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

**Christoph Jansen**[1] **Georg Schollmeyer**[1] **Hannah Blocher**[1] **Julian Rodemann**[1] **Thomas Augustin**[1]

[1]Department of Statistics, Ludwig-Maximilians-Universität, Munich, Bavaria, Germany

## Abstract

Spaces with locally varying scale of measurement, like multidimensional structures with differently scaled dimensions, are pretty common in statistics and machine learning. Nevertheless, it is still understood as an open question how to exploit the entire information encoded in them properly. We address this problem by considering an order based on (sets of) expectations of random variables mapping into such non-standard spaces. This order contains stochastic dominance and expectation order as extreme cases when no, or respectively perfect, cardinal structure is given. We derive a (regularized) statistical test for our proposed generalized stochastic dominance (GSD) order, operationalize it by linear optimization, and robustify it by imprecise probability models. Our findings are illustrated with data from multidimensional poverty measurement, finance, and medicine.

## 1 INTRODUCTION

Numerous challenges in statistics and machine learning can – at least theoretically – be broken down to comparing random variables $X, Y : \Omega \to A$ mapping between measurable spaces $(\Omega, \mathcal{S}_1)$ and $(A, \mathcal{S}_2)$. Consequently, much attention has been paid to find and apply well-founded *stochastic orderings* enabling such comparison. Examples range from classifier comparisons (e.g., Demsar [2006], Corani et al. [2017], or Blocher et al. [2023]) over ranking risky assets (e.g., Chang et al. [2015]) to deriving optimal (generalized) Neyman-Pearson tests (e.g., [Augustin et al., 2014b, §7.4]).

In the traditional case where the context allows to specify both a probability $\pi$ on $\mathcal{S}_1$, and a *cardinal* scale $u : A \to \mathbb{R}$ representing the structure on $A$, a common order $\succsim_{E(u)}$ on $\left\{X \in A^\Omega : u \circ X \in \mathcal{L}^1(\Omega, \mathcal{S}_1, \pi)\right\}$ is obtained by setting $(X, Y) \in \succsim_{E(u)}$ if and only if

$$\mathbb{E}_\pi(u \circ X) = \int_\Omega u \circ X d\pi \geq \int_\Omega u \circ Y d\pi = \mathbb{E}_\pi(u \circ Y). \quad (1)$$

Here, random variables are ranked according to the expectations of their numerical equivalents induced by the scale $u$. We take the following perspective: This order $\succsim_{E(u)}$ would be the desired order if we were confronted with a problem under pure *aleatoric* uncertainty where an (objective) probability measure $\pi$ and a cardinal scale $u$ *were* available.[1]

Our paper addresses all situations where, in addition, *epistemic uncertainty* has to be taken into account. Then, such single $\pi$ and $u$ (and consequently the expectations in (1)) are not available, rendering a comparison by $\succsim_{E(u)}$ impossible. This non-availability corresponds to two facets (e.g. Hüllermeier and Waegeman [2021]) of epistemic uncertainty: Referring to $\pi$, *approximation* uncertainty arises since – as common in statistics – only samples of the considered variables are available.[2] Concerning $u$, on the other hand, *model* uncertainty is assumed to occur from weakly structured order information, making a non-singleton *set* $\mathcal{U}$ of candidate scales compatible with the structure on $A$.

Naturally, such situations can be approached in two steps: Focusing– in the first step – on model uncertainty, and thus assuming $\pi$ still to be known, the order $\succsim_{E(u)}$ can be weakened to a *preorder* $\succsim_{(\mathcal{U},\pi)}$ on

$$\left\{X \in A^\Omega : u \circ X \in \mathcal{L}^1(\Omega, \mathcal{S}_1, \pi)\ \forall u \in \mathcal{U}\right\}$$

by setting $(X, Y) \in \succsim_{(\mathcal{U},\pi)}$ if and only if Inequality (1) holds for all candidate scales $u \in \mathcal{U}$. Depending on the concrete choice of the set $\mathcal{U}$, the relation $\succsim_{(\mathcal{U},\pi)}$ has some

[1]The term "aleatoric uncertainty" seems adequate only when $\pi$ refers to a stochastic phenomenon. However, $\pi$ might as well represent subjective beliefs which can be formalized by a probability measure such as, e.g., in the Bayesian school of thought.

[2]In Section 6 we go beyond approximation uncertainty and consider robustification by a candidate set of probabilities.

*Accepted for the 39th Conference on Uncertainty in Artificial Intelligence* (UAI 2023).

prominent special cases: If $A$ is equipped with a preorder, and $\mathcal{U}$ is the set of all functions that are bounded and isotone w.r.t. this preorder, then $\succsim_{(\mathcal{U},\pi)}$ is (essentially) equivalent to (first-order) stochastic dominance. In contrast, if $(A, \mathcal{S}_2) = (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ and $\mathcal{U}$ consists of all bounded and *concave* functions, then $\succsim_{(\mathcal{U},\pi)}$ (essentially) corresponds to second-order stochastic dominance.

If – in a second step – information about $\pi$ comes only from samples from the distributions of $X$ and $Y$, then, instead of the order $\succsim_{(\mathcal{U},\pi)}$, one has to rely on the corresponding empirical version. Then, a statistical test is needed to control the probability of wrong conclusions from the data.

**Motivation of our work:** The main goal of the present work is to provide scientists from different fields of application with an inference methodology for the robust analysis of systematic distributional differences within a population. On the one hand, it is important to go beyond a simple comparison of location measures, similar to the case of classical stochastic dominance. On the other hand, we want to take into account the fact that classical (first-order) stochastic dominance systematically ignores potentially available metric information. We achieve this by a generalized stochastic dominance ordering (GSD), which is based on the flexible concept of preference systems. Specifically, we propose a nonparametric permutation test for subgroup comparison that robustifies (therefore further weakening the already parsimonious assumptions) towards the often-criticized assumption of exactly representative sampling.

**Our contribution**: We consider generalized stochastic dominance (GSD) that ensures exploiting the entire information encoded in data with locally varying scale of measurement. For that purpose, we (primarily) focus, technically speaking, on that specific class of preorders $\succsim_{(\mathcal{U},\pi)}$ where $\mathcal{U}$ is the set of representations of a *preference system* (cf. Sections 2 to 4). Then, using linear optimization, we derive a corresponding (regularized) test (cf. Section 5) and robustify it relying on imprecise probabilities (cf. Section 6). Particularly, our framework allows handling multidimensional structures with differently scaled dimensions in an information-efficient way (cf. Section 7). We illustrate this with data from multidimensional poverty measurement, finance, and medicine (cf. Section 8 and Supp. D) and conclude with a brief discussion (cf. Section 9). The proofs of Propositions 1 to 8, and Corollary 1 can be found in the supplementary material (cf., Supp. A). Our code is available under: `https://github.com/hannahblo/Robust_GSD_Tests`

**Related work:** Work on tests and/or checking algorithms for stochastic dominance (SD) outside preference systems includes McFadden [1989], Mosler and Scarsini [1991], Mosler [1995], Barrett and Donald [2003], Schollmeyer et al. [2017], Range and Østerdal [2019], Chetverikov et al. [2021]. Optimization under SD constraints was recently considered by, e.g., Dai et al. [2023]. Preference systems and related structures are discussed in a decision theoretic context in Pivato [2013] and Jansen et al. [2018, 2022a]. A test for GSD in the special case of a preference system arising from multiple quality metrics in classifier comparison is discussed in Jansen et al. [2022b].

Neighborhood models that are used to robustify tests are studied in e.g., Destercke et al. [2022], Augustin and Schollmeyer [2021], Montes et al. [2020]. Among others, Maua and de Campos [2021], Cabanas et al. [2020], Maua and Cozman [2020] study credal networks as robustifications of Bayesian networks, and, e.g., Utkin and Konstantinov [2022], Rodemann and Augustin [2022], Carranza and Destercke [2021], Utkin [2020], Abellan et al. [2018] have proposed robustifications and extensions of other machine learning procedures like forests or discriminant analyses by imprecise probabilities.

Accounting for both approximation uncertainty and model uncertainty is in line with recent deliberations in uncertainty quantification (e.g., Malinin and Gales [2018], Hüllermeier and Waegeman [2021], Bengs et al. [2022], Hüllermeier et al. [2022]).

## 2 BACKGROUND & PRELIMINARIES

We will consider *binary relations* at several points, relying on the following notation and terminology: A binary relation $R$ on a set $M \neq \emptyset$ is a subset of the Cartesian product of $M$ with itself, i.e. $R \subseteq M \times M$. $R$ is called *reflexive*, if $(a, a) \in R$, *transitive*, if $(a, b), (b, c) \in R \Rightarrow (a, c) \in R$, *antisymmetric*, if $(a, b), (b, a) \in R \Rightarrow a = b$, *complete*, if $(a, b) \in R$ or $(b, a) \in R$ (or both) for arbitrary elements $a, b, c \in M$. A *preference relation* is a binary relation that is complete and transitive; a *preorder* is a binary relation that is reflexive and transitive; a *linear order* is a preference relation that is antisymmetric; a *partial order* is a preorder that is antisymmetric. If $R$ is a preorder, we denote by $P_R \subseteq M \times M$ its *strict part* and by $I_R \subseteq M \times M$ its *indifference part*, defined by $(a, b) \in P_R \Leftrightarrow (a, b) \in R \wedge (b, a) \notin R$, and $(a, b) \in I_R \Leftrightarrow (a, b) \in R \wedge (b, a) \in R$.

This leads us to the central ordering structure under consideration in the present paper, namely *preference systems*. These formalize the idea of spaces with locally varying scale of measurement and were introduced in Jansen et al. [2018].[3]

**Definition 1** *Let $A \neq \emptyset$ be a set, $R_1 \subseteq A \times A$ a preorder on $A$, and $R_2 \subseteq R_1 \times R_1$ a preorder on $R_1$. The triplet $\mathcal{A} = [A, R_1, R_2]$ is then called a* ***preference system*** *on $A$. We call $\mathcal{A}$* ***bounded****, if there exist $a_*, a^* \in A$ such that $(a^*, a) \in R_1$, and $(a, a_*) \in R_1$ for all $a \in A$, and $(a^*, a_*) \in P_{R_1}$. Moreover, the preference system $\mathcal{A}' = [A', R_1', R_2']$ is called*

[3]For a study on representation results of the related concept of *incomplete difference preorders* see, e.g., Pivato [2013].

***subsystem*** *of $\mathcal{A}$ if $A' \subseteq A$, $R_1' \subseteq R_1$, and $R_2' \subseteq R_2$. In this case, we call $\mathcal{A}$ a* ***supersystem*** *of $\mathcal{A}'$.*

The concrete definition of a preference system now also makes it possible to concretize the idea of a space with *locally varying scale of measurement*: While the relation $R_1$ formalizes the available ordinal information, i.e. information about the arrangement of the elements of $A$, the relation $R_2$ describes the cardinal part of the information in the sense that pairs standing in relation are ordered with respect to the intensity of the relation. Thus, intuitively speaking, the set $A$ is locally almost cardinally ordered on subsets where $R_1$ and $R_2$ are very dense, while on subsets where $R_2$ is sparse or even empty, locally at most an ordinal scale of measurement can be assumed. A natural example is multi-dimensional structures with differently scaled dimensions, such as those that appear in the poverty analysis application discussed in Section 8: While variables like education can be assumed to have only ordinal scale of measurement, a variable like income is rather metrically scaled.

To ensure that $R_1$ and $R_2$ are compatible, we use a consistency criterion for preference systems relying on the idea that both relations should be simultaneously representable.

**Definition 2** *The preference system $\mathcal{A} = [A, R_1, R_2]$ is* ***consistent*** *if there exists a* ***representation*** *$u : A \to \mathbb{R}$ such that for all $a, b, c, d \in A$ we have:*

*i) If we have that $(a, b) \in R_1$, then it holds that $u(a) \geq u(b)$, where equality holds if and only if $(a, b) \in I_{R_1}$.*

*ii) If we have that $((a, b), (c, d)) \in R_2$, then it holds that $u(a) - u(b) \geq u(c) - u(d)$, where equality holds if and only if $((a, b), (c, d)) \in I_{R_2}$.*

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

Especially when regularizing our test statistic in Section 5, normalized versions of the set $\mathcal{U}_\mathcal{A}$ play a crucial role.

**Definition 3** *Let $\mathcal{A} = [A, R_1, R_2]$ be a consistent and bounded preference system with $a_*, a^*$ as before. Then*

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

*is called the* ***normalized representation set*** *of $\mathcal{A}$. Further, for $\delta \in [0, 1)$, we denote by $\mathcal{N}_\mathcal{A}^\delta$ the set of all $u \in \mathcal{N}_\mathcal{A}$ with*

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

*for all $(a, b) \in P_{R_1}$ and for all $((c, d), (e, f)) \in P_{R_2}$. We call $\mathcal{A}$ $\delta$-****consistent*** *if $\mathcal{N}_\mathcal{A}^\delta \neq \emptyset$.*

We conclude the section with an immediate observation of the connection between consistency and 0-consistency.

**Proposition 1** *Let $\mathcal{A} = [A, R_1, R_2]$ be a bounded preference system. Then $\mathcal{A}$ is consistent if and only if it is 0-consistent.*

# 3 REGULARIZATION

We now discuss some thoughts on regularization in preference systems. Since our later considerations primarily concern statistical testing, regularization then aims at making the test statistic more sensitive, i.e., to increase discriminative power. In contrast to the usually advocated Thikonov-type regularization, here we think in terms of Ivanov-type regularization that constraints the space of functions (in our case $\mathcal{N}_\mathcal{A}$) over which later our optimization is done (cf., Section 5.1 where our test statistic is introduced as an infimum type test statistic). Beyond the different more or less equivalent ways of representing regularization in a Thikonov-, Ivanov- or in a Morozov-type style (cf. Oneto et al. [2016]), here additionally, two different ways of implementing regularization are conceivable: On the one hand, an *order-theoretic* regularization could be carried out by extending the considered preference system by additional comparable pairs (or pairs of pairs) to a consistent super system. On the other hand, a *parameter-driven* regularization could be performed to reduce the set of representations of the preference system. Both ways are schematically compared in Figure 1.

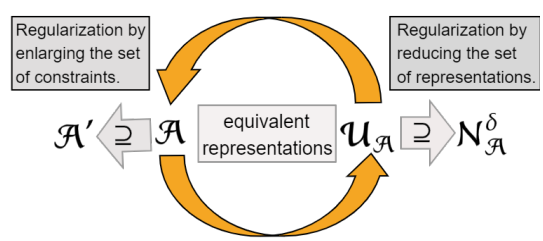


Figure 1: Two ways for regularizing a preference system.

Both approaches have their own strengths and weaknesses: In the case of order-theoretic regularization, the influence of the regularization on the content-related question can be controlled very precisely. However, this comes at the price that the concrete mathematical influence of the regularization can only be characterized with difficulty. The situation tends to be reversed in the case of parameter-driven regularization: Here, it is straightforward – by choosing larger and larger parameter values – to control the mathematical influence of the regularization. However, an interpretation of the regularization in the context of the content-related question is less direct than in the first case. Nevertheless, a possible interpretation in a decision-theoretic context is given in Jansen et al. [2018] by establishing a connection to Luce's *just noticeable differences* [Luce, 1956]. In this paper, we focus on parameter-driven regularization since, for regularization of the test statistic used later, the interpretation of the parameter is of secondary importance.

# 4 GENERALIZED DOMINANCE

As indicated at the outset, we now turn to a stochastic order between random variables with values in a preference system. This order rigorously generalizes stochastic dominance in the sense that it optimally exploits also the partial cardinal information encoded in these spaces. Therefore, it is neither limited to a purely ordinal analysis as first-order stochastic dominance nor requires perfect cardinal information as second-order stochastic dominance. Consequently, in cases without any cardinal information, i.e., where $R_2$ is the trivial preorder, the considered order reduces back to the first-order stochastic dominance.

We start by introducing some additional notation: For $\pi$ a probability measure on $(\Omega, \mathcal{S}_1)$ and $\mathcal{A}$ a consistent preference system, we define by $\mathcal{F}_{(\mathcal{A},\pi)}$ the set

$$\Big\{X \in A^\Omega : u \circ X \in \mathcal{L}^1(\Omega, \mathcal{S}_1, \pi)\ \forall u \in \mathcal{U}_\mathcal{A}\Big\}.$$

We then can define the following preorder on $\mathcal{F}_{(\mathcal{A},\pi)}$.

**Definition 4** *Let $\mathcal{A} = [A, R_1, R_2]$ be consistent. For $X, Y \in \mathcal{F}_{(\mathcal{A},\pi)}$, we say $Y$ is $(\mathcal{A}, \pi)$**-dominated** by $X$ if*

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

*for all $u \in \mathcal{U}_\mathcal{A}$. The induced relation is denoted by $R_{(\mathcal{A},\pi)}$ and called **generalized stochastic dominance (GSD)**.*

We have the following immediate simplification if the underlying preference system $\mathcal{A}$ is additionally bounded.

**Proposition 2** *If $\mathcal{A}$ is consistent and bounded with $a_*, a^*$ as before, then $(X, Y) \in R_{(\mathcal{A},\pi)}$ iff*

$$\forall u \in \mathcal{N}_\mathcal{A} : \mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y). \tag{2}$$

# 5 TESTING FOR DOMINANCE

Throughout this section, let $\mathcal{A} = [A, R_1, R_2]$ be *consistent* and *bounded* with $a_*, a^* \in A$ as in Definition 1.

We now turn to the statistical version of our investigation, where we do not know the underlying probability $\pi$ but *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. The fundamental question now is when we can, with a certain error probability, conclude from this information that $X, Y \in \mathcal{F}_{(\mathcal{A},\pi)}$ are in relation with respect to the GSD-relation $R_{(\mathcal{A},\pi)}$. Constructing a corresponding test, we first need to be clear about appropriate statistical hypotheses. Ideally, we would be interested in the following pair of hypotheses:

$$H_0^{id} : (X, Y) \notin R_{(\mathcal{A},\pi)} \quad \textbf{vs.} \quad H_1^{id} : (X, Y) \in R_{(\mathcal{A},\pi)} \tag{3}$$

In the pair $(H_0^{id}, H_1^{id})$ of hypotheses – as intended in a statistical test – the question actually of interest would be formulated as the alternative hypothesis. Then, the probability of falsely assuming it to be true could be controlled by the significance level. Unfortunately, similar to the situation of classical stochastic dominance as described, e.g., in Barrett and Donald [2003] and further investigated in Shaked and Shanthikumar [2007], or generally in the context of bioequivalence testing (e.g., Brown et al. [1997]), the hypothesis $H_0^{id}$ seems to be too broad for a meaningful analysis, in the sense that the most conservative scenario under $H_0^{id}$ is not clearly specifiable.[4] For this reason, we choose a pair of alternatives that deviates slightly from the actual question of interest and afterwards try to make the deviation from the actual pair of hypotheses of interest assessable by testing with the variables in reversed roles. The modified pair of hypotheses looks as follows:

$$H_0 : (Y, X) \in R_{(\mathcal{A},\pi)} \quad \textbf{vs.} \quad H_1 : (Y, X) \notin R_{(\mathcal{A},\pi)} \tag{4}$$

The advantage of the pair $(H_0, H_1)$ is that a worst-case analysis of the distribution of a suitable test statistic under $H_0$ is possible: The test statistic would have to be analyzed under the most conservative case within $H_0$, namely $\pi_X = \pi_Y$, with $\pi_X$ and $\pi_Y$ the image measures of $X$ and $Y$ under $\pi$. The drawback to the pair $(H_0, H_1)$ is that in the case of rejection of $H_0$ we can only control the erroneous conclusion on $(Y, X) \notin R_{(\mathcal{A},\pi)}$ (and not the one actually of interest on $(X, Y) \in R_{(\mathcal{A},\pi)}$) in its probability by the significance level. To mitigate this effect, we can test with the pair $(H_0, H_1)$ of hypotheses additionally with $X$ and $Y$ in reversed roles.

## 5.1 THE CHOICE OF THE TEST STATISTIC

For defining an adequate test statistic, we first note that – due to the boundedness of $\mathcal{A}$ and Proposition 2 – it holds $(X, Y) \in R_{(\mathcal{A},\pi)}$ if and only if

$$D(X, Y) := \inf_{u \in \mathcal{N}_\mathcal{A}} (\mathbb{E}_\pi(u \circ X) - \mathbb{E}_\pi(u \circ Y)) \geq 0, \tag{5}$$

i.e., if the infimal expectation difference with respect to the available information is at least zero. Thus, a straightforward test statistic is the empirical version of $D(X, Y)$, i.e.,

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

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

with, for $\omega \in \Omega$ fixed, $\hat{\pi}_X^\omega(\cdot) := \frac{1}{n}|\{i : X_i(\omega) \in \cdot\}|$ and $\hat{\pi}_Y^\omega(\cdot) := \frac{1}{m}|\{i : Y_i(\omega) \in \cdot\}|$ 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 $(\mathbf{XY})_\omega$. If $d_{\mathbf{X},\mathbf{Y}}(\omega_0) \geq 0$ holds for some $\omega_0 \in \Omega$, we say there is

[4] The problem is due to the fact that the relation $R_{(\mathcal{A},\pi)}$ is a *partial* order. Compare also [Schollmeyer et al., 2017, p. 24-25].

*in-sample GSD* of $X$ over $Y$ in the sample induced by $\omega_0$. If the underlying space $\mathcal{A}$ is not too complex[5] (under *i.i.d.* within every subgroup) this test statistic converges to the true value of $D(X,Y)$ and is therefore an adequate test statistic for our test.

As a further test statistic, we consider a regularized version $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$ of $d_{\mathbf{X},\mathbf{Y}}$: The infimum in the definition of $d_{\mathbf{X},\mathbf{Y}}$ is now only computed among $[0,1]$-normalized representations of $\mathcal{A}_\omega$ that distinguish between strictly related alternatives over some prespecified threshold value. In this way, the regularized test statistic is also sensitive for distinguishing situations under dominance regarding their *extent* of dominance: While in-sample GSD (essentially) implies $d_{\mathbf{X},\mathbf{Y}}(\omega_0) = 0$, it often holds $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) > 0$. Thus, for $V, W$ with $(\mathbf{VW})_{\omega_0} = (\mathbf{XY})_{\omega_0}$ it might be that $d_{\mathbf{V},\mathbf{W}}(\omega_0) = 0$ and $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) > d^{\varepsilon}_{\mathbf{V},\mathbf{W}}(\omega_0) > 0$ and, hence, that under regularization $X$ (empirically) dominates $Y$ more strongly than $V$ dominates $W$.[6]

Formally, the regularized test statistic looks as follows:

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

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

with $\varepsilon \in [0,1]$ and $\delta_\varepsilon(\omega) := \varepsilon \cdot \sup\{\xi : \mathcal{N}^{\xi}_{\mathcal{A}_\omega} \neq \emptyset\}$. Observe that $d_{\mathbf{X},\mathbf{Y}} = d^0_{\mathbf{X},\mathbf{Y}}$, i.e., the unregularized test statistic equals the regularized one if $\varepsilon = 0$.

### 5.2 A PERMUTATION-BASED TEST

As the distribution of $d_{\mathbf{X},\mathbf{Y}}$ and $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$ can not be straightforwardly analyzed, we utilize that under the above i.i.d.-assumption a permutation-based test (see, e.g., Pratt and Gibbons [2012]) can be performed. For this, we assume we made observations of the i.i.d. variables, i.e., we observed

$$\mathbf{x} := (x_1, \ldots, x_n) := (X_1(\omega_0), \ldots, X_n(\omega_0)) \quad (6)$$
$$\mathbf{y} := (y_1, \ldots, y_m) := (Y_1(\omega_0), \ldots, Y_m(\omega_0)) \quad (7)$$

for some $\omega_0 \in \Omega$. The resampling scheme for analyzing the distributions of $d_{\mathbf{X},\mathbf{Y}}$ and $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$, respectively, can then be described by the following steps:

[5] A concrete sufficient condition for consistency of $d_{\mathbf{X},\mathbf{Y}}$ is a finite VC dimension of the class of all indicator functions of the form $\{a \mid u(a) \geq c\}$ with $u \in \mathcal{N}_{\mathcal{A}}$. This property is usually given, for example if we have finitely many dimensions which have itself a finite VC dimension. Therefore, especially in our applications of Section 8 consistency is guaranteed.

[6] As an example, in the situation of a preference system guaranteeing a totally ordered space (i.e., $R_2$ is the trivial preorder, $R_1$ is complete) where the laws of the random variables build a location family $\{f(\cdot + c) \mid c \in \mathbb{R}\}$, the regularized statistic (with appropriately chosen $\delta$) will capture the difference $\Delta = c - \tilde{c}$ between two populations distributed according to $f(\cdot + c)$ and $f(\cdot + \tilde{c})$, respectively, whereas the non-regularized test will not.

**Step 1:** Take the pooled 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 $n$. Evaluate $d_{\mathbf{X},\mathbf{Y}}$ resp. $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$ for $(w_i)_{i \in I}$ and $(w_i)_{i \in \{1,\ldots,n+m\} \setminus I}$ instead of $\mathbf{x}$ and $\mathbf{y}$. Denote the evaluations by $d_I$ resp. $d^{\varepsilon}_I$.

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

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

Note that, for large $\binom{n+m}{n}$, we can approximate the above resampling scheme by computing $d_I$ resp. $d^{\varepsilon}_I$ only for a large number $N$ of randomly drawn $I$. Moreover, note that only the *i.i.d.* assumption is needed for the above test to be valid. (Precisely, it would already be enough to assume *exchangeable* observations of both variables.)

### 5.3 COMPUTATION OF $\mathbf{d_{X,Y}}$ AND $\mathbf{d^{\varepsilon}_{X,Y}}$

We show how the test statistics $d_{\mathbf{X},\mathbf{Y}}$ and $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$ can be computed in concrete cases. For that, we consider samples $\mathbf{x}$ and $\mathbf{y}$ of the form (6) and (7), and we assume w.l.o.g. that

$$(\mathbf{XY})_{\omega_0} = \{z_1 = a_*, z_2 = a^*, z_3, \ldots, z_s\}$$

Further, we denote by $C(\mathbf{x}, \mathbf{y})$ the set of all vectors $(v_1, \ldots, v_s, \xi) \in [0,1]^{s+1}$ such that $v_1 = 0$ and $v_2 = 1$ and for which it holds that

- $v_i = v_j$ if $(z_i, z_j) \in I_{R_1}$,
- $v_i - v_j \geq \xi$ if $(z_i, z_j) \in P_{R_1}$,
- $v_k - v_l = v_r - v_t$ if $((z_k, z_l), (z_r, z_t)) \in I_{R_2}$ and
- $v_k - v_l - v_r + v_t \geq \xi$ if $((z_k, z_l), (z_r, z_t)) \in P_{R_2}$.

Moreover, for $\xi_0 \in [0,1]$ fixed, we define $C_{\xi_0}(\mathbf{x}, \mathbf{y})$ as $\{(v_1, \ldots, v_s) \in [0,1]^s : (v_1, \ldots, v_s, \xi_0) \in C(\mathbf{x}, \mathbf{y})\}$, i.e., the set of all sample weights that respect the observed preference system and distinguish the strict part of its relations above a threshold of $\xi_0$. Both $C(\mathbf{x}, \mathbf{y})$ and $C_{\xi_0}(\mathbf{x}, \mathbf{y})$ are described by finitely many linear inequalities on $(v_1, \ldots, v_s, \xi)$ resp. $(v_1, \ldots, v_s)$. This allows to formulate Propositions 3 and 4. The first one demonstrates how to compute the maximum regularization threshold, whereas the second one captures the computation of $d_{\mathbf{X},\mathbf{Y}}$ and $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$.

**Proposition 3** *For samples* $\mathbf{x}$ *and* $\mathbf{y}$ *of the form (6) and (7) and* $\varepsilon \in [0,1]$*, we consider the linear program (LP)*

$$\xi \longrightarrow \max_{(v_1, \ldots, v_s, \xi)} \quad (8)$$

*with constraints* $(v_1, \ldots, v_s, \xi) \in C(\mathbf{x}, \mathbf{y})$*. Denote by* $\xi^*$ *its optimal value. It then holds* $\delta_\varepsilon(\omega_0) = \varepsilon \cdot \xi^*$*.*

**Proposition 4** *For samples* $\mathbf{x}$ *and* $\mathbf{y}$ *of the form (6) and (7) and* $\varepsilon \in [0,1]$*, we consider the following LP*

$$\sum_{\ell=1}^{s} v_\ell \cdot \left( \frac{|\{i : x_i = z_\ell\}|}{n} - \frac{|\{i : y_i = z_\ell\}|}{m} \right) \longrightarrow \min_{(v_1,\ldots,v_s)} \tag{9}$$

*with* $(v_1, \ldots, v_s) \in C_{\varepsilon\xi^*}(\mathbf{x}, \mathbf{y})$*, where* $\xi^*$ *is the optimal value of (8). Denote by* $opt_\varepsilon(\mathbf{x}, \mathbf{y})$ *its optimal value. Then:*

*i)* $opt_\varepsilon(\mathbf{x}, \mathbf{y}) = d^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0)$.

*ii) It holds in-sample GSD of* $X$ *over* $Y$ *iff* $opt_0(\mathbf{x}, \mathbf{y}) \geq 0$.

# 6 ROBUSTIFIED TESTING USING IP

Our test for GSD relies on i.i.d. samples of the populations of actual interest. It thus can be based directly on the observed empirical distributions. We now show how *imprecise probabilities (IP)* and *credal sets* (e.g., Walley [1991], Augustin et al. [2014a]) can be used to robustify our test towards deviations of its assumptions. Credal sets – and generally imprecise probabilities – form a consequent generalization of classical probability theory, which also accounts for partial probabilistic knowledge. Indeed, there are various reasons why the i.i.d. assumption can be violated, ranging from unobserved heterogeneity to dependencies arising from data collection. The latter reason is particularly prevalent in surveys, where the survey mode (e.g., phone, web, in-person) often results in unequal, and even outcome-dependent, chances of the units to be sampled. Although methods exist to tackle this, such as reweighting schemes or random routing, most of them come with flaws of their own kind. For example, Bauer [2014, 2016] shows that random routing may be substantially biased, leading to informatively distorted selection probabilities, hence non i.i.d. data.

## 6.1 THE ROBUSTIFIED TESTING FRAMEWORK

The rough idea of our robustification is to not analyze the test statistic based on $\hat{\pi}_X$ and $\hat{\pi}_Y$ alone, but use neighbourhood models or, more generally, *credal sets* $\mathcal{M}_X \ni \hat{\pi}_X$ and $\mathcal{M}_Y \ni \hat{\pi}_Y$ of candidate probability measures instead. Credal sets – introduced in Levi [1974] – model partial probabilistic information by the set of all non-contradictory probabilities and have gained popularity in machine learning (e.g., Corani and Zaffalon [2008], Lienen and Hüllermeier [2021], Shaker and Hüllermeier [2021], Jansen et al. [2022c], Rodemann et al. [2023], see also the corresponding literature referenced as related work in Section 1).

The concrete idea behind our robustification is that we allow our samples to be (potentially) biased. We assume that these biased samples are similar to the true ones in the sense that the associated true empirical laws are contained in the credal sets $\mathcal{M}_X$ and $\mathcal{M}_Y$ around the biased empirical laws, respectively. We start by only assuming both $\mathcal{M}_X$ and $\mathcal{M}_Y$ to be (random) convex polyhedra with extreme points collected in the finite sets $\mathcal{E}(\mathcal{M}_X)$ and $\mathcal{E}(\mathcal{M}_Y)$.

Now, we again want to test $H_0$ from Eq. (4), however, under the difficulty that the samples are biased. In the spirit of the concept of *cautious data completion* (see, e.g., [Augustin et al., 2014b, p. 181] or also Schollmeyer [2019] for the connections with stochastic dominance), one actually would adapt the resampling scheme discussed before by performing the test under all pairs of laws in the corresponding credal sets $\mathcal{M}_X$ and $\mathcal{M}_Y$. The null hypothesis $H_0$ from Eq. (4) would then be rejected whenever it is rejected for all such pairs. Since this adapted resampling scheme is computationally cumbersome, we instead look at the corresponding *lower envelopes* $\underline{d}_{\mathbf{X},\mathbf{Y}} : \Omega \to \mathbb{R}$ and $\underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}} : \Omega \to \mathbb{R}$, respectively, given by

$$\begin{aligned} \omega &\mapsto \inf_{(\pi_1,\pi_2,u)\in\mathcal{D}} \sum_{z\in(\mathbf{XY})_\omega} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \\ \omega &\mapsto \inf_{(\pi_1,\pi_2,u)\in\mathcal{D}_\varepsilon} \sum_{z\in(\mathbf{XY})_\omega} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \end{aligned}$$

with $\mathcal{D} = \mathcal{M}^\omega_X \times \mathcal{M}^\omega_Y \times \mathcal{N}_{\mathcal{A}_\omega}$, $\mathcal{D}_\varepsilon = \mathcal{M}^\omega_X \times \mathcal{M}^\omega_Y \times \mathcal{N}^{\delta_\varepsilon(\omega)}_{\mathcal{A}_\omega}$ and $\mathcal{M}^\omega_X$ resp. $\mathcal{M}^\omega_Y$ the empirical credal sets given $\omega \in \Omega$. We compare these lower envelopes with the distribution (in the resamples) of the corresponding upper envelopes, $\overline{d}_{\mathbf{X},\mathbf{Y}}$ and $\overline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}$, that are obtained by replacing the part of $\inf$ concerning $\mathcal{M}^\omega_X \times \mathcal{M}^\omega_Y$ with the respective $\sup$ in the above definitions. This gives a conservative yet valid statistical test.

## 6.2 COMPUTATION OF $\underline{\mathbf{d}}_{\mathbf{X},\mathbf{Y}}$ AND $\underline{\mathbf{d}}^\varepsilon_{\mathbf{X},\mathbf{Y}}$

We now give an algorithm for the robustified test statistics.

**Proposition 5** *For* $\mathbf{x}$ *and* $\mathbf{y}$ *of form (6) and (7),* $\varepsilon \in [0,1]$*, and* $(\pi_1, \pi_2) \in \mathcal{E}(\mathcal{M}^{\omega_0}_X) \times \mathcal{E}(\mathcal{M}^{\omega_0}_Y)$*, consider the LP*

$$\sum_{\ell=1}^{s} v_\ell \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \longrightarrow \min_{(v_1,\ldots,v_s)} \tag{10}$$

*with* $(v_1, \ldots, v_s) \in C_{\varepsilon\xi^*}(\mathbf{x}, \mathbf{y})$ *and* $\xi^*$ *the optimum of (8). Call* $opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2)$ *its optimum and* $\underline{opt}_\varepsilon(\mathbf{x}, \mathbf{y})$ *the minimal optimum over* $(\pi_1, \pi_2) \in \mathcal{E}(\mathcal{M}^{\omega_0}_X) \times \mathcal{E}(\mathcal{M}^{\omega_0}_Y)$*. Then:*

*i)* $\underline{opt}_\varepsilon(\mathbf{x}, \mathbf{y}) = \underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0)$.

*ii) There is in-sample GSD of* $X$ *over* $Y$ *for any* $\pi$ *with* $\hat{\pi}^{\omega_0}_X \in \mathcal{M}^{\omega_0}_X$ *and* $\hat{\pi}^{\omega_0}_Y \in \mathcal{M}^{\omega_0}_Y$ *if* $\underline{opt}_0(\mathbf{x}, \mathbf{y}) \geq 0$.

Proposition 5 requires to solve $|\mathcal{E}(\mathcal{M}^{\omega_0}_X)| \cdot |\mathcal{E}(\mathcal{M}^{\omega_0}_Y)|$ linear programs. Depending on the concrete neighbourhood models, this is obviously limited: The number of programs is simply too large. A common strategy in such a case is to additionally assume 2-monotonicity of the considered credal sets, since this allows us – at least for $R_1$ complete

– to give closed formulas for the upper and lower expectations. Unfortunately, this is not so simple in the case of a partially ordered $R_1$: since the representation via the Choquet integral (e.g., Denneberg [1994]) depends on the order of elements of $A$, an optimum over all linear extensions of $R_1$ is needed to determine the most extreme Choquet integrals. In the worst case, this would lead to optimizing a non-convex function and thus hardly simplify the original problem (see Timonin [2012]).

Another strategy is restricting to credal sets with moderately many extreme points. We now consider one such possibility, namely the the $\gamma$- *contamination model* (or *linear-vacuous model*, see, e.g., [Walley, 1991, p. 147]). Here, we assume that for $\omega \in \Omega$, $\gamma \in [0,1]$, and $Z \in \{X, Y\}$ fixed, we have

$$\mathcal{M}_Z^\omega = \Big\{\pi : \pi \geq (1-\gamma) \cdot \hat{\pi}_Z^\omega\Big\}, \tag{11}$$

where $\geq$ is understood event-wise. For $\gamma$-contamination models there are exactly as many extreme points as there are observed distinct data points, concretely given by

$$\mathcal{E}(\mathcal{M}_Z^\omega) = \Big\{\gamma\delta_z + (1-\gamma)\hat{\pi}_Z^\omega : \exists j \text{ s.t. } Z_j(\omega) = z\Big\}, \tag{12}$$

where $\delta_z$ denotes the Dirac-measure in $z$ (see again Walley [1991, p. 147]). Proposition 6 states that if the credal sets are both $\gamma$-contamination models, then a least favorable pair of extreme points can a priori be specified. The test statistics thus can be computed by solving one linear program.

**Proposition 6** *Consider again the situation of Proposition 5, where additionally $\mathcal{M}_X^{\omega_0}$ and $\mathcal{M}_Y^{\omega_0}$ are of the form (11) with extreme points as in (12). It then holds:*

$$\underline{opt}_\varepsilon(\mathbf{x}, \mathbf{y}) = opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_*, \pi^*), \textit{where}$$

$$\pi_* = \gamma\delta_{a_*} + (1-\gamma)\hat{\pi}_X^{\omega_0} \;\; \textit{and} \;\; \pi^* = \gamma\delta_{a^*} + (1-\gamma)\hat{\pi}_Y^{\omega_0}.$$

# 7 MULTIDIMENSIONAL SPACES WITH DIFFERENTLY SCALED DIMENSIONS

We now turn to a special case that is very common in applied research: multidimensional spaces whose dimensions may be of different scale of measurement.[7] While traditional empirical research and policy support (e.g., European Commission [2023]) summarizes such situations by indices/indicators that suffer eo ipso from "the subjectivity of choices associated with them" ([UNECE, 2019, p. 11]), the embedding into the framework considered here allows a faithful representation of the entire underlying information.

Concretely, we address $r \in \mathbb{N}$ dimensional spaces for which we assume – w.l.o.g. – that the first $0 \leq z \leq r$ dimensions are of cardinal scale (implying that differences of elements may be interpreted as such), while the remaining ones are purely ordinal (implying differences to be meaningless apart from the sign). Specifically, we consider (bounded subsystems of) the preference system[8]

$$\textit{pref}(\mathbb{R}^r) = [\mathbb{R}^r, R_1^*, R_2^*] \tag{13}$$

where

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

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

While $R_1^*$ can be interpreted as a simple component-wise dominance relation, $R_2^*$ deserves some more explanation: One pair of consequences is preferred to another one if it is ensured in the ordinal dimensions that the exchange associated with the first pair is not a deterioration to the exchange associated with the second pair and, in addition, there is component-wise dominance of the differences of the cardinal dimensions. The following proposition lists some important results for a more precise characterization of the GSD-relation on multidimensional structures.

**Proposition 7** *Let $\pi$ be a probability measure on $(\Omega, \mathcal{S}_1)$, and $X = (\Delta_1, \ldots, \Delta_r), Y = (\Lambda_1, \ldots, \Lambda_r) \in \mathcal{F}_{(\textit{pref}(\mathbb{R}^r),\pi)}$. Then, the following holds:*

*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. $\pi$ 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 $j = 1, \ldots, r$, then*

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

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

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

Part iii) of Proposition 7 is complete in the sense that the addition actually holds only under stochastic independence.

**Remark 1** *The addition to iii) does not generally hold. A counterexample is $z = 1$, $r = 2$, $\Omega = \{\omega_1, \ldots, \omega_4\}$, and $\pi$ the uniform distribution over $\Omega$. Then, for $\Delta_1(\omega) = 1,1,2,2$, $\Delta_2(\omega) = 1,1,2,2$, $\Lambda_1(\omega) = 1,1,2,2$, and $\Lambda_2(\omega) = 1,2,1,2$ for $\omega = \omega_1, \ldots, \omega_4$, it holds that $\mathbb{E}_\pi(\Delta_1) = \mathbb{E}_\pi(\Lambda_1)$. In fact, the first components are equivalent with respect to first order stochastic dominance. The same holds for the second components. However, the whole*

[7] For recent applications of such special preference systems to classifier comparison or multi-target decision making see Jansen et al. [2022b], Jansen and Augustin [2022] and Jansen et al. [2023].

[8] One easily verifies that $R_1^*$ and $R_2^*$ are preorders.

*vectors are incomparable with respect to first order stochastic dominance, since there is no corresponding mass transport from higher values to lower (or equal) values possible. Additionally, for $u(x,y) := x \cdot y$, we have that $u \in \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$, $\mathbb{E}_\pi(u \circ \Delta) = 10/4$, whereas $\mathbb{E}_\pi(u \circ \Lambda) = 9/4$. Thus, $\Delta$ and $\Lambda$ can not be equivalent with respect to GSD.*

As an immediate consequence of Proposition 7, we have the following corollary for bounded subsystems of $\mathsf{pref}(\mathbb{R}^r)$.

**Corollary 1** *If $\mathcal{C} = [C, R_1^c, R_2^c]$ is a bounded subsystem of $\mathsf{pref}(\mathbb{R}^r)$ and $X, Y \in \mathcal{F}_{(\mathcal{C},\pi)}$, then $\mathcal{C}$ is 0-consistent and ii) and iii) from Prop. 7 hold, if we replace $R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ by $R_{(\mathcal{C},\pi)}$, $FSD(R_1^*, \pi)$ by $FSD(R_1^c, \pi)$, and $(X,Y) \in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ by $\forall u \in \mathcal{N}_\mathcal{C} : \mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$.*

Finally, we give a characterization of the set of all representations of $\mathsf{pref}(\mathbb{R}^r)$ if only one dimension is cardinal.

**Proposition 8** *Let $z = 1$ and denote by $\mathcal{U}_{sep}$ the set of all $u : \mathbb{R}^r \to \mathbb{R}$ such that, for $(x_2, \ldots, x_r) \in \mathbb{R}^{r-1}$ fixed, the function $u(\cdot, x_2, \ldots, x_r)$ is strictly increasing and (affine) linear and such that, for $x_1 \in \mathbb{R}$ fixed, the function $u(x_1, \cdot, \ldots, \cdot)$ is strictly isotone w.r.t. the the componentwise partial order on $\mathbb{R}^{r-1}$. Then $\mathcal{U}_{sep} = \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$.*

# 8 APPLICATIONS

We now apply our framework on three examples: dermatological symptoms, credit approval data, and multidimensional poverty measurement. Results from the former two applications are presented in Supp. D, while Section 8.2 discusses results from poverty analysis. Before that, some details on the concrete implementation are given.

## 8.1 IMPLEMENTATION

To compute the test statistics for sample size $s$, we use a LP with constraints given by $C(x,y)$ (Section 5.3). The computation of the test statistics and the maximum regularization strength $\xi^*$, see Proposition 4 and 3, are LPs based on this same constraint matrix. The robustified statistics under $\gamma$-contamination are shifted versions of the original ones. Here, we utilize the linear connection between $\underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0)$ and $d^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0)$, $\overline{d}^\varepsilon_I$ and $d^\varepsilon_I$, respectively, for fixed $\epsilon$ (see Supp. C).

Although one only needs to compute the constraint matrix once, the worst-case complexity of the computation is $\mathcal{O}(s^4)$. In the implementation, we focused on the case of two ordinal variables and only one numerical variable, using the preference system (13). We exploit the fact that sorting the data set allows some comparisons to be skipped immediately by considering only the ordinal components. In particular, if the ordinal variables have a small number of categories compared to the sample size $s$, this can lead to a large proportion of comparisons being skipped. In the most cases, this reduces the computational cost of computing the constraint matrix compared to a naive implementation. Of course, in the worst case, the computation time cannot be drastically reduced in this way. For further details on the implementation, see Supp. B.

## 8.2 EXAMPLE: POVERTY ANALYSIS

At least since the capability approach by Sen [1985], there is mostly consensus that poverty has more facets than income or wealth. It is perceived as multidimensional concept, involving variables that are often ordinally scaled, e.g., level of education. One common task in poverty analysis is to compare subgroups like men and women. Stochastic dominance is a popular way of comparing such subpopulations, see e.g. Garcia-Gomez et al. [2019]. Excitingly, our approach allows us to extend this to multidimensional poverty measurement with any kind (of scales) of dimensions.

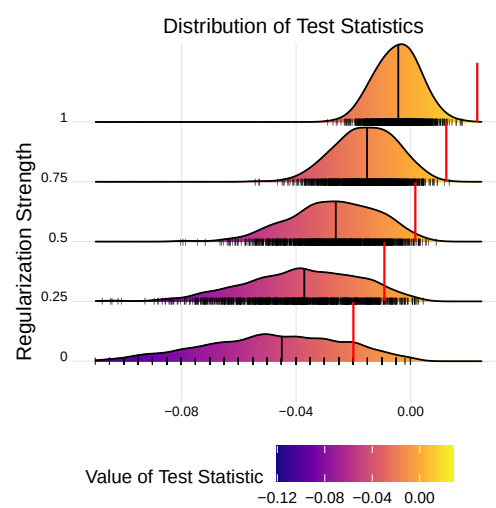


Figure 2: Distributions of $d^\varepsilon_I$ with $\varepsilon \in \{0, 0.25, 0.5, 0.75, 1\}$ obtained from $N = 1000$ resamples of ALLBUS data. Black stripes show exact positions of $d^\varepsilon_I$ values. Vertical black line marks median. Red line shows value of the respective observed test statistics $d^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega)$.

In the following, we will use data from the German General Social Survey (ALLBUS) GESIS [2018] that accounts for three dimensions of poverty: income (numeric), health (ordinal, 6 levels) and education (ordinal, 8 levels), see also Breyer and Danner [2015]. We are using the 2014 edition and focus on a subsample with $n = m = 100$ men and women each. We are interested in the hypothesis that women are dominated by men with respect to GSD – differently put, that women are poorer than men regarding any compatible utility representation of income, health and education.

As discussed in Section 5, we test the hypotheses (4), where

$X$ resp. $Y$ correspond to the subpopulation of men resp. women. We deploy our test with varying regularization strength $\varepsilon$. Figure 2 displays the distribution of the test statistics obtained trough $N = 1000$ resamples (cf. Section 5.3). It becomes evident that our proposed regularization serves its purpose: As $\varepsilon$ increases, the distribution of tests statistics becomes both more centered and closer to zero. Moreover, we reject for higher shares of the test statistics, see the position of $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega)$ (red line) compared to $d^{\varepsilon}_{I}$ (black stripes). For $\varepsilon \in \{0.5, 0.75, 1\}$ we reject for the common significance level of $\alpha \approx 0.05$.

As touched upon in Section 8.1, the robustified versions of the test statistic under the linear-vacuous model are shifted versions of the regular test statistics, i.e., they do not have to be computed explicitly. Exploiting this fact, we visualize the share of regularized test statistics for which we do not reject the null hypothesis (black stripes right of red line in Figure 2), depending on the contamination parameter $\gamma$ of the underlying linear-vacuous model, see Figure 3 (and Supp. C for details on computing the shares). It should be mentioned that these shares correspond to p-values telling at which significance levels $\alpha$ the test would be marginally rejected. Generally, it becomes apparent that even for small

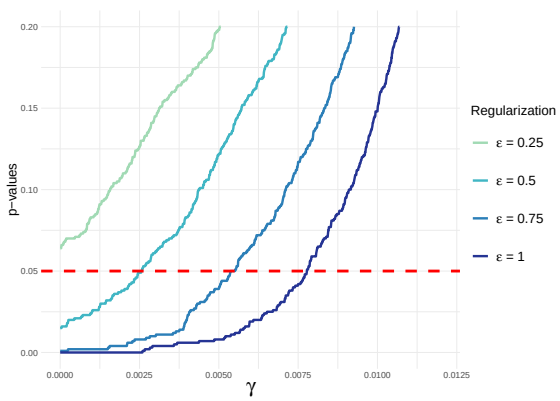


Figure 3: P-values as function of the contamination $\gamma$ (see Supp. C) for tests with different regularization strength $\varepsilon$. Dotted red line marks significance level $\alpha = 0.05$.

values of $\gamma$ the test statistics can be severely corrupted. If we allow more than $1\%$ ($\gamma > 0.01$) of the data (2 observations) to be redistributed in any manner, the shares of rejections drop drastically. Therefore, ignoring an (even very tiny) contamination $\gamma$ of the underlying distributions leads to a seriously inflated type I error. Remarkably, our regularization hedges against this to some extent: Given a significance level $\alpha = 0.05$, the fully regularized version (i.e., $\varepsilon = 1$) of our robustified test (cf., Section 6) comes to the same decision for $\gamma$ up to $0.075$. As explained in Section 5, rejecting $H_0$ does not necessarily mean that women are dominated by men; they could also be incomparable. However, our tests with reversed variables give no evidence of incomparability: all their observed p-values are above $0.95$.

**Further Applications:** We also analyzed a dermatology data set that contains variables on symptoms of the eryhemato-squamous disease, see Demiroz et al. [1998] accessed via Dua and Graff [2017], as well as the German credit data set that consists of variables on credit applicants, see Dua and Graff [2017]. In case of the credit data, we reject the hypothesis that high-risk applicants are dominated by low-risk applicants w.r.t. GSD for a common significance level of $\alpha \approx 0.05$. In the first application we are interested in the hypothesis that patients without a family history of the disease are dominated by patients without a family history with respect to GSD. We reject again for $\alpha \approx 0.05$. However, the p-values are much higher than in the other two applications. For detailed results as well as more information on the data sets, we refer to the supplement.

## 9 CONCLUDING REMARKS

**Summary:** We have further explored a generalized stochastic dominance (GSD) order among random variables with locally varying scale of measurement. We focused on four aspects: First, the investigation of (regularized) statistical tests for GSD when only samples of the variables are available. Second, robustifications of these tests w.r.t. their underlying assumptions using ideas from imprecise probabilities. Third, a detailed investigation of our ordering for preference systems arising from multidimensional structures with differently scaled dimensions. Finally, applications to examples from poverty measurement, finance, and medicine.

**Limitations and future research:** Two particular limitations offer promising opportunities for future research.

*Extending robust testing to belief function:* In Section 6, we have focused – for computational complexity – to linear-vacuous models. However, the idea of identifying least favorable extreme points seems to generalize to any credals sets induced by belief functions in the sense of Shafer [1976].

*Improving computational complexity:* The LPs for checking in-sample GSD become computer intensive for larger amounts of data. Although complexity reduces for the special case of preference systems discussed in Section 7 (cf. Section 8.1), Proposition 8 suggests that a further drastic reduction can be expected for only one cardinal dimension.


### Acknowledgements

We thank all reviewers for constructive comments that helped to improve the presentation of the paper. JR and TA gratefully acknowledge support by the Federal Statistical Office of Germany within the project "Machine Learning in Official Statistics". HB, JR, and GS gratefully acknowledge the financial and general support of the LMU Mentoring program. HB sincerely thanks Evangelisches Studienwerk e.V. for funding and supporting her PhD project.

## References


J. Abellan, C. Mantas, J. Castellano, and S. Moral-Garcia. Increasing diversity in random forest learning algorithm via imprecise probabilities. *Expert Syst Appl*, 97:228–243, 2018.

T. Augustin and G. Schollmeyer. Comment: On focusing, soft and strong revision of Choquet capacities and their role in statistics. *Stat Sci*, 36(2):205–209, 2021.

T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, editors. *Introduction to Imprecise Probabilities*. Wiley, 2014a.

T. Augustin, G. Walter, and F. Coolen. Statistical inference. In T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 135–189. Wiley, 2014b.

G. Barrett and S. Donald. Consistent tests for stochastic dominance. *Econometrica*, 71(1):71–104, 2003.

J. Bauer. Selection errors of random route samples. *Sociol Method Res*, 43(3):519–544, 2014.

J. Bauer. Biases in random route surveys. *Journal of Survey Statistics and Methodology*, 4(2):263–287, 2016.

V. Bengs, E. Hüllermeier, and W. Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Advances in Neural Information Processing Systems*, 2022.

H. Blocher, G. Schollmeyer, C. Jansen, and M. Nalenz. Depth functions for partial orders with a descriptive analysis of machine learning algorithms. In *International Symposium on Imprecise Probabilities: Theories and Applications*, 2023. PMLR (to appear).

B. Breyer and D. Danner. Skala zur Erfassung des Lebenssinns (ALLBUS). In *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS) (GESIS – Leibniz-Institut für Sozialwissenschaften)*, volume 10, 2015.

L. Brown, J. Hwang, and A. Munk. An unbiased test for the bioequivalence problem. *Ann Stat*, 25(6):2345 – 2367, 1997.

R. Cabanas, A. Antonucci, D. Huber, and M. Zaffalon. CREDICI: A Java library for causal inference by credal networks. In M. Jaeger and T. Nielsen, editors, *International Conference on Probabilistic Graphical Models*, volume 138 of *PMLR*, pages 597–600, 2020.

Y. Carranza and S. Destercke. Imprecise Gaussian discriminant classification. *Pattern Recogn*, 112:107739, 2021.

C. Chang, J. Jimenez-Martin, E. Maasoumi, and T. Perez-Amaral. A stochastic dominance approach to financial risk management strategies. *J Econometrics*, 187:472–485, 2015.

D. Chetverikov, D. Wilhelm, and D. Kim. An adaptive test of stochastic monotonicity. *Econometric Theory*, 37(3): 495–536, 2021.

G. Corani and M. Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *J Mach Learn Res*, 9(4), 2008.

G. Corani, A. Benavoli, J. Demsar, F. Mangili, and M. Zaffalon. Statistical comparison of classifiers through Bayesian hierarchical modelling. *Mach Learn*, 106(11): 1817–1837, 2017.

H. Dai, Y. Xue, N. He, Y. Wang, N. Li, D. Schuurmans, and B. Dai. Learning to optimize for stochastic dominance constraints. In F. Ruiz, J. Dy, and J. van de Meent, editors, *Artificial Intelligence and Statistics*, volume 206 of *PMLR*, pages 8991–9009, 2023.

G. Demiroz, H. Govenir, and N. Ilter. Learning differential diagnosis of Eryhemato-Squamous diseases using voting feature intervals. *Artif Intell Med*, 13:147–165, 1998.

J. Demsar. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res*, 7:1–30, 2006.

D. Denneberg. *Non-additive Measure and Integral*. Kluwer Academic Publishers, 1994.

S. Destercke, I. Montes, and E. Miranda. Processing distortion models: A comparative study. *Int J Approx Reason*, 145:91–120, 2022.

D. Dua and C. Graff. UCI machine learning repository, 2017. `http://archive.ics.uci.edu/ml`.

European Commission. Knowledge service: Competence centre on composite indicators and scoreboards, 2023. URL `https://knowledge4policy.ec.europa.eu/composite-indicators_en`. (Febr. 16, 2023).

C. Garcia-Gomez, A. Perez, and M. Prieto-Alaiz. A review of stochastic dominance methods for poverty analysis. *J Econ Surv*, 33(5):1437–1462, 2019.

GESIS. Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2014. GESIS Datenarchiv, Köln. ZA5240 Datenfile Version 2.2.0, https://doi.org/10.4232/1.13141, 2018.

E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach Learn*, 110(3):457–506, 2021.

E. Hüllermeier, S. Destercke, and M. Shaker. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In J. Cussens and K. Zhang, editors, *Uncertainty in Artificial Intelligence*, volume 180 of *PMLR*, pages 548–557, 2022.

C. Jansen and T. Augustin. Decision making with state-dependent preference systems. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 729–742. Springer, 2022.

C. Jansen, G. Schollmeyer, and T. Augustin. Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences. *Int J Approx Reason*, 98:112–131, 2018.

C. Jansen, H. Blocher, T. Augustin, and G. Schollmeyer. Information efficient learning of complexly structured preferences: Elicitation procedures and their application to decision making under uncertainty. *Int J Approx Reason*, 144:69–91, 2022a.

C. Jansen, M. Nalenz, G. Schollmeyer, and T. Augustin. Statistical comparisons of classifiers by generalized stochastic dominance, 2022b. URL `https://arxiv.org/abs/2209.01857`. arXiv preprint.

C. Jansen, G. Schollmeyer, and T. Augustin. Quantifying degrees of E-admissibility in decision making with imprecise probabilities. In T. Augustin, F. Cozman, and G. Wheeler, editors, *Reflections on the Foundations of Probability and Statistics: Essays in Honor of Teddy Seidenfeld*, pages 319–346. Springer, 2022c.

C. Jansen, G. Schollmeyer, and T. Augustin. Multi-target decision making under conditions of severe uncertainty. In V. Torra and Y. Narukawa, editors, *Modeling Decisions for Artificial Intelligence*, pages 45–57. Springer, 2023.

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

J. Lienen and E. Hüllermeier. Credal self-supervised learning. *Advances in Neural Information Processing Systems*, 34:14370–14382, 2021.

R. Luce. Semiorders and a theory of utility discrimination. *Econometrica*, 24:178–191, 1956.

A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. *Advances in Neural Information Processing Systems*, 31, 2018.

D. Maua and F. Cozman. Thirty years of credal networks: Specification, algorithms and complexity. *Int J Approx Reason*, 126:133–157, 2020.

D. Maua and C. de Campos. Editorial to: Special issue on robustness in probabilistic graphical models. *Int J Approx Reason*, 137:113, 2021.

D. McFadden. Testing for stochastic dominance. In T. Fomby and T. Seo, editors, *Studies in the Economics of Uncertainty*, pages 113–134. Springer, 1989.

I. Montes, E. Miranda, and S. Destercke. Unifying neighbourhood and distortion models: Part II – new models and synthesis. *Int J Gen Syst*, 49:636–674, 2020.

K. Mosler. Testing whether two distributions are stochastically ordered or not. In H. Rinne, B. Rüger, and H. Strecker, editors, *Grundlagen der Statistik und ihre Anwendungen: Festschrift für Kurt Weichselberger*, pages 149–155. Physica-Verlag, 1995.

K. Mosler and M. Scarsini. Some theory of stochastic dominance. *Lecture Notes-Monograph Series*, 19:261–284, 1991.

L. Oneto, S. Ridella, and D. Anguita. Tikhonov, Ivanov and Morozov regularization for support vector machine learning. *Mach Learn*, 103:103–136, 2016.

M. Pivato. Multiutility representations for incomplete difference preorders. *Math Sco Sci*, 66:196–220, 2013.

J. Pratt and J. Gibbons. *Concepts of Nonparametric Theory*. Springer, 2012.

T. Range and L. Østerdal. First-order dominance: stronger characterization and a bivariate checking algorithm. *Math Program*, 173:193—219, 2019.

J. Rodemann and T. Augustin. Accounting for Gaussian process imprecision in Bayesian optimization. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM)*, pages 92–104. Springer, 2022.

J. Rodemann, C. Jansen, G. Schollmeyer, and T. Augustin. In all likelihoods: Robust selection of pseudo-labeled data. In *International Symposium on Imprecise Probabilities: Theories and Applications*, 2023. PMLR (to appear).

G. Schollmeyer. A short note on the equivalence of the ontic and the epistemic view on data imprecision for the case of stochastic dominance for interval-valued data. In *International Symposium on Imprecise Probabilities: Theories and Applications*, pages 330–337. PMLR, 2019.

G. Schollmeyer, C. Jansen, and T. Augustin. Detecting stochastic dominance for poset-valued random variables as an example of linear programming on closure systems, 2017. URL `https://epub.ub.uni-muenchen.de/40416/13/TR_209.pdf`. Technical Report 209, Department of Statistics, LMU Munich.

A. Sen. *Commodities and Capabilities*. Elsevier, 1985.

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

M. Shaked and G. Shanthikumar. *Stochastic orders*. Springer, 2007.

M. Shaker and E. Hüllermeier. Ensemble-based uncertainty quantification: Bayesian versus credal inference, 2021. URL https://arxiv.org/abs/2107.10384. arXiv preprint.

M. Timonin. Maximization of the Choquet integral over a convex set and its application to resource allocation problems. *Ann Oper Res*, 196:543–579, 2012.

UNECE. Guidelines on producing leading, composite and sentiment indicators, 2019. URL https://unece.org/DAM/stats/publications/2019/ECECESSTAT20192.pdf. (Febr. 16, 2023).

L. Utkin. An imprecise deep forest for classification. *Expert Syst Appl*, 141:112978, 2020.

L. Utkin and A. Konstantinov. Attention-based random forest and contamination model. *Neural Networks*, 154: 346–359, 2022.

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