Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement

Christoph Jansen, Georg Schollmeyer , Hannah Blocher, Julian Rodemann, Thomas Augustin

Department of Statistics, Ludwig-Maximilians-Universität München (LMU Munich)

abstraction | concretization | exemplification

Locally Varying Scale of Measurement

Motivation: What if the codomain of a random variable of interest is not of standard scale of measurement (e.g., ordinal or cardinal) but its structure *varies along its subsets*?

GENERAL ↓ SITUATION

Formalization: Preference Systems

Notation: For a preorder R, denote by P_R its *strict* and by I_R its *indifference* part.

Definition 1. *Let $A \neq \emptyset$ be a set. Let $R_1 \subseteq A \times A$ be a preorder on A and $R_2 \subseteq R_1 \times R_1$ be a preorder 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 $\exists\, u : A \to [0,1]$ s.t. for all $a,b,c,d \in A$:*

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

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

Definition 2. *A consistent preference system $\mathcal{A}$ is* ***bounded****, if $\exists\, a_*, a^* \in A$ such that $(a^*, a) \in R_1$, and $(a, a_*) \in R_1$ for all $a \in A$, and $(a^*, a_*) \in P_{R_1}$. In this case, for $\delta \in [0,1)$, denote by $\mathcal{N}^{\delta}_{\mathcal{A}}$ the set of all $u \in \mathcal{U}_{\mathcal{A}}$ with $u(a_*) = 0$, $u(a^*) = 1$, and*

$$u(a) - u(b) \geq \delta \quad \wedge \quad u(c) - u(d) - u(e) + u(f) \geq \delta$$

for all $(a,b) \in P_{R_1}$ and for all $((c,d),(e,f)) \in P_{R_2}$.

GENERAL ↓ SITUATION

Generalized Stochastic Dominance (GSD)

For π a probability measure on $(\Omega, \mathcal{S})$ and $\mathcal{A}$ a consistent preference system, set

$$\mathcal{F}_{(\mathcal{A},\pi)} := \left\{ X \in A^{\Omega} : u \circ X \in \mathcal{L}^1(\Omega, \mathcal{S}, \pi)\ \forall u \in \mathcal{U}_{\mathcal{A}} \right\}.$$

For $X, Y \in \mathcal{F}_{(\mathcal{A},\pi)}$, say Y is $(\mathcal{A},\pi)$**-dominated** by X, formally $(X,Y) \in R_{(\mathcal{A},\pi)}$, if

$$\forall\, u \in \mathcal{U}_{\mathcal{A}} : \quad \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y).$$

The preorder $R_{(\mathcal{A},\pi)}$ on $\mathcal{F}_{(\mathcal{A},\pi)}$ called **generalized stochastic dominance (GSD)**.

GENERAL ↓ SITUATION

Testing for GSD

Assume *i.i.d.* samples $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_m)$ of X and Y.

Hypotheses:

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

Test Statistic:

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

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

with, for $\omega \in \Omega$ and $\varepsilon \in [0,1]$ fixed, and

- $\hat{\pi}^{\omega}_{X}$ and $\hat{\pi}^{\omega}_{Y}$ the observed empirical image measures of X and Y,
- $(\mathbf{XY})_{\omega} = \{X_i(\omega) : i \leq n\} \cup \{Y_i(\omega) : i \leq m\} \cup \{a_*, a^*\}$, and
- $\mathcal{A}_{\omega}$ the subsystem of $\mathcal{A}$ restricted to $(\mathbf{XY})_{\omega}$, and
- $\delta_{\varepsilon}(\omega) := \varepsilon \cdot \sup\{\xi : \mathcal{N}^{\xi}_{\mathcal{A}_{\omega}} \neq \emptyset\}$.

Computation: $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$ can be computed by solving one single *linear program.*

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

$$\mathbf{x} := (x_1, \dots, x_n) := (X_1(\omega_0), \dots, X_n(\omega_0)) \ , \quad \mathbf{y} := (y_1, \dots, y_m) := (Y_1(\omega_0), \dots, Y_m(\omega_0))$$

As the worst case of H_0 is $\pi_X = \pi_Y$, we can perform a *permutation test*:

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

Step 2: For all $I \subseteq \{1, \dots, n+m\}$ with $|I| = n$, compute $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$ for $(w_i)_{i \in I}$ and $(w_i)_{i \in \{1,\dots,n+m\} \setminus I}$ instead of $\mathbf{x}/\mathbf{y}$ to get d^{ε}_{I}. Sort all d^{ε}_{I} increasingly to get $d^{\varepsilon}_{(1)}, \dots, d^{\varepsilon}_{(k)}$.

Step 3: Reject H_0 if $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) > d^{\varepsilon}_{(\ell)}$, with $\ell := \lceil (1-\alpha)k \rceil$ and α the significance level.

GENERAL ↓ SITUATION

Robustifying the Test

Idea: Use *credal sets* to robustify the permutation test. Concretely, allow the 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 test scheme: Replace

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

Results in: Valid (yet conservative) statistical test!

SPECIAL CASE →

Spaces with Differently Scaled Dimensions

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 dimensions are purely ordinal.

Utilize the cardinal information *only* on parts of A where there is *no possible conflict* with the ordinal one. Consider A to be a subsystem of $\textit{pref}(\mathbb{R}^r) = [\mathbb{R}^r, R^*_1, R^*_2]$, where

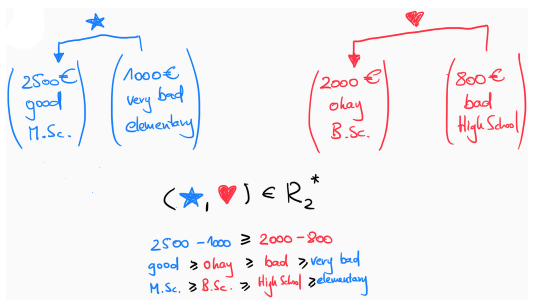
$$R^*_1 = \left\{ (x,y) : x_j \geq y_j \ \forall j \leq r \right\}$$

$$R^*_2 = \left\{ ((x,y),(x',y')) : \begin{array}{l} x_j - y_j \geq x'_j - y'_j \ \ \forall j \leq z \\ x_j \geq x'_j \geq y'_j \geq y_j \ \forall j > z \end{array} \right\}.$$

CONCRETE ↓ SDSD

Example: Poverty Analysis

We use the ALLBUS data and account for three dimensions of poverty: `income` (numeric), `health` (ordinal, 6 levels) and `education` (ordinal, 8 levels). E.g., for the following two pairs of vectors we can utilize the cardinal dimensions:



FOR SDSDs

Some Properties of SDSDs

Theorem 1. *Let $X = (\Delta_1, \dots, \Delta_r), Y = (\Lambda_1, \dots, \Lambda_r) \in \mathcal{F}_{(\textit{pref}(\mathbb{R}^r),\pi)}$. Then:*

i) $\textit{pref}(\mathbb{R}^r)$ *is consistent.*

ii) *If $z = 0$, then $R_{(\textit{pref}(\mathbb{R}^r),\pi)}$ equals (first-order) stochastic dominance w.r.t. π and R^*_1 (short: $FSD(R^*_1, \pi)$).*

iii) *If $(X,Y) \in R_{(\textit{pref}(\mathbb{R}^r),\pi)}$ and $\Delta_j, \Lambda_j \in \mathcal{L}^1(\Omega, \mathcal{S}_1, \pi)$ for all $j = 1, \dots, r$, then*

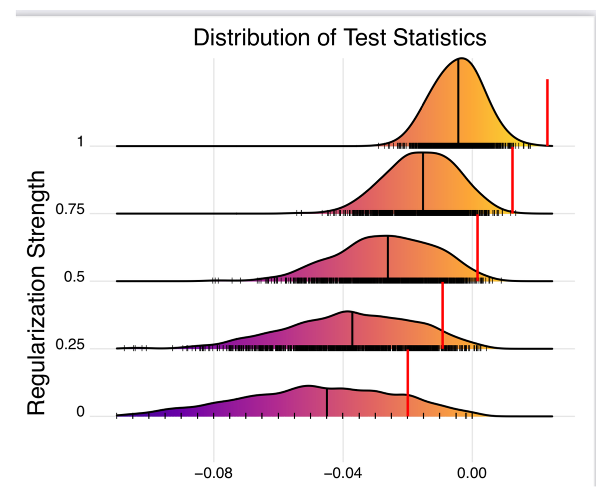
I. $\mathbb{E}_{\pi}(\Delta_j) \geq \mathbb{E}_{\pi}(\Lambda_j)$ *for all $j = 1, \dots, r$, and*

II. $(\Delta_j, \Lambda_j) \in FSD(\geq, \pi)$ *for all $j = z+1, \dots, r$.*

If all components of X are jointly independent and all components of Y are jointly independent, I. and II. imply $(X,Y) \in R_{(\textit{pref}(\mathbb{R}^r),\pi)}$.

Test in the Example

For the ALLBUS data, we focus on subsamples with $n = m = 100$ men and women.



Results: All tests significant for $\alpha = 0.05$. P-values decrease with increasing regularization strength ε of the test statistic.

Credal Sets in Example

A special class of credal sets are *γ-contamination models.* For $\omega \in \Omega$, $\gamma \in [0,1]$, and $Z \in \{X, Y\}$, we set

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

Interpretation: The contamination parameter γ can be interpreted as the share of data that can deviate from the *i.i.d.* sampling assumption.

Results Robust Testing