Thomas Augustin

University of Munich

Robust Neyman Pearson theory & summary view on imprecise probabilities

1. Some preliminaries

2. Robust Testing (Neighborhood Models)

3. Decision Making

4. Some Additional Remarks

5. Some Challenges

---

6. Summary of Previous Talks

7. Participants' Research Directions and Ideas

1. Some preliminaries

Classical probability and statistics

Def. *Given a sample-space Ω and a σ-field $\mathcal{A}$ of random events in Ω, a set function $p(\cdot)$ defined on $\mathcal{A}$ is called a* **classical probability**, *if it satisfies the following three axioms:*

I. $$p(A) \geq 0, \quad \forall A \in \mathcal{A}. \qquad (1)$$

II. $$p(\Omega) = 1. \qquad (2)$$

III. $\forall A_i,\ A_j \in \mathcal{A}$ *with* $A_i \cap A_j = \emptyset$, *if* $i \neq j$:

$$p(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} p(A_i). \qquad (3)$$

□

- Kolmogorov's axioms; σ additivity

- $(\Omega, \mathcal{A}, p)$ probability space (mathematical model of a probability experiment)

- Specification of probability

  * Ω is countable: Specify *probability mass function* $p(\{\omega\})$, then

$$p(A) = \sum_{\omega \in A} p(\{\omega\}), \quad \forall A \in \mathcal{P}(\Omega)$$

  * Else, for instance, when $\Omega = \mathbb{R}$, specify a *density function* $f : \Omega \to \mathbb{R}^+$ with $\int_\Omega f(\omega) d\omega = 1$, then

$$p(A) = \int_A f(\omega) d\omega, \forall A \in \mathcal{A} \subsetneq \mathcal{P}(\Omega)$$

- Most often: *parametric models*, basic form of the density $f_\vartheta(\cdot)$ is known up to a parameter ϑ of finite dimension, for instance: normal distribution with mean value μ and variance σ^2.

- *basic model of statistical inference*

  * Given a set of probability spaces which are potentially true $\Rightarrow (\Omega, \mathcal{A}, (p_\vartheta)_{\vartheta\in\Theta})$

  * One of these is true, i.e. it generates the observations.

  * Find the true model on the basis of a *sample* consisting of values $\omega^{(1)}, \ldots, \omega^{(n)}$ where $\omega^{(i)}$ is the outcome of the i-th repetition of the probability experiment $(\Omega, \mathcal{A}, (p_\vartheta)_{\vartheta\in\Theta})$
    (Survey; n persons drawn randomly, $\omega^{(i)}$ value of the i-th person)

  * estimation, testing, decision making

1.2 Interval Probability

R-probability, F-probability, structure

The Need of a New Calculus of Probability:

Probability and uncertainty as two dimensional phenomenon

	uncertainty	
	=	
ideal randomness	+	ambiguity

Peirce (1878, p.421): "...[T]o express the proper state of our belief, not one number but two are requisite, the first depending on the inferred probability, the second on the amount of knowledge on which that probability is based"

Two main approaches

- sets of classical probabilities
- interval

  $$[L(A); U(A)]$$

  consisting of non-additive set functions $L(\cdot)$ and $U(\cdot)$.

  The width reflects the degree of ambiguity

  * $P(A) = [a; a]$: classical probability, situation of ideal randomness

    $\vdots$ increasing ambiguity

  * $P(A) = [0; 1]$: complete ignorance

Axiomatization of interval probability
Weichselberger (2000, IntJApproxReas; 2001, Physika):

Look at the *relation* between the non-additive set functions $L(\cdot)$ and $U(\cdot)$ and the structure $\mathcal{M}$, i.e the set of all compatible, classical probabilities

$$\mathcal{M} := \{p(\cdot) \mid L(A) \leq p(A) \leq U(A), \quad \forall A \in \mathcal{A}\},$$

$$\Downarrow$$

several levels of quality

1. $\mathcal{M}$ is empty: Contradictory assignment in the probabilistic sense.

2. $\mathcal{M}$ is not empty: Interval probability in the narrow sense

$$\swarrow \qquad\qquad \searrow$$

$[L(\cdot), U(\cdot)] \to \mathcal{M}$	$[L(\cdot), U(\cdot)] \leftrightarrow \mathcal{M}$
R-probabilty	*F-probability*
not contradictory, but possibly inhomogeneous	not contradictory, homogeneous $\mathcal{F} = (\Omega, \mathcal{A}, L(\cdot))$

Basic Definitions

- generalization of Kolmogorov's axioms
- $\mathcal{K}(\Omega, \mathcal{A})$: the set of all classical probabilities on a measurable space $(\Omega, \mathcal{A})$
- $P(\cdot)$ is **R-probability** with **structure** $\mathcal{M}$, if

  1. $P(\cdot)$ is of the form

  $$\begin{aligned} P(\cdot)\colon \mathcal{A} &\to \mathcal{Z}_0 := \{[L,U] \,|\, 0 \le L \le U \le 1\} \\ A &\mapsto \qquad P(A) = [L(A), U(A)]\,. \end{aligned}$$

  2. The set

  $$\begin{aligned} \mathcal{M} := \{p(\cdot) \in \mathcal{K}(\Omega,\mathcal{A}) \,|\, \\ L(A) \le p(A) \le U(A),\ \forall A \in \mathcal{A}\} \end{aligned}$$

  is not empty.

- $P(\cdot)$ is **F-probability** with structure $\mathcal{M}$, if $P(\cdot)$ is R-probability with structure $\mathcal{M}$ and

$$\left.\begin{array}{l} \inf\limits_{p(\cdot)\in\mathcal{M}} p(A) = L(A) \\ \\ \sup\limits_{p(\cdot)\in\mathcal{M}} p(A) = U(A) \end{array}\right\} \quad \forall A \in \mathcal{A}\,.$$

- R-probability:
  - * Not-contradictory from the probabilistic point of view, but (eventually) not fully homogeneous
  - * Strongly related to *avoiding sure loss* in Walley's theory (Walley (1991)) (but σ-additivity of classical probabilities required)
- F-probability:
  - * fully homogeneous assignment, one-to-one correspondence between interval-limits and the structure
  - * $L(A) = 1 - U(A^C)$, $\forall A \in \mathcal{A}$.
  - * $\mathcal{F} = (\Omega, \mathcal{A}, L(\cdot))$: **F-probability field**.
  - * Corresponds to *lower probability* in the sense of Huber & Strassen
    Strong relation to *lower envelopes* (Fine and students) (e.g. Papamarcou, A. & Fine, T.L. (1991)) and to *coherence* (Walley (1991))

- more general: Walley (1991, Chapm. & Hall): imprecise previsions obtained from interval-valued expectations = linear partial information (Kofler & Menges (1976, SpringerLN Econ); Huschens (1985, R.G. Fischer); Kofler (1989, Campus))

- The concept 'structure' establishes a strong relation between interval-probability and sets of classical probability measures. It serves as a **guiding principle** for generalizing classical probability theory to interval probability:

  * expectation
  * independent product of F-prob. fields
  * conditional probability
  * law of large number

  – ...

  "inside the structure": *strict uncertainty*!

1.2.2 Two-monotone Probability

- **Special Cases: Capacities of Higher Order**
  * Belief-functions (totally monotone probabilities), corresponding to a basic probability assignment (Shafer (1976, Princeton UP), Yager, Fedrizzi und Kacprzyk (1994, Wiley))
  * Neighborhood models in robust statistics (pseudo capacities, Choquet-capacities)(Huber (1981, Wiley), for a survey (and extensions): Augustin (2002a, JStatPlanInf.))
  * Probability intervals (PRI) (Weichselberger & Pöhlmann (1990; Springer LN AI)), Campos, Huete and Moral (1994, IJUncFuzzKbS)

* Other common names 'supermodular' (Denneberg (1994; Kluwer)) or 'convex' (Jaffray (1989, OR Letters))

Two-monotone probability (field): Special case of F-probability (field), where $L(\cdot)$ is *two-monotone* (*(strong) superadditive, supermodular, convex*) i.e.,

$$L(A \cup B) + L(A \cap B) \geq L(A) + L(B), \quad \forall A, B \in \mathcal{A}\,.$$

classical probability field

$\Downarrow$ $\not\Uparrow$

totally monotone two-monotone probability field

$\Downarrow$ $\not\Uparrow$

two-monotone probability field

$\Downarrow$ $\not\Uparrow$

F-probability field

$\Downarrow$ $\uparrow$ Derived F-pr./F-Hull

R-probability field

$\Downarrow$ $\uparrow$

lower previsions avoiding sure loss

Some Aspects of the Insufficiency of Two-monotone Probabilities

- *General neighbourhood-models*
- *Why should the distortion of a classical probability always be convex?*
- *Problems of interpretation*

  There is neither a subjectivist's nor a frequentist's operational definition or interpretation for two-monotone probabilities.
- *No closeness by working with partial determinated probability*
- *The integrative character of interval probability would be lost*

  There is – at least up to now – no proper way to come from an arbitary set of classical probabilities to a corresponding two-monotone probability.

- *Parametric models*

  Usually, parametrically constructed F-probability fields (e.g. the F-normaldistributions) are not two monotone probabilities.

Conclusion: The restriction of the calculus to two-monotone probability would lead to a serious reduction of the expressive power of the concept of interval probability.

There isn't "[...] any 'rationality' argument for 2-monotonicity, beyond its computational convenience" (P. Walley, 1981)

1.2.3 Interval-valued Expectation and/or Choquet-Integral

X random variable: (measurable) mapping from Ω to $\mathbb{R}$

Classical expectation

- For discrete spaces

$$\mathbb{E}X = \sum_x x \cdot P(X(\omega) = x)$$

If X takes only values in $\mathbb{N}$, then

$$\mathbb{E}X = \sum_{x \in I\!N} P(X(\omega) \geq x)$$

- In the continuous case with density f:

$X(\cdot) \geq 0$:

$$\mathbb{E}X = \int x \cdot f(x) dx = \int p(\{\omega \mid X(\omega) > t\})\, dt$$

$$\mathbb{E}X = \int x \cdot f(x)dx = \int p(\{\omega \mid X(\omega) > t\})\, dt$$

Two possible ways to generalize this to F-probability $P(\cdot) = [L(\cdot), U(\cdot)]$ with structure $\mathcal{M}$

- "outer method": substitute $p(\cdot)$ by $L(\cdot)$ and $U(\cdot)$ (*Choquet integral, fuzzy integral*) In general:

  $$\mathbb{E}_L X := \int_0^\infty L(\{\omega \mid X(\omega) > t\})\, dt\,.$$

  For $X \in \mathbb{N}$:

  $$\mathbb{E}_L X := \sum_{x \in I\!N} L(X(\omega) \geq x)$$

- "inner method": refers to the structure; considers $\inf_{p(\cdot)\in\mathcal{M}}$ and $\sup_{p(\cdot)\in\mathcal{M}}$ (here in what follows, closely related to Walley's *natural extension)*

Def. (Integrability, interval-valued expectation)

* Random variable X
* $\mathcal{M}$-integrable: p-integrable for all $p(\cdot) \in \mathcal{M}$.

$$\begin{aligned} \mathbb{E}_{\mathcal{M}} X &:= \left[{}^{\text{L}}\mathbb{E}_{\mathcal{M}} X, {}^{\text{U}}\mathbb{E}_{\mathcal{M}} X \right] \\ &:= \left[\inf_{p(\cdot)\in\mathcal{M}} \mathbb{E}_p X \,, \sup_{p(\cdot)\in\mathcal{M}} \mathbb{E}_p X \right] \\ &\subseteq [-\infty, \infty] \end{aligned}$$

Theorem (e.g, Denneberg (1994, Kluwer, Prop. 10.3)):
In the case of **two-monotone** probability both definitions **coincide.**

Therefore: In the case of two-monotonicity everything said here is also valid for the Choquet integral. Often Choquet-type form easier to calculate.

1.2.4 A Closer Look at the Structure

Prop. (Properties of the structure)

- $\mathcal{M}$ is convex.
- In the case of a finite sample space: $\mathcal{M}$ is a convex polyhedron.
  - $\mathcal{M}$ is closed.
  - The set $\mathcal{E}(\mathcal{M})$ of the **extreme points (vertices)** is non-empty, finite, and it uniquely determines $\mathcal{M}$.
  - $|\mathcal{E}(\mathcal{M})| \leq k!$ for $k := |\Omega|$.

Treatment of typical problems of interval probability with linear programming: Weichselberger (1996); see also later today.

Calculation of $\mathcal{E}(\mathcal{M})$:

- Algorithm from the theory of convex polyhedra. (Intersection of k hyperplains)
- For two-monotone and totally monotone probability closed form available:

  $$\mathcal{E}(\mathcal{M}) = \{p_\varsigma(\cdot) \mid \varsigma \in \Upsilon\}$$

  with

  $$p_\varsigma(E_i) = L\left(\bigcup_{j=1}^{i} E_{\varsigma(j)}\right) - L\left(\bigcup_{j=1}^{i-1} E_{\varsigma(j)}\right),$$

  for all $i = 1, \ldots, k$ and Υ as the set of all permutations of $\{1, \ldots, k\}$. Via Möbius inversion: also explicite formula using the basic probability assignments available.

Vertice Reduction Lemma (VRL): Extreme Points of the Structure and Calculation of Interval-Valued Expectation $\mathcal{F} = (\Omega; \mathcal{A}; L(\cdot))$ on finite Ω with structure $\mathcal{M}$ and extreme points $\mathcal{E}(\mathcal{M})$. Then

- $$\llcorner\!\mathbb{E}_{\mathcal{M}} X = \left[\min_{p(\cdot)\in\mathcal{E}(\mathcal{M})} I\!E_p X \ ; \ \max_{p(\cdot)\in\mathcal{E}(\mathcal{M})} I\!E_p X \right] .$$

- For every real g

  $$\llcorner\!\mathbb{E}_{\mathcal{M}} X \geq g \Longleftrightarrow I\!E_p X \geq g, \quad \forall p(\cdot) \in \mathcal{E}(\mathcal{M})$$

- For infinite spaces: Vertex reduction possible for continuous F-probability (Augustin, 2004c, Manuscript)

1.2.5 Prestructure; Parametric Models and Independence

Prestructures:

- Let a set $\mathcal{V}$ of classical probabilities be given.

- Construct the (unique) narrowest F-probability field $\mathcal{F}_{\mathcal{V}} = (\Omega, \mathcal{A}, L_{\mathcal{V}}(\cdot))$, whose structure $\mathcal{M}_{\mathcal{V}}$ contains $\mathcal{V}$:

  $$L_{\mathcal{V}}(A) := \inf_{p(\cdot)\in\mathcal{V}} p(A) \quad \wedge \quad U_{\mathcal{V}}(A) := \sup_{p(\cdot)\in\mathcal{V}} p(A)$$

- Then $\mathcal{V}$ is called **prestructure** of $\mathcal{F}_{\mathcal{V}}$ and of $\mathcal{M}_{\mathcal{V}}$. $L_{\mathcal{V}}(\cdot)$ is then the **lower envelope** of $\mathcal{V}$.

- Important applications

  * **Independent product of F-prob. fields** (strong independence):
    $\mathcal{F}_i$, $i \in I \subseteq I\!N$ F-probability fields with structure $\mathcal{M}_i$. Then the independent product

    $$\otimes_{i \in I} \mathcal{F}_i$$

    is defined as that F-probability field, which has

    $$\times_{i \in I} \mathcal{M}_i$$

    as a prestructure.

  * **Parametric way to construct F-probability**
    Take a set of parametric classical probabilities as a prestructure
    E.g. F-normal-distribution with parameter $[\underline{\mu}, \overline{\mu}]$.

  * Robustification of the classical concepts.

2. Robust Testing / Neighborhood Models

2.1 Testing Statistical Hypotheses

- Basic Situation: Comparison of the means μ_A, μ_B of a certain variable X in two samples A and B

- Typical examples

  a) net income
  b) aggressive behavior
  c) duration of unemployment
  d) decrease of blood pressure

X	A	B
a)	male	female
b)	upper class	lower class
c)	special training	none
d)	standard treatment treatment	new hypotensive treatment

- Distinguish between the two hypotheses

$$H_0 : \mu_A \geq \mu_B \qquad H_1 : \mu_A < \mu_B$$

by means of two samples $X_1, \ldots, X_n$ taken from group A and $Y_1, \ldots, Y_n$ taken from group B

- Here assume $X_1, \ldots, X_n \sim N(\mu_A, \sigma_A^2)$; $Y_1, \ldots, Y_n \sim N(\mu_B, \sigma_B^2)$

- Basic procedure: needs

* a summary measure T from the samples which is sensitive whether H_0 or H_1 is true

* a 'critical region' C such that f the concrete observation of T falls into C then the decision is in favor of H_1

- choose C such that it satisfies two conflicting aims

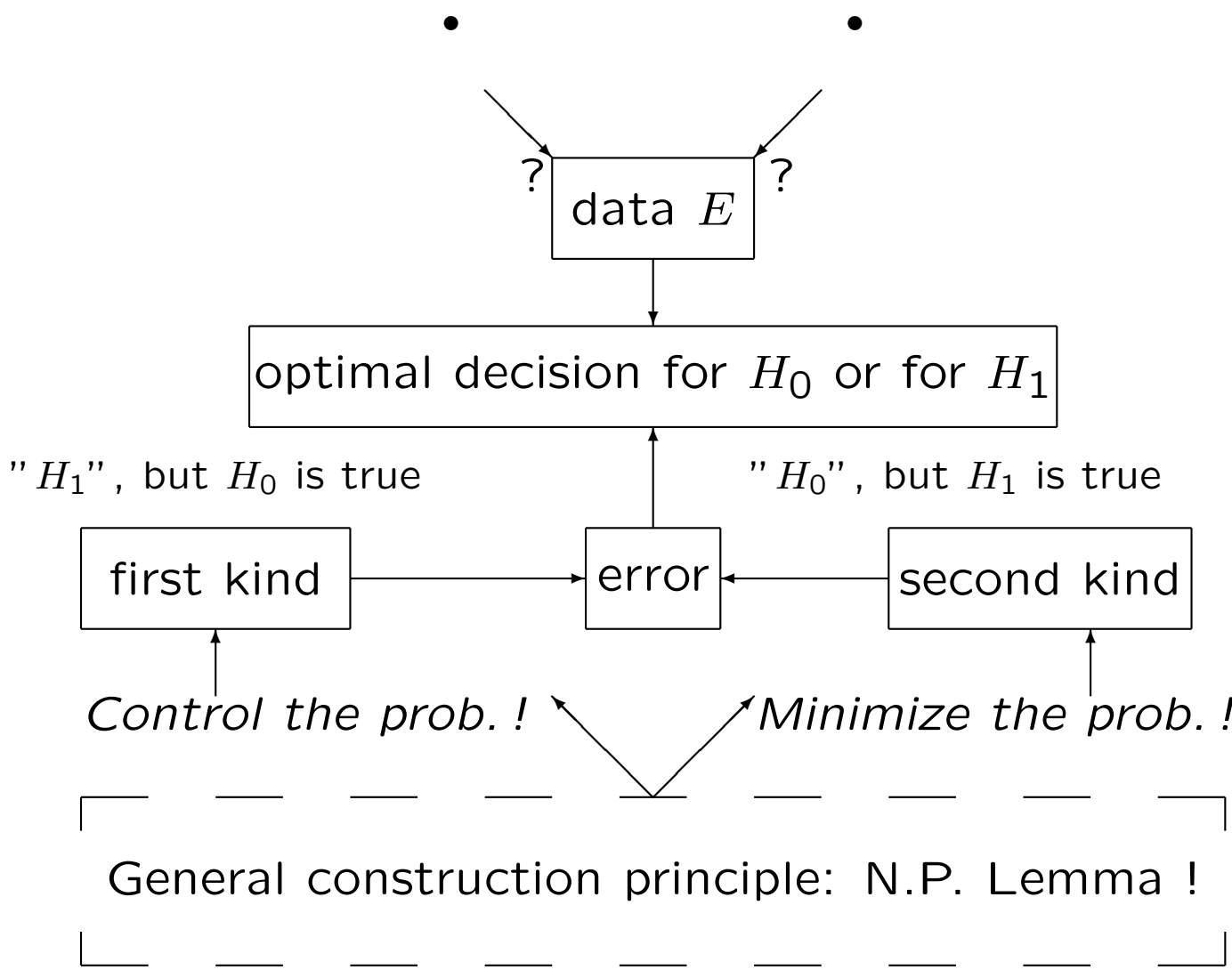
  a) probability of observing C is very small, when H_0 is true (small probability of the error of first kind)

  b) probability of observing C is as high as possible when H_1 is true (small probability of the error of second kind)

- Neyman Pearson theory:
  asymmetric: firstly satisfy a) by stating a small upper bound α (level of significance), then, among all procedures respecting this bound, take b) into account

The Neyman-Pearson Problem

$H_0 : \{ P_0(\cdot) \}$ $H_1 : \{ P_1(\cdot) \}$



- Neyman Pearson's (fundamental) lemma for simple hypothesis:

$$H_0 : p(\cdot) = p_0(\cdot) \qquad H_1 : p(\cdot) = p_1(\cdot)$$

General construction principle: optimal critical region $\tilde{C}$ can be obtained from the *likelihood ratio* $\pi(\cdot)$ between p_1 and p_0

$$\tilde{C} := \left\{ \omega | \pi(\omega) = \frac{p_1(\omega)}{p_0(\omega)} > \tau \right\}$$

where τ is such that

$$p_0(\tilde{C}) \leq \alpha \,.$$

- From that also optimal T and C for the situation above can be derived.

2.2 Robust Statistics and Neighborhood Models

- General problem: Many standard procedures (typically based on normal distributions) may show disastrous behavior even under small deviations from the true model. In particular high sensitivity to single outlying observations.

- Idea: Protect yourself by an insurance contract:

  * cost of premium: some small loss in efficiency if the model is completely correct,

  * but in the case of an 'accident' (=model is wrong): still acceptable behavior of the procedures

- Look at the sample mean $\bar{X}$

  a) If $X_1, \ldots, X_n \sim N(\mu, 1)$ (normal distribution), then

  $$\overline{X} \sim N(\mu, \frac{1}{n})$$

  We can learn the true mean μ from a sample; the larger the sample, the higher the precision of the estimator $\bar{X}$.

  b) If $X_1, \ldots, X_n \sim \mathcal{C}(\mu, 1)$ (Cauchy distribution), then

  $$\overline{X} \sim N(\mu, 1)$$

  Interpretation: impossibility to learn. Even millions of observations do not yield any gain in precision of the estimation based on the sample mean.

Two approaches

- continuity of functionals considered (Hampel)

- work with neighborhood models (Huber, Strassen)

  * Imprecise/interval probability provides a powerful superstructure upon these model

  * huge area of potential applications, far beyond testing

Neighborhood models (e.g. Huber (1981, Robust Statistics))

- Central idea: develop a theory of 'approximately true models'
- Instead of $p_0(\cdot)$ only, consider the set of all distributions "close" to $p_0(\cdot)$

$$U_\epsilon(p) := \{p(\cdot) \in \mathcal{K} | d(p_0, p) \leq \epsilon\}$$

  ϵ-neighborhood of $p_0(\cdot)$

- Different models, depending on the norm in which the distance d is measured
  * total-variation norm
  * Levy-Prokorov norm
  * Kolmogorov norm
  * Levy norm

- most common model: ϵ-**contamination model** Huber (1965), Rieder (1977, 1978), Walley: linear-vacuous mixture

  * unobserved heterogenity:
  * $(1-\epsilon)\cdot 100\%$ of the observations are distributed according to $p_0(\cdot)$
  * $\epsilon \cdot 100\%$, however, come from any distribution
  * sensitivity analysis application (Rieder). How large can ϵ be chosen such that the essential conclusions do not change qualitatively?

- All these models lead to two-monotone probabilities

- In the case of the ϵ-contamination model

$$\begin{aligned} L(A) &= (1-\epsilon)\cdot p_0(A), & \forall A \in \mathcal{A} \\ U(A) &= (1-\epsilon)\cdot p_0(A) + \epsilon, & \forall A \in \mathcal{A}. \end{aligned}$$

General form of neighborhood models

- **Distorted probabilities** (pseudo capacities, special capacities)

$$L(A) = f(p_0(A)), \quad \forall A \in \mathcal{A}.$$

With $p_0(\cdot)$ classical probability (central distribution) and f distortion function:

$$f : [0,1] \to [0,1]; \quad f(1) = 1.$$

- If f convex, then $L(\cdot)$ is two-monotone.

- Quite general: $f(x) \leq x, \forall x \in [0,1]$ yields R-probability, i.e. non-empty structure.

- Wide class of F-probabilities:
$f(\cdot)$ bi-elastic: Wallner (2003, ISIPTA)

- Robust premium principles (Denneberg)

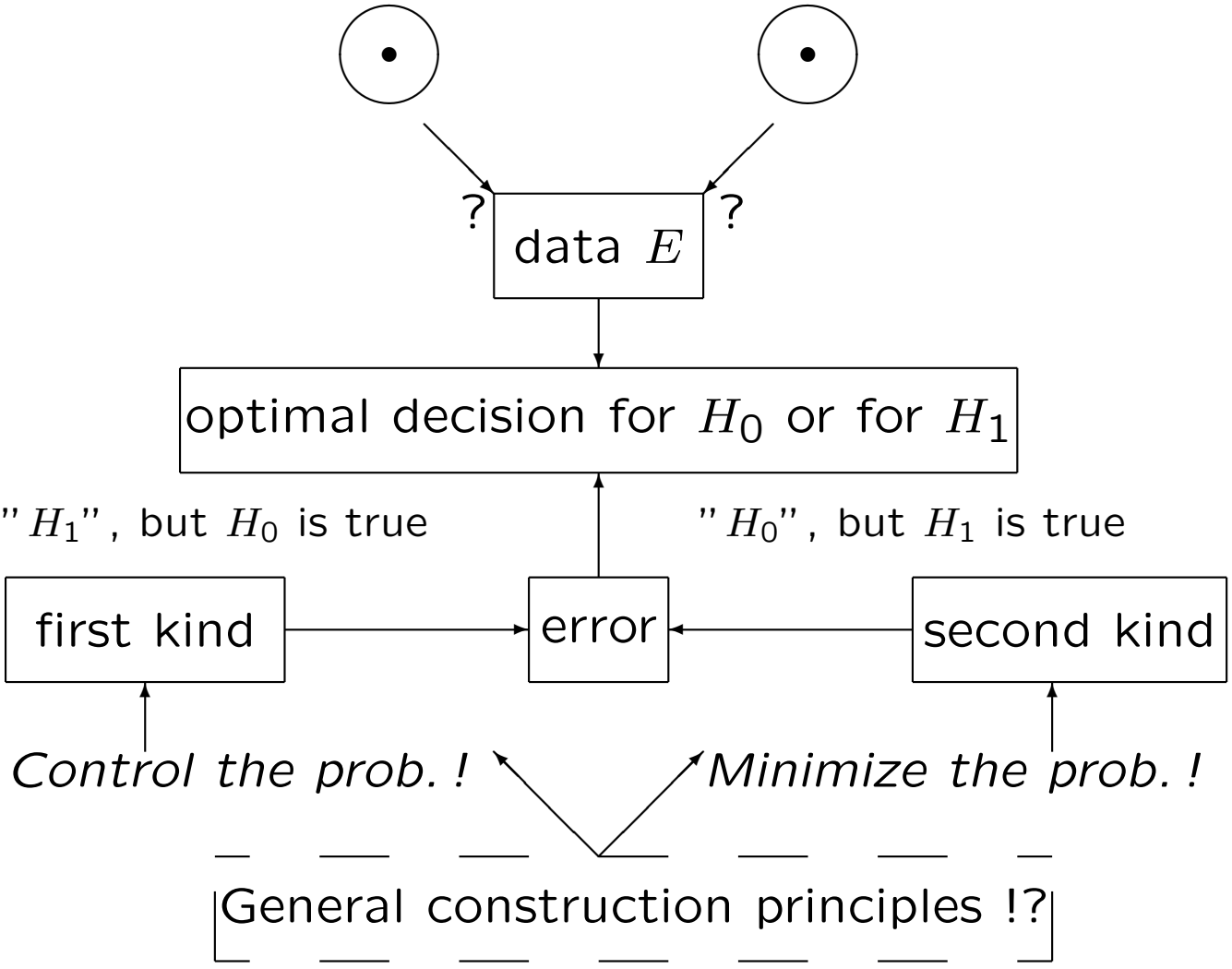
Generalized neighborhood models Augustin (2002a, JStatPlanInf), Wallner (2003, ISIPTA)

- Neighborhood models for interval probability
- Use F-probability field $(\Omega, \mathcal{A}, L_0(\cdot))$ as central distribution
- $L(A) = f(L_0(A)),\ \forall A \in \mathcal{A}$

2.3 Huber Strassen Theory

The generalized Neyman Pearson problem

$H_0 : \{ P_0(\cdot) \}$ $H_1 : \{ P_1(\cdot) \}$



Neyman-Pearson testing between interval prob.

- 'sample size 1'. Situations with sample size n are included by considering the independent products.
- Decision via test(-functions) $\varphi : \Omega \to [0, 1]$: Only classical (and not interval-valued) probabilities of rejecting hypotheses are allowed. (Idea of randomization based on an idealized random-experiment without any non-probabilistic uncertainty)

Optimality criteria

- Minimize the probability of the error of the second kind $\mathbb{E}_{\mathcal{M}_1}(1-\varphi)$
  while controlling for the error of the first kind $\mathbb{E}_{\mathcal{M}_0}\varphi$
- Conservative view in testing: interval ordering by the upper bound of the errors only
- φ^* **level-α-maximin-test** (**LAMT**) (for $\mathcal{F}_0$ versus $\mathcal{F}_1$) , if
  * $\varphi^*(\cdot)$ respects the given level of significance $\alpha \in (0,1)$, i.e.

$$\overline{\mathbb{E}}_{\mathcal{M}_0}\varphi^* \leq \alpha\,.$$

  * $\varphi^*(\cdot)$ has maximal power among all tests under consideration, i.e.

$$\forall \psi \in \Phi \left[\overline{\mathbb{E}}_{\mathcal{M}_0}\psi \leq \alpha \Rightarrow \underline{\mathbb{E}}_{\mathcal{M}_1}\psi \leq \underline{\mathbb{E}}_{\mathcal{M}_1}\varphi^*\right].$$

(Globally) least favorable pairs

Main paper: **Huber & Strassen (1973)**

Construction of LAMTs?

- Idea: 'If one succeeds in convincing the hardliner of two parties one has convinced all their members'. Therefore,
- search for two elements $q_0(\cdot)$ and $q_1(\cdot)$ of the structures, where the testing is most difficult.

- A **Globally least favorable pair (GLFP))** is a pair $(q_0(\cdot), q_1(\cdot))$ of classical probabilities with the following properties:

  * $(q_0(\cdot), q_1(\cdot)) \in \mathcal{M}_0 \times \mathcal{M}_1$
  * The critical region based on the likelihood-ratio $\pi(\cdot)$ of $q_0(\cdot)$ and $q_1(\cdot)$ lies least favorable, i.e.

$$\forall t \geq 0, \forall p_0(\cdot) \in \mathcal{M}_0 : \\ p_0(\{\omega \,|\pi(\omega) > t\}) \leq q_0(\{\omega \,|\pi(\omega) > t\})$$

$$\forall t \geq 0, \forall p_1(\cdot) \in \mathcal{M}_1 : \\ p_1(\{\omega \,|\pi(\omega) > t\}) \geq q_1(\{\omega \,|\pi(\omega) > t\}) \,.$$

- In Huber-Strassen theory: 'least favorable pair'

- Note: the reduction to the classical probabilities $q_0(\cdot)$ and $q_1(\cdot)$ does not mean that the ambiguity is cancelled out!
  The interval probability field $\mathcal{F}_i$ is only represented in *this concrete* testing problem by $q_i(\cdot)$.

Properties of globally least favorable pairs

- Globally least favorable pairs indeed lead to level-α-maximin-tests
- Globally least favorable pairs do not depend on the sample size ($\rightarrow$ very convenient for calculation)
- **Huber-Strassen theorem** (Huber, Strassen (1973), Buja (1986))
  If $\mathcal{F}_0$ and $\mathcal{F}_1$ are two-monotone probability fields with

$$(A_n)_{n\in\mathbb{N}} \uparrow A,\ A_n \text{ open}, n \in \mathbb{N} \Longrightarrow \lim_{n\to\infty} L(A_n) = L(A)\,,$$

  then there exists a globally least favorable pair for $\mathcal{F}_0$ versus $\mathcal{F}_1$.

- construction algorithms mainly for neighborhood models (e.g. Rieder (1977), Österreicher (1978), Bednarski (1981), Hafner (1992))

- 'necessity theorem'

Necessity theorem Let $\mathcal{F}_0$ be an F-probability field with structure $\mathcal{M}_0$. If there exists for any K-probability $p_1(\cdot)$ with $p_1(\cdot) \notin \mathcal{M}_0$ a K-probability $p_0(\cdot) \in \mathcal{M}_0$ such that $(p_0(\cdot), p_1(\cdot))$ is a globally least favorable pair for $\mathcal{F}_0$ versus $\mathcal{F}_1 := (\Omega, \mathcal{A}, p_1(\cdot))$, then $\mathcal{F}_0$ must be a two-monotone probability field.

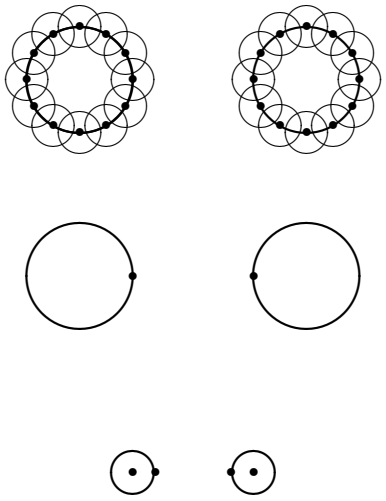
2.5 Extension to General Interval Probability

- nevertheless **generalization** to general interval probability possible (Augustin (1998, 1999a,b, 2002a oder b?***))

- Existence of GLFP in particular in the case of

  * generalizations of parametric distributions with monotone density ratio to interval prob. (e.g. F-normal distribution)
  * generalized neighborhood models (neighborhood models of interval probabilities)

- generalization of construction algorithms by looking at least favorable pairs of interval probabilities



Idea:

Least favorable pseudo-capacity

=

(Usual) pseudo-capacity around the GLFP
for the central distributions ?

Sufficient conditions available.

- **locally least favorable pairs** (least favorable position only for a certain value of the level of significance α (adopting Baumann (1968), Plachky & Rüschendorf (1984), see Augustin (1998, 1999b)).

  * $(q_0(\cdot), q_1(\cdot)) \in \mathcal{M}_0 \times \mathcal{M}_1$,
  * and there exists a best test $\varphi^*(\cdot)$ for $\{q_0(\cdot)\}$ versus $\{q_1(\cdot)\}$ with $\sqcup\!\!\mathbb{E}_{\mathcal{M}_0}\varphi^* \leq \alpha$ and $I\!E_{q_1}\varphi^* = \sqcup\!\!\mathbb{E}_{\mathcal{M}_1}\varphi^*$.

- Also sufficient to obtain LAMTS
- Sufficient for the existence: F-probability fields with

$$(A_n)_{n\in I\!N} \uparrow A,\ n \in I\!N \Longrightarrow$$

$$\lim_{n\to\infty} L_i(A_n) = L_i(A)\,, \quad i \in \{0,1\}\,.$$

- On finite spaces: Reformulation as a linear optimization problem
- There dualization yields a generalization of the Neyman Pearson Lemma.

3. Decision Making

3.1 Basic Concepts, No-data Problem

	Actions
a) Walk to the mountain	different types to dress
b) Investment problem	Investment No Investment
c) medical checkup	different treatment

	Actions	States of Nature
a)	different types to dress	hot / warm / rain / snow
b)	Investment No Investment	economic situation (upswing, neutral, downswing)
c)	different treatments	different diseases no disease

Classical Decision Theory

no-data problem (on finite spaces)

- set $\mathcal{A} = \{a_1, \ldots, a_s, \ldots, a_n\}$ of *actions*,
- set $\Theta = \{\vartheta_1, \ldots, \vartheta_j, \ldots, \vartheta_m\}$ of *states* of nature
- precise *loss function* (or utility)

$$\begin{array}{rccc} l & : & (\mathcal{A} \times \Theta) & \to & \mathbb{R} \\ & & (a, \vartheta) & \mapsto & l(a, \vartheta) \end{array} ,$$

- represented in an *loss table*

	ϑ_1	ϑ_j	ϑ_m
a_1	$l(a_1, \vartheta_1) \ldots$	$l(a_1, \vartheta_j) \ldots$	$l(a_1, \vartheta_m)$
	$\vdots$	$\vdots$	$\vdots$
a_s	$l(a_s, \vartheta_1) \ldots$	$l(a_s, \vartheta_j) \ldots$	$l(a_s, \vartheta_m)$
	$\vdots$	$\vdots$	$\vdots$
a_n	$l(a_n, \vartheta_1) \ldots$	$l(a_n, \vartheta_j) \ldots$	$l(a_n, \vartheta_m)$

- associated random variable $\mathbf{l}(a)$ on $(\Theta, \mathcal{P}o(\Theta))$
- **Aim:** Choose an optimal action a^*!

randomized actions

Often more general approach: extension of $\mathbb{A}$ by considering *randomized action*

$$\tilde{a} = (\lambda(a_1), \ldots, \lambda(a_s), \ldots, \lambda(a_n))$$

- with $\lambda(a_s)$ as the classical probability to choose action a_s
  $(s = 1, \ldots, n; \sum_{s=1}^{n} \lambda(a_s) = 1,\ \lambda(a_s) \geq 0\ )$
- $\Lambda(\mathbb{A})$ set of all randomized actions from $\mathbb{A}$
- Such mixed actions need not be senseless (e.g. mixture of investment strategies).
- Simplifies formal treatment.
- Utility of a randomized action $\tilde{a}$ at state θ_j:

$$u(\tilde{a}; \theta_j) = \sum_{s=1}^{n} \Big(u(a_s; \theta_j) \cdot \lambda(a_s) \Big)$$

Optimality criteria

1) Minimax optimality

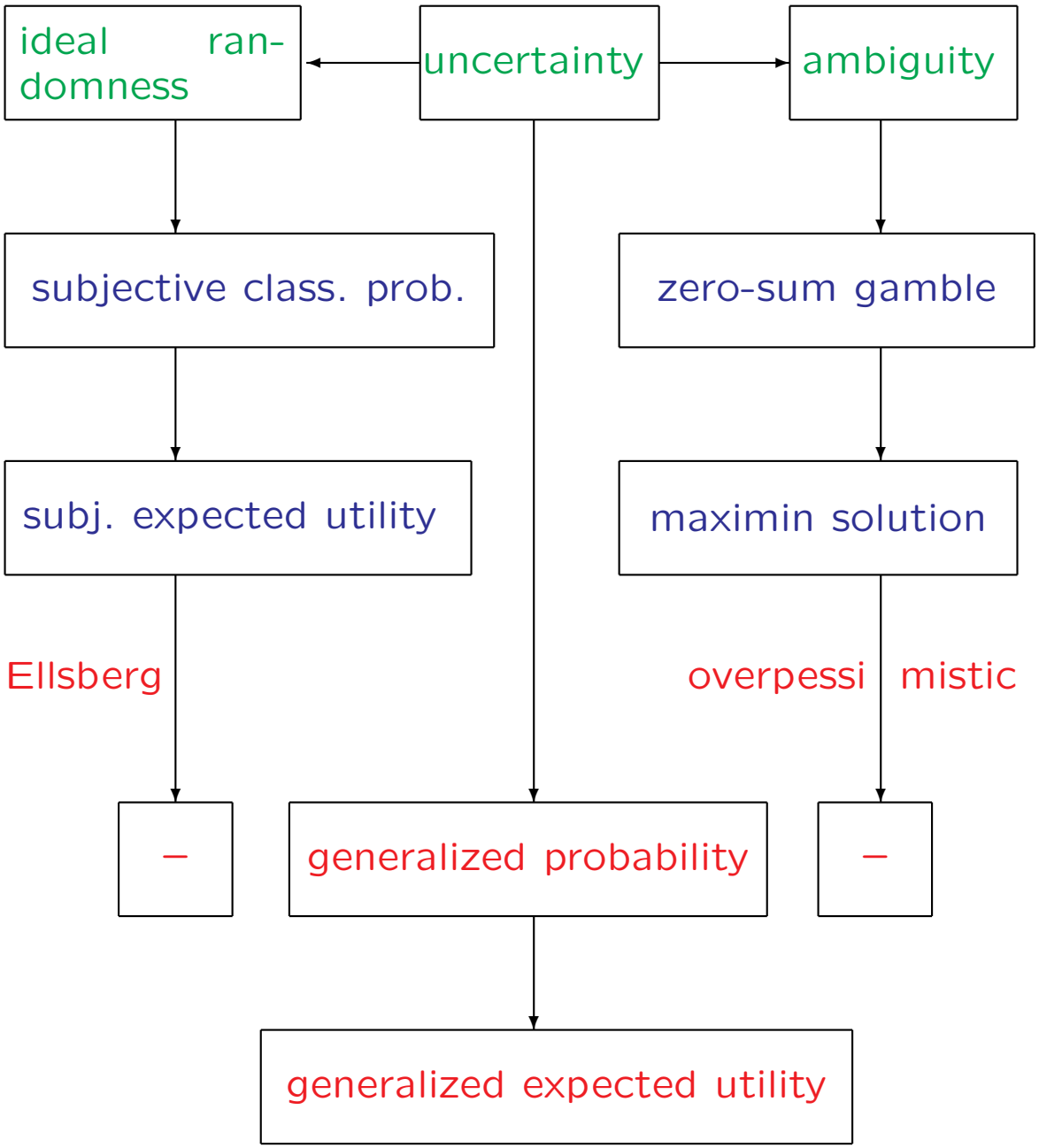
- Wald: decision problems as gambles against nature; zero-sum game; concentrate on the worst states

- In the no-data problem: $\max_{\vartheta \in \Theta} l(a, \vartheta) \rightarrow \min$

2) Bayes optimality with respect to prior $\pi(\cdot)$ on $(\Theta, \mathcal{P}o(\Theta))$. Maximize **expected utility**

[P1] Every uncertainty can adequately be described by a classical probability distribution $\Rightarrow$ prior distribution $\pi(\cdot)$ for a parameter

- In the no-data problem: $\mathbb{E}_\pi(l(a, \vartheta)) \rightarrow \min$


ideal randomness
uncertainty
ambiguity
subjective class. prob.
zero-sum gamble
subj. expected utility
maximin solution
Ellsberg
overpessimistic
–
generalized probability
–
generalized expected utility

Ellsberg's Experiments

- Ellsberg (1961, Quart. J. Econ.)
- Ellsberg (2001, Series of most influential Harvard theses)
- Does the difference between an ideal lottery situation and the general decision situation under uncertainty matter?
- Urn with balls of three different colours: one with known proportion, two with partially unknown proportions
- participants express preferences which can not be modelled by any classical probability measure
- deliberate (not only empirical!) violations of the axioms of (classical) probability!
- Conclusion: (Classical) probability is insufficient to adequately model ambiguous uncertainty.

Ambiguity

- Ellsberg (1961, Quart. J. Econ.)
  Ellsberg (2001, Series of most influential Harvard theses)
- in psychology, management science and economics
  * bibliography: Smithson (1999, tech. report)
  * collection of important papers: Hamouda & Rowley (1997, Edward Elgar)
  * in principle even in Knight (1921) & Keynes (1921)
- in statistics
  * Walley (1991, Chap. & Hall, Ch. 5)
  * Weichselberger (2001, Physica, Ch. 1, 2.6)
  * ISIPTA Proceedings (1999, 2001, 2003)
  * Special volumes Statistical Papers (2002), J.Stat.Plan.Inf. (2002)
- in artificial intelligence
  * Uncertainty in Artificial Intelligence Proceedings (Annual)

3.2 Generalized Expected Utility/Loss

- **Def.: Generalized expected loss**

  * basic decision problem $(\mathbb{A}, \Theta, l(\cdot))$

  * F-probability $\Pi(\cdot)$ on $(\Theta, \mathcal{P}_o(\Theta))$ with structure $\mathcal{M}$.

  Then, for every pure action $a \in \mathbb{A}$ and for every randomized action $a \in \Lambda(\mathbb{A})$, resp.,

  $$\mathbb{E}_{\mathcal{M}} \mathrm{l}(a)$$

  is the *generalized expected loss* (with respect to the prior $\Pi(\cdot)$).

- Notice: $\mathbb{E}_{\mathcal{M}} \mathrm{l}(a)$ is an interval-valued quantity. If a linear ordering is desired $\longrightarrow$ *representation*

Interval ordering

When is $[\underline{\mathbb{E}}_{\mathcal{M}}X\,;\,\overline{\mathbb{E}}_{\mathcal{M}}X] \preceq [\underline{\mathbb{E}}_{\mathcal{M}}Y\,;\,\overline{\mathbb{E}}_{\mathcal{M}}Y]$?

- partial ordering, e.g. $\mathbb{E}_{\mathcal{M}}X \preceq \mathbb{E}_{\mathcal{M}}Y$ iff
  - $\overline{\mathbb{E}}_{\mathcal{M}}X \leq \underline{\mathbb{E}}_{\mathcal{M}}Y$ or
  - $\underline{\mathbb{E}}_{\mathcal{M}}X \leq \underline{\mathbb{E}}_{\mathcal{M}}Y \wedge \overline{\mathbb{E}}_{\mathcal{M}}X \leq U\mathbb{E}_{\mathcal{M}}Y$
- linear (complete) ordering by real-valued representations (here: **pignistic view**)
  - in general $\mathbb{E}_{\mathcal{M}}X \preceq \mathbb{E}_{\mathcal{M}}Y$ iff

    $f(\underline{\mathbb{E}}_{\mathcal{M}}X;\,\overline{\mathbb{E}}_{\mathcal{M}}X) \leq f(\underline{\mathbb{E}}_{\mathcal{M}}Y\,;\,\overline{\mathbb{E}}_{\mathcal{M}}Y)$

    for an appropriate function $f(\cdot)$
  - usual representation: $f(\mathbb{E}_{\mathcal{M}}X) := \overline{\mathbb{E}}_{\mathcal{M}}X$ (upper bound, Γ-Minimax principle, Max-E-Min principle, (e.g. Berger (1990), Kofler (1989))
  - Ellsberg (1961), Jaffray (1989), Weichselberger & A. (1998), Weichselberger (2000, Chapter 2.6) linear representation with *caution* η

$$f(\underline{\mathbb{E}}_{\mathcal{M}}X;\,\overline{\mathbb{E}}_{\mathcal{M}}X) = \eta\cdot\underline{\mathbb{E}}_{\mathcal{M}}X + (1-\eta)\cdot\overline{\mathbb{E}}_{\mathcal{M}}X\,.$$

- easiest choice: $\mathbb{E}_{\mathcal{M}} \mathrm{l}(a) \mapsto {}^{\cup}\mathbb{E}_{\mathcal{M}} \mathrm{l}(a)$
  - a^* is optimal iff

$$ {}^{\cup}\mathbb{E}_{\mathcal{M}} \mathrm{l}(a^*) \leq {}^{\cup}\mathbb{E}_{\mathcal{M}} \mathrm{l}(a) \,, \quad \forall a \in \Lambda(\mathbb{A}) \,. $$

  - strict ambiguity aversion
  - Gamma-Minimax criterion (e.g., Berger (1984, Springer, Section 4.7.6), Vidakovic (2000, in Rios-Insua & Ruggeri (eds.)),
  - Maxmin expected utility model (Gilboa & Schmeidler (1989, J. Math. Econ.))
  - MaxEMin criterion (Kofler & Menges (1976, SpringerLN Econ); Kofler (1989, Campus))
  - For two-monotone capacities: Choquet expected utility (e.g., Chateauneuf, Cohen & Meilijson (1997, Finance))

!!

The two classical decision criteria are contained as border cases:

* perfect probabilistic information, no ambiguity: $\mathcal{M} = \{\pi(\cdot)\} \longrightarrow$ Bayes optimality with respect to $\pi(\cdot)$.

* Completely lacking information, $\Pi(B) = [0;1]$, for every $B \in \mathcal{P}o(\Omega) \setminus \{\emptyset, \Theta\}$, (‘*non-selective* or *vacuous* prior’); leads to the maximin criterion

Calculating the optimal randomized action
(Augustin, 2002b, Stat. Pap.)

$$\mathbb{E}_{\mathcal{M}} l(\tilde{a}) \to \min_{a \in \Lambda(\mathbb{A})},$$

$$\iff \max_{\pi(\cdot)\in\mathcal{M}} \sum_{j=1}^{m} \Big(l(\tilde{a};\vartheta_j) \cdot \pi(\{\vartheta_j\}) \Big) \to \min_{a \in \Lambda(\mathbb{A})}$$

$$\iff \max_{\pi(\cdot)\in\mathcal{M}} \sum_{j=1}^{m} \left(\sum_{s=1}^{n} \Big(l(a_s;\theta_j)\lambda(a_s) \Big) \cdot \pi(\{\vartheta_j\}) \right) \to \min_{\lambda(\cdot)}$$

subject to the constraints

$$\begin{aligned} \sum_{s=1}^{n} \lambda(a_s) &= 1 \\ \lambda(a_s) &\geq 0, \qquad s = 1, \ldots, n\,. \end{aligned}$$

This problem can be transformed to become a linear problem.

Auxiliary variable g for $\mathbb{UE}(\tilde{a})$ leading to the problem:

$$g \to \min_{\lambda(\cdot)}$$

subject to the constraints

$$\begin{aligned}
\sum_{j=1}^{m} \left(\sum_{s=1}^{n} \Big(l(a_s, \vartheta_j) \cdot \lambda(a_s) \Big) \cdot \pi(\{\vartheta_j\}) \right) &\leq g, \\
&\forall \pi(\cdot) \in \mathcal{M}. \\
\lambda(a_s) &\geq 0\,, \\
s = 1, \ldots, n\,, & \\
\sum_{s=1}^{n} \lambda(a_s) &= 1\,.
\end{aligned}$$

This problem is linear in the variables, but uncountably many constraints!

But remember the vertex reduction lemma:

$$\sum_{j=1}^{m}\left(\sum_{s=1}^{n}\Big(l(a_s,\vartheta_j)\cdot\lambda(a_s)\Big)\cdot\pi(\{\vartheta_j\})\right) \leq g, \quad \forall\pi(\cdot)\in\mathcal{M}.$$

$$\Longleftrightarrow^{VRL}$$

$$\sum_{j=1}^{m}\left(\sum_{s=1}^{n}\Big(l(a_s,\vartheta_j)\cdot\lambda(a_s)\Big)\cdot\pi(\{\vartheta_j\})\right) \leq g, \quad \forall\pi(\cdot)\in\mathcal{E}(\mathcal{M})\,.$$

- Transformed into a single linear programming problem
- Easy proof of existence of solutions and of the convexity of the set of solutions
- Easy proof of existence and calculation of a least favorable prior (slack to g equals 0)
- further insights by dualization?
  * theoretical issues ($\leftarrow$ generalized Neyman Pearson Lemma for testing with interval probability (Augustin (1998, Chapter 5))
  * computational issues (Utkin): calculation of all vertices may be circumvented

Now, for the next few slides, change to a utility table:

Corollary Consider a basic decision problem $(\mathbb{A}, \Theta, u(\cdot))$, and an F-probability $\Pi(\cdot) = [L(\cdot), U(\cdot)]$ on $(\Theta, \mathcal{P}o(\Theta))$ with structure $\mathcal{M}$ where $L(\cdot)$ is a two-monotone capacity. Let Υ be the set of all permutations of $\{1, \ldots, m\}$.

a) The following three statements are equivalent:

   i) $\lambda^*(\cdot)$ is maximizing generalized expected utility with respect to the prior $\Pi(\cdot)$.

   ii) $\lambda^*(\cdot)$ is maximizing Choquet expected utility with respect to $L(\cdot)$, i.e., $\mathbb{E}_L\mathbf{u}(\lambda^*) \geq \mathbb{E}_L\mathbf{u}(\lambda)$, for all $\lambda \in \Lambda(\mathbb{A})$.

   iii) $\lambda^*(\cdot)$ corresponds to an optimal solution $(\lambda^*(\{a_1\}), \ldots, \lambda^*(\{a_n\}, g^*)$ of the linear programming problem

$$g \rightarrow \max$$

subject to the constraints $\lambda(a_s) \geq 0\,, s = 1, \ldots, n\,,\ \sum_{s=1}^{n} \lambda(a_s) = 1$ and

$$\sum_{j=1}^{m} \sum_{s=1}^{n} u(a_s, \vartheta_j) \cdot \lambda(a_s) \cdot$$

$$\left(L\Big(\bigcup_{l=1}^{j} \{\vartheta_{\varsigma(l)}\} \Big) - L\Big(\bigcup_{l=1}^{j-1} \{\vartheta_{\varsigma(l)}\} \Big) \right) \ \geq \ g\,,\ \forall \varsigma \in \Upsilon\,.$$

b) If $\mu(\cdot)$ is the Moebius inverse of $L(\cdot)$, then the relation above can be written as

$$\sum_{j=1}^{m} \sum_{s=1}^{n} u(a_s, \vartheta_j) \cdot \lambda(a_s) \cdot \sum_{A \subseteq \cup_{l=1}^{j-1} \vartheta_{\varsigma(l)}} \mu(A \cup \vartheta_{\varsigma(j)}) \geq g\,,\ \forall \varsigma \in \Upsilon\,.$$

Other interval orderings: linear combinations

Ellsberg (1963), Jaffray (1989), Weichselberger & A. (1998), Weichselberger (2000, Chapter 2.6): linear representation

Define

$$\mathbb{E}_{\mathcal{M}}\, u(a^*, \theta) \succeq \mathbb{E}_{\mathcal{M}}\, u(a, \theta)\,,$$

iff

$$\eta \cdot L\mathbb{E}_{\mathcal{M}} u(a^*, \theta) + (1-\eta) \cdot U\mathbb{E}_{\mathcal{M}} u(a^*, \theta) \geq$$
$$\eta \cdot L\mathbb{E}_{\mathcal{M}} u(a, \theta) + (1-\eta) \cdot U\mathbb{E}_{\mathcal{M}} u(a, \theta)$$

a^* optimal action with respect to imprecise prior $\Pi(\cdot)$ and *caution* η.

After introducing again randomized actions $\lambda(a_s)$ one has to consider the maximization of

$$\eta \cdot L\!I\!E_{\mathcal{M}}(\tilde{a}) + (1-\eta) \cdot U\!I\!E_{\mathcal{M}}(\tilde{a})$$

$$\iff \quad \eta \cdot \left(\min_{\pi(\cdot)\in\mathcal{M}} \sum_{j=1}^{m} \Big(u(\tilde{a};\vartheta_j) \cdot \pi(\{\vartheta_j\}) \Big) \right) + + (1-\eta) \cdot \left(\max_{\pi(\cdot)\in\mathcal{M}} \sum_{j=1}^{m} \Big(u(\tilde{a};\vartheta_j) \cdot \pi(\{\vartheta_j\}) \Big) \right)$$

$$\iff \quad \eta \cdot \left(\min_{\pi(\cdot)\in\mathcal{M}} \sum_{j=1}^{m} \left(\sum_{s=1}^{n} \Big(u(a_s;\theta_j)\lambda(a_s) \Big) \cdot \pi(\{\vartheta_j\}) \right) \right) - + (1-\eta) \cdot \left(\max_{\pi(\cdot)\in\mathcal{M}} \sum_{j=1}^{m} \left(\sum_{s=1}^{n} \Big(u(a_s;\theta_j)\lambda(a_s) \Big) \cdot \pi(\{\vartheta_j\}) \right) \right)$$

subject to the constraints

$$\begin{aligned} \sum_{s=1}^{n} \lambda(a_s) &= 1 \\ \lambda(a_s) &\geq 0, \qquad s = 1, \ldots, n\,. \end{aligned}$$

Try to make this problem linear:

- auxiliary variable h for $U\mathbb{E}u(\tilde{a})$:

$$\sum_{j=1}^{m}\left(\sum_{s=1}^{n}\Big(u(a_s,\vartheta_j)\cdot\lambda(a_s)\Big)\,\pi(\{\vartheta_j\})\right)=h$$

  subject to the constraint

$$\pi(\cdot)\in\mathcal{M}\,,$$

  with $\pi(\cdot)\in\mathcal{M}$ which is equivalent to

$$L\left(\bigcup_{j\in\mathcal{J}}\{\vartheta_j\}\right)\leq\sum_{j\in\mathcal{J}}\pi(\{\vartheta_j\})\leq U\left(\bigcup_{j\in\mathcal{J}}\{\vartheta_j\}\right)\,.$$

$$\forall\mathcal{J}\subseteq\{1,\ldots,m\}$$

- auxiliary variable g for $L\mathbb{E}u(\tilde{a})$:

$$\sum_{j=1}^{m}\left(\sum_{s=1}^{n}\Big(u(a_s,\vartheta_j)\cdot\lambda(a_s)\Big)\,\pi(\{\vartheta_j\})\right)\;\geq\;g,$$

$$\forall\pi(\cdot)\in\mathcal{M}\text{ resp.}\qquad\forall\pi(\cdot)\in\mathcal{E}(\mathcal{M})\,.$$

$$\eta \cdot g + (1-\eta) \cdot h \to \max_{\lambda(\cdot) \in \Lambda(\mathbb{A})}$$

subject to the constraints

$$\left(\sum_{j=1}^{m} \left(\sum_{s=1}^{n} \left(u(a_s, \vartheta_j) \cdot \lambda(a_s) \right) \right) \pi(\{\vartheta_j\}) \right) \geq g\,, \quad \forall \pi(\cdot) \in \mathcal{E}(\mathcal{M})$$

$$\left(\sum_{j=1}^{m} \left(\sum_{s=1}^{n} \left(u(a_s, \vartheta_j) \cdot \lambda(a_s) \right) \right) \bar{\pi}(\{\vartheta_j\}) \right) = h$$

$$U\left(\bigcup_{j \in \mathcal{J}} \{\vartheta_j\} \right) \geq \sum_{j \in \mathcal{J}} \bar{\pi}(\{\vartheta_j\}) \geq L\left(\bigcup_{j \in \mathcal{J}} \{\vartheta_j\} \right), \quad \forall \mathcal{J} \subseteq \{1, \ldots, m\}\,,$$

$$\sum_{s=1}^{n} \lambda(a_s) = 1, \quad \lambda(a_s) \geq 0, \; s = 1, \ldots$$

- This problem is linear only for $\eta = 1$. For $\eta < 1$ bilinear optimization problem
- Many results still hold.
- Software available, or split the problem into several linear problems.
- Also neat results available by appropriate dualization theory????

3.3 Utilizing Additional Information (Data)

	Actions	States of Nature	
a)	different types to dress	hot / warm / rain / snow	
b)	Investment No Investment	economic situation (upswing, neutral, downswing)	
c)	different treatment	different diseases no disease	

	Additional Inform
a)	looking at the barometer
b)	Expert judgement
c)	X-ray

Data problem

- Incorporate additional information from a sample !
- Choose an optimal strategy !
- What is the value of a certain information ?

Information on ϑ_j from an experiment where the probability depends on ϑ_j:

For every j a classical probability $p_j(\cdot)$ is given

	$x_1 \dots$	$x_i \dots$	x_k
ϑ_1	$p_1(\{x_1\}) \dots$	$p_1(\{x_i\}) \dots$	$p_1(\{x_k\})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
ϑ_j	$p_j(\{x_1\}) \dots$	$p_j(\{x_i\}) \dots$	$p_j(\{x_k\})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
ϑ_m	$p_m(\{x_1\}) \dots$	$p_m(\{x_i\}) \dots$	$p_m(\{x_k\})$

Often $p_j(\{x_i\})$ is interpreted as $p(\{x_j\}|\{\vartheta_j\})$.

decision functions (strategies)

- describing randomized action in dependence on the observation $\{x_i\}$

$$\begin{array}{rccl} d: & \{x_1, \ldots, x_k\} & \to & \Lambda(\mathbb{A}) \\ & x_i & \mapsto & d(x_i) = a \ . \end{array}$$

- randomized decision functions $d(x_i, a_s)$; classical probability to choose a_s if $\{x_i\}$ occurs.
- $\mathcal{D}$ set of all decision functions
- associated random variable $\mathrm{l}(d, \vartheta_j)$ on $(\Omega, \mathcal{A})$
- risk of $d(\cdot)$

$$R(d, \vartheta_j) := \mathbb{E}_{p_j}\left(\mathrm{l}(d, \vartheta_j)\right).$$

- New decision problem $(\mathcal{D}, \Theta, R(\cdot, \cdot))$.

The value of the information experiment

loss of the optimal action
in the no-data problem
—
risk of the optimal decision function
in the data problem
=
value of information

Always nonnegative.

Optimality criteria

1) Minimax optimality

- In the no-data problem: $\max_{\vartheta \in \Theta} l(a, \vartheta) \rightarrow \min$
- In the data problem: $\max_{\vartheta \in \Theta} R(d, \vartheta) \rightarrow \min$

2) Bayes optimality with respect to prior $\pi(\cdot)$ on $(\Theta, \mathcal{P}o(\Theta))$.

- In the no-data problem: $\mathbb{E}_\pi(l(a, \vartheta)) \rightarrow \min$
- In the data problem: $\mathbb{E}_\pi(R(d, \vartheta)) \rightarrow \min$

"Main theorem of Bayesian decision analysis"

- Optimal $d^*(\cdot)$ can be obtained by solving, for every observation $\{x\}$, the no-data problem with the posterior $\pi(\cdot|x)$ as the 'updated prior'.

  optimality with respect to prior risk

  $=$

  optimality with respect to posterior loss
- For maximin solutions NO reduction of the data problem to no-data problems possible.

- Two paradigms underlying the classical Bayesian view

  [P1] Every uncertainty can adequately be described by a single classical probability distribution $\Rightarrow$ prior distribution $\pi(\cdot)$ for a parameter

  [P2] After having observed the sample $\{x\}$, the posterior $\pi(\cdot|x)$ contains all the relevant information. Every inference procedure depends on $\pi(\cdot|x)$, and only on $\pi(\cdot|x)$.

Data problem under interval probability:

- Now, for every j, an F-probability field

$$P_j(\cdot) = [L_j(\cdot), U_j(\cdot)]$$

  with structure $\mathcal{M}_j$ is given.
- risk of the decision function $d(\cdot)$

  * Given $\vartheta_j : \mathbb{E}_{\mathcal{M}_j}\left(\mathrm{l}(d, \vartheta_j)\right)$ represented by

$$\mathbf{R}(d) := {}^{\cup}\mathbb{E}_{\mathcal{M}_j}\left(\mathrm{l}(d, \vartheta_j)\right)$$

  * with prior structure $\mathcal{M}$ look at

$$\mathbb{E}_{\mathcal{M}}\left(\mathbf{R}(d)\right)$$

    represented by

$${}^{\cup}\mathbb{E}_{\mathcal{M}}\left(\mathbf{R}(d)\right)$$

The value of the information experiment

loss of the optimal action
in the no-data problem
—
risk of the optimal decision function
in the data problem
=
value of information

Still always nonnegative.

3.4 Robust Bayesian Procedures Inference by the Generalized Bayes Rule

- classical statistics:

  data problem with prior $\pi(\cdot)$

  $\equiv$

  no-data problem with updated prior $\pi(\cdot|x)$

  $\Rightarrow$ posterior contains full information

- Generalization: Robust Bayesian Inference (Survey: Wasserman (1997, Enc. Stat. Sc., Update 1))

  $$\begin{array}{rl} & \text{prior structure } \mathcal{M} \\ + & \text{observation } x \\ \hline & \text{posteriori structures } \mathcal{M}_{|x} \end{array}$$

  with

  $$\mathcal{M}_{|x} = \{\pi(\cdot|x) | \pi(\cdot) \in \mathcal{M}\}$$

  and $\Pi(\cdot|x) = [\underline{\pi}(\cdot|x), \overline{\pi}(\cdot|x)]$ derived from it.

- Used in Kofler & Menges' (1976) theory of partial information

- Strong justification by coherence axioms (Walley (1991): Generalized Bayes Rule)

- algorithms by Cozman (1999, J. Comp. Graph. Stat.; 2000, Int. J. Approx. Reas.)

- intuitively very plausible

- elegant modelling of prior-data conflict (Walley (1991, Ch. 1))

- successive updating: use $\Pi(\cdot|x)$ as a new prior in handling new observations

BUT

- Decision theoretic justification is lost.

- Decision functions constructed via the posterior structure may have higher risk.

- optimality with respect to imprecise prior risk

$$\neq$$

optimality with respect to imprecise posterior loss

- $\overset{?}{\Longrightarrow}$ The imprecise posterior does not contain all the relevant information !?!?

WHY?

Decision functions constructed via the posterior structure may have higher risk

- Vacuous prior("$\Pi(\cdot) = [0, 1]$")
  - * Minimax decision function $d^*(\cdot)$ minimizes prior risk.
  - * Vacuous posterior for every observation (we do not learn from the data!); minimax action a^* minimizes posterior loss for every observation
  - * Usually $d^* > (a^*, \ldots, a^*)$
- Non-degenerated counterexample: toy example below
- Representation theorem: Optimal decision functions with respect to an imprecise prior $\pi(\cdot)$ are always minimax solutions (in a different decision problem)
- Imprecise posteriors may be dilated (Seidenfeld & Wasserman (1993, Ann. Statist.); Wasserman & Seidenfeld (1994, J. Stat. Plann. Inf.))
  $[\underline{\pi}(\cdot|x), \overline{\pi}(\cdot|x)] \supset [\underline{\pi}(\cdot), \overline{\pi}(\cdot)]\,, \ \forall x$
  Often to a negative value of information.

Representation Theorem: Optimal decision functions with respect to an imprecise prior $\pi(\cdot)$ are always minimax solutions (in a different decision problem): Consider

- a basic decision problem $(\mathbb{A}, \Theta, l(\cdot,\cdot))$ with
- prior structure $\mathcal{M}$ and
- (precise) sampling information $(p_\vartheta(\cdot))_{\vartheta\in\Theta}$

i) An action a^* is optimal optimal with respect to the prior structure $\mathcal{M}$

iff

it is maximin action in the decision problem $(\mathbb{A}, \mathcal{M}, \tilde{l}(\cdot,\cdot))$ with

$$\begin{array}{rccl} \tilde{l} & : & (I\!A \times \mathcal{M}) & \to & \mathbb{R} \\ & & (a,\, \pi) & \mapsto & \tilde{l}(a,\, \pi) := \mathbb{E}_\pi(l(a, \vartheta)) \end{array}$$

ii) A decision function $d^*(\cdot)$ is optimal

iff

$d^*(\cdot)$ is maximin decision function in the decision problem $(\mathcal{D}, \mathcal{M}, \tilde{R}(\cdot,\cdot))$ with

$$\begin{array}{rcccl} \tilde{R} & : & (\mathcal{D} \times \mathcal{M}) & \to & \mathbb{R} \\ & & (d,\, \pi) & \mapsto & \tilde{R}(d,\, \pi) := \mathbb{E}_\pi(R(d, \vartheta))\,. \end{array}$$

Proof: $\max_{\underbrace{\pi(\cdot) \in \mathcal{M}}_{\uparrow}} \underbrace{\mathbb{E}_\pi(l(a, \vartheta))}_{\uparrow}$

new states of nature

$\tilde{l}(a,\, \pi)$

Remarks

- Optimal decision functions have all the ((un)pleasant) properties of minimax solutions.
- Neither

  * equivalence of posterior loss and prior risk

  nor

  * essentially completeness of unrandomized actions (also for robust Bayesian solutions!)

  can be expected.
- Representation similar to Schneeweiß's (1964) representation of a no-data problem.
- Extensions to interval-valued sampling model and Hurwicz-like criterion.
- Framework for decision making with second order probabilities.

3.5 Calculating Decision Functions Minimizing Prior Risk

- Vidakovic (2000, in Rios-Insua & Ruggeri (eds.))
- Noubiap & Seidel (2001, Comp. Stat. & Data Anal.), (2001, Ann. Stat)
- On finite parameter spaces solution via a single linear programming problem available (Augustin (2001, ISIPTA01), (2004a, ZAMM))

Consider finite sample spaces. Minimize

$$\overline{\mathbb{E}}_{\mathcal{M}}\left(\overline{\mathbb{E}}_{\mathcal{M}_j}(l(d,\vartheta_j))\right) = \max_{\pi\in\mathcal{M}} \left(\sum_{j=1}^{m} \max_{p_j(\cdot)\in\mathcal{M}_j} \left(\sum_{i=1}^{k} \left(\sum_{s=1}^{n} \underbrace{l(a_s;\vartheta_j)}_{\text{given}} \cdot \underbrace{d(x_i;a_s)}_{\text{unknown}} \right) \cdot \underbrace{p_j(\{x_i\})}_{\in\mathcal{M}_j} \right) \right) \cdot \underbrace{\pi(\{\vartheta_j\})}_{\in\mathcal{M}}$$

constr. constr.

w.r.t. $\sum_{j=1}^{n} d(x_i, a_s) = 1\,, d(x_i, a_s) \geq 0\ \forall i, s.$

- Make this problem linear: auxiliary variables g for $\mathbb{L}\mathbb{E}_{\mathcal{M}}$, as well as g_j for $\mathbb{L}\mathbb{E}_{\mathcal{M}_j}$.
- Use the vertice reduction lemma to reformulate the task as a linear optimization problem:

$$g \longrightarrow \min$$

subject to the constraints

$$\sum_{i=1}^{k} \left(\sum_{s=1}^{n} l(a_s, \vartheta_j) \cdot d(x_i, a_s) \right) \cdot p_j(\{x_i\}) \leq g_j$$
$$\forall p_j(\cdot) \in \mathcal{E}(\mathcal{M}_j);\ \forall j \in \{1, \ldots, m\}$$
$$\sum_{j=1}^{m} g_j \cdot \pi(\{\vartheta_j\}) \leq g$$
$$\forall \pi(\cdot) \in \mathcal{E}(\mathcal{M})$$
$$\sum_{j=1}^{n} d(x_i, a_s) = 1\,,\ d(x_i, a_s) \geq 0\ \forall i, s\,.$$

- Single linear programming problem
- Easy proof of existence of solutions and of the convexity of the set of optimal decision functions
- further insights by dualization!
- also for Choquet Expected Utility in case of two-monotone capacities or belief functions
- optimal unrandomized actions by integer programming (Boolean optimization) (but not essentially complete)

Further Work in That Direction

- Extension to infinite sample spaces using results by Rüger, Utkin
- Apply dualization:
  * $\rightarrow$ least favorable constellations
  * for hypothesis testing: Generalization of the Generalized Neyman-Pearson Lemma (Augustin (1998, Vandenh. & R., Ch. 5))

- Use more sophisticated interval ordering to model general ambiguity attitudes (cf. above for the no-data problem)

- sequential decision making, but be careful!
  * backward induction: Hermanec (2002, J. Stat. Plan. Inf.)
  * 'sophisticated versus step by step optimal': Jaffray (1999, ISIPTA99)
  * for sequential testing: Augustin & Pöhlmann (2004, J. Stat. Plan. Inf., in Revision.)
  * Teddy's talk

3.6 Concluding Remarks

How to live with the observed discrepancy between posterior loss optimality and prior risk optimality?

- Fundamental questions still debatable (Augustin (2003, ISIPTA03)):
  * What is updating?
  * How to learn from data? (inference)
  * How to make optimal decisions?

  * Does $[\underline{\pi}(\cdot|x), \overline{\pi}(\cdot|x)]$ really deserve to be called posterior, since it
    - does not contain the full information from a sample and
    - leads to suboptimal decision functions ?
  * Implicit definition of posterior ??

    or
  * Separate updating/inference and decision in an uncompromising way !
  * But check for potential paradoxes (statistical estimating and testing problem can be formulated as inference as well as decision problems.)

- Old debate between Bayesians and frequentists is back with higher importance than ever; now the standpoint matters.