integral can be considerably simplified: $\int l \,\mathrm{d}(f \circ LR) = \sup\,(f \circ \overline{lik})\, l$. This simple form allows its application in difficult problems. Choosing a function f different from the identity means stressing or smoothing the differences of likelihood; if we use the identity we obtain the following criterion.

Shilkret Integral

The Shilkret integral $\int l \, \mathrm{d}(f \circ LR) = \sup_{x>0} x \, f[LR(\{l > x\})]$ is not translation invariant, but in **statistical decision problems** the loss function L is usually not considered equivalent to $L' = L + c$ (for a constant $c > 0$).

If f is continuous, the Shilkret integral can be considerably simplified: $\int l \, \mathrm{d}(f \circ LR) = \sup \, (f \circ \overline{lik}) \, l$. This simple form allows its application in difficult problems. Choosing a function f different from the identity means stressing or smoothing the differences of likelihood; if we use the identity we obtain the following criterion.

MPL criterion: minimize $\sup lik \, l_d = \sup_{P \in \mathcal{P}} lik(P) \, L(P, d)$

Shilkret Integral

The Shilkret integral $\int l \,\mathrm{d}(f \circ LR) = \sup_{x>0} x\, f[LR(\{l > x\})]$ is not translation invariant, but in **statistical decision problems** the loss function L is usually not considered equivalent to $L' = L + c$ (for a constant $c > 0$).

If f is continuous, the Shilkret integral can be considerably simplified: $\int l \,\mathrm{d}(f \circ LR) = \sup\,(f \circ \overline{lik})\, l$. This simple form allows its application in difficult problems. Choosing a function f different from the identity means stressing or smoothing the differences of likelihood; if we use the identity we obtain the following criterion.

MPL criterion: minimize $\sup lik\, l_d = \sup_{P \in \mathcal{P}} lik(P)\, L(P, d)$

If lik is constant (that is, we have no information about the models in $\mathcal{P}$), the MPL criterion reduces to the conditional minimax criterion.
But if lik is not constant, it is used as a weighting of l_d (MPL means **Minimax Plausibility-weighted Loss**).

Example: Mixed Effects Models

Estimation of the variance components in the 3×3 random effect one-way layout, under normality assumptions and weighted squared error loss.

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \forall\, i, j \in \{1, 2, 3\}$$

Example: Mixed Effects Models

Estimation of the variance components in the 3×3 random effect one-way layout, under normality assumptions and weighted squared error loss.

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \forall\, i, j \in \{1, 2, 3\}$$

Normality assumptions:

$$\alpha_i \sim \mathcal{N}(0, v_a), \;\; \varepsilon_{ij} \sim \mathcal{N}(0, v_e), \;\; \text{all independent}$$

$$\Rightarrow \;\; X_{ij} \sim \mathcal{N}(\mu, v_a + v_e) \;\text{dependent}, \;\; \mu \in (-\infty, \infty), \;\; v_a, v_e \in (0, \infty)$$

Example: Mixed Effects Models

The estimates $\widehat{v_e}$ and $\widehat{v_a}$ of the variance components v_e and v_a are functions of

$$SS_e = \sum_{i=1}^{3}\sum_{j=1}^{3}(x_{ij} - \bar{x}_{i\cdot})^2 \quad \text{and} \quad SS_a = 3\,\sum_{i=1}^{3}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2\ ,$$

where

$$\bar{x}_{i\cdot} = \frac{1}{3}\,\sum_{j=1}^{3} x_{ij}\ , \quad \bar{x}_{\cdot\cdot} = \frac{1}{9}\,\sum_{i=1}^{3}\sum_{j=1}^{3} x_{ij}\ ,$$

$$\frac{SS_e}{v_e} \sim \chi^2_6 \quad \text{and} \quad \frac{\frac{1}{3}\,SS_a}{v_a + \frac{1}{3}\,v_e} \sim \chi^2_2\ .$$

Example: Mixed Effects Models

The estimates $\widehat{v_e}$ and $\widehat{v_a}$ of the variance components v_e and v_a are functions of

$$SS_e = \sum_{i=1}^{3} \sum_{j=1}^{3} (x_{ij} - \bar{x}_{i\cdot})^2 \quad \text{and} \quad SS_a = 3 \sum_{i=1}^{3} (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 \ ,$$

where

$$\bar{x}_{i\cdot} = \frac{1}{3} \sum_{j=1}^{3} x_{ij} \ , \quad \bar{x}_{\cdot\cdot} = \frac{1}{9} \sum_{i=1}^{3} \sum_{j=1}^{3} x_{ij} \ ,$$

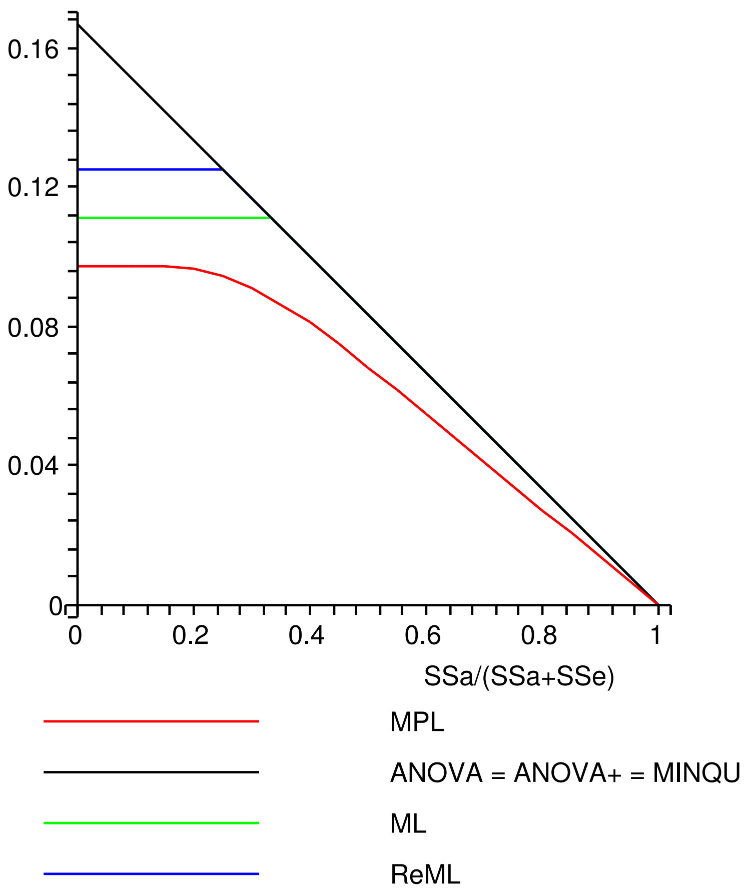
$$\frac{SS_e}{v_e} \sim \chi_6^2 \quad \text{and} \quad \frac{\frac{1}{3}\, SS_a}{v_a + \frac{1}{3}\, v_e} \sim \chi_2^2 \ .$$

The considered loss functions are

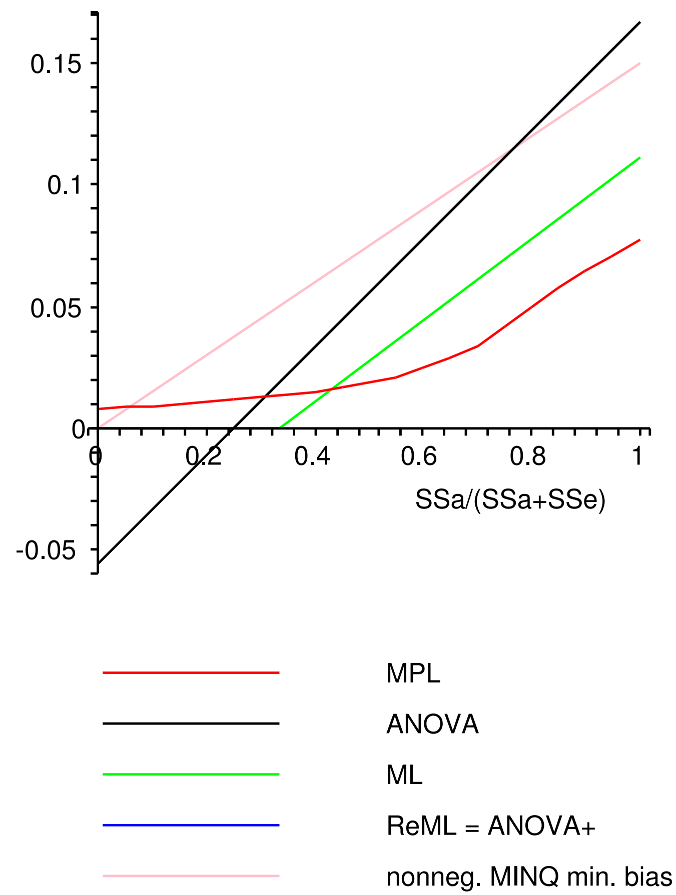
$$3\, \frac{(\widehat{v_e} - v_e)^2}{{v_e}^2} \quad \text{and} \quad \frac{(\widehat{v_a} - v_a)^2}{(v_a + \frac{1}{3}\, v_e)^2} \ .$$

Example: Mixed Effects Models

$$\frac{\widehat{v_e}}{SS_a+SS_e}$$

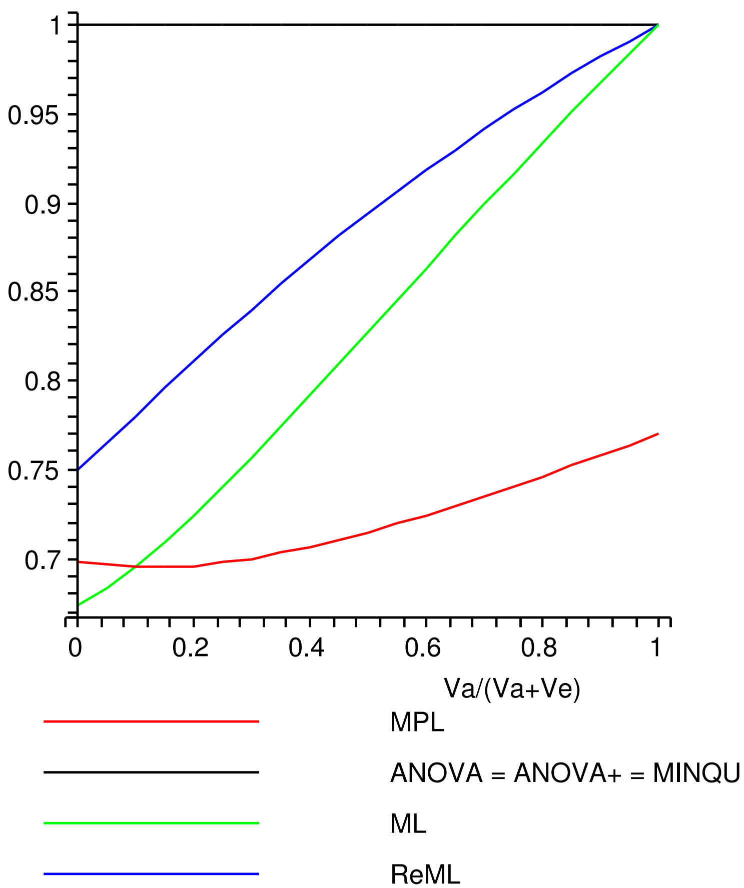


$$\frac{\widehat{v_a}}{SS_a+SS_e}$$



Example: Mixed Effects Models

$$3\,\frac{E[(\widehat{v_e}-v_e)^2]}{{v_e}^2}$$



$$\frac{E[(\widehat{v_a}-v_a)^2]}{(v_a+\frac{1}{3}\,v_e)^2}$$