Decision Making under Weakly Structured Information

with Applications to Robust Statistics and Machine Learning

SIPTA Seminar, April 30, 2024

Christoph Jansen, Department of Statistics, LMU Munich

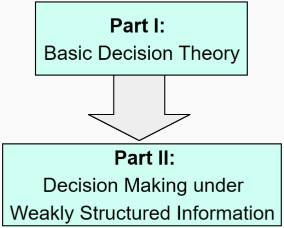
Introducing Myself

- Christoph Jansen
- Member of Thomas Augustin's *Foundations of Statistics* group
- Affiliated with the Department of Statistics of LMU Munich
- About to finish my habilitation in statistics
- Assistant professor at Lancaster University Leipzig starting June 2024
- First contact with SIPTA at the 2014 Summerschool in Montpellier

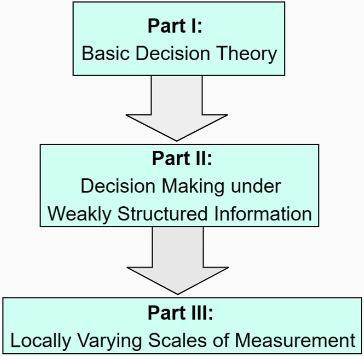
Outline

Part I:
Basic Decision Theory

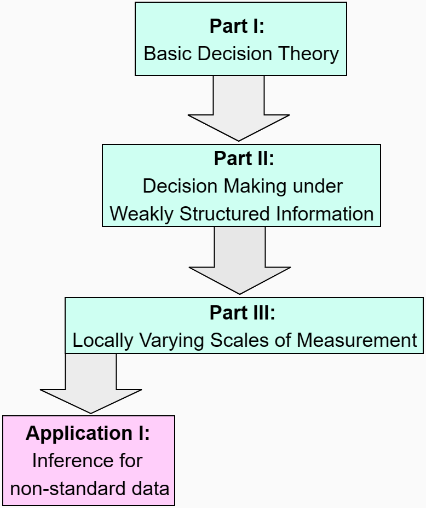
Outline


Part I:
Basic Decision Theory
Part II:
Decision Making under
Weakly Structured Information

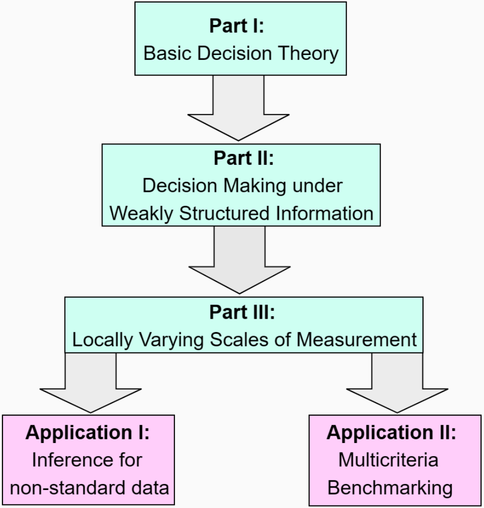
Outline


Part I:
Basic Decision Theory
Part II:
Decision Making under
Weakly Structured Information
Part III:
Locally Varying Scales of Measurement

Outline


Part I:
Basic Decision Theory
Part II:
Decision Making under
Weakly Structured Information
Part III:
Locally Varying Scales of Measurement
Application I:
Inference for
non-standard data

Outline


Part I:
Basic Decision Theory
Part II:
Decision Making under
Weakly Structured Information
Part III:
Locally Varying Scales of Measurement
Application I:
Inference for
non-standard data
Application II:
Multicriteria
Benchmarking

Basic Decision Theory

Classical Decision Theory

Informal description of the model:

- An agent has to choose among different acts X from a set $\mathcal{G}$.
- The consequence that choosing X yields depends on which state of nature s from a set S is the true one.

Classical Decision Theory

Informal description of the model:

- An agent has to choose among different acts X from a set $\mathcal{G}$.
- The consequence that choosing X yields depends on which state of nature s from a set S is the true one.

Formal description of the model:

- Let A denote some non-empty set of consequences.
- Each act X corresponds to a mapping $X : S \to A$.
- The set $\mathcal{G}$ is a subset of $A^S = \{X : S \to A\}$.

Classical Decision Theory

Informal description of the model:

- An agent has to choose among different acts X from a set $\mathcal{G}$.
- The consequence that choosing X yields depends on which state of nature s from a set S is the true one.

Formal description of the model:

- Let A denote some non-empty set of consequences.
- Each act X corresponds to a mapping $X : S \to A$.
- The set $\mathcal{G}$ is a subset of $A^S = \{X : S \to A\}$.

Goal: Determining a choice function

$$ch : 2^{\mathcal{G}} \to 2^{\mathcal{G}} \quad \text{with} \quad ch(\mathcal{D}) \subseteq \mathcal{D} \quad \text{for all} \quad \mathcal{D} \in 2^{\mathcal{G}}$$

that best possibly utilizes the available information.

Interpreting Choice Functions

Depending on the quality of the underlying information, the choice sets $ch(\mathcal{D})$ can be given two different interpretations:

Interpreting Choice Functions

Depending on the quality of the underlying information, the choice sets $ch(\mathcal{D})$ can be given two different interpretations:

Strong interpretation:

$ch(\mathcal{D})$ is the set of equally optimal acts from $\mathcal{D}$.
The agent is indifferent between these acts.

Interpreting Choice Functions

Depending on the quality of the underlying information, the choice sets $ch(\mathcal{D})$ can be given two different interpretations:

Strong interpretation:

$ch(\mathcal{D})$ is the set of equally optimal acts from $\mathcal{D}$.
The agent is indifferent between these acts.

Weak interpretation:

$ch(\mathcal{D})$ is the set of all non-neglectable acts from $\mathcal{D}$ given the information.
These acts are incomparable for the agent.

Interpreting Choice Functions

Depending on the quality of the underlying information, the choice sets $ch(\mathcal{D})$ can be given two different interpretations:

Strong interpretation:

$ch(\mathcal{D})$ is the set of equally optimal acts from $\mathcal{D}$.
The agent is indifferent between these acts.

Weak interpretation:

$ch(\mathcal{D})$ is the set of all non-neglectable acts from $\mathcal{D}$ given the information.
These acts are incomparable for the agent.

Obvious comment:

If only weakly structured information is available, we often have to work with weakly interpretable choice functions.

Two Classical Choice Functions under Risk

Expected utility:

If a probability π on S and a cardinal scale $u : A \to [0, 1]$ are available, set

$$ch_{u,\pi}(\mathcal{D}) = \Big\{ Y \in \mathcal{D} : \mathbb{E}_\pi(u \circ Y) \geq \mathbb{E}_\pi(u \circ X) \text{ for all } X \in \mathcal{D} \Big\},$$

and choose that acts from $\mathcal{D}$ that maximize expected utility.

Two Classical Choice Functions under Risk

Expected utility:

If a probability π on S and a cardinal scale $u : A \to [0,1]$ are available, set

$$ch_{u,\pi}(\mathcal{D}) = \Big\{Y \in \mathcal{D} : \mathbb{E}_\pi(u \circ Y) \geq \mathbb{E}_\pi(u \circ X) \text{ for all } X \in \mathcal{D}\Big\},$$

and choose that acts from $\mathcal{D}$ that maximize expected utility.

First-Order Stochastic Dominance:

If a probability π on S and a preorder $\succsim$ on A are available, set

$$ch_{\succsim,\pi}(\mathcal{D}) = \left\{ Y : \nexists X \text{ s.t. } \begin{array}{lll} \mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y) & \text{for all} & u \in \mathcal{U}_\succsim \\ \mathbb{E}_\pi(u \circ X) > \mathbb{E}_\pi(u \circ Y) & \text{for some} & u \in \mathcal{U}_\succsim \end{array} \right\}$$

where $\mathcal{U}_\succsim$ is the set of all $\succsim$-isotone $u : A \to [0,1]$. Choose acts that are not excluded by every compatible EU-maximizer.

A Toy Example

An agent wants to invest in exactly one of the stocks in $\mathcal{G} = \{X_1, X_2, X_3\}$.

The consequence depends on the true economic scenario from $S = \{s_1, s_2, s_3\}$.

Suppose we have the following consequence table, where $A = \{a_1, \ldots, a_9\}$:

	s_1	s_2	s_3
X_1	a_1	a_4	a_7
X_2	a_2	a_5	a_8
X_3	a_3	a_6	a_9

Moreover, assume π is the uniform distribution on S.

A Toy Example, continued: Cardinal preferences

Assume, the agent's preferences allow for a description via a cardinal utility $u : A \to \mathbb{R}$ on A (i.e., u is unique up to plts).

A Toy Example, continued: Cardinal preferences

Assume, the agent's preferences allow for a description via a cardinal utility $u : A \to \mathbb{R}$ on A (i.e., u is unique up to plts).

$\Rightarrow$ Consequence table can be transformed in utility table, e.g.:

	s_1	s_2	s_3
$u \circ X_1$	6000	3000	-2000
$u \circ X_2$	8000	1000	-3000
$u \circ X_3$	5000	4000	0

Due to uniqueness of u, maximizing expected utility is well-defined.

A Toy Example, continued: Cardinal preferences

Assume, the agent's preferences allow for a description via a cardinal utility $u : A \to \mathbb{R}$ on A (i.e., u is unique up to plts).

$\Rightarrow$ Consequence table can be transformed in utility table, e.g.:

	s_1	s_2	s_3
$u \circ X_1$	6000	3000	-2000
$u \circ X_2$	8000	1000	-3000
$u \circ X_3$	5000	4000	0

Due to uniqueness of u, maximizing expected utility is well-defined.

$\Rightarrow$ We can apply

$$\text{ch}_{u,\pi}(\mathcal{G}) = \text{argmax}_{X \in \mathcal{G}} \mathbb{E}_\pi(u \circ X) = \{X_3\}$$

A Toy Example, continued: Cardinal preferences

Assume, the agent's preferences allow for a description via a cardinal utility $u : A \to \mathbb{R}$ on A (i.e., u is unique up to plts).

$\Rightarrow$ Consequence table can be transformed in utility table, e.g.:

	s_1	s_2	s_3
$u \circ X_1$	6000	3000	-2000
$u \circ X_2$	8000	1000	-3000
$u \circ X_3$	5000	4000	0

Due to uniqueness of u, maximizing expected utility is well-defined.

$\Rightarrow$ We can apply

$$\mathrm{ch}_{u,\pi}(\mathcal{G}) = \mathrm{argmax}_{X \in \mathcal{G}} \mathbb{E}_\pi(u \circ X) = \{X_3\}$$

$\Rightarrow$ X_3 is the unique optimal stock. (Strong interpretation!)

A Toy Example, continued: (P)ordinal preferences

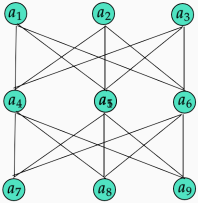
Now assume, the preferences allow only for a preorder $\succsim$ on A.

A Toy Example, continued: (P)ordinal preferences

Now assume, the preferences allow only for a preorder $\succsim$ on A.

Then, the situation looks for instance like this:

	s_1	s_2	s_3
X_1	a_1	a_4	a_7
X_2	a_2	a_5	a_8
X_3	a_3	a_6	a_9



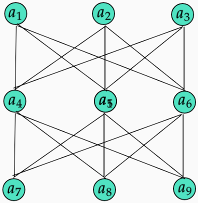
Every $\succsim$-isotone function $u : A \to \mathbb{R}$ is a compatible scale.

A Toy Example, continued: (P)ordinal preferences

Now assume, the preferences allow only for a preorder $\succsim$ on A.

Then, the situation looks for instance like this:

	s_1	s_2	s_3
X_1	a_1	a_4	a_7
X_2	a_2	a_5	a_8
X_3	a_3	a_6	a_9



Every $\succsim$-isotone function $u : A \to \mathbb{R}$ is a compatible scale.

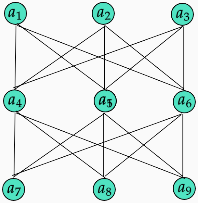
$\Rightarrow$ Maximizing expected utility is not-well-defined!

A Toy Example, continued: (P)ordinal preferences

Now assume, the preferences allow only for a preorder $\succsim$ on A.

Then, the situation looks for instance like this:

	s_1	s_2	s_3
X_1	a_1	a_4	a_7
X_2	a_2	a_5	a_8
X_3	a_3	a_6	a_9



Every $\succsim$-isotone function $u : A \to \mathbb{R}$ is a compatible scale.

$\Rightarrow$ Maximizing expected utility is not-well-defined!

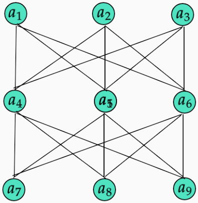
$\Rightarrow$ We can still apply FSD, but now $\mathrm{ch}_{\succsim,\pi}(\mathcal{G}) = \mathcal{G}$, since for every stock there exists a compatible scale making it the unique EU-maximizer.

A Toy Example, continued: (P)ordinal preferences

Now assume, the preferences allow only for a preorder $\succsim$ on A.

Then, the situation looks for instance like this:

	s_1	s_2	s_3
X_1	a_1	a_4	a_7
X_2	a_2	a_5	a_8
X_3	a_3	a_6	a_9



Every $\succsim$-isotone function $u : A \to \mathbb{R}$ is a compatible scale.

$\Rightarrow$ Maximizing expected utility is not-well-defined!

$\Rightarrow$ We can still apply FSD, but now $\mathrm{ch}_{\succsim,\pi}(\mathcal{G}) = \mathcal{G}$, since for every stock there exists a compatible scale making it the unique EU-maximizer.

$\Rightarrow$ None of the stocks can be excluded. (Weak interpretation!)

Weakly structured Information

Common Assumptions in Classic Decision Theory

Classical assumptions:

(I) The agent's preferences among the elements of A are characterized by a cardinal utility function $u : A \to \mathbb{R}$.

(II) The uncertainty among the states from S is described by some classical probability measure π.

Recall:

Expected utility rule $\mathrm{ch}_{u,\pi}(\cdot)$ relies on both (I) and (II).

Stochastic dominance rule $\mathrm{ch}_{\succsim,\pi}(\cdot)$ relies on (II) but not on (I).

Challenging the Classical Assumptions

Problem: Both (I) and (II) rely on **strong axiomatic assumptions**.

(e.g., [von Neumann et al., 1944, Savage, 1954]))

Together, these assumptions explicitly dismiss:

- Purely ordinal or partial preferences.

  (e.g., [Seidenfeld et al., 1995, Nau, 2006]))

- Agents with partial probabilistic beliefs.

  (e.g., [Levi, 1974, Walley, 1991, Kikuti et al., 2011])

- Problems of group decision making.

  (e.g., [Bacharach, 1975, Bradley, 2019]))

These are **highly relevant situations** to investigate!

Weakly structured Information

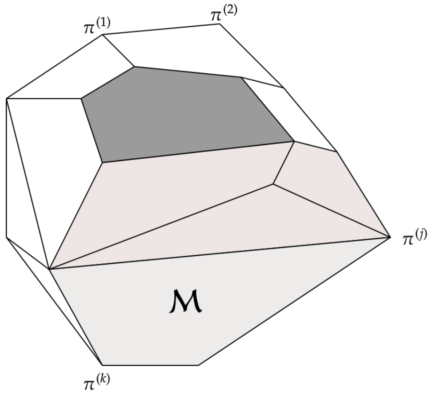
Two different sources of complexity:

Weakly structured Information

Two different sources of complexity:

(I)$'$ Imprecise probabilistic models: If it isn't possible to specify **one** probability on S, we still can work with the *set* $\mathcal{M}$ of all probabilities **compatible with the information**.

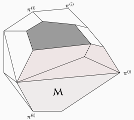
Weakly structured Information



Weakly structured Information

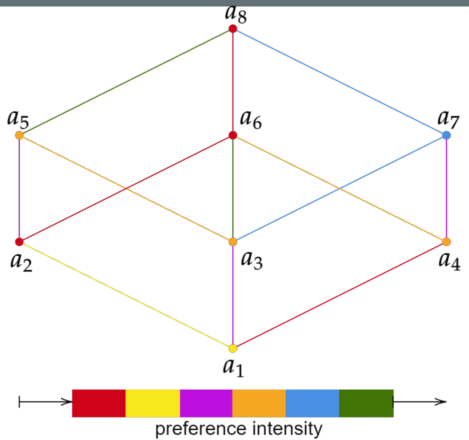
Two different sources of complexity:

$(\mathrm{I})'$ Imprecise probabilistic models: If it isn't possible to specify **one** probability on S, we still can work with the *set* $\mathcal{M}$ of all probabilities **compatible with the information**.



$(\mathrm{II})'$ Complexly ordered consequences: A cardinal utility demands the agent to satisfy **very restrictive axioms**. If these are too restrictive, we still can work with the **set** $\mathcal{U}$ of all utilities **compatible with the information**.

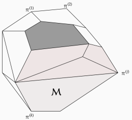
Weakly structured Information



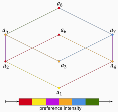
Weakly structured Information

Two different sources of complexity:

(I)′ Imprecise probabilistic models: If it isn't possible to specify **one** probability on S, we still can work with the *set* $\mathcal{M}$ of all probabilities **compatible with the information**.



(II)′ Complexly ordered consequences: A cardinal utility demands the agent to satisfy **very restrictive axioms**. If these are too restrictive, we still can work with the **set** $\mathcal{U}$ of all utilities **compatible with the information**.



Modelling $\mathcal{U}$: Preference Systems

Notation: Binary relation R has **strict part** P_R and **indifference part** I_R.

Preference system & Consistency

Let A denote a set of consequences. Let further

- $R_1 \subseteq A \times A$ be a binary relation on A
- $R_2 \subseteq R_1 \times R_1$ be a binary relation on R_1

The triplet $\mathcal{A} = [A, R_1, R_2]$ is called a **preference system** on A.

We call $\mathcal{A}$ **consistent** if there is $u : A \to [0, 1]$ with for all $a, b, c, d \in A$:

$$(a, b) \in R_1 \Rightarrow u(a) \geq u(b) \qquad (\text{with} = \textit{iff} \in I_{R_1}).$$

$$((a, b), (c, d)) \in R_2 \Rightarrow u(a) - u(b) \geq u(c) - u(d) \qquad (\text{with} = \textit{iff} \in I_{R_2}).$$

The set of all representations u of $\mathcal{A}$ is denoted by $\mathcal{U}_{\mathcal{A}}$.

Interpretation of the components of $\mathcal{A}$:

- $(a, b) \in R_1$: *"a is at least as desirable as b"*
- $((a, b), (c, d)) \in R_2$: *"exchanging b by a is at least as desirable as d by c"*

Modelling $\mathcal{U}$: Preference Systems

Notation: Binary relation R has **strict part** P_R and **indifference part** I_R.

Preference system & Consistency

Let A denote a set of consequences. Let further

$R_1 \subseteq A \times A$ be a binary relation on A

$R_2 \subseteq R_1 \times R_1$ be a binary relation on R_1

The triplet $\mathcal{A} = [A, R_1, R_2]$ is called a **preference system** on A.

We call $\mathcal{A}$ **consistent** if there is $u : A \to [0, 1]$ with for all $a, b, c, d \in A$:

- $(a, b) \in R_1 \Rightarrow u(a) \geq u(b) \qquad$ (with $=$ *iff* $\in I_{R_1}$).
- $((a, b), (c, d)) \in R_2 \Rightarrow u(a) - u(b) \geq u(c) - u(d) \qquad$ (with $=$ *iff* $\in I_{R_2}$).

The set of all representations u of $\mathcal{A}$ is denoted by $\mathcal{U}_{\mathcal{A}}$.

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \ldots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \ldots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Then $A \subseteq \mathbb{R}^+ \times 2^{\mathcal{X}}$ and $\mathcal{A} = [A, R_1, R_2]$ is given by

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \ldots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Then $A \subseteq \mathbb{R}^+ \times 2^{\mathcal{X}}$ and $\mathcal{A} = [A, R_1, R_2]$ is given by

$$R_1 = \Big\{ \big((y_1, B_1), (y_2, B_2)\big) \colon y_1 \geq y_2 \wedge B_1 \supseteq B_2 \Big\}$$

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \dots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Then $A \subseteq \mathbb{R}^+ \times 2^{\mathcal{X}}$ and $\mathcal{A} = [A, R_1, R_2]$ is given by

$$R_1 = \Big\{ ((y_1, B_1), (y_2, B_2)) \colon y_1 \geq y_2 \wedge B_1 \supseteq B_2 \Big\}$$

$$R_2 = \left\{ \Big(\big((y_1, B_1), (y_2, B_2) \big), \big((y_3, B_3), (y_4, B_4) \big) \Big) : \begin{array}{l} y_1 - y_2 \geq y_3 - y_4 \wedge \\ B_1 \supseteq B_3 \supseteq B_4 \supseteq B_2 \end{array} \right\}$$

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \dots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Then $A \subseteq \mathbb{R}^+ \times 2^{\mathcal{X}}$ and $\mathcal{A} = [A, R_1, R_2]$ is given by

$$R_1 = \Big\{ ((y_1, B_1), (y_2, B_2)) \colon y_1 \geq y_2 \wedge B_1 \supseteq B_2 \Big\}$$

$$R_2 = \left\{ \Big(\big((y_1, B_1), (y_2, B_2) \big), \big((y_3, B_3), (y_4, B_4) \big) \Big) : \begin{array}{l} y_1 - y_2 \geq y_3 - y_4 \wedge \\ B_1 \supseteq B_3 \supseteq B_4 \supseteq B_2 \end{array} \right\}$$

“$(y_1, B_1) \leftarrow (y_2, B_2)$ is more desirable than $(y_3, B_3) \leftarrow (y_4, B_4)$ if it yields a higher increase of salary and a superset increase of benefits.”

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \dots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Then $A \subseteq \mathbb{R}^+ \times 2^{\mathcal{X}}$ and $\mathcal{A} = [A, R_1, R_2]$ is given by

$$R_1 = \Big\{ \big((y_1, B_1), (y_2, B_2)\big) \colon y_1 \geq y_2 \wedge B_1 \supseteq B_2 \Big\}$$

$$R_2 = \left\{ \Big(\big((y_1, B_1), (y_2, B_2)\big), \big((y_3, B_3), (y_4, B_4)\big) \Big) : \begin{array}{l} y_1 - y_2 \geq y_3 - y_4 \wedge \\ B_1 \supseteq B_3 \supseteq B_4 \supseteq B_2 \end{array} \right\}$$

*"$(y_1, B_1) \leftarrow (y_2, B_2)$ is more desirable than $(y_3, B_3) \leftarrow (y_4, B_4)$ if it yields a **higher increase of salary** and a superset increase of benefits."*

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \dots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Then $A \subseteq \mathbb{R}^+ \times 2^{\mathcal{X}}$ and $\mathcal{A} = [A, R_1, R_2]$ is given by

$$R_1 = \Big\{ \big((y_1, B_1), (y_2, B_2)\big) \colon y_1 \geq y_2 \wedge B_1 \supseteq B_2 \Big\}$$

$$R_2 = \left\{ \Big(\big((y_1, B_1), (y_2, B_2)\big), \big((y_3, B_3), (y_4, B_4)\big) \Big) : \begin{array}{l} y_1 - y_2 \geq y_3 - y_4 \wedge \\ B_1 \supseteq B_3 \supseteq B_4 \supseteq B_2 \end{array} \right\}$$

"$(y_1, B_1) \leftarrow (y_2, B_2)$ is more desirable than $(y_3, B_3) \leftarrow (y_4, B_4)$ if it yields a higher increase of salary and a ***superset increase of benefits****."*

Modelling $\mathcal{M}$: Credal sets

Credal set

The uncertainty among the elements of S is described by a polyhedral *credal set* of probability measures of the form

$$\mathcal{M} = \left\{ \pi \in \mathcal{P} : \underline{b}_\ell \leq \mathbb{E}_\pi(f_\ell) \leq \overline{b}_\ell \text{ for } \ell = 1, \ldots, r \right\}$$

where $\mathcal{P}$ is the set of all probability measures on $(S, \sigma(S))$ and

- $f_1, \ldots, f_r : S \to \mathbb{R}$ are real-valued mappings and
- $\underline{b}_\ell \leq \overline{b}_\ell$, $\ell = 1, \ldots, r$, are lower and upper expectation bounds.

Modelling $\mathcal{M}$: Credal sets

Credal set

The uncertainty among the elements of S is described by a polyhedral *credal set* of probability measures of the form

$$\mathcal{M} = \left\{\pi \in \mathcal{P} : \underline{b}_\ell \leq \mathbb{E}_\pi(f_\ell) \leq \overline{b}_\ell \text{ for } \ell = 1, \ldots, r\right\}$$

where $\mathcal{P}$ is the set of all probability measures on $(S, \sigma(S))$ and

- $f_1, \ldots, f_r : S \to \mathbb{R}$ are real-valued mappings and
- $\underline{b}_\ell \leq \overline{b}_\ell$, $\ell = 1, \ldots, r$, are lower and upper expectation bounds.

Special cases: *Classical probability – Probability intervals – Interval probability – Linear partial information – (Finitely generated) Lower previsions*

Generalized Choice Functions and Elicitation

Choice functions for decision making based on the sets $\mathcal{U}_\mathcal{A}$ and $\mathcal{M}$ as well as efficient computation algorithms have been developed in:


Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

ELSEVIER

Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences ☆,☆☆

C. Jansen *, G. Schollmeyer, T. Augustin

[Jansen et al., 2018]

Information-efficient procedures for eliciting optimal decisions according to these criteria are discussed in:


Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

ELSEVIER

Information efficient learning of complexly structured preferences: Elicitation procedures and their application to decision making under uncertainty

C. Jansen *, H. Blocher, T. Augustin, G. Schollmeyer

Department of Statistics, LMU Munich, Ludwigsstr. 33, 80539 Munich, Germany

[Jansen et al., 2022]

Generalized Stochastic Dominance

Today, we focus on only one choice function from these papers, based on:

Generalized Stochastic Dominance Relation (GSD-Relation)

Let $\mathcal{A} = [A, R_1, R_2]$ be consistent and $\mathcal{M}$ a credal set on $(S, \mathcal{S})$.

For $X, Y \in \mathcal{F}_{(\mathcal{A},S)}$,[1] we say that Y is $(\mathcal{A}, \mathcal{M})$**-dominated** by X if

$$\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$$

for all $u \in \mathcal{U}_\mathcal{A}$ and $\pi \in \mathcal{M}$. The induced relation is denoted by $\geq_{(\mathcal{A},\mathcal{M})}$ and called Generalized Stochastic Dominance Relation (**GSD-Relation**).

[1] $\mathcal{F}_{(\mathcal{A},S)} := \left\{ X \in A^S : u \circ X \text{ is } \mathcal{S}\text{-}\mathcal{B}_\mathbb{R}([0,1])\text{-measurable for all } u \in \mathcal{U}_\mathcal{A} \right\}$.

Generalized Stochastic Dominance

Today, we focus on only one choice function from these papers, based on:

Generalized Stochastic Dominance Relation (GSD-Relation)

Let $\mathcal{A} = [A, R_1, R_2]$ be consistent and $\mathcal{M}$ a credal set on $(S, \mathcal{S})$.

For $X, Y \in \mathcal{F}_{(\mathcal{A},S)}$,[1] we say that Y is **$(\mathcal{A}, \mathcal{M})$-dominated** by X if

$$\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$$

for all $u \in \mathcal{U}_\mathcal{A}$ and $\pi \in \mathcal{M}$. The induced relation is denoted by $\geq_{(\mathcal{A},\mathcal{M})}$ and called Generalized Stochastic Dominance Relation (**GSD-Relation**).

The GSD-relation now directly induces the GSD choice function by setting

$$\mathrm{ch}_{\mathcal{A},\mathcal{M}}(\mathcal{D}) := \left\{ X \in \mathcal{D} : \nexists Y \in \mathcal{D} \text{ such that } (Y, X) \in >_{(\mathcal{A},\mathcal{M})} \right\}$$

[1] $\mathcal{F}_{(\mathcal{A},S)} := \left\{ X \in A^S : u \circ X \text{ is } \mathcal{S}\text{-}\mathcal{B}_\mathbb{R}([0,1])\text{-measurable for all } u \in \mathcal{U}_\mathcal{A} \right\}$.

Some Special Cases of GSD

The GSD-relation $\geq_{(\mathcal{A},\mathcal{M})}$ has some prominent special cases.

For ...

Some Special Cases of GSD

The GSD-relation $\geq_{(\mathcal{A},\mathcal{M})}$ has some prominent special cases.

For . . .

- ... and $\mathcal{M} = \{\pi\}$ and R_2 trivial

  $\rightarrow$ Reduction to (first-order) **stochastic dominance**

  (see, e.g., [Mosler and Scarsini, 1991]))

Some Special Cases of GSD

The GSD-relation $\geq_{(\mathcal{A},\mathcal{M})}$ has some prominent special cases.

For . . .

- ... and $\mathcal{M} = \{\pi\}$ and R_2 trivial

  $\rightarrow$ Reduction to (first-order) **stochastic dominance**

  (see, e.g., [Mosler and Scarsini, 1991]))

- ... and $\mathcal{M} = \{\pi\}$ and R_1 and R_2 guaranteeing utility unique up to plts

  $\rightarrow$ Reduction to comparing **expected utilities**.

  (see, e.g., [Krantz et al., 1971]))

Some Special Cases of GSD

The GSD-relation $\geq_{(\mathcal{A},\mathcal{M})}$ has some prominent special cases.

For . . .

- ... and $\mathcal{M} = \{\pi\}$ and R_2 trivial

  $\rightarrow$ Reduction to (first-order) **stochastic dominance**

  (see, e.g., [Mosler and Scarsini, 1991]))

- ... and $\mathcal{M} = \{\pi\}$ and R_1 and R_2 guaranteeing utility unique up to plts

  $\rightarrow$ Reduction to comparing **expected utilities**.

  (see, e.g., [Krantz et al., 1971]))

- ... $\mathcal{M}$ non-trivial and R_1 and R_2 guaranteeing utility unique up to plts

  $\rightarrow$ Reduction to **Bewley dominance**.

  (see, e.g., [Bewley, 2002, Troffaes, 2007, Etner et al., 2012]))

Locally Varying Scales of Measurement

Group and collaborators

Most of the following is joint work with (in alphabetic order):

- Thomas Augustin,
- Hannah Blocher,
- Malte Nalenz,
- Julian Rodemann,
- Georg Schollmeyer,

and mainly based on the following three papers: ([Jansen et al., 2023])

C. Jansen, G. Schollmeyer, H. Blocher, J. Rodemann and T. Augustin (2023): Robust statistical comparison of random variables with locally varying scale of measurement. In: Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (**UAI 2023**). Proceedings of Machine Learning Research, vol. 216.

C. Jansen, M. Nalenz, G. Schollmeyer and T. Augustin (2023): Statistical comparisons of classifiers by generalized stochastic dominance. Journal of Machine Learning Research (**JMLR**), 24 (231): 1 - 37.

C. Jansen, G. Schollmeyer, J. Rodemann, H. Blocher and T. Augustin (2024): Statistical multicriteria benchmarking via the GSD-front. Under review.

Motivation

1.) Statistical methods are usually tailored for data situations that can be clearly assigned to a standard scale of measurement.

2.) Non-standard data can often not clearly be assigned to a standard scale.

1.)+2.) $\Rightarrow$ Statistical methods are often not well-suited for analyzing non-standard data!

Idea: Use the notion of a preference system to model data with scales of measurement which not correspon to one of these extreme poles.

Standard Scales of Measurement

Consider some random variable $X : \Omega \to A$ mapping to some set A.

Standard Scales of Measurement

Consider some random variable $X : \Omega \to A$ mapping to some set A.

- If A is structured only by a preorder $\succsim$, we call A of ordinal scale.

  $\Rightarrow$ The set $\mathcal{U}_{all}$ of all $\succsim$-isotone **candidate scales** $u : A \to \mathbb{R}$ as a **whole** represents the structural information on A.

  $\Rightarrow$ Any analysis of the variable X should be **invariant** under the choice of the candidate scale $u \in \mathcal{U}_{all}$.

Standard Scales of Measurement

Consider some random variable $X : \Omega \to A$ mapping to some set A.

- If A is structured only by a preorder $\succsim$, we call A of ordinal scale.

  $\Rightarrow$ The set $\mathcal{U}_{all}$ of all $\succsim$-isotone **candidate scales** $u : A \to \mathbb{R}$ as a **whole** represents the structural information on A.

  $\Rightarrow$ Any analysis of the variable X should be **invariant** under the choice of the candidate scale $u \in \mathcal{U}_{all}$.

- If the order on A is induced by some metric d, we call A of cardinal scale.

  $\Rightarrow$ There exists a scale $u^* : A \to \mathbb{R}$ that is **unique** (up to irrelevant trafos).

  $\Rightarrow$ Any analysis of the variable X can be based on u^* **alone**.

Preference Systems in Statistics

Question: What if the structure on A does not belong to either extreme pole?

In other words: What if the structuredness of A varies along its subsets?

Preference Systems in Statistics

Question: What if the structure on A does not belong to either extreme pole?

In other words: What if the structuredness of A varies along its subsets?

A preference system $\mathcal{A} = [A, R_1, R_2]$ helps to formalize this intuition:

- R_1 formalizes the available ordinal information, i.e. information about the arrangement of the elements of A.
- R_2 describes the available cardinal information, i.e. pairs standing in relation are ordered with respect to the intensity of the relation.
- A is locally almost cardinal on subsets where R_1 and R_2 are very dense.
- A is locally at most ordinal on subsets where R_2 is sparse or even empty.

Regularization and Preference Systems

Opportunity: Preference systems offer a nice way for regularization by excluding those $u \in \mathcal{U}_{\mathcal{A}}$ that are too extreme (in some sense).

Regularization and Preference Systems

Opportunity: Preference systems offer a nice way for regularization by excluding those $u \in \mathcal{U}_{\mathcal{A}}$ that are too extreme (in some sense).

Simple idea: If $\mathcal{A}$ has R_1-minimal/maximal elements a_*, a^*, define

$$\mathcal{N}_{\mathcal{A}} := \left\{ u \in \mathcal{U}_{\mathcal{A}} : u(a_*) = 0 \ \wedge \ u(a^*) = 1 \right\}$$

$$\mathcal{N}_{\mathcal{A}}^{\delta} := \left\{ u \in \mathcal{N}_{\mathcal{A}} : u(c) - u(d) - u(e) + u(f) \geq \delta \ \ \forall ((c,d),(e,f)) \in P_{R_2} \right\}$$

Regularization and Preference Systems

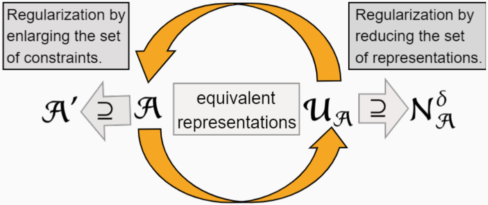
Opportunity: Preference systems offer a nice way for regularization by excluding those $u \in \mathcal{U}_{\mathcal{A}}$ that are too extreme (in some sense).

Simple idea: If $\mathcal{A}$ has R_1-minimal/maximal elements a_*, a^*, define

$$\mathcal{N}_{\mathcal{A}} := \Big\{u \in \mathcal{U}_{\mathcal{A}} : u(a_*) = 0 \ \wedge \ u(a^*) = 1\Big\}$$

$$\mathcal{N}_{\mathcal{A}}^{\delta} := \Big\{u \in \mathcal{N}_{\mathcal{A}} : u(c) - u(d) - u(e) + u(f) \geq \delta \ \ \forall ((c,d),(e,f)) \in P_{R_2}\Big\}$$

Two ways for regularization:



Random Variables Mapping Into Preference Systems

Goal: We now want to address the problem of comparing random variables $X, Y : \Omega \to A$ that map into a preference system.

Challenge: We have epistemic uncertainty in form of

- **Approximation uncertainty:** Only samples of the considered variables (rather than π itself) are available.
- **Model uncertainty:** The weakly structured order information makes a set of candidate scales compatible with the structure on A.

Addressing Model Uncertainty via GSD

Idea: Weaken $\succsim_{E(u)}$ to a preorder by demanding expectation dominance for all scales u compatible with the preference system $\mathcal{A}$.

$\Rightarrow$ This idea leads to a "precise" version of GSD.

Recall:

Precise GSD

Let $\mathcal{A}$ be consistent and π be a probability measure on $(S, \mathcal{S})$.

For $X, Y \in \mathcal{F}_{(\mathcal{A},S)}$, we call Y **$(\mathcal{A}, \{\pi\})$-dominated** by X if

$$\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$$

for all $u \in \mathcal{U}_\mathcal{A}$. This induces preorder $R_{(\mathcal{A},\pi)}$ on $\mathcal{F}_{(\mathcal{A},\{\pi\})}$ which is called the **precise GSD-relation**.

Obviously, precise GSD is invariant under the scale.

Addressing Approximation Uncertainty

Practical Problem: Usually, we do not know π but only *i.i.d.* samples $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)$ of X and Y are available.

Approach: Perform a statistical test for GSD.

Ideal Hypotheses:

$$H_0^{id} : (X, Y) \notin R_{(\mathcal{A},\pi)} \quad \textbf{vs.} \quad H_1^{id} : (X, Y) \in R_{(\mathcal{A},\pi)}$$

Pragmatic Hypotheses:

$$H_0 : (Y, X) \in R_{(\mathcal{A},\pi)} \quad \textbf{vs.} \quad H_1 : (Y, X) \notin R_{(\mathcal{A},\pi)}$$

Addition: To mitigate the effect of the reversed hypotheses, we can additionally test with the variables X and Y in reversed roles.

The Choice of the Test Statistic

Observation: It holds $(X, Y) \in R_{(\mathcal{A},\pi)}$ if and only if

$$D(X, Y) := \inf_{u \in \mathcal{N}_{\mathcal{A}}} \left(\mathbb{E}_\pi(u \circ X) - \mathbb{E}_\pi(u \circ Y)\right) \geq 0.$$

Consequence: A natural test statistic is the empirical version of $D(X, Y)$, i.e.,

$$d_{\mathrm{X,Y}} : \Omega \to \mathbb{R}$$

$$\omega \mapsto \inf_{u \in \mathcal{N}_{\mathcal{A}_\omega}} \sum_{z \in (\mathrm{XY})_\omega} u(z) \cdot \left(\hat{\pi}_X^\omega(\{z\}) - \hat{\pi}_Y^\omega(\{z\})\right)$$

with, for $\omega \in \Omega$ fixed,

- $\hat{\pi}_X^\omega$ and $\hat{\pi}_Y^\omega$ the observed empirical image measures of X and Y,
- $(\mathrm{XY})_\omega = \{X_i(\omega) : i \leq n\} \cup \{Y_i(\omega) : i \leq m\} \cup \{a_*, a^*\}$, and
- $\mathcal{A}_\omega$ the subsystem of $\mathcal{A}$ restricted to $(\mathrm{XY})_\omega$.

Regularization of the Test Statistic

Observation: $d_{\mathsf{X},\mathsf{Y}}$ cannot measure extent of GSD in the sample. Thus, $d_{\mathsf{X},\mathsf{Y}}$ may be too little sensitive.

Idea: Regularize $d_{\mathsf{X},\mathsf{Y}}$ so that it can also account for the extent of GSD.

Formally: The regularized test statistic looks as follows:

$$d^{\varepsilon}_{\mathsf{X},\mathsf{Y}} : \Omega \to \mathbb{R}$$

$$\omega \mapsto \inf_{u \in \mathcal{N}^{\delta_\varepsilon(\omega)}_{\mathcal{A}_\omega}} \sum_{z \in (\mathsf{XY})_\omega} u(z) \cdot (\hat{\pi}^{\omega}_{X}(\{z\}) - \hat{\pi}^{\omega}_{Y}(\{z\}))$$

with $\varepsilon \in [0,1]$ and

$$\delta_\varepsilon(\omega) := \varepsilon \cdot \sup\{\xi : \mathcal{N}^{\xi}_{\mathcal{A}_\omega} \neq \emptyset\}.$$

Computation: Both test statistics $d_{\mathsf{X},\mathsf{Y}}$ and $d^{\varepsilon}_{\mathsf{X},\mathsf{Y}}$ can be computed by solving one single linear programming problem.

A Permutation Test

Assumption: We made observations of the i.i.d. variables, i.e., we observed:

$$\begin{aligned} \mathbf{x} &:= (x_1, \ldots, x_n) := (X_1(\omega_0), \ldots, X_n(\omega_0)) \\ \mathbf{y} &:= (y_1, \ldots, y_m) := (Y_1(\omega_0), \ldots, Y_m(\omega_0)) \end{aligned}$$

A Permutation Test

Assumption: We made observations of the i.i.d. variables, i.e., we observed:

$$\begin{aligned} \mathbf{x} &:= (x_1, \ldots, x_n) := (X_1(\omega_0), \ldots, X_n(\omega_0)) \\ \mathbf{y} &:= (y_1, \ldots, y_m) := (Y_1(\omega_0), \ldots, Y_m(\omega_0)) \end{aligned}$$

Good News: As the worst case of the null hypothesis H_0 is $\pi_X = \pi_Y$, performing a permutation test is a valid level α test.

The resampling scheme then looks:

Step 1: Pool data sample: $\mathbf{w} := (w_1, \ldots, w_{n+m}) := (x_1, \ldots, x_n, y_1, \ldots, y_m)$

Step 2: Take all $k := \binom{n+m}{n}$ index sets $I \subseteq \{1, \ldots, n+m\}$ of size n. Compute $d_{\mathsf{X},\mathsf{Y}}$ resp. $d^{\varepsilon}_{\mathsf{X},\mathsf{Y}}$ for $(w_i)_{i \in I}$ and $(w_i)_{i \in \{1,\ldots,n+m\} \setminus I}$ instead of $\mathbf{x}/\mathbf{y}$ to get d_I resp. d^{ε}_I.

Step 3: Sort all d_I resp. d^{ε}_I in increasing order to get $d_{(1)}, \ldots, d_{(k)}$ resp. $d^{\varepsilon}_{(1)}, \ldots, d^{\varepsilon}_{(k)}$.

Step 4: Reject H_0 if $d_{\mathsf{X},\mathsf{Y}}(\omega_0)$ resp. $d^{\varepsilon}_{\mathsf{X},\mathsf{Y}}(\omega_0)$ is greater than $d_{(\ell)}$ resp. $d^{\varepsilon}_{(\ell)}$, with $\ell := \lceil (1-\alpha) \cdot k \rceil$ and α the significance level.

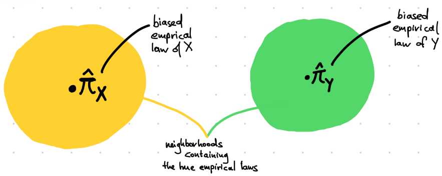
Credal Sets For Robustification

Rough Idea: Use credal sets to robustify the permutation test to small deviations from the *i.i.d.* assumption.

Credal Sets For Robustification

Rough Idea: Use credal sets to robustify the permutation test to small deviations from the *i.i.d.* assumption.

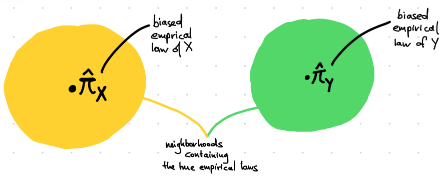
More concrete: We allow our samples to be (potentially) biased in the sense that we only assume the true empirical laws to lie in some credal neighborhoods $\mathcal{M}_X$ and $\mathcal{M}_Y$ around the biased empirical laws.



Credal Sets For Robustification

Rough Idea: Use credal sets to robustify the permutation test to small deviations from the *i.i.d.* assumption.

More concrete: We allow our samples to be (potentially) biased in the sense that we only assume the true empirical laws to lie in some credal neighborhoods $\mathcal{M}_X$ and $\mathcal{M}_Y$ around the biased empirical laws.



Adapted Resampling Scheme: Replace

- $d^{\varepsilon}_{X,Y}(\omega_0)$ by $\inf_{(\pi_1,\pi_2)\in\mathcal{M}_X^{\omega_0}\times\mathcal{M}_Y^{\omega_0}} \tilde{d}^{\varepsilon}_{X,Y}(\omega_0)$
- $d^{\varepsilon}_{I}(\omega_0)$ by $\sup_{(\pi_1,\pi_2)\in\mathcal{M}_X^{\omega_0}\times\mathcal{M}_Y^{\omega_0}} \tilde{d}^{\varepsilon}_{I}(\omega_0)$

Results in: Valid (yet conservative) statistical test!

γ-Contamination Model

A special class of credal sets with a very intuitive interpretation are

γ-contamination models

For $\omega \in \Omega$, $\gamma \in [0,1]$, and $Z \in \{X, Y\}$ fixed, we set

$$\mathcal{M}_Z^\omega = \left\{\pi : \pi \geq (1-\gamma) \cdot \hat{\pi}_Z^\omega\right\}$$

or equivalently

$$\mathcal{M}_Z^\omega = \left\{\gamma \cdot \nu + (1-\gamma) \cdot \hat{\pi}_Z^{\omega_0} : \nu \text{ probability measure}\right\}.$$

γ-Contamination Model

A special class of credal sets with a very intuitive interpretation are

γ-contamination models

For $\omega \in \Omega$, $\gamma \in [0,1]$, and $Z \in \{X, Y\}$ fixed, we set

$$\mathcal{M}_Z^\omega = \Big\{ \pi : \pi \geq (1-\gamma) \cdot \hat{\pi}_Z^\omega \Big\}$$

or equivalently

$$\mathcal{M}_Z^\omega = \Big\{ \gamma \cdot \nu + (1-\gamma) \cdot \hat{\pi}_Z^{\omega_0} : \nu \text{ probability measure} \Big\}.$$

The observed p-values of the robustified test can then be computed as a function of the contamination size γ:

$$f_\varepsilon(\gamma) := 1 - \frac{1}{N} \cdot \sum_{I \in \mathcal{I}_N} \mathbb{1}_{\left\{ d^\varepsilon_{\mathrm{X},\mathrm{Y}}(\omega_0) - d^\varepsilon_I > \frac{2\gamma}{(1-\gamma)} \right\}}$$

Application I

Spaces with Differently Scaled Dimensions (SDSDs)

Situation: Consider an r-dimensional space $A \subseteq \mathbb{R}^r$ and assume that

- the first $0 \leq z \leq r$ dimensions are of cardinal scale and
- the remaining dimensiones are purely ordinal.

Question: How can we utilize the cardinal dimensions without making unjustified assumptions about the ordinal ones?

Idea: Utilize the cardinal information only on those parts of A where there is no possible conflict with the ordinal information.

Formalization: Consider A to be a subsystem of $\mathcal{P} = [\mathbb{R}^r, R_1^*, R_2^*]$, where

$$R_1^* = \Big\{(x,y) : x_j \geq y_j \; \forall j \leq r\Big\}$$

$$R_2^* = \left\{((x,y),(x',y')) : \begin{array}{l} x_j - y_j \geq x'_j - y'_j \;\; \forall j \leq z \\ x_j \geq x'_j \geq y'_j \geq y_j \; \forall j > z \end{array}\right\}.$$

A Characterization Theorem in SDSDs

For the special case of A being a multidimensional space with differently scaled dimensions, the GSD-relation can be neatly characterized.

Theorem

Let $X = (\Delta_1, \ldots, \Delta_r), Y = (\Lambda_1, \ldots, \Lambda_r) \in \mathcal{F}_{(\mathcal{P},\pi)}$. Then:

i) $\mathcal{P}$ is consistent.

ii) If $z = 0$, then $R_{(\mathcal{P},\pi)}$ equals (first-order) stochastic dominance w.r.t. π and R_1^* (short: $\mathrm{FSD}(R_1^*, \pi)$).

iii) If $(X, Y) \in R_{(\mathcal{P},\pi)}$ and $\Delta_j, \Lambda_j \in \mathcal{L}^1(\Omega, \mathcal{S}_1, \pi)$ for all $j = 1, \ldots, r$, then

   I. $\mathbb{E}_\pi(\Delta_j) \geq \mathbb{E}_\pi(\Lambda_j)$ for all $j = 1, \ldots, r$, and

   II. $(\Delta_j, \Lambda_j) \in \mathrm{FSD}(\geq, \pi)$ for all $j = z+1, \ldots, r$.

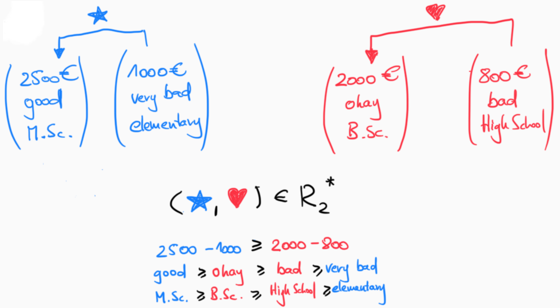
   If all components of X are jointly independent and all components of Y are jointly independent, I. and II. imply $(X, Y) \in R_{(\mathcal{P},\pi)}$.

Multidimensional Poverty Analysis

Capability Approach: Poverty is a multidimensional concept with more facets than just income or wealth ([Sen, 1985]).

Exemplary operationalization: We use the ALLBUS data and account for three dimensions of poverty: *`income`* (numeric), *`health`* (ordinal, 6 levels) and *`education`* (ordinal, 8 levels)

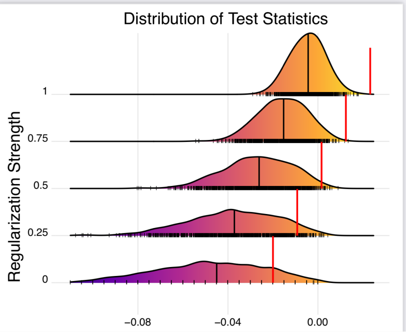
Example:



Multidimensional Poverty Analysis, cont.

For the ALLBUS data, we focus on a subsample with $n = m = 100$ men and women each. Again, we operationalize poverty by the variables ***income*** (numeric), ***health*** (ordinal, 6 levels) and ***education*** (ordinal, 8 levels)

Test results:

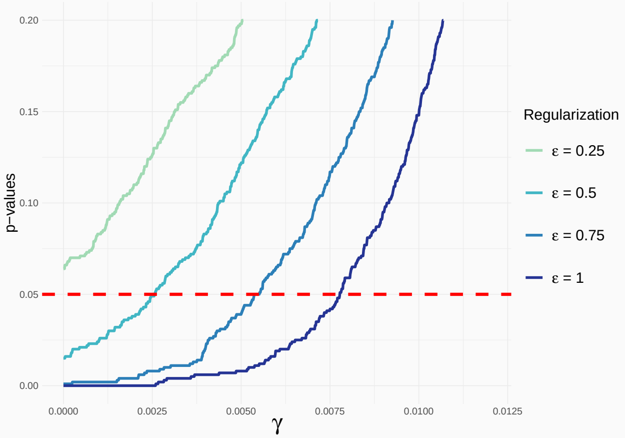


Results: All tests significant for $\alpha = 0.05$.

Reversed test: No evidence for incomparability: All reversed p-values ≥ 0.95.

Multidimensional Poverty Analysis, cont.

Results of the robustified test:



Application II

Statistical Multicriteria Comparison of Classifiers

Question of interest: How to utilize our decision-theoretical approach for comparing classifiers under multiplicity of quality criteria and data sets?

Setup: Let

- $\mathcal{D}$ denote the set of all relevant data sets,
- $\mathcal{C}$ denote the finite set of all relevant classifiers,
- $\left(\phi_i : \mathcal{C} \times \mathcal{D} \to [0,1]\right)_{i \in \{1,\ldots,r\}}$ denote a family of quality criteria,
- $\phi := (\phi_1, \ldots, \phi_r) : \mathcal{D} \times \mathcal{C} \to [0,1]^r$ be a mulidimensional quality criterion.

Statistical Multicriteria Comparison of Classifiers

Question of interest: How to utilize our decision-theoretical approach for comparing classifiers under multiplicity of quality criteria and data sets?

Setup: Let

- $\mathcal{D}$ denote the set of all relevant data sets,
- $\mathcal{C}$ denote the finite set of all relevant classifiers,
- $\left(\phi_i : \mathcal{C} \times \mathcal{D} \to [0,1]\right)_{i \in \{1,\dots,r\}}$ denote a family of quality criteria,
- $\phi := (\phi_1, \dots, \phi_r) : \mathcal{D} \times \mathcal{C} \to [0,1]^r$ be a mulidimensional quality criterion.

Assumptions:

- For $0 \leq z \leq r$, the criteria $\phi_1, \dots, \phi_z$ are of cardinal scale (differences may be interpreted)
- The remaining criteria are purely ordinal (differences are meaningless apart from sign).

Statistical Multicriteria Comparison of Classifiers

Three levels of problems when comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously.

data sets / classifier	D_1	$\dots$	D_s
C_1	$\begin{pmatrix} \phi_1(C_1, D_1) \\ \vdots \\ \phi_n(C_1, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} \phi_1(C_q, D_1) \\ \vdots \\ \phi_n(C_q, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

Statistical Multicriteria Comparison of Classifiers

Three levels of problems when comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously.

classifier \ data sets	D_1	$\ldots$	D_s
C_1	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.7 \end{pmatrix}$	$\ldots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.8 \end{pmatrix}$	$\ldots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

Level 1: On a **fixed** data set D it may hold

$$\phi_1(C_1, D) > \phi_1(C_2, D) \;\wedge\; \phi_2(C_1, D) < \phi_2(C_2, D).$$

Statistical Multicriteria Comparison of Classifiers

Three levels of problems when comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously.

classifier \ data sets	D_1	$\dots$	D_s
C_1	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.6 \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.9 \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

Level 2: Even if, for all $i \in \{1, \dots, n\}$, we have

$$\phi_i(C_1, D_1) > \phi_i(C_2, D_1)$$

there may exists some $i_0 \in \{1, \dots, n\}$ such that

$$\phi_{i_0}(C_1, D_2) < \phi_{i_0}(C_2, D_2).$$

Statistical Multicriteria Comparison of Classifiers

Three levels of problems when comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously.

classifier \ data sets	D_1	$\dots$	D_s
C_1	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$

Level 3: Even if a decision can be made for a sample $(D_1, \dots, D_s)$ of data sets,

Statistical Multicriteria Comparison of Classifiers

Three levels of problems when comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously.

classifier \ data sets	D_1^*	$\dots$	D_s^*
C_1	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.9 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.75 \\ \vdots \\ 0.4 \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} 0.85 \\ \vdots \\ 0.67 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.33 \\ \vdots \\ 0.98 \end{pmatrix}$

Level 3: Even if a decision can be made for a sample $(D_1, \dots, D_s)$ of data sets, no clear decision might be possible for a different sample $(D_1^*, \dots, D_s^*)$.

Transferring GSD to Classifier Comparison

Idea: Embed the range $\Phi(\mathcal{C} \times \mathcal{D})$ of Φ in the following preference system $\mathcal{P} = [\mathbb{R}^r, R_1^*, R_2^*]$ from before.

Then:

- To transfer the GSD-relation, interpret the data sets in $\mathcal{D}$ as realizations of a random variable $T : \Omega \to \mathcal{D}$ on some probability space $(\Omega, \mathcal{S}, \pi)$.
- Associate each $C \in \mathcal{C}$ with the variable $\Phi_C := \Phi(C, T(\cdot))$ on Ω and compare classifiers by comparing the associated random variables by precise GSD.

Formally:

GSD for Classifier Comparison

Let $\mathcal{P}_\Phi$ be the preference system obtained by restricting $\mathcal{P}$ to $\Phi(\mathcal{C} \times \mathcal{D})$.

Further, let $\mathcal{C}$ be such that $\{\Phi_C : C \in \mathcal{C}\} \subseteq \mathcal{F}_{(\mathcal{P}_\Phi, \pi)}$.

For $C, C' \in \mathcal{C}$, say that C **dominates** C', abbreviated $C \succsim C'$, whenever

$$(\Phi_C, \Phi_{C'}) \in R_{(\mathcal{P}_\Phi, \pi)}.$$

Theoretical and Empirical GSD-Front

We associate the following two sets to the relation $\succsim$:

The GSD-Front

Let $\mathcal{C}$ be such that $\{\Phi_C : C \in \mathcal{C}\} \subseteq \mathcal{F}_{(\mathcal{P}_\Phi, \pi)}$ and $T_1, \ldots, T_s$ be *i.i.d.* copies of T.

i) The **GSD-front** is the set

$$\mathrm{gsd}(\mathcal{C}) := \{C \in \mathcal{C} : \nexists C' \in \mathcal{C} \text{ s.t. } C' \succ C\},$$

where $\succ$ denotes the strict part of $\succsim$.

ii) Let $\rho \in [0, 1]$. The ρ-**empirical GSD-front** is the (random) subset of $\mathcal{C}$ defined by

$$\mathrm{egsd}_s^\rho(\mathcal{C}) = \left\{ C : \nexists C' \in \mathcal{C} \text{ s.t. } \begin{array}{l} d_{(\Phi_{C'}, \Phi_C)} \geq -\rho \\ d_{(\Phi_C, \Phi_{C'})} < 0 \end{array} \right\}.$$

Consistent Estimability of the GSD-Front

The following theorem on the consistent estimability of the GSD-front holds:

Estimating the GSD-Front

Denote by $\mathcal{I}_\Phi$ the set of all sets $\{a : u(a) \geq c\}$, where $c \in [0, 1]$ and $u \in \mathcal{U}_{\mathcal{P}_\Phi}$.

Assume that $\precsim$ is antisymmetric.

If the VC-dimension[2] of $\mathcal{I}_\Phi$ is finite and $\rho : \mathbb{N} \to [0, 1]$ converges to 0 with at most $\Theta(1/\sqrt[4]{s})$, then $(\mathrm{egsd}_s^{\rho(s)}(\mathcal{C}))_{s\in\mathbb{N}}$ is consistent, i.e.,

$$\pi\left(\left\{\omega \in \Omega : \lim_{s\to\infty} \mathrm{egsd}_s^{\rho(s)}(\mathcal{C}) = \mathrm{gsd}(\mathcal{C})\right\}\right) = 1,$$

where set convergence is defined via the trivial metric.

[2]The VC-dimension of a family of sets $\mathcal{S}$ is the largest possible cardinality of a set A, such that $2^A = \{A \cap S : S \in \mathcal{S}\}$, i.e., A can be shattered by $\mathcal{S}$.

Consistent Tests for the GSD-Front

Goal: Compare the (multivariate, mixed-scaled) quality of a newly developed classifier C with a set $\mathcal{C}$ of state-of-the-art classifiers.

Consistent Tests for the GSD-Front

Goal: Compare the (multivariate, mixed-scaled) quality of a newly developed classifier C with a set $\mathcal{C}$ of state-of-the-art classifiers.

How to proceed? Develop a statistical test for the pair

$$H_0 : C \notin \text{gsd}(\mathcal{C}) \ \ \textbf{vs.} \ \ H_1 : C \in \text{gsd}(\mathcal{C})$$

Consistent Tests for the GSD-Front

Goal: Compare the (multivariate, mixed-scaled) quality of a newly developed classifier C with a set $\mathcal{C}$ of state-of-the-art classifiers.

How to proceed? Develop a statistical test for the pair

$$H_0 : C \notin \text{gsd}(\mathcal{C}) \;\; \textbf{vs.} \;\; H_1 : C \in \text{gsd}(\mathcal{C})$$

How exactly? Note that H_0 can be rewritten as:

$$H_0 : \exists C' \in \mathcal{C} \setminus \{C\} : C' \succsim C.$$

Thus, H_0 is false iff the hypothesis $H_0^{C'} : C' \succsim C$ is false *for every* $C' \in \mathcal{C} \setminus \{C\}$.

Consistent Tests for the GSD-Front

Goal: Compare the (multivariate, mixed-scaled) quality of a newly developed classifier C with a set $\mathcal{C}$ of state-of-the-art classifiers.

How to proceed? Develop a statistical test for the pair

$$H_0 : C \notin \mathrm{gsd}(\mathcal{C}) \ \ \textbf{vs.} \ \ H_1 : C \in \mathrm{gsd}(\mathcal{C})$$

How exactly? Note that H_0 can be rewritten as:

$$H_0 : \exists C' \in \mathcal{C} \setminus \{C\} : C' \succsim C.$$

Thus, H_0 is false iff the hypothesis $H_0^{C'} : C' \succsim C$ is false *for every* $C' \in \mathcal{C} \setminus \{C\}$.

Good news:

- The pairs $(H_0^{C'}, \neg H_0^{C'})$ can be tested using the test from Application I.
- Thus, $(H_0, \neg H_0)$ can (essentially) be tested by running these tests multiple times, while rejecting H_0 if all $H_0^{C'}$ are rejected.
- This even allows to construct consistent tests.

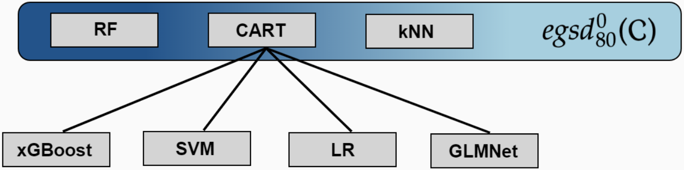
OpenML Benchmarking Experiments: Setup

- We use 80 binary classification datasets from the Open Multimedia Library (OpenML) [Van Rijn et al., 2013].
- We compare the performance of *Support Vector Machine* (SVM) with
  - *Random Forest* (RF),
  - *Decision Tree* (CART),
  - *Logistic Regression* (LR),
  - *Generalized Linear Model with Elastic net* (GLMNet),
  - *Extreme Gradient Boosting* (xGBoost), and
  - *k-Nearest Neighbors* (kNN).
- Comparison is based on the multivariate metric Φ composed of
  - *predictive accuracy*,
  - *computation time on the test data*, and
  - *computation time on the training data*.

  Since computation time strongly depends on the computing environment used, we treat the time-related metrics as purely ordinal.

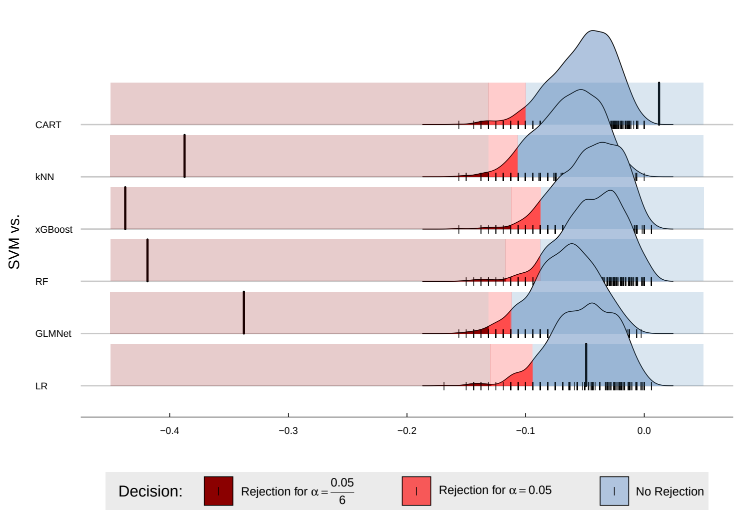
OpenML Benchmarking Experiments: Empirical GSD-Front

The Hasse graph of the empirical GSD relation:



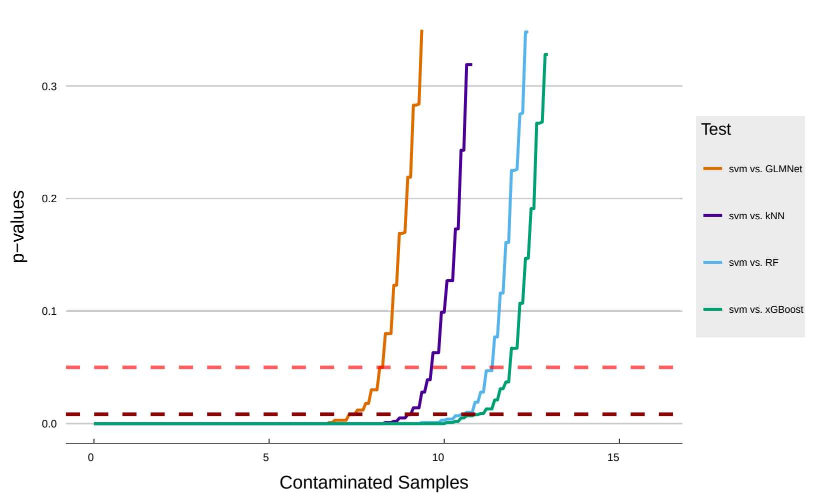
The blue shaded region symbolizes the 0-empirical GSD-front.

Results of the GSD-front test:



OpenML Benchmarking Experiments: Robustness

Robustness of test decision under contamination of the benchmark suite:



Summary and Outlook

Summary:

- Presented a framwork for decision making under weakly structured information
- Demonstrated two applications of this framework in problems of robust statistics and machine learning

What is next?

- Exploit other problems/fileds where a decision-theoretic perspective might be fruitful

Thank you very much for your attention!

References i

Bacharach, M. (1975).
Group decisions in the face of differences of opinion.
Management Science, 22:182–191.

Bewley, T. F. (2002).
Knightian decision theory. part i.
Decisions in economics and finance, 25:79–110.

Bradley, S. (2019).
Aggregating belief models.
In *Proceedings of ISIPTA 2019*, Proceedings of Machine Learning Research.

References ii

Etner, J., Jeleva, M., and Tallon, J.-M. (2012).
Decision theory under ambiguity.
Journal of Economic Surveys, 26(2):234–270.

Jansen, C., Blocher, H., Augustin, T., and Schollmeyer, G. (2022).
Information efficient learning of complexly structured preferences: Elicitation procedures and their application to decision making under uncertainty.
International Journal of Approximate Reasoning, 144:69–91.

Jansen, C., Schollmeyer, G., and Augustin, T. (2018).
Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences.
International Journal of Approximate Reasoning, 98:112–131.

Jansen, C., Schollmeyer, G., Blocher, H., Rodemann, J., and Augustin, T. (2023).
Robust statistical comparison of random variables with locally varying scale of measurement.
In *Uncertainty in Artificial Intelligence (UAI)*. PMLR.
To appear.

Kikuti, D., Cozman, F., and Filho, R. (2011).
Sequential decision making with partially ordered preferences.
Artificial Intelligence, 175:1346 – 1365.

Krantz, D., Luce, R., Suppes, P., and Tversky, A. (1971).
Foundations of Measurement. Volume I: Additive and Polynomial Representations.
Academic Press, San Diego and London.

Levi, I. (1974).
On indeterminate probabilities.
The Journal of Philosophy, 71:391–418.

Mosler, K. and Scarsini, M. (1991).
Some theory of stochastic dominance.
In Mosler, K. and Scarsini, M., editors, *Stochastic Orders and Decision under Risk*, pages 203–212. Institute of Mathematical Statistics, Hayward, CA.

Nau, R. (2006).
The shape of incomplete preferences.
Annals of Statistics, 34:2430––2448.

Savage, L. (1954).
The Foundations of Statistics.
Wiley.

Seidenfeld, T., Kadane, J., and Schervish, M. (1995).
A representation of partially ordered preferences.
Annals of Statistics, 23:2168–2217.

Sen, A. (1985).
Commodities and Capabilities.
Elsevier.

Troffaes, M. (2007).
Decision making under uncertainty using imprecise probabilities.
International Journal of Approximate Reasoning, 45:17–29.

Van Rijn, J. N., Bischl, B., Torgo, L., Gao, B., Umaashankar, V., Fischer, S., Winter, P., Wiswedel, B., Berthold, M. R., and Vanschoren, J. (2013).
Openml: A collaborative science platform.
In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 645–649. Springer.

von Neumann, J., Morgenstern, O., Kuhn, H., and Rubinstein, A. (1944).
Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition).
Princeton University Press.

Walley, P. (1991).
Statistical Reasoning with Imprecise Probabilities.
Chapman and Hall, London.