# 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 such generalized stochastic dominance, 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], Eugster et al. [2012], or Corani et al. [2017]) 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 $\big\{X \in A^\Omega : u \circ X \in \mathcal{L}^1(\Omega, \mathcal{S}_1, \pi)\big\}$ 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 (and *only* if) we were confronted with a problem under pure *aleatoric* uncertainty where a probability $\pi$ and a cardinal scale $u$ *were* available.

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.[1] 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 *pre-order* $\succsim_{(\mathcal{U},\pi)}$ on

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

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 prominent special cases: If $A$ is equipped with a pre-order, and $\mathcal{U}$ is the set of all functions that are bounded and isotone w.r.t. this pre-order, then $\succsim_{(\mathcal{U},\pi)}$ is (essentially) equivalent to (first-order) stochastic dominance. In contrast, if

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

$(A, \mathcal{S}_2) = (\mathbb{R}, \mathcal{B}_\mathbb{R})$ 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.

**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 pre-orders $\succsim_{(\mathcal{U},\pi)}$ 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 Observations 1 and 2, Propositions 1 to 6, and Corollary 1 can be found in the supplementary material (cf., Supp. A). Our code is available under: `https://anonymous.4open.science/r/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]. 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]. 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. [2022a].

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], 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 *pre-order* is a binary relation that is reflexive and transitive; a *linear order* is a preference relation that is antisymmetric; a *partial order* is a pre-order that is antisymmetric. If $R$ is a pre-order, 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].[2]

**Definition 1** *Let $A \neq \emptyset$ be a set, $R_1 \subseteq A \times A$ a pre-order on $A$, and $R_2 \subseteq R_1 \times R_1$ a pre-order 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* ***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.

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 $(a,b) \in I_{R_1}$.*

[2] For a study on representation results of the related concept of *incomplete difference pre-orders* see, e.g., Pivato [2013].

*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*

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

*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.

**Observation 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 principle, two different ways of 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.

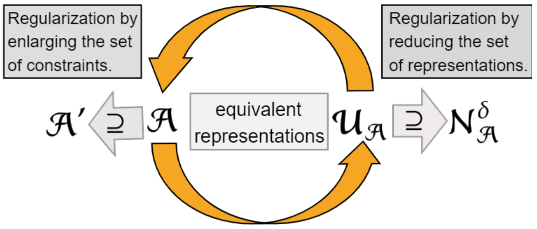
Regularization by enlarging the set of constraints.
Regularization by reducing the set of representations.
$\mathcal{A}'$
$\supseteq$
$\mathcal{A}$
equivalent representations
$\mathcal{U}_\mathcal{A}$
$\supseteq$
$N^\delta_\mathcal{A}$

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 concrete interpretation of the parameter is only 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 pre-order, 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 pre-order 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.

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

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

# 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 $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_m)$ 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:

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

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], or generally in the context of bio-equivalence 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.[3] 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:

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

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 Observation 2 – it holds $(X,Y) \in R_{(\mathcal{A},\pi)}$ if and only if

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

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, $\hat{\pi}_X^\omega$ and $\hat{\pi}_Y^\omega$ the observed empirical image measures of $X$ and $Y$,

$$(\mathbf{XY})_\omega = \{X_i(\omega) : i \leq n\} \cup \{Y_i(\omega) : i \leq n\} \cup \{a_*, a^*\},$$

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 *in-sample GSD* of $X$ over $Y$ in the sample induced by $\omega_0$.

As a further test statistic, we consider a regularized version 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 pre-specified threshold value. In this way, the regularized test statistic is also sensitive for distinguishing situations under dominance regarding their *extent* of dominance.[4] Formally, the regularized test statistic looks as follows:

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

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

with $\varepsilon \in [0,1]$ and $\delta_\varepsilon(\omega) := \varepsilon \cdot \sup\{\xi : \mathcal{N}_{\mathcal{A}_\omega}^{\xi} \neq \emptyset\}$. Observe that $d_{\mathbf{X},\mathbf{Y}} = d_{\mathbf{X},\mathbf{Y}}^0$, 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 $d_{\mathbf{X},\mathbf{Y}}^{\varepsilon}$ 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

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

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

[3] 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].

[4] While in-sample GSD (essentially) implies $d_{\mathbf{X},\mathbf{Y}}(\omega_0) = 0$, it often holds $d_{\mathbf{X},\mathbf{Y}}^{\varepsilon}(\omega_0) > 0$. 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 $d_{\mathbf{X},\mathbf{Y}}^{\varepsilon}(\omega_0) > d_{\mathbf{V},\mathbf{W}}^{\varepsilon}(\omega_0) > 0$ and, hence, that under regularization $X$ (empirically) dominates $Y$ more strongly than $V$ dominates $W$.

**Step 1:** Take the pooled data sample:

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

**Step 2:** Take all $I \subseteq \{1, \dots, n+m\}$ of size $n$. Evaluate $d_{\mathbf{X},\mathbf{Y}}$ resp. $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$ for $(w_i)_{i \in I}$ and $(w_i)_{i \in \{1,\dots,n+m\} \setminus I}$ 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.

**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 the $\lceil (1-\alpha) \cdot \binom{n+m}{n} \rceil$-th value of the increasingly ordered values $d_I$ resp. $d^{\varepsilon}_I$, where $\alpha$ is the significance level.

For large $\binom{n+m}{n}$, approximate the above by computing $d_I$ resp. $d^{\varepsilon}_I$ only for a large number $N$ of randomly drawn $I$.

### 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

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

Further, we denote by $C(\mathbf{x},\mathbf{y})$ the set of all vectors $(v_1, \dots, v_s, \xi) \in [0,1]^{s+1}$ 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 $\{(v_1, \dots, v_s) \in [0,1]^s : (v_1, \dots, v_s, \xi_0) \in C(\mathbf{x},\mathbf{y})\}$, 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 $(v_1, \dots, v_s, \xi)$ resp. $(v_1, \dots, v_s)$. This allows to formulate Propositions 1 and 2. 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 1** *For samples $\mathbf{x}$ and $\mathbf{y}$ of the form (6) and (7) and $\varepsilon \in [0,1]$, we consider the linear program (LP)*

$$\xi \longrightarrow \max_{(v_1, \dots, v_s, \xi)} \tag{8}$$

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

**Proposition 2** *For samples $\mathbf{x}$ and $\mathbf{y}$ of the form (6) and (7) and $\varepsilon \in [0,1]$, we consider the following LP*

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

*with $(v_1, \dots, v_s) \in C_{\varepsilon \xi^*}(\mathbf{x},\mathbf{y})$, 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. 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 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 problem, 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], see also the corresponding literature referenced as related work in Section 1.

The concrete idea behind our robustification is that we allow our samples (potentially) not to come from the populations of actual interest, but instead from some biased populations. We assume that these biased populations are similar to the true ones in the sense that they are contained in the credal sets $\mathcal{M}_X$ and $\mathcal{M}_Y$, respectively. We start by only assuming both $\mathcal{M}_X$ and $\mathcal{M}_Y$ to be convex polyhedra with extreme points collected in the finite sets $\mathcal{E}(\mathcal{M}_X)$ and $\mathcal{E}(\mathcal{M}_Y)$. Most naturally, the test statistics $d_{\mathbf{X},\mathbf{Y}}$ and $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$ can then be replaced by their *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

$$\omega \quad \mapsto \quad \inf_{(\pi_1, \pi_2, u) \in \mathcal{D}} \sum_{z \in (\mathbf{XY})_\omega} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\}))$$

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

with $\mathcal{D} = \mathcal{M}_X^\omega \times \mathcal{M}_Y^\omega \times \mathcal{N}_{\mathcal{A}_\omega}$, $\mathcal{D}_\varepsilon = \mathcal{M}_X^\omega \times \mathcal{M}_Y^\omega \times \mathcal{N}_{\mathcal{A}_\omega}^{\delta_\varepsilon(\omega)}$ and $\mathcal{M}_X^\omega$ resp. $\mathcal{M}_Y^\omega$ the empirical credal sets given $\omega \in \Omega$.

Now, $\underline{d}_{\mathbf{X},\mathbf{Y}}$ (or $\underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}$) can be used to test the same $H_0$ as before, however, under the additional difficulty that the samples are drawn from a biased population of which we only know it is contained in some neighborhood around the true population. To appropriately adapt the resampling scheme, one must perform the test under all laws within the corresponding credal sets. Since this is computationally cumbersome, we instead compare the obtained lower envelope $\underline{d}_{\mathbf{X},\mathbf{Y}}$ (or $\underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}$) with the distribution (in the resamples) of the corresponding upper envelope, $\overline{d}_{\mathbf{X},\mathbf{Y}}$ (or $\overline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}$) that is obtained by replacing the part of $\inf$ concerning $\mathcal{M}_X^\omega \times \mathcal{M}_Y^\omega$ with the respective $\sup$ in the above definitions. This gives a conservative yet valid statistical test.[5]

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

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

*with* $(v_1, \dots, v_s) \in C_{\varepsilon\xi^*}(\mathbf{x},\mathbf{y})$ *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* $(\pi_1,\pi_2) \in \mathcal{E}(\mathcal{M}_X^{\omega_0}) \times \mathcal{E}(\mathcal{M}_Y^{\omega_0})$*. 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* $\hat{\pi}_X^{\omega_0} \in \mathcal{M}_X^{\omega_0}$ *and* $\hat{\pi}_Y^{\omega_0} \in \mathcal{M}_Y^{\omega_0}$ *if* $\underline{opt}_0(\mathbf{x},\mathbf{y}) \geq 0$.

Proposition 3 requires to solve $|\mathcal{E}(\mathcal{M}_X^{\omega_0})| \cdot |\mathcal{E}(\mathcal{M}_Y^{\omega_0})|$ 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

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

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

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

where $\delta_z$ denotes the Dirac-measure in $z$ (see again Walley [1991, p. 147]). Proposition 4 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 4** *Consider again the situation of Proposition 3, 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.[6] 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[7]

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

where

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

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

[5] Actually, we must assume the true empirical laws to lie in a neighborhood of the biased empirical laws almost surely (or with arbitrarily high probability) to get an approximate test.

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

[7] One easily verifies that $R_1^*$ and $R_2^*$ are pre-orders.

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

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

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

*iii) If $(X, Y) \in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ 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 $(X, Y) \in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$.*

Part iii) of Proposition 5 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 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 5, 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. 5 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 6** *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 2 and 1, are LPs based on this same constraint matrix. The robustified statistics under $\gamma$-contamination are shifted versions of the original ones (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). Note that a small number of ordinal variables with a small number of categories, compared to the sample size $s$, already leads to many incomparable observations. This can be used to reduce the computation time of the constraint matrix. 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.

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

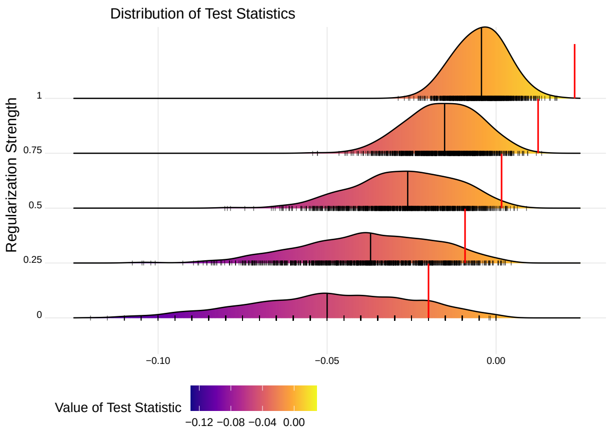

Figure 2: Distributions of $d_I^\varepsilon$ with $\varepsilon \in \{0, 0.25, 0.5, 0.75, 1\}$ obtained from $N = 1000$ resamples of ALLBUS data. Black stripes show exact positions of $d_I^\varepsilon$ values. Vertical black line marks median. Red line shows value of the respective observed test statistics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

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 $d_I^\varepsilon$ (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 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 $0.95$.

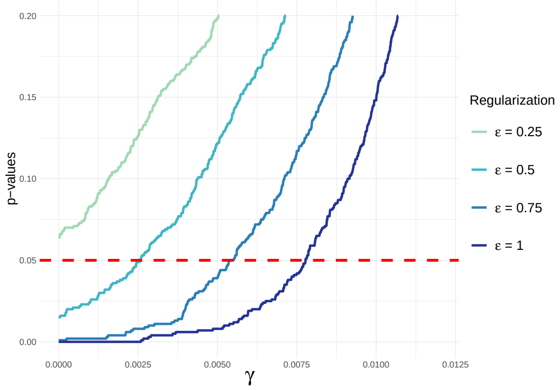

Figure 3: P-values as function of the contamination $\gamma$ (see Supp. C) for tests with different regularization strength $\varepsilon$. Dotted red line marks significance level $\alpha = 0.05$.

## 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 6 suggests that a further drastic reduction can be expected for only one cardinal dimension.

## Acknowledgements

Thomas Augustin gratefully acknowledges support by the Federal Statistical Office of Germany within the cooperation project "Machine Learning in Official Statistics".

## References

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

T. Augustin and G. Schollmeyer. Comment: On focusing, soft and strong revision of Choquet capacities and their role in statistics. *Statistical Science*, 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. *Sociological Methods & Research*, 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.

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. *The Annals of Statistics*, 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, *Proceedings of the 10th International Conference on Probabilistic Graphical Models*, volume 138 of *Proceedings of Machine Learning Research*, pages 597–600. PMLR, 2020.

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

C. Chang, J.-A. Jimenez-Martin, E. Maasoumi, and T. Perez-Amaral. A stochastic dominance approach to financial risk management strategies. *Journal of Econometrics*, 187(2):472–485, 2015.

G. Corani and M. Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *Journal of Machine Learning Research*, 9(4), 2008.

G. Corani, A. Benavoli, J. Demsar, F. Mangili, and M. Zaffalon. Statistical comparison of classifiers through Bayesian hierarchical modelling. *Machine Learning*, 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 *Artificial Intelligence and Statistics*, 2023. to appear.

J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 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. *International Journal of Approximate Reasoning*, 145:91–120, 2022.

M. Eugster, T. Hothorn, and F. Leisch. Domain-based benchmark experiments: Exploratory and inferential analysis. *Austrian Journal of Statistics*, 41(1):5–26, 2012.

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. *Journal of Economic Surveys*, 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. *Machine Learning*, 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, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume

180 of *Proceedings of Machine Learning Research*, pages 548–557. PMLR, 2022.

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

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

C. Jansen, G. Schollmeyer, and T. Augustin. Multi-target decision making under conditions of severe uncertainty, 2022b. URL `https://arxiv.org/abs/2212.06832`. arXiv preprint.

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. *International Journal of Approximate Reasoning*, 126:133–157, 2020.

D. Maua and C. de Campos. Editorial to: Special issue on robustness in probabilistic graphical models. *International Journal of Approximate Reasoning*, 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. *International Journal of General Systems*, 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.

M. Pivato. Multiutility representations for incomplete difference preorders. *Mathematical Social Sciences*, 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. *Mathematical Programming*, 173:193—219, 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. 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. *Annals of Operations Research*, 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 Systems with Applications*, 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.

# Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement (Supplementary Material)

**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

In the following, we give supplementary information and material to the main paper. This includes all mathematical proofs of the propositions, corollaries, and observations established in the main paper (Part A), further details on the implementation and reproducibility (Part B), further calculations for the robustified test statistics (Part C), and further analyses of the applications in the main paper (Part D). If not explicitly stated otherwise, from now on, all references to equations, propositions, etc. refer to the main part of the paper.

## A PROOFS OF THE RESULTS IN THE MAIN PAPER

### A.1 PROOFS FOR OBSERVATIONS 1 AND 2: BOUNDED PREFERENCE SYSTEMS

We start by proving Observations 1 resp. 2 from Sections 2 resp. 4 that state that checking consistency resp. GSD simplifies if the underlying preference system is bounded.

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

**Proof.** If $\mathcal{A}$ is 0-consistent, then it is obviously also consistent, since every normalized representation is in particular a representation. For the other direction, assume $\mathcal{A}$ to be consistent. Choose $u \in \mathcal{U}_{\mathcal{A}}$ arbitrarily and denote by $a_*, a^*$ the $R_1$-minimal resp. $R_1$-maximal elements satisfying $(a^*, a_*) \in P_{R_1}$. From the latter we know that $u(a^*) > u(a_*)$. Thus, the function

$$\tilde{u} : A \to [0,1] \ , \ \ a \mapsto \frac{u(a) - u(a_*)}{u(a^*) - u(a_*)}$$

is well-defined. Moreover, one easily verifies that $\tilde{u} \in \mathcal{U}_{\mathcal{A}}$, and $u(a_*) = 0$, and $u(a^*) = 1$. Thus, we can conclude that $\tilde{u} \in \mathcal{N}_{\mathcal{A}}$, which – by definition – implies 0-consistency. □

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

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

**Proof.** The direction $\Rightarrow$ follows trivially by observing $\mathcal{N}_{\mathcal{A}} \subseteq \mathcal{U}_{\mathcal{A}}$. For the direction $\Leftarrow$, assume that it holds $\forall u \in \mathcal{N}_{\mathcal{A}}$ : $\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$. Choose $u \in \mathcal{U}_{\mathcal{A}}$ arbitrarily. With the same argument as given in the proof of Observation 1, we know that then $\tilde{u} \in \mathcal{N}_{\mathcal{A}}$, where $\tilde{u}$ is defined as in the proof of Observation 1. Since $\tilde{u}$ is a positive (affine) linear transformation of $u$, we know that $\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$ if and only if $\mathbb{E}_\pi(\tilde{u} \circ X) \geq \mathbb{E}_\pi(\tilde{u} \circ Y)$. Since the latter is true by assumption (utilizing $\tilde{u} \in \mathcal{N}_{\mathcal{A}}$), the first also is true. As $u$ was chosen arbitrarily, this completes the proof. □

### A.2 PROOFS OF PROPOSITIONS 1 AND 2: COMPUTATIONS FOR THE PERMUTATION TEST

We now give proofs for Propositions 1 resp. 2 from Section 5.3 that concern the computation of the maximum regularization strength resp. the computation of the (regularized) test statistic for the permutation-test.

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

$$\xi \longrightarrow \max_{(v_1,\dots,v_s,\xi)}$$

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

**Proof.** The Proposition follows from standard results on linear optimization and the fact that $C(\mathbf{x},\mathbf{y})$ is compact. Set $I := \big\{\ell : (z_\ell, a^*) \in I_{R_1}\big\}$ and define the vector $\underline{v} := (0,1,v_3,\dots,v_s,0) \in [0,1]^{s+1}$ by $v_\ell = 1$ if $\ell \in I$ and $v_\ell = 0$ otherwise. One then easily verifies that $\underline{v}$ is an admissible solution to the above linear program. Since $C(\mathbf{x},\mathbf{y})$ is compact, this implies the existence of an optimal solution. Denote thus by $\underline{v}^* := (0,1,v_3^*,\dots,v_s^*,\xi^*)$ an arbitrary optimal solution. We have to show that

$$\xi^* = \sup\big\{\xi : \mathcal{N}^{\xi}_{\mathcal{A}_{\omega_0}} \neq \emptyset\big\} =: c.$$

Assume, for contradiction, the above equality does not hold. We distinguish two cases:

*Case 1:* $\xi^* < c$. Then, one easily verifies that for any function $u \in \mathcal{N}^{c}_{\mathcal{A}_{\omega_0}}$ the vector $(u(z_1),\dots,u(z_s),c)$ defines an admissible solution to the above linear program with an objective value of $c$. This constradicts the optimality of $\underline{v}^*$.

*Case 2:* $\xi^* > c$. Then, setting $u : (\mathbf{XY})_{\omega_0} \to [0,1]$ with $u(z_\ell) := v_\ell^*$ defines an element of $\mathcal{N}^{\xi^*}_{\mathcal{A}_{\omega_0}}$, contradicting that $c$ is the largest number for which $\mathcal{A}_{\omega_0}$ is $c$-consistent.

Thus, we have that $c = \xi^*$, implying $\delta_\varepsilon(\omega_0) = \varepsilon \cdot \xi^*$. □

**Proposition 2** *For samples $\mathbf{x}$ and $\mathbf{y}$ of the form (6) and (7) and $\varepsilon \in [0,1]$, we consider the following linear program*

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

*with constraints $(v_1,\dots,v_s) \in C_{\varepsilon\xi^*}(\mathbf{x},\mathbf{y})$, where $\xi^*$ denotes the optimal value of (8). Denote by $opt_\varepsilon(\mathbf{x},\mathbf{y})$ its optimal value. It then holds:*

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

*ii) There is in-sample GSD of $X$ over $Y$ if and only if $opt_0(\mathbf{x},\mathbf{y}) \geq 0$.*

**Proof.** i) By definition and Proposition 2 , we know that $\mathcal{N}^{\xi^*}_{\mathcal{A}_{\omega_0}} \neq \emptyset$. As these sets are nested with decreasing $\xi$-value and we have $\varepsilon\xi^* \leq \xi^*$, this implies that also $\mathcal{N}^{\varepsilon\xi^*}_{\mathcal{A}_{\omega_0}} \neq \emptyset$. Hence, we can choose $u \in \mathcal{N}^{\varepsilon\xi^*}_{\mathcal{A}_{\omega_0}}$. One then easily verifies that the vector $(u(z_1),\dots,u(z_s))$ defines an admissible solution to the above linear program. Since $C_{\varepsilon\xi^*}(\mathbf{x},\mathbf{y})$ is compact, this implies the existence of an optimal solution. Thus, denote by $\underline{v}^* := (v_1^*,\dots,v_s^*)$ an arbitrary such optimal solution. If we then define $u : (\mathbf{XY})_{\omega_0} \to [0,1]$ with $u(z_\ell) := v_\ell^*$, then one easily verifies that $u \in \mathcal{N}^{\varepsilon\xi^*}_{\mathcal{A}_{\omega_0}}$ and that

$$opt_\varepsilon(\mathbf{x},\mathbf{y}) = \sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot\left(\hat{\pi}_X^{\omega_0}(\{z\}) - \hat{\pi}_Y^{\omega_0}(\{z\})\right) \tag{1}$$

(to see this, note that the right side of the equation is a simple reformulation of the objective function with $\underline{v}^*$ plugged-in).

We have to show that

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

Assume, for contradiction, the above equality does not hold. We distinguish two cases:

*Case 1: $opt_\varepsilon(\mathbf{x},\mathbf{y}) > d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0)$.* This would imply that there exists an $u' \in \mathcal{N}^{\varepsilon\xi^*}_{\mathcal{A}_{\omega_0}}$ that – if it was set in the right-hand side of the above Equation (1) (in the supplementary material) instead of $u$ – would produce a value strictly smaller than $opt_\varepsilon(\mathbf{x},\mathbf{y})$. This contradicts the optimality of $\underline{v}^*$, since every $u' \in \mathcal{N}^{\varepsilon\xi^*}_{\mathcal{A}_{\omega_0}}$ produces an admissible solution to the linear program with objective value given by the right-hand side of the above Equation (1).

*Case 2: $opt_\varepsilon(\mathbf{x},\mathbf{y}) < d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0)$.* This would be an immediate contradiction to the above Equation (1) (in the supplementary material), since $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0)$ is by definition the infimum over all the expressions on the equation's right-hand side.

This completes the proof of i). To see ii), note that i) implies $opt_0(\mathbf{x},\mathbf{y}) = d^{0}_{\mathbf{X},\mathbf{Y}}(\omega_0)$. Thus, we have $opt_0(\mathbf{x},\mathbf{y}) \geq 0$ if and only if $d^{0}_{\mathbf{X},\mathbf{Y}}(\omega_0) \geq 0$, which – by definition – is true if and only if there is in-sample GSD of $X$ over $Y$. □

### A.3 PROOFS OF PROPOSITION 3 AND 4: COMPUTATIONS FOR ROBUSTIFIED TESTING

We now give proofs of Proposition 3 resp. 4 from Section 6 concerning the computation of the robustified test statistic resp. its simplification under the special case of a $\gamma$-contamination model (with $\gamma \in [0,1]$).

**Proposition 3** *For samples* $\mathbf{x}$ *and* $\mathbf{y}$ *of the form (6) and (7),* $\varepsilon \in [0,1]$, *and* $(\pi_1, \pi_2) \in \mathcal{E}(\mathcal{M}_X^{\omega_0}) \times \mathcal{E}(\mathcal{M}_Y^{\omega_0})$, *we consider the following linear program:*

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

*with constraints* $(v_1, \dots, v_s) \in C_{\varepsilon\xi^*}(\mathbf{x}, \mathbf{y})$, *where* $\xi^*$ *denotes the optimal value of (8). Denote by* $opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2)$ *its optimal value and by* $\underline{opt}_\varepsilon(\mathbf{x}, \mathbf{y})$ *the minimal optimum over all combinations of* $(\pi_1, \pi_2) \in \mathcal{E}(\mathcal{M}_X^{\omega_0}) \times \mathcal{E}(\mathcal{M}_Y^{\omega_0})$. *It then holds:*

*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* $\hat{\pi}_X^{\omega_0} \in \mathcal{M}_X^{\omega_0}$ *and* $\hat{\pi}_Y^{\omega_0} \in \mathcal{M}_Y^{\omega_0}$ *if* $\underline{opt}_0(\mathbf{x}, \mathbf{y}) \geq 0$.

**Proof.** i) Since nothing in the proof of Proposition 2 hinges on the concrete structure of the involved empirical image measures, Proposition 2 is still valid if we replace $\hat{\pi}_X^{\omega_0}$ and $\hat{\pi}_Y^{\omega_0}$ by arbitrary $\pi_1 \in \mathcal{M}_X^{\omega_0}$ and $\pi_2 \in \mathcal{M}_Y^{\omega_0}$, respectively. This specifically implies

$$opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2) = \inf_{u \in \mathcal{N}_{\mathcal{A}_{\omega_0}}^{\delta_\varepsilon(\omega_0)}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})). \tag{2}$$

In order to show i), we now need to verify that

$$\inf_{(\pi_1,\pi_2) \in \mathcal{E}(\mathcal{M}_X^{\omega_0}) \times \mathcal{E}(\mathcal{M}_Y^{\omega_0})} opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2) = \underline{d}^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0).$$

Due to the above Equation (2) (in the supplementary material) and the fact that iterated infima can be equivalently replaced by one global infimum, we know that

$$\inf_{(\pi_1,\pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}} opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2) = \underline{d}^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0). \tag{3}$$

We then can compute:

$$\begin{aligned}
\underline{d}^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) &\overset{(3)}{=} \inf_{(\pi_1,\pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}} opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2) \\
&= \inf_{(\pi_1,\pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}} \inf_{u \in \mathcal{N}_{\mathcal{A}_\omega}^{\delta_\varepsilon(\omega)}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \\
&= \inf_{u \in \mathcal{N}_{\mathcal{A}_\omega}^{\delta_\varepsilon(\omega)}} \inf_{(\pi_1,\pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \\
&\overset{(\star)}{=} \inf_{u \in \mathcal{N}_{\mathcal{A}_\omega}^{\delta_\varepsilon(\omega)}} \left( \inf_{\pi_1 \in \mathcal{M}_X^{\omega_0}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \pi_1(\{z\}) - \sup_{\pi_2 \in \mathcal{M}_Y^{\omega_0}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \pi_2(\{z\}) \right) \\
&\overset{(\star\star)}{=} \inf_{u \in \mathcal{N}_{\mathcal{A}_\omega}^{\delta_\varepsilon(\omega)}} \left( \inf_{\pi_1 \in \mathcal{E}(\mathcal{M}_X^{\omega_0})} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \pi_1(\{z\}) - \sup_{\pi_2 \in \mathcal{E}(\mathcal{M}_Y^{\omega_0})} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \pi_2(\{z\}) \right) \\
&= \inf_{(\pi_1,\pi_2) \in \mathcal{E}(\mathcal{M}_X^{\omega_0}) \times \mathcal{E}(\mathcal{M}_Y^{\omega_0})} \inf_{u \in \mathcal{N}_{\mathcal{A}_\omega}^{\delta_\varepsilon(\omega)}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \\
&= \inf_{(\pi_1,\pi_2) \in \mathcal{E}(\mathcal{M}_X^{\omega_0}) \times \mathcal{E}(\mathcal{M}_Y^{\omega_0})} opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2)
\end{aligned}$$

Here, $(\star)$ follows since – for $u$ fixed – the infimum of the differences of the two sums is attained if the first sum is smallest possible and the second sum is largest possible (note that all sums involved are finite). Further, $(\star\star)$ follows since – again for $u$ fixed – the sums are linear functions on the compact sets $\mathcal{M}_X^{\omega_0}$ resp. $\mathcal{M}_Y^{\omega_0}$ and, therefore, attain their optima on $\mathcal{E}(\mathcal{M}_X^{\omega_0})$ resp. $\mathcal{E}(\mathcal{M}_Y^{\omega_0})$. The fith and sixth equalities are just reversing the computation done in the first three equalities.

To see ii), note that i) implies $\underline{opt}_0(\mathbf{x}, \mathbf{y}) = \underline{d}^0_{\mathbf{X},\mathbf{Y}}(\omega_0)$. Thus, $\underline{opt}_0(\mathbf{x}, \mathbf{y}) \geq 0$ if and only if $\underline{d}^0_{\mathbf{X},\mathbf{Y}}(\omega_0) \geq 0$. But – by definition – the latter is true if and only if

$$\inf_{u \in \mathcal{N}^0_{\mathcal{A}_{\omega_0}}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \geq 0$$

for all $(\pi_1, \pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}$. This obviously implies in-sample GSD of $X$ over $Y$ for any $\pi$ with $\hat{\pi}_X^{\omega_0} \in \mathcal{M}_X^{\omega_0}$ and $\hat{\pi}_Y^{\omega_0} \in \mathcal{M}_Y^{\omega_0}$, since $\mathcal{N}^0_{\mathcal{A}_{\omega_0}} = \mathcal{N}_{\mathcal{A}_{\omega_0}}$. □

**Proposition 4** *Consider again the situation of Proposition 3 with the additional assumption that $\mathcal{M}_X^{\omega_0}$ and $\mathcal{M}_Y^{\omega_0}$ are of the form (11) with extreme points as in (12). It then holds:*

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

*where*

$$\pi_* = \gamma \delta_{a_*} + (1-\gamma)\hat{\pi}_X^{\omega_0}$$

*and*

$$\pi^* = \gamma \delta_{a^*} + (1-\gamma)\hat{\pi}_Y^{\omega_0}.$$

**Proof.** By again utilizing Equation (2) (of the supplementary material), the claim modifies to showing that

$$\underline{opt}_\varepsilon(\mathbf{x}, \mathbf{y}) = \inf_{u \in \mathcal{N}^{\delta_\varepsilon(\omega_0)}_{\mathcal{A}_{\omega_0}}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\pi_*(\{z\}) - \pi^*(\{z\})).$$

Since, by Proposition 2, we know that $\underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0) = \underline{opt}_\varepsilon(\mathbf{x}, \mathbf{y})$ and $\underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0)$ is by definition the infimum over all the expressions on the right-hand side, the direction $\leq$ is immediate. So, it remains to show the direction $\geq$. To do so, choose $(\pi_1, \pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}$ arbitrarily. Since both $\mathcal{M}_X^{\omega_0}$ and $\mathcal{M}_Y^{\omega_0}$ are of the form (11), we then know that there exist probability measures $\nu_1$ and $\nu_2$ such that

$$\pi_1 = \gamma \cdot \nu_1 + (1-\gamma) \cdot \hat{\pi}_X^{\omega_0}$$

and

$$\pi_2 = \gamma \cdot \nu_2 + (1-\gamma) \cdot \hat{\pi}_Y^{\omega_0}.$$

Here, we utilized the fact that credal sets of the form (11) can be equivalently characterized as

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

For $u \in \mathcal{N}^{\delta_\varepsilon(\omega_0)}_{\mathcal{A}_{\omega_0}}$ fixed (but arbitrary), we then can compute:

$$\begin{aligned}
\sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \pi_1(\{z\}) &= \gamma \cdot \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \nu_1(\{z\}) + (1-\gamma) \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \hat{\pi}_X^{\omega_0}(\{z\}) \\
&\geq \gamma \cdot u(a_*) + (1-\gamma) \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \hat{\pi}_X^{\omega_0}(\{z\}) \\
&= \gamma \cdot \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \delta_{a_*}(\{z\}) + (1-\gamma) \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \hat{\pi}_X^{\omega_0}(\{z\}) \\
&= \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \pi_*(\{z\})
\end{aligned}$$

Analogous reasoning yields:

$$\sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot\pi_2(\{z\}) \leq \sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot\pi^*(\{z\})$$

Putting the two together, we arrive at:

$$\sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot(\pi_1(\{z\})-\pi_2(\{z\})) \geq \sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot(\pi_*(\{z\})-\pi^*(\{z\}))$$

As $\pi_1, \pi_2$, and $u$ were chosen arbitrarily, the inequality remains valid for the infimum, i.e.

$$\inf_{(\pi_1,\pi_2,u)\in\mathcal{M}_X^{\omega_0}\times\mathcal{M}_Y^{\omega_0}\times\mathcal{N}_{\mathcal{A}_{\omega_0}}^{\delta_\varepsilon(\omega_0)}} \sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot(\pi_1(\{z\})-\pi_2(\{z\})) \geq \inf_{u\in\mathcal{N}_{\mathcal{A}_{\omega_0}}^{\delta_\varepsilon(\omega_0)}} \sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot(\pi_*(\{z\})-\pi^*(\{z\}))$$

Observing that the left side of this inequality by definition equals $\underline{d}^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0)$ and, therefore, by Proposition 2, also $\underline{opt}_{\varepsilon}(\mathbf{x},\mathbf{y})$ completes the direction $\geq$ and thus the proof. □

### A.4 PROOFS OF PROPOSITIONS 5 AND 6: MULTI-DIMENSIONAL SPACES

Finally, we give proofs of Propositions 5 and 6 from Section 7 concerning several different characterizing properties of the GSD-order for the special case of preferences systems arising from multi-dimensional spaces with differently scaled dimensions. For this, recall that in Section 4 for a preference system $\mathcal{A}$ and a probability measure $\pi$ we defined

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

This definition is needed for stating the next proposition.

**Proposition 5** *Let $\pi$ be a probability measure on $(\Omega,\mathcal{S}_1)$, and $X=(\Delta_1,\dots,\Delta_r), Y=(\Lambda_1,\dots,\Lambda_r)\in\mathcal{F}_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$, where the first $0\leq z\leq r$ dimensions of $\mathsf{pref}(\mathbb{R}^r)$ are of cardinal scale. Then, the following holds:*

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

*ii) If $z=0$, then $R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ coincides with (first-order) stochastic dominance w.r.t. $\pi$ and $R_1^*$ (short: FSD$(R_1^*,\pi)$).*

*iii) If $(X,Y)\in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ 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 $j=z+1,\dots,r$.*

*Additionally, in the special case where all components of $X$ are jointly independent and all components of $Y$ are jointly independent, properties I. and II. imply $(X,Y)\in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ (i.e. also the converse implication holds).*

**Proof.** i) Let $\alpha_1,\dots,\alpha_r\in\mathbb{R}^+$ and $\phi_{z+1},\dots,\phi_r:\mathbb{R}\to\mathbb{R}$ strictly isotone functions. Define $u:\mathbb{R}^r\to\mathbb{R}$ by setting

$$u(x) := \sum_{s=1}^{z}\alpha_s\cdot x_s + \sum_{s=z+1}^{r}\alpha_s\cdot\phi_s(x_s).$$

Then one easily verifies that $u$ defines a representation of $\mathsf{pref}(\mathbb{R}^r)$, proving its consistency.

ii) Assume $z=0$, i.e. all considered dimensions are purely ordinal. We claim that for $\mathcal{A}_0 := [\mathbb{R}^r, R_1^*, \emptyset]$ it holds $\mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}=\mathcal{U}_{\mathcal{A}_0}$. The direction $\subseteq$ is trivial, so assume $u\in\mathcal{U}_{\mathcal{A}_0}$ arbitrary. It suffices to show that $u$ represents arbitrary pairs of pairs in $R_2^*$. As $R_2^*$ is antisymmetric for $z=0$, this reduces to show that $u$ strictly represents arbitrary pairs of pairs in $P_{R_2^*}$. So, let $((v,w),(x,y))\in P_{R_2^*}$. This means that for all $j\in\{1,\dots,r\}$ we have $v_j\geq x_j\geq y_j\geq w_j$ and that there is $j_0\in\{1,\dots,r\}$ such that either $v_{j_0}>x_{j_0}$ or $y_{j_0}>w_{j_0}$. Together, this implies $u(v)>u(x)\geq u(y)\geq u(w)$ or $u(v)\geq u(x)\geq u(y)>u(w)$, either way implying $u(v)-u(w)>u(x)-u(y)$. Thus $u\in\mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$. As $R_{(\mathcal{A}_0,\pi)}$ coincides with (first-order) stochastic dominance by definition and we have $\mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}=\mathcal{U}_{\mathcal{A}_0}$ also $R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ coincides with (first-order) stochastic dominance.

iii) Let $(X,Y) \in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$. We start by showing I, so choose $j \in \{1,\dots,r\}$ arbitrary. By part i) of the proof, for every $n \in \mathbb{N}$, the function $u_n : \mathbb{R}^r \to \mathbb{R}$ defined by

$$u_n(x) := x_j + \frac{1}{n} \cdot \sum_{s \neq j} x_s$$

is a representation of $\mathsf{pref}(\mathbb{R}^r)$, that is $u_n \in \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$. Thus, by our assumption $(X,Y) \in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$, we know that we have $\mathbb{E}_\pi(u_n \circ X) \geq \mathbb{E}_\pi(u_n \circ Y)$. This implies (by the linearity of the expectation operator)

$$\mathbb{E}_\pi(\Delta_j) + \frac{1}{n} \cdot \sum_{s \neq j} \mathbb{E}_\pi(\Delta_s) \geq \mathbb{E}_\pi(\Lambda_j) + \frac{1}{n} \cdot \sum_{s \neq j} \mathbb{E}_\pi(\Lambda_s).$$

Letting $n \to \infty$ on both sides gives $\mathbb{E}_\pi(\Delta_j) \geq \mathbb{E}_\pi(\Lambda_j)$.

We use a very similar argument to see II: Choose $j \in \{z+1,\dots,r\}$ arbitrarily and let $\phi : \mathbb{R} \to \mathbb{R}$ be strictly isotone. By part i) of the proof, for every $n \in \mathbb{N}$, the function $u'_n : \mathbb{R}^r \to \mathbb{R}$ defined by

$$u'_n(x) := \phi(x_j) + \frac{1}{n} \cdot \sum_{s \neq j} x_s$$

is a representation of $\mathsf{pref}(\mathbb{R}^r)$, that is $u_n \in \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$. Thus, by our assumption $(X,Y) \in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$, we know that we have $\mathbb{E}_\pi(u_n \circ X) \geq \mathbb{E}_\pi(u_n \circ Y)$. This implies (by the linearity of the expectation operator)

$$\mathbb{E}_\pi(\phi \circ \Delta_j) + \frac{1}{n} \cdot \sum_{s \neq j} \mathbb{E}_\pi(\Delta_s) \geq \mathbb{E}_\pi(\phi \circ \Lambda_j) + \frac{1}{n} \cdot \sum_{s \neq j} \mathbb{E}_\pi(\Lambda_s).$$

Letting $n \to \infty$ gives $\mathbb{E}_\pi(\phi \circ \Delta_j) \geq \mathbb{E}_\pi(\phi \circ \Lambda_j)$. As $\phi$ was chosen arbitrarily, this implies $(\Delta_j, \Lambda_j) \in$FSD$(\geq, \pi)$.

To see the addition to part iii), let $X = (\Delta_1, \dots \Delta_r)$ and $Y = (\Lambda_1, \dots, \Lambda_r)$ have both jointly independent components, respectively, and let I. and II. of iii) be true. Let furthermore $u \in \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$ be an arbitrary utility function that represents the preference system $\mathsf{pref}(\mathbb{R}^r)$. We now show that $\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$ holds: Because of independence we can compute the expectations of $u \circ X$ and $u \circ Y$ by using Fubini's theorem. To prove the inequality, we first integrate over the ordinal part and use isotonicity of $u$ in every integration. Then we integrate over the cardinal parts and iteratively use the fact that the corresponding functions are representing the corresponding cardinal subsystem built by the components we did not integrate over before. Formally, we arrive at:

$$\begin{aligned}
\mathbb{E}_\pi(u \circ X) &= \int_\Omega u \circ X d\pi \\
&\overset{(ind.)}{=} \int_{\Delta_1(\Omega)} \cdots \int_{\Delta_r(\Omega)} u(\delta_1, \dots, \delta_z, \delta_{z+1}, \dots \delta_r) d\pi_{\Delta_r} \dots d\pi_{\Delta_{z+1}} d\pi_{\Delta_z} \dots d\pi_{\Delta_1} \\
&\overset{(\star)}{\geq} \int_{\Delta_1(\Omega)} \cdots \int_{\Lambda_r(\Omega)} u(\delta_1, \dots, \delta_z, \lambda_{z+1}, \dots \lambda_r) d\pi_{\Lambda_r} \dots d\pi_{\Lambda_{z+1}} d\pi_{\Delta_z} \dots d\pi_{\Delta_1} \\
&\overset{(\star\star)}{\geq} \int_{\Lambda_1(\Omega)} \cdots \int_{\Lambda_r(\Omega)} u(\lambda_1, \dots, \lambda_z, \lambda_{z+1}, \dots \lambda_r) d\pi_{\Lambda_r} \dots d\pi_{\Lambda_{z+1}} d\pi_{\Lambda_z} \dots d\pi_{\Lambda_1} \\
&\overset{(ind.)}{=} \mathbb{E}_\pi(u \circ Y)
\end{aligned}$$

Here, $(\star)$ is valid because, for fixed cardinal components, $u$ is isotone in every ordinal component and we have first order stochastic dominance, which means that the iterated integrals gets smaller if one switches from $\pi_{\Delta_k}$ to $\pi_{\Lambda_k}$.

Similarly, $(\star\star)$ is valid because e.g., for the mapping

$$\psi : \mathbb{R}^{z-1} \to \mathbb{R} \quad , \quad (\delta_1, \dots, \delta_{z-1}) \mapsto \int_{\Delta_z(\Omega)} u(\delta_1, \dots, \delta_r) d\pi_{\Delta_z}$$

is a positive (affine) linear transformation w.r.t. the corresponding subsystem. □

**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. 5 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 $\forall u \in \mathcal{N}_{\mathcal{C}} : \mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$.*

**Proof.** As, according to Proposition 5 i), we know that $\mathsf{pref}(\mathbb{R}^r)$ is consistent, the same holds true for all of its subsystems. Hence, $\mathcal{C}$ is consistent. Since $\mathcal{C}$ is assumed to be bounded, it then is 0-consistent by Observation 1. The rest of the Corollary follows, since – by Observation 2 – for bounded preference systems it suffices to check for dominance only over all normalized representations. □

**Proposition 6** *Let $z = 1$ and denote by $\mathcal{U}_{sep}$ the set of all $u : \mathbb{R}^r \to \mathbb{R}$ such that, for $(x_2, \dots, x_r) \in \mathbb{R}^{r-1}$ fixed, the function $u(\cdot, x_2, \dots, x_r)$ is strictly increasing and (affine) linear and such that, for $x_1 \in \mathbb{R}$ fixed, the function $u(x_1, \cdot, \dots, \cdot)$ is strictly isotone w.r.t. the the componentwise partial order on $\mathbb{R}^{r-1}$. Then $\mathcal{U}_{sep} = \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$.*

**Proof.** First, let $u \in \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$. One easily verifies that, for $x_- := (x_2, \dots, x_r) \in \mathbb{R}^{r-1}$ fixed, the preference system $Z := [\mathbb{R}, R_1^{x-}, R_2^{x-}]$, where $R_1^{x-} :=\geq$ and $R_2^{x-}$ is defined by

$$\left\{((t,u),(v,w)) : \left(\left(\begin{pmatrix} t \\ x_- \end{pmatrix}, \begin{pmatrix} u \\ x_- \end{pmatrix}\right), \left(\begin{pmatrix} v \\ x_- \end{pmatrix}, \begin{pmatrix} w \\ x_- \end{pmatrix}\right)\right) \in R_2^*\right\}$$

is a complete positive-difference structure in the sense of Krantz et al. [1971, Definition 1, p. 147]. According to Krantz et al. [1971, Theorem 1, p. 147] this implies that any two representations of $Z$ are positive (affine) linear transformations of each other. But it is immediate that both $u(\cdot, x_2, \dots, x_r)$ and $id_{\mathbb{R}}(\cdot)$ are representations of $Z$. Thus, $u(\cdot, x_2, \dots, x_r) = \alpha \cdot id_{\mathbb{R}}(\cdot) + \beta$ for some $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}$, proving the first claim of this direction. The second claim – i.e., the strict isotony of the function $u(x_1, \cdot, \dots, \cdot)$ w.r.t. the the componentwise partial order on $\mathbb{R}^{r-1}$ for fixed $x_1 \in \mathbb{R}$ – is also immediate. Thus, $u \in \mathcal{U}_{sep}$.

For the other direction, assume that $u \in \mathcal{U}_{sep}$. It follows directly from the assumptions that $u$ is strictly isotone w.r.t. $R_1^*$. To see that $u$ also strictly represents $R_2^*$, choose $((x,y),(x',y')) \in R_2^*$ arbitrary. We have two cases:

*Case 1:* $((x,y),(x',y')) \in I_{R_2^*}$. This implies that $x_1 - y_1 = x_1' - y_1'$ and therefore also $x_1 - x_1' = y_1 - y_1'$. Moreover, one easily verifies that the restriction of $R_2^*$ to the ordinal dimensions is antisymmetric . Since we have that $x_-$ componentwise dominates $x_-'$ and vice versa and that $y_-$ componentwise dominates $y_-'$ and vice versa, this antisymmetry then implies that $x_- = x_-'$ and $y_- = y_-'$. Therefore, there are common $\alpha_1, \alpha_2 \in \mathbb{R}^+$ and $\beta_1, \beta_2 \in \mathbb{R}$ such that

$$u(x) = \alpha_1 \cdot x_1 + \beta_1 \quad , \quad u(x') = \alpha_1 \cdot x_1' + \beta_1$$
$$u(y) = \alpha_2 \cdot y_1 + \beta_2 \quad , \quad u(y') = \alpha_2 \cdot y_1' + \beta_2$$

Moreover, observe that $\alpha_1 = \alpha_2$, since otherwise there wolud be $x^* \in \mathbb{R}$ with $u(x^*, x_-) < u(x^*, y_-)$, which is not possible, since $u$ is strictly isotone w.r.t. $R_1^*$. Define

$$D := (u(x) - u(y)) - (u(x') - u(y')).$$

Simple computations then yield

$$D = \alpha_1 \cdot (x_1 - x_1') - \alpha_2 \cdot (y_1 - y_1') = (x_1 - x_1') \cdot (\alpha_1 - \alpha_2)$$

which, as $\alpha_1 = \alpha_2$, implies $D = 0$.

*Case 2:* $((x,y),(x',y')) \in P_{R_2^*}$. This implies $x_- \geq x_-' \geq y_-' \geq y_-$, where $\geq$ is to be understood componentwise. Using the same argument as seen before, this implies that there exists a $\alpha \in \mathbb{R}^+$ and $\beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}$ such that

$$u(x) = \alpha \cdot x_1 + \beta_1 \quad , \quad u(x') = \alpha \cdot x_1' + \beta_3$$
$$u(y) = \alpha \cdot y_1 + \beta_2 \quad , \quad u(y') = \alpha \cdot y_1' + \beta_4$$

Thus, computing $D$ defined as above yields:

$$D = \alpha \cdot ((x_1 - y_1) - (x_1' - y_1')) + \beta_1 - \beta_2 - \beta_3 + \beta_4$$

*Sub-Case 2.1:* $x_1 - y_1 > x_1' - y_1'$. Observe that, as $u$ is isotone w.r.t. $R_1^*$, we have that $u(y_1', y_-') \geq u(y_1', y)$. However, this implies $\beta_4 \geq \beta_2$. Analogous reasoning yields $\beta_1 \geq \beta_3$. Using the assumptions of the sub-case, this implies $D > 0$.

*Sub-Case 2.2:* $x_1 - y_1 = x_1' - y_1'$. Using the case assumption, this implies that either $x_- > x_-'$ or $y_-' > y_-$, where the $>$ is to be understood as the strict part of the componentwise $\geq$. As $u$ is strictly isotone w.r.t. $R_1^*$, this implies that either $u(y_1', y_-') > u(y_1', y)$ or $u(x_1', x_-) > u(x_1', x_-')$, which itself implies either $\beta_4 > \beta_2$ or $\beta_1 > \beta_3$. As we know $\beta_4 \geq \beta_2$ and $\beta_1 \geq \beta_3$, this, together with the sub-case assumption, implies $D > 0$. □

# B DETAILS ON IMPLEMENTATION AND REPRODUCIBILITY

In Section 8.1 we stated that the implementation of the constraint matrix has worst-case complexity $\mathcal{O}(s^4)$. This worst case occurs when everything in $R_1^*$ and $R_2^*$ is comparable and then

$$s \cdot (s-1) + (s \cdot (s-1)) \cdot ((s \cdot (s-1)) - 1) = s^4 - 2s^3 + s^2$$

many pairwise comparisons have to be considered. Note that we omit the reflexive part of the pre-orders $R_1^*$ and $R_2^*$.

In implementing the constraint matrix, 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, if the observations grouped by their ordinal components are highly skewed and the largest ordinal components correspond to the largest group, the computation time cannot be drastically reduced in this way.

We are interested in the non-regularized test statistic as well as the regularized test statistic with $\varepsilon \in \{0.25, 0.5, 0.75, 1\}$, see Section 8. For all these cases, we compute the test statistics based on the sample, as well as 1000 times on a permuted version of that sample. Note that the linear programs for computing the test statistics based on the permuted data are identical to that for the non-permuted data except for the objective function, see Section 5.2. In Section C (in the supplementary material), we prove that the robustified test statistics are a shift of the non-robustified test statistic. Thus, the robustified test statistics are immediately given.

The simulation is based on a random sample of the data set. Two of the data sets and the corresponding R-code can be found here:

`https://anonymous.4open.science/r/Robust_GSD_Tests`

The data set used for the poverty analysis (ALLBUS) is freely accessible, but registration in the corresponding online portal is needed.[1]

For the computation of the linear programs, we used the R interface of Gurobi optimizer, which is documented in Gurobi Optimization, LLC [2020]. This is a commercial solver that offers free academic licenses[2]. In particular, the computation of linear programs is faster than using the free and open source solvers known to us, see Meindl and Templ [2012]. We also used the R-packages *purrr*, *dplyr*, *slam*, *readr*, *tidyr, forcast, ggplot2, reshape2, tidyverse, ggridges, latex2exp, RColorBrewer*, *rcartocolor* and *foreign* for our implementation, see Mailund [2022], Yarberry and Yarberry [2021], Wickham et al. [2022], Hornik et al. [2022], Wickham et al. [2023], Hyndman et al. [2023], Wickham and Chang [2014], Wickham [2022], Wickham and RStudio [2022], Wilke [2022], Meschiari [2022], Neuwirth [2022], Nowosad [2022], R Core Team et al. [2022].

The computation was done for

- ALLBUS data set, see GESIS [2018], on a commodity desktop laptop with a 8-core Intel(R) Core(TM) i7-8665U CPU @ 1.90GHz processor and 16 GB RAM in R version 4.2.2.
- dermatology data set, see Demiroz et al. [1998] accessed via Dua and Graff [2017], on a commodity desktop computer with a 32-core Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz processor and 64 GB RAM in R version 4.2.1
- German credit data set, see Dua and Graff [2017], on a commodity desktop laptop with a 8-core Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz processor and 16 GB RAM in R version 4.2.2.

# C CALCULATIONS FOR ROBUSTIFIED TEST STATISTICS

In Section 8 we show a graph visualizing the fraction of resamples in favor of **non**-rejection of $H_0$ (i.e., the p-values) as a function of the size of the contamination $\gamma$ of the underlying linear-vacuous model (see Figure 3). We will briefly show here how the exact function is calculated. For general (polyhedral) credal sets, a resample $I$ is in favor of rejection of $H_0$ under

[1] Further information on the survey and the data set itself can be found here: `https://search.gesis.org/research_data/ZA5240` (accessed: Febr 16, 2023)

[2] Further details can be found here: `https://www.gurobi.com/academia/academic-program-and-licenses/` (accessed: Febr 16, 2023)

the robustified resampling scheme, if $\underline{d}^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) > \overline{d}^{\varepsilon}_{I}$. Hence, the fraction of resamples in favor of rejection of $H_0$ is given by

$$\frac{1}{N} \cdot \sum_{I \in \mathcal{I}_N} \mathbb{1}_{\left\{\underline{d}^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) > \overline{d}^{\varepsilon}_{I}\right\}}$$

where $N$ denotes the number of resamples and $\mathcal{I}_N$ is the corresponding set of resamples. In the special case that the credal sets involved are $\gamma$-contamination models, we can use Proposition 4 (and a slight variation of it with $\pi_*$ and $\pi^*$ in reversed roles) to obtain

$$\underline{d}^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) = (1-\gamma) \cdot d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) - \gamma$$

and

$$\overline{d}^{\varepsilon}_{I} = (1-\gamma) \cdot d^{\varepsilon}_{I} + \gamma$$

and, therefore, the condition in the indicator above is satisfied if and only if

$$d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) - d^{\varepsilon}_{I} > \frac{2\gamma}{(1-\gamma)}.$$

Finally, if we interpret $\varepsilon$ as a function parameter, then we can write the fraction of resamples in favor of **non**-rejection of $H_0$ (i.e., the observed p-values) as a function of the size $\gamma$ of the contamination of the underlying linear-vacuous model:

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

# D FURTHER DETAILS ON THE APPLICATIONS

## D.1 DATA SETS

We applied our analysis to three different data sets:

- For the poverty analysis, see Section 8, we used the ALLBUS data set. The data set is described by GESIS [2018] and Breyer and Danner [2015]. As mentioned already in the previous section, the data set is freely accessible, but only after registration in the corresponding online portal: `https://search.gesis.org/research_data/ZA5240` (accessed: 08.02.2023). Please download the file ZA5240_v2-2-0.sav (5.31MB) there.

  The analysis was done on a sample consisting of 100 female and 100 male observations.
- We analyzed the dermatology data set, see Demiroz et al. [1998] accessed via Dua and Graff [2017].

  The analysis was performed on a sample of 46 individuals with family history of eryhemato-squamous disease and 100 individuals without.
- We analyzed the German credit data set, see Dua and Graff [2017].

  The analysis was performed on a sample of 100 credit risks classified as good and 100 credit risks classified as poor individuals.

## D.2 APPLICATION ON CREDIT DATA

We focus on three variables (features) in the German credit data set Dua and Graff [2017]: credit amount (numeric), credit history (ordinal, $5$ levels ranging from “delay in paying off in the past” to “all credits paid back duly”) and employment status (ordinal, $5$ levels ranging from “unemployed” to “present employment longer than 7 years”). We use a subsample with $n = m = 100$ high-risk applicants and low-risk applicants each. We are interested in the hypothesis that high-risk applicants are dominated by low-risk applicants w.r.t. GSD. The test results (see Figures 1 and 2 in the supplementary material) can be interpreted analogously to Section 8: For $\varepsilon \in \{0.75, 1\}$ we reject for the common significance level of $\alpha \approx 0.05$. This time, we do not reject in case of $\varepsilon = 0.5$.

Similar to the example of poverty analysis in Section 8, rejecting $H_0$ does not necessarily mean that high-risk applicants are dominated by low-risk applicants. They could also be incomparable, see also Section 5. However, our tests with reversed variables give no evidence of incomparability: The observed p-values for all these reversed tests are all $1$.

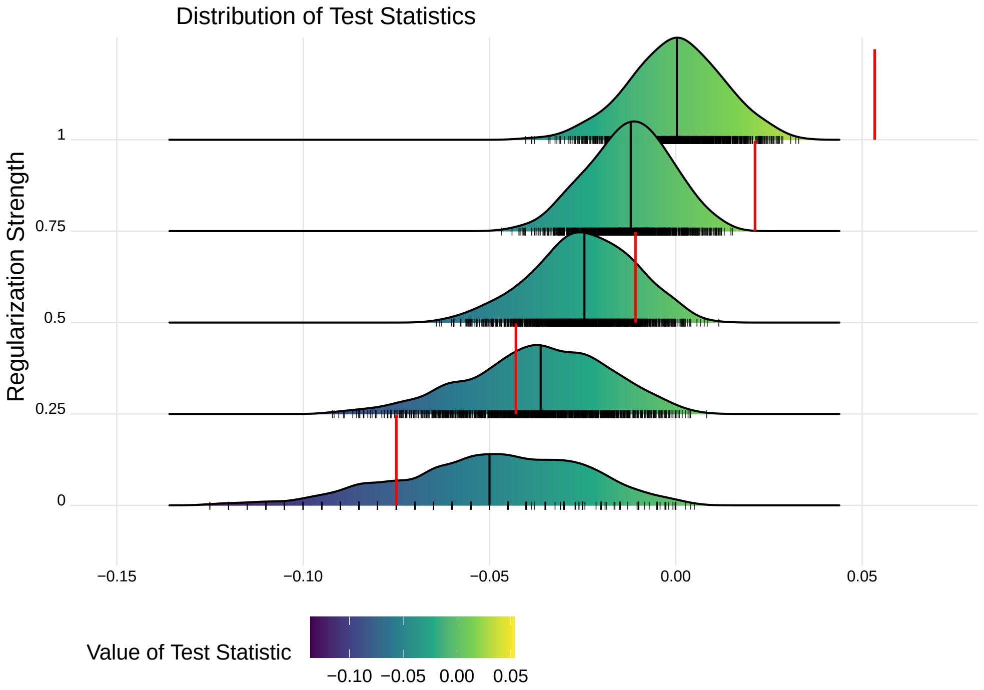

Figure 1: Distributions of $d_I^\varepsilon$ with $\varepsilon \in \{0, 0.25, 0.5, 0.75, 1\}$ obtained from $N = 1000$ resamples of Credit data. Black stripes show exact positions of $d_I^\varepsilon$ values. Vertical black line marks median. Red line shows value of the respective observed test statistics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

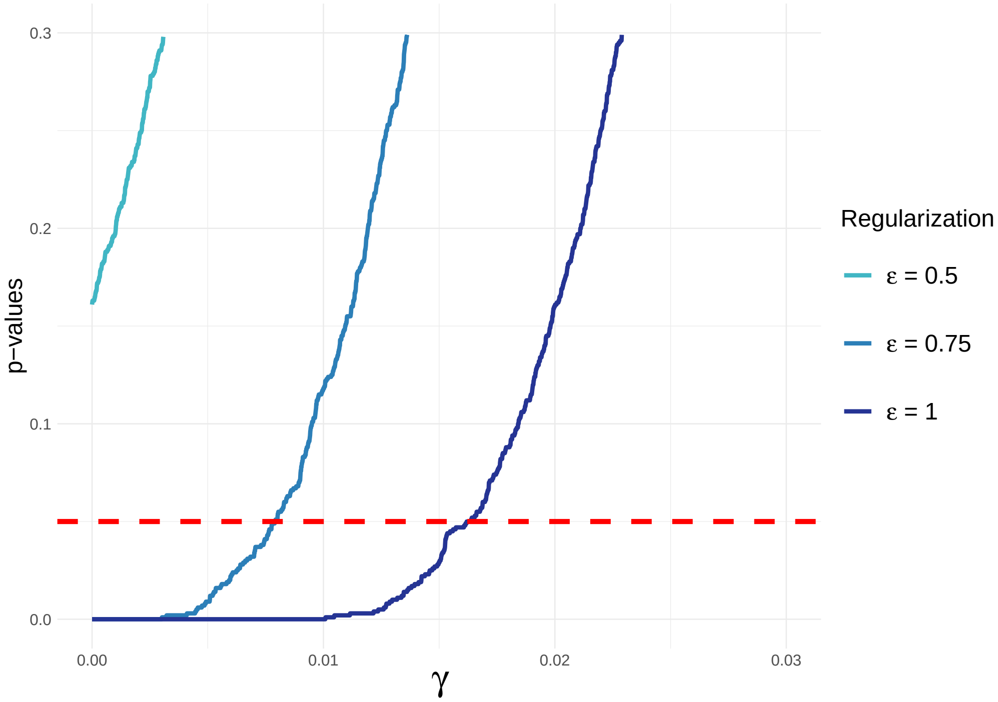

Figure 2: P-values as function of the contamination $\gamma$ (see Supp. C) for tests with different regularization strength $\varepsilon$ performed nd on credit data set. Dotted red line marks significance level $\alpha = 0.05$.

## D.3 APPLICATION ON DERMATOLOGICAL DATA

We focus on three variables (features) in the dermatology data set Demiroz et al. [1998], Dua and Graff [2017]: age of skin (numeric), the intensity of itching (ordinal, 4 levels ranging from “no itching” to “strong itching”) and erythema (redness of skin) (ordinal, 4 levels again ranging from no to highest intensity). We use a subsample with $n = 46$ patients with a family history of eryhemato-squamous disease and $m = 100$ without. 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. The test results (see Figures 3 and 4 in the supplementary material) can be interpreted analogously to Section 8: For $\varepsilon \in \{0.75, 1\}$ we again reject for the common significance level of $\alpha \approx 0.05$. However, the p-values are much higher than in the other two applications, see also Figure 4 (in the supplementary material).

Similar to the example of poverty analysis in Section 8, rejecting $H_0$ does not necessarily mean that patients with a family history of eryhemato-squamous disease are dominated by patients without. They could also be incomparable; see also Section 5. However, our tests with reversed variables give no evidence of incomparability: The observed p-values for all these reversed tests are all 1.

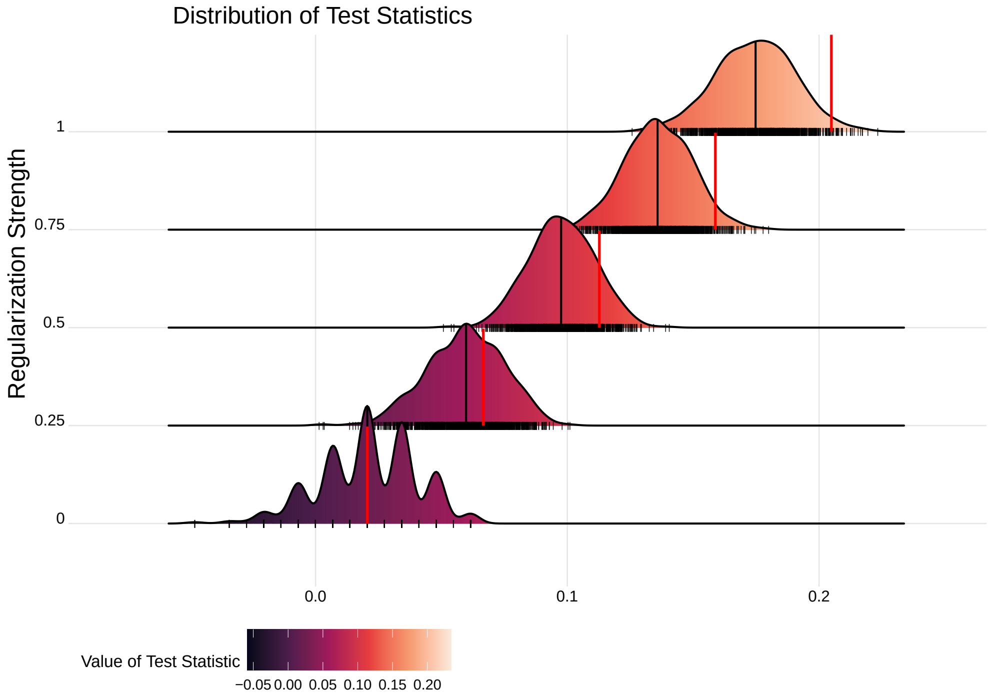

Figure 3: Distributions of $d_I^\varepsilon$ with $\varepsilon \in \{0, 0.25, 0.5, 0.75, 1\}$ obtained from $N = 1000$ resamples of dermatology data. Black stripes show exact positions of $d_I^\varepsilon$ values. Vertical black line marks median. Red line shows value of the respective observed test statistics $d^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega)$.

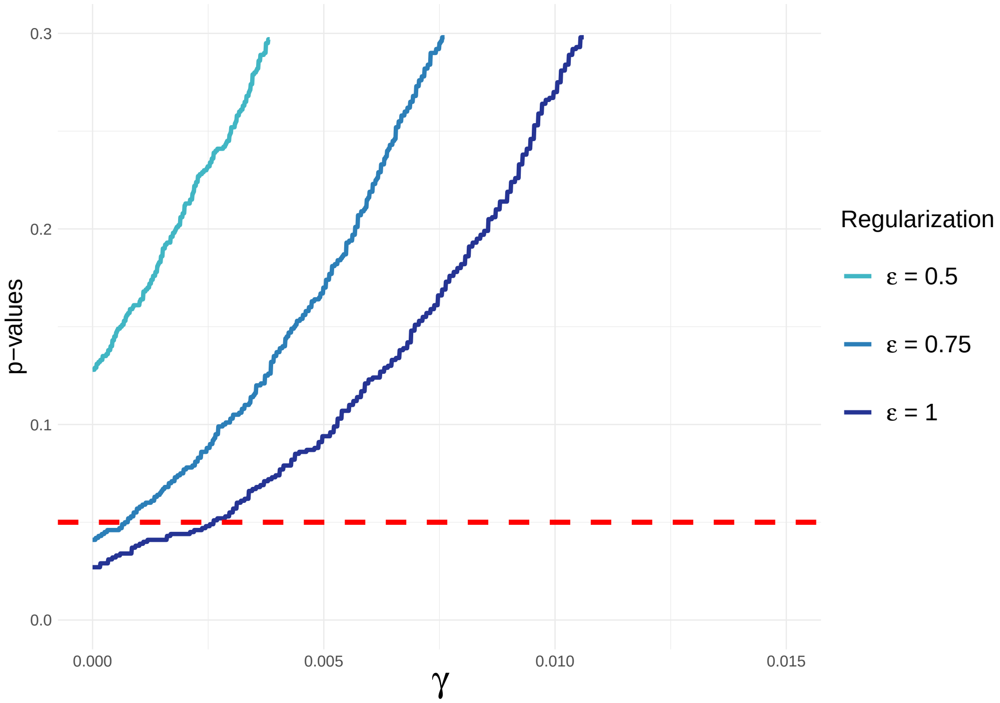

Figure 4: P-values as function of the contamination $\gamma$ (see Supp. C) for tests with different regularization strength $\varepsilon$ performed on Dermatology data set. The dotted red line marks significance level $\alpha = 0.05$.

## References

Gurobi Optimization, LLC. Gurobi optimizer reference manual, 2020. URL `https://www.gurobi.com/wp-content/plugins/hd_documentations/documentation/9.0/refman.pdf`. [Online; Accessed 08.02.2023].

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.

G. Demiroz, H. Govenir, and N. Ilter. Learning differential diagnosis of eryhemato-squamous diseases using voting feature intervals. *Artificial Intelligence in Medicine*, 13(3):147–165, 1998.

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

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

K. Hornik, D. Meyer, and C. Buchta. Package 'slam', October 2022. URL `https://cran.r-project.org/web/packages/slam/slam.pdf`. [Online; Accessed 08.02.2023].

R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, K. Kuroptev, M. O'Hara-Wild, F. Petropoulos S. Razbash, E. Wang, F. Yasmeen, F. Garza, D. Girolimetto, R. Ihaka, R Core Team, D. Reid, D. Shaub, Y. Tang, X. Wang, and Z. Zhou. Package 'forcast', January 2023. URL `https://cran.r-project.org/web/packages/forecast/forecast.pdf`. [Online: Accessed 09.02.2023].

D. Krantz, R. Luce, P. Suppes, and A. Tversky. *Foundations of Measurement. Volume I: Additive and Polynomial Representations*. Academic Press, 1971.

T. Mailund. Functional programming: purrr. In *R 4 Data Science Quick Reference: A Pocket Guide to APIs, Libraries, and Packages*, pages 89–110. Springer, 2022.

B. Meindl and M. Templ. Analysis of commercial and free and open source solvers for linear optimization problems. *Eurostat and Statistics Netherlands within the project ESSnet on common tools and harmonised methodology for SDC in the ESS*, 20, 2012.

S. Meschiari. Package 'latex2exp', November 2022. URL `https://cran.r-project.org/web/packages/latex2exp/latex2exp.pdf`. [Online: Accessed 09.02.2023].

E. Neuwirth. Package 'rcolorbrewer', October 2022. URL `https://cran.r-project.org/web/packages/RColorBrewer/RColorBrewer.pdf`. [Online: Accessed 09.02.2023].

J. Nowosad. Package 'rcartocolor', October 2022. URL `https://cran.r-project.org/web/packages/rcartocolor/rcartocolor.pdf`. [Online: Accessed 09.02.2023].

R Core Team, R. Bivand, V. Carey, S. DebRoy, S. Eglen, R. Guha, S. Herbrandt, N. Lewin-Koh, M. Myatt, M. Nelson, B. Pfaff, B. Quistorff, F. Warmerdam, S. Weigand, and Inc. Free Software Foundation. Package 'foreign', December 2022. URL `https://cran.r-project.org/web/packages/foreign/foreign.pdf`. [Online: Accessed 09.02.2023].

H. Wickham. Package 'reshape', October 2022. URL `https://cran.r-project.org/web/packages/reshape/reshape.pdf`. [Online: Accessed 09.02.2023].

H. Wickham and W. Chang. Package 'ggplot2', December 2014. URL `https://cran.microsoft.com/snapshot/2015-01-06/web/packages/ggplot2/ggplot2.pdf`. [Online: Accessed 09.02.2023].

H. Wickham and RStudio. Package 'tidyverse', October 2022. URL `https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf`. [Online: Accessed 09.02.2023].

H. Wickham, J. Hester, R. Francois, J. Bryan, and S. Bearrows. Package 'readr', October 2022. URL `https://cran.r-project.org/web/packages/readr/readr.pdf`. [Online; Accessed 08.02.2023].

H. Wickham, D. Vaughan, M. Girlich, K. Ushey, and PBC Posit. Package 'tidyr', January 2023. URL `https://www.vps.fmvz.usp.br/CRAN/web/packages/tidyr/tidyr.pdf`. [Online: Accessed 09.02.2023].

C. O. Wilke. Package 'ggridges', October 2022. URL `https://cran.r-project.org/web/packages/ggridges/ggridges.pdf`. [Online: Accessed 09.02.2023].

W. Yarberry and W. Yarberry. Dplyr. *CRAN Recipes: DPLYR, Stringr, Lubridate, and RegEx in R*, pages 1–58, 2021.