The likelihood approach to statistics

Marco Cattaneo
Department of Statistics, LMU Munich
cattaneo@stat.uni-muenchen.de

December 9, 2008

my research

- PhD with Frank Hampel at ETH Zurich
  (November 2002 – March 2007):

  Statistical Decisions Based Directly on the Likelihood Function

my research

- PhD with Frank Hampel at ETH Zurich
  (November 2002 – March 2007):

  Statistical Decisions Based Directly on the Likelihood Function

- Postdoc with Thomas Augustin at LMU Munich
  (SNSF Research Fellowship, October 2007 – March 2009):

  Decision making on the basis of a probabilistic-possibilistic hierarchical description of uncertain knowledge

the likelihood function

- $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ set of probabilistic models:
  each $P_\theta \in \mathcal{P}$ is a probability measure on $(\Omega, \mathcal{A})$

the likelihood function

- $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ set of probabilistic models:
  each $P_\theta \in \mathcal{P}$ is a probability measure on $(\Omega, \mathcal{A})$

- when data $A \in \mathcal{A}$ are observed, the **likelihood function**

  $$lik(\theta) \propto P_\theta(A)$$

  describes the *relative* ability of the models in $\mathcal{P}$ to forecast the observed data

the likelihood function

- $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ set of probabilistic models: each $P_\theta \in \mathcal{P}$ is a probability measure on $(\Omega, \mathcal{A})$

- when data $A \in \mathcal{A}$ are observed, the **likelihood function**

$$lik(\theta) \propto P_\theta(A)$$

  describes the *relative* ability of the models in $\mathcal{P}$ to forecast the observed data

- the likelihood function is a central concept in statistics (both frequentist and Bayesian)

maximum likelihood

- the **maximum likelihood** estimate of θ is

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} lik(\theta)$$

maximum likelihood

- the **maximum likelihood** estimate of θ is

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} lik(\theta)$$

- a statistical decision problem is described by a **loss function**

$$L : \Theta \times \mathcal{D} \to [0, \infty],$$

  where $L(\theta, d)$ is the loss incurred by making decision d, according to model P_θ

maximum likelihood

- the **maximum likelihood** estimate of θ is

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} lik(\theta)$$

- a statistical decision problem is described by a **loss function**

$$L : \Theta \times \mathcal{D} \to [0, \infty],$$

  where $L(\theta, d)$ is the loss incurred by making decision d, according to model P_θ
- for instance, estimation of $\theta \in \mathbb{R}^n$ with squared error:
  $\mathcal{D} = \Theta \subseteq \mathbb{R}^n$ and $L(\theta, d) = \|\theta - d\|^2$

maximum likelihood

- the **maximum likelihood** estimate of θ is

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} lik(\theta)$$

- a statistical decision problem is described by a **loss function**

$$L : \Theta \times \mathcal{D} \to [0, \infty],$$

  where $L(\theta, d)$ is the loss incurred by making decision d, according to model P_θ

- for instance, estimation of $\theta \in \mathbb{R}^n$ with squared error: $\mathcal{D} = \Theta \subseteq \mathbb{R}^n$ and $L(\theta, d) = \|\theta - d\|^2$

- **MLD** criterion: make decision

$$d_{MLD} = \arg\min_{d \in \mathcal{D}} L(\hat{\theta}_{ML}, d)$$

maximum likelihood

- the **maximum likelihood** estimate of θ is

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} lik(\theta)$$

- a statistical decision problem is described by a **loss function**

$$L : \Theta \times \mathcal{D} \to [0, \infty],$$

  where $L(\theta, d)$ is the loss incurred by making decision d, according to model P_θ
- for instance, estimation of $\theta \in \mathbb{R}^n$ with squared error:
  $\mathcal{D} = \Theta \subseteq \mathbb{R}^n$ and $L(\theta, d) = \|\theta - d\|^2$
- **MLD** criterion: make decision

$$d_{MLD} = \arg\min_{d \in \mathcal{D}} L(\hat{\theta}_{ML}, d)$$

- the MLD criterion is very often implicitly used, but disregards the model uncertainty

the hierarchical model

- **Bayesian** criterion with prior probability measure π: make decision

$$d_\pi = \arg\min_{d \in \mathcal{D}} \int L(\theta, d)\, lik(\theta)\, \mathrm{d}\pi(\theta)$$

the hierarchical model

- **Bayesian** criterion with prior probability measure π: make decision

$$d_\pi = \arg\min_{d \in \mathcal{D}} \int L(\theta, d)\, lik(\theta)\, \mathrm{d}\pi(\theta)$$

- the set $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probabilistic models and the likelihood function *lik* on Θ are the two levels of a **hierarchical model**

the hierarchical model

- **Bayesian** criterion with prior probability measure π: make decision

$$d_\pi = \arg\min_{d \in \mathcal{D}} \int L(\theta, d)\, lik(\theta)\, \mathrm{d}\pi(\theta)$$

- the set $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probabilistic models and the likelihood function *lik* on Θ are the two levels of a **hierarchical model**
- when data $B \in \mathcal{A}$ are observed, P_θ is updated to $P_\theta(\cdot|B)$, and $lik(\theta)$ is updated to $lik(\theta)\, P_\theta(B)$

the hierarchical model

- **Bayesian** criterion with prior probability measure π: make decision

$$d_\pi = \arg\min_{d \in \mathcal{D}} \int L(\theta, d)\, lik(\theta)\, \mathrm{d}\pi(\theta)$$

- the set $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probabilistic models and the likelihood function *lik* on Θ are the two levels of a **hierarchical model**
- when data $B \in \mathcal{A}$ are observed, P_θ is updated to $P_\theta(\cdot|B)$, and $lik(\theta)$ is updated to $lik(\theta)\, P_\theta(B)$

- before observing data, prior information ca be described by a (subjective) **prior** likelihood function

the hierarchical model

- **Bayesian** criterion with prior probability measure π: make decision

$$d_\pi = \arg\min_{d \in \mathcal{D}} \int L(\theta, d)\, lik(\theta)\, \mathrm{d}\pi(\theta)$$

- the set $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probabilistic models and the likelihood function *lik* on Θ are the two levels of a **hierarchical model**
- when data $B \in \mathcal{A}$ are observed, P_θ is updated to $P_\theta(\cdot|B)$, and $lik(\theta)$ is updated to $lik(\theta)\, P_\theta(B)$

- before observing data, prior information ca be described by a (subjective) **prior** likelihood function
- prior **ignorance** is described by a constant likelihood function

nonadditive measures and integrals

- the nonadditive measure on 2^Θ defined by

$$\lambda(\mathcal{H}) = \sup_{\theta \in \mathcal{H}} \mathit{lik}(\theta) \quad \text{for all } \mathcal{H} \subseteq \Theta$$

is used in particular in the **likelihood ratio** test

nonadditive measures and integrals

- the nonadditive measure on 2^Θ defined by

$$\lambda(\mathcal{H}) = \sup_{\theta \in \mathcal{H}} lik(\theta) \quad \text{for all } \mathcal{H} \subseteq \Theta$$

  is used in particular in the **likelihood ratio** test

- likelihood-based decision criterion: make decision

$$d' = \arg\min_{d \in \mathcal{D}} \int L(\theta, d) \, \mathrm{d}\lambda(\theta)$$

nonadditive measures and integrals

- the nonadditive measure on 2^Θ defined by

$$\lambda(\mathcal{H}) = \sup_{\theta \in \mathcal{H}} lik(\theta) \quad \text{for all } \mathcal{H} \subseteq \Theta$$

  is used in particular in the **likelihood ratio** test

- likelihood-based decision criterion: make decision

$$d' = \arg\min_{d \in \mathcal{D}} \int L(\theta, d) \, \mathrm{d}\lambda(\theta)$$

- the resulting decision functions satisfy in particular asymptotic optimality (consistency), parametrization invariance, equivariance, and asymptotic efficiency

MPL decision criterion

- λ is completely maxitive:

$$\lambda\left(\bigcup_{\mathcal{H}\in\mathcal{S}}\mathcal{H}\right) = \sup_{\mathcal{H}\in\mathcal{S}}\lambda(\mathcal{H}) \quad \text{for all } \mathcal{S} \subseteq 2^{\Theta}$$

MPL decision criterion

- λ is completely maxitive:

$$\lambda\left(\bigcup_{\mathcal{H}\in\mathcal{S}} \mathcal{H}\right) = \sup_{\mathcal{H}\in\mathcal{S}} \lambda(\mathcal{H}) \quad \text{for all } \mathcal{S} \subseteq 2^{\Theta}$$

- hence the **Shilkret integral** is

$$\int^{\mathrm{S}} L(\theta, d)\,\mathrm{d}\lambda(\theta) = \sup_{\theta\in\Theta} \mathit{lik}(\theta)\, L(\theta, d)$$

MPL decision criterion

- λ is completely maxitive:

$$\lambda\left(\bigcup_{\mathcal{H}\in\mathcal{S}}\mathcal{H}\right)=\sup_{\mathcal{H}\in\mathcal{S}}\lambda(\mathcal{H})\quad\text{for all } \mathcal{S}\subseteq 2^{\Theta}$$

- hence the **Shilkret integral** is

$$\int^{\mathrm{S}} L(\theta,d)\,\mathrm{d}\lambda(\theta)=\sup_{\theta\in\Theta} lik(\theta)\,L(\theta,d)$$

- **MPL** criterion: make decision

$$d_{MPL}=\arg\min_{d\in\mathcal{D}}\sup_{\theta\in\Theta} lik(\theta)\,L(\theta,d)$$

MPL decision criterion

- λ is completely maxitive:

$$\lambda\left(\bigcup_{\mathcal{H}\in\mathcal{S}} \mathcal{H}\right) = \sup_{\mathcal{H}\in\mathcal{S}} \lambda(\mathcal{H}) \quad \text{for all } \mathcal{S} \subseteq 2^{\Theta}$$

- hence the **Shilkret integral** is

$$\int^{\mathrm{S}} L(\theta, d)\,\mathrm{d}\lambda(\theta) = \sup_{\theta\in\Theta} lik(\theta)\, L(\theta, d)$$

- **MPL** criterion: make decision

$$d_{MPL} = \arg\min_{d\in\mathcal{D}} \sup_{\theta\in\Theta} lik(\theta)\, L(\theta, d)$$

- the MPL criterion is the only likelihood-based decision criterion satisfying the **sure-thing principle**: if d is optimal with respect to the set $\mathcal{P}$ of probabilistic models, and d is optimal also with respect to $\mathcal{P}'$, then d is optimal with respect to $\mathcal{P} \cup \mathcal{P}'$

example: estimation of variance components

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

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{for all } i, j \in \{1, 2, 3\}$$

example: estimation of variance components

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

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{for all } i, j \in \{1, 2, 3\}$$

- 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: estimation of variance components

- estimates $\widehat{v_e}$ and $\widehat{v_a}$ of 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: estimation of variance components

- estimates $\widehat{v_e}$ and $\widehat{v_a}$ of 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

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

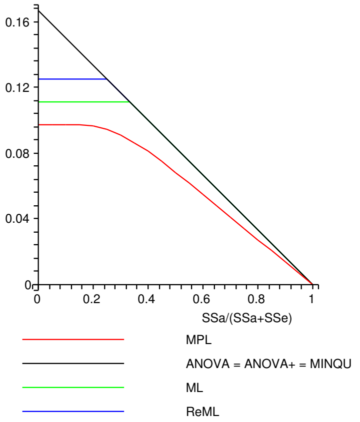
$$\frac{SS_e}{v_e} \sim \chi^2_6, \quad \text{and} \quad \frac{\frac{1}{3}\, SS_a}{v_a + \frac{1}{3}\, v_e} \sim \chi^2_2$$

- invariant loss functions:

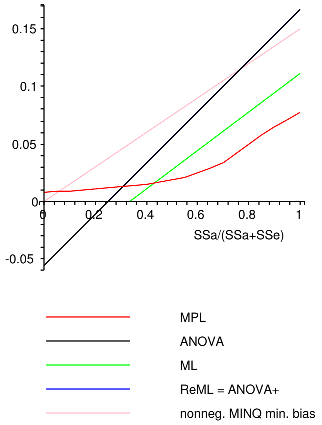
$$L(v_e, \widehat{v_e}) = 3\, \frac{(v_e - \widehat{v_e})^2}{{v_e}^2} \quad \text{and} \quad L(v_a, \widehat{v_a}) = \frac{(v_a - \widehat{v_a})^2}{(v_a + \frac{1}{3}\, v_e)^2}$$

example: estimation of variance components

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

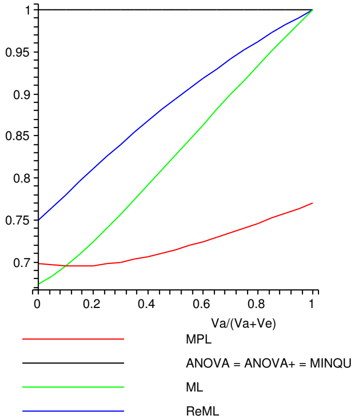


$$\frac{\widehat{v_a}}{SS_a+SS_e}$$

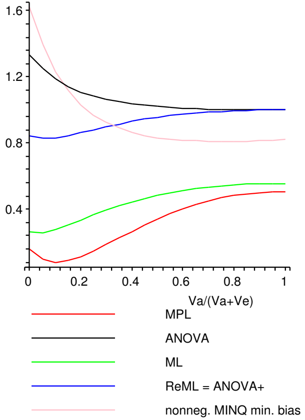


example: estimation of variance components

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



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



present and future research

- comparison/combination with other approaches in various applications

present and future research

- comparison/combination with other approaches in various applications
- application to robustness problems
  (relaxation of i.i.d. assumption, robust likelihood)

present and future research

- comparison/combination with other approaches in various applications
- application to robustness problems
  (relaxation of i.i.d. assumption, robust likelihood)
- application to imprecise probability theory
  (better updating, possibilistic previsions, convex set of non-normalized measures)

present and future research

- comparison/combination with other approaches in various applications
- application to robustness problems
  (relaxation of i.i.d. assumption, robust likelihood)
- application to imprecise probability theory
  (better updating, possibilistic previsions, convex set of non-normalized measures)
- application to graphical models
  (probabilistic and non-probabilistic aspects of uncertainty)

present and future research

- comparison/combination with other approaches in various applications
- application to robustness problems
  (relaxation of i.i.d. assumption, robust likelihood)
- application to imprecise probability theory
  (better updating, possibilistic previsions, convex set of non-normalized measures)
- application to graphical models
  (probabilistic and non-probabilistic aspects of uncertainty)
- application to financial risk measures
  (derivation and interpretation of convex risk measures)