LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Julia Plass, Marco Cattaneo, Georg Schollmeyer, Thomas Augustin

On the testability of coarsening assumptions: A hypothesis test for subgroup independence

Technical Report Number 201, 2017
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

On the testability of coarsening assumptions: A hypothesis test for subgroup independence

J. Plass[a,*], M. Cattaneo[b], G. Schollmeyer[a], T. Augustin[a]

[a] *Department of Statistics, LMU Munich, Ludwigsstr. 33, 80539 Munich, Germany*
[b] *School of Mathematics & Physical Sciences, University of Hull, Hull, HU6 7RX, UK*

Abstract

Since coarse(ned) data naturally induce set-valued estimators, analysts often assume coarsening at random (CAR) to force them to be single-valued. Focusing on a coarse categorical response variable and a precisely observed categorical covariate, we re-illustrate the impossibility to test CAR and contrast it to another type of coarsening called subgroup independence (SI), using the data of the German Panel Study "Labour Market and Social Security" as an example. It turns out that – depending on the number of subgroups and categories of the response variable – SI can be point-identifying as CAR, but testable unlike CAR. A main goal of this paper is the construction of the likelihood-ratio test for SI. All issues are similarly investigated for the here proposed generalized versions, gCAR and gSI, thus allowing a more flexible application of this hypothesis test.

Keywords: coarse data, missing data, coarsening at random (CAR), likelihood-ratio test, partial identification, sensitivity analysis

1. Introduction: The problem of testing coarsening assumptions

Traditional statistical methods dealing with missing data (e.g. EM algorithm or imputation techniques) require identifiability of parameters, which frequently tempts analysts to make the *missing at random* (MAR) assumption (cf. [24]) simply for pragmatic reasons without justifications in substance (cf. [12]). Since MAR is not testable (e.g. [16]) and wrongly including this assumptions may induce a substantial bias, this way to proceed is especially alarming.
Looking at the problem in a more general way, incomplete observations may occur not only in the sense of missing, but also coarse(ned) data. In this way, additionally to fully observed and unobserved, also partially observed values are

☆A preliminary version of this paper was presented at the 8th Conference on Soft Methods in Probability and Statistics (SMPS) in Rome, September, 12-14, 2016 [22].
*Corresponding author
Email address: julia.plass@stat.uni-muenchen.de (J. Plass)

considered.[1] In the context of coarse data, the *coarsening at random* (CAR) (cf. [9]) assumption is the analogue of MAR. Although the impossibility of testing CAR is already known from literature (cf. e.g. [11]), providing an intuitive insight into this point will be a first goal. Apart from CAR, we focus on another, in a sense dual, assumption that we called *subgroup independence* (SI) in [20] and elaborate the substantial difference between CAR and SI with regard to testability.

Our argumentation is based on the maximum likelihood estimators obtained under the specific assumptions in focus. There is already a variety of maximum likelihood approaches for incomplete data. While some rely on optimization strategies, as for instance maximax or maximin, to force a single-valued result (cf. e.g. [8], [10]), others end up with set-valued results (cf. e.g. [2], [13], [20], [32]). A general view is given by Couso and Dubois [5], distinguishing between different types of likelihoods, the visible, the latent and the total likelihood. Here, we use the cautious approach developed in [20], which refers to the latent likelihood and is – just as e.g. [17] (in the context of misclassification) and [25] – strongly influenced by the methodology of *partial identification* (cf. [16]). Thus, according to the spirit of partial identification, instead of being forced to make untenable, strict assumptions, as CAR or SI, to give an answer to the research question at all, we can explicitly make use of in practice more realistic partial knowledge about the incompleteness, which would have to be left out of considerations if a traditional approach were used. For this purpose, we use an observation model as a powerful medium to include the available knowledge into the estimation problem. By considering generalized versions of the strict assumptions in focus, which we call gCAR and gSI, we can express this knowledge in a flexible and careful way. This means, we are no longer restricted to formalize the very specific types of coarsening assumptions, but can incorporate (even partial) knowledge about arbitrary dependencies of the coarsening on the values of some variables, which turns out to be also beneficial in context of testing.

Throughout the paper, we refer to the case of a coarse categorical response variable Y and one precisely observed categorical covariate X, but the results may be easily formulated in terms of cases with more than one categorical covariate. For sake of conciseness, the example refers to the case of a binary Y, where coarsening corresponds to missingness, but all results are applicable in a general categorical setting.

For this categorical setting, we characterize cases where SI may not only make parameters identifiable, but is also testable. Besides the investigation of the testability of SI, a main contribution of this paper is the construction of the likelihood-ratio test for this assumption. For this purpose, we give different representations of the hypotheses, illustrate the sensitivity of the test statistic

[1] When dealing with coarse data, it is important to distinguish *epistemic data imprecision* considered here, i.e. incomplete observations due to an imperfect measurement process, from *ontic data imprecision* (cf. [4]).

with regard to the deviation from the null hypothesis and study the asymptotic distribution of the test statistic to obtain a decision rule in dependence of the significance level. Straightforwardly, a test for a specific pattern of gSI is constructed.
Our paper is structured as follows: In Section 2 we introduce the technical framework and the running example based on the German Panel Study "'Labour Market and Social Security", which we also use for the illustration of both assumptions, CAR and SI, as well as gCAR and gSI, in Section 3. After sketching the crucial argument of identifiability issues and our estimation method as well as showing how the generally set-valued estimators are refined by implying CAR/gCAR or SI/gSI in Section 4, the obtained estimators are used to discuss the testability of both assumptions in Section 5. The likelihood-ratio test for SI is developed and then illustrated for the running example in Section 6, where the generalized view on subgroup independence is used to extend this hypothesis test to a more flexible version, including a test on partial information, in Section 7. Finally, Section 8 concludes with a summary and some additional remarks.

2. Coarse data: The basic viewpoint

Before we discuss the running example, let us explicitly formulate the technical framework in which our discussion of the coarsening assumptions, the estimation of parameters and the construction of the likelihood-ratio test is embedded. We approach the problem of coarse data in our categorical setting by distinguishing between a latent and an observed world: Let $(x_1,\ y_1),\dots,(x_n,\ y_n)$ be a sample of n independent realizations of a pair (X,Y) of categorical random variables with sample space $\Omega_X \times \Omega_Y$. Our basic goal consists of estimating the probabilities $\pi_{xy} = P(Y = y|X = x)$, where Y is regarded as response variable and X as covariate. Since the values of Y unfavorably can only be observed partially, i.e. subsets of Ω_Y instead of single elements may be observed, this variable is part of the latent world. Instead, we only observe a sample $(x_1,\ \mathfrak{y}_1),\dots,(x_n,\ \mathfrak{y}_n)$ of n independent realizations of the pair $(X,\mathcal{Y})$, where the random object $\mathcal{Y}$ with sample space $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$ constitutes the observed world. A connection between both worlds, and thus between π_{xy} and $p_{x\mathfrak{y}}$, is established via an observation model governed by the coarsening parameters $q_{\mathfrak{y}|xy} = P(\mathcal{Y} = \mathfrak{y}|X = x, Y = y)$ with $\mathfrak{y} \in \Omega_{\mathcal{Y}}$, $x \in \Omega_X$ and $y \in \Omega_Y$. Throughout the paper, we not only assume that the coarsening depends on individual i $(i = 1,\dots,n)$ via the values x and y exclusively, but also require error-freeness[2], i.e. $\mathfrak{y} \ni y$, and distinct parameters in the sense of Rubin [24]. An essential part of our argumentation is based on comparing the dimensions of the parameter space of the observed world Θ_{obs} and the parameter space of the latent world Θ_{lat} referring to parameters $p_{x\mathfrak{y}}$ as well as π_{xy} and $q_{\mathfrak{y}|xy},\ \mathfrak{y} \in \Omega_{\mathcal{Y}},\ x \in \Omega_X,\ y \in \Omega_Y$,

[2]This implies that Y is an almost sure selector of $\mathcal{Y}$ (in the sense of e.g. [18]).

Table 1: Data of the PASS example

UBII (X)	Income ($\mathcal{Y}$)	observed counts	total counts
0	$\{a\}$	$n_{0\{a\}} = 38$	$n_0 = 518$
	$\{b\}$	$n_{0\{b\}} = 385$	
	$\{a, b\}$	$n_{0\{a,b\}} = 95$	
1	$\{a\}$	$n_{1\{a\}} = 36$	$n_1 = 87$
	$\{b\}$	$n_{1\{b\}} = 42$	
	$\{a, b\}$	$n_{1\{a,b\}} = 9$	

respectively. Thereby, we choose the minimal possible parametrizations, in order to be precise about the dimension of the parameter space.
As the number of the coarsening parameters $l = k \cdot m \cdot (2^{m-1} - 1)$ increases considerably with $k = |\Omega_X|$ and $m = |\Omega_Y|$, for reasons of conciseness, we start by mainly confining ourselves to the discussion of the following running example, while the main contributions of this paper, considerations regarding identifiability and testability as well as the proposed hypothesis test, refer to the general categorical setting. In the example we consider a situation with $\Omega_X = \{0,\ 1\}$, $\Omega_Y = \{a,\ b\}$, and thus $\Omega_\mathcal{Y} = \{\{a\},\ \{b\},\ \{a, b\}\}$, where "$\{a, b\}$" denotes the only coarse observation, which corresponds to a missing one in this case. It is introduced in the following box:

Running example:
The German Panel Study "Labour Market and Social Security" (PASS, [28], wave 5, 2011) deals with the expected low response to the income question by follow-up questions for non-respondents, starting from providing rather large income classes that are then narrowed step by step. In this way, answers with different levels of coarseness are received by simultaneously respecting privacy. For convenience, we consider only that income question where respondents are required to report if their income is < 1000€ (a) or ≥ 1000€ (b) ($y \in \{a, b\} = \Omega_Y$). Some respondents gave no suitable answer, such that only values of $\mathcal{Y}$ are observable ($\mathfrak{y} \in \{\{a\},\ \{b\},\ \{a, b\}\} = \Omega_\mathcal{Y}$). The receipt of the so-called Unemployment Benefit II (UBII) is used as covariate with $x \in \{0\ (\text{no}), 1\ (\text{yes})\}$. A summary of the data is given in Table 1.

In the sequel, we distinguish between the case of making no assumptions beyond the basic setting formulated in this section and assuming CAR/gCAR and SI/gSI, both strongly restricting the coarsening.

3. Coarsening models

3.1. Coarsening at random and its generalized version

Heitjan and Rubin ([9]) consider maximum likelihood estimation in coarse data situations by deriving assumptions simplifying the likelihood. These assumptions – CAR and distinct parameters – make the coarsening *ignorable* (e.g.

[15]). The CAR assumption requires constant coarsening parameters $q_{\mathfrak{y}|xy}$, regardless which true value y is underlying subject to the condition that it matches with the fixed observed value $\mathfrak{y}$. In this way, the coarsening mechanism is "uninformative" about the true underlying value of Y. Referring to the case where the information of a covariate is available, we consider a slightly changed notion of the CAR assumption by additionally conditioning on the value of the covariate. Since this covariate might generally have an influence on the coarsening process, we assume CAR for each subgroup. A geometric representation and an appealing way to model CAR, also in case of a large $|\Omega_{\mathcal{Y}}|$, is given in [7].
The strong limitation of the CAR assumption is also evident in the running example. Under CAR, which coincides here with MAR, the probability of giving no suitable answer is taken to be independent of the true income category in both subgroups split by the receipt of UBII, i.e.

$$q_{\{a,b\}|0a} = q_{\{a,b\}|0b} \quad \text{and} \quad q_{\{a,b\}|1a} = q_{\{a,b\}|1b}.$$

Generally, CAR could be quite problematic in this context, as practical experiences show that reporting missing or coarsened answers is notably common in specific income groups (e.g. [14], [27]).

A generalization (extending Nordheim's [19] proposals for MAR to CAR) of the CAR assumption, allows a more flexible incorporation of coarsening assumptions. We refer to this generalization as *generalized CAR* (gCAR) and express assumptions about the ratios of coarsening parameters for given subgroups, choosing these ratios as

$$R_{0,a,b,\{a,b\}} = \frac{q_{\{a,b\}|0a}}{q_{\{a,b\}|0b}} \quad \text{and} \quad R_{1,a,b,\{a,b\}} = \frac{q_{\{a,b\}|1a}}{q_{\{a,b\}|1b}}$$

in our missing data situation, where $R_{0,a,b,\{a,b\}} = R_{1,a,b,\{a,b\}} = 1$ represents the special case of CAR/MAR. More generally, $k \cdot (m \cdot (2^{m-1} - 1) - 1)$ ratios of coarsening parameters are defined, for each subgroup one ratio less than the general number of coarsening parameters (cf. Section 2). In most cases, it might be difficult to justify knowledge about the exact value of the ratios, but former studies or material considerations might provide a rough evaluation of their magnitude. In this way, for a given subgroup partial assumptions as "respondents from the high income class tend to give a coarse answer more likely" may be expressed by choosing $R_{0,a,b,\{a,b\}},\ R_{1,a,b,\{a,b\}} \in [0, 1[$, which can be covered in a powerful way in the likelihood approach (cf. [20]) also underlying our paper.

3.2. Subgroup independence and its generalized version

If the data are missing not at random (MNAR) [24], commonly the missingness process is modelled by including parametric assumptions (e.g. [9]), or a cautious procedure is chosen ending up in set-valued estimators (cf. e.g. [6], [31]). For the categorical case, it turns out that there is a special case of MNAR,

in which single-valued estimators can be obtained without parametric assumptions. For motivating this case, one can further differentiate MNAR, distinguishing between the situation where missingness depends on both the values of the response Y and the covariate X and the situation where it depends on the values of Y only. Referring to the related coarsening case, the latter case corresponds to SI sketched in [20], and studied in detail here. This independence from the covariate value shows, beside CAR, an alternative kind of coarsening. Again, one should use this assumption cautiously: Under SI, in our example giving a coarse answer is then taken to be independent of the receipt of UBII given the value of Y, i.e.

$$q_{\{a,b\}|0a} = q_{\{a,b\}|1a} \quad \text{and} \quad q_{\{a,b\}|0b} = q_{\{a,b\}|1b}.$$

In practice, a different coarsening behaviour with regard to the income question is expected from respondents receiving and not receiving UBII, such that this assumption turns out to be doubtful.

Again, a generalization, in the following called *generalized subgroup independence* (gSI), is conceivable by now considering the ratios

$$R_{0,1,a,\{a,b\}} = \frac{q_{\{a,b\}|0a}}{q_{\{a,b\}|1a}} \quad \text{and} \quad R_{0,1,b,\{a,b\}} = \frac{q_{\{a,b\}|0b}}{q_{\{a,b\}|1b}}, \tag{1}$$

where assuming $R_{0,1,a,\{a,b\}} = R_{0,1,b,\{a,b\}} = 1$ corresponds to SI. In the general case, gSI relies on $(k-1) \cdot m \cdot (2^{m-1} - 1)$ ratios

$$R_{x,x',y,\mathfrak{y}} = \frac{q_{\mathfrak{y}|xy}}{q_{\mathfrak{y}|x'y}}\,, \qquad x, x' \in \Omega_X, y \in \Omega_Y, \mathfrak{y} \in \Omega_{\mathcal{Y}}, \tag{2}$$

i.e. for each compatible pair of $\mathfrak{y}$ and y one ratio less than the number of subgroups is needed. By e.g. selecting $R_{0,1,a,\{a,b\}},\ R_{0,1,b,\{a,b\}} \in]1, \infty[$ for a given true income group, partial information in the sense that "respondents who do not receive UBII tend to give coarse answers more likely" can be expressed, which again can be included into the likelihood-based approach explained in the next section. These ratios will be the starting point for the generalized hypothesis test in Section 7.

4. Identifiability and estimation: General case, CAR/gCAR and SI/gSI

This section recalls some important aspects of our approach developed in [20] by sketching the basic idea of the therein considered cautious, likelihood-based estimation technique and giving the obtained estimators with and without the assumptions in focus. Beyond that, we confirm that CAR/gCAR is point-identifying and elaborate a criterion that may allow point-identified parameters under SI/gSI .

4.1. Basic argument of the estimation method

To estimate $(\pi_{xy})_{x\in\Omega_X, y\in\Omega_Y}$ of the latent world, basically three steps are accomplished. Firstly, we determine the maximum likelihood estimator (MLE) $(\hat{p}_{x\mathfrak{y}})_{x\in\Omega_X,\mathfrak{y}\in\Omega_\mathcal{Y}}$ in the observed world based on all $n = \sum_{x\in\Omega_X} n_x$ observations with $n_x > 0,\ x \in \Omega_X$. Since the counts $(n_{x\mathfrak{y}})_{x\in\Omega_X,\mathfrak{y}\in\Omega_\mathcal{Y}}$ are multinomially distributed, under some regularity conditions satisfied here the MLE is uniquely obtained by the relative frequencies of the respective categories (cf. [23]), coarse categories treated as own categories. Secondly, we connect the parameters of both worlds by a mapping

$$\begin{aligned} \Phi:\ \Theta_{lat} &\to \Theta_{obs}, \\ \theta_{lat} &\to \theta_{obs} \end{aligned} \tag{3}$$

expressing the observation process, where Θ_{lat} and Θ_{obs} are the parameter space of the latent and the observed world, respectively. In the situation of the example we consider

$$\begin{aligned} \theta_{lat} &= (\pi_{0a},\ q_{\{a,b\}|0a},\ q_{\{a,b\}|0b},\ \pi_{1a},\ q_{\{a,b\}|1a},\ q_{\{a,b\}|1b})^T \quad \text{and} \\ \theta_{obs} &= (p_{0\{a\}},\ p_{0\{b\}},\ p_{1\{a\}},\ p_{1\{b\}})^T\ , \end{aligned} \tag{4}$$

thus obtaining $dim(\Theta_{lat}) = 6$ and $dim(\Theta_{obs}) = 4$ as dimensions of the respective parameter spaces. Both dimensions will govern our argumentation with regard to identifiability and testability later on. The mapping Φ can be seperated into components Φ_x corresponding to subgroup x, $x \in \Omega_X$. For our example, we obtain

$$\Phi_x \begin{pmatrix} \pi_{xa} \\ q_{\{a,b\}|xa} \\ q_{\{a,b\}|xb} \end{pmatrix} = \begin{pmatrix} \pi_{xa} \cdot (1 - q_{\{a,b\}|xa}) \\ (1 - \pi_{xa}) \cdot (1 - q_{\{a,b\}|xb}) \end{pmatrix} = \begin{pmatrix} p_{x\{a\}} \\ p_{x\{b\}} \end{pmatrix}\ , \tag{5}$$

$x \in \{0, 1\}$, determined by utilizing the law of total probability. Thirdly, by the invariance of the likelihood under parameter transformations, we may incorporate the parametrization in terms of π_{xy} and $q_{\mathfrak{y}|xy}$ into the likelihood of the observed world. Since the mapping Φ is generally not injective, we obtain multiple combinations of estimated latent variable distributions and estimated coarsening parameters, all leading to the same maximum value of the likelihood. These set-valued estimators

$$\hat{\Gamma} = \{\hat{\theta}_{lat} \mid \Phi(\hat{\theta}_{lat}) = \hat{\theta}_{obs}\}, \tag{6}$$

with $\hat{\theta}_{lat}$ and $\hat{\theta}_{obs}$ as the MLE's of θ_{lat} and θ_{obs}, respectively, are here illustrated by building the one dimensional projections, which are represented as intervals, in the example leading to

$$\hat{\pi}_{xa} \in \left[\frac{n_{x\{a\}}}{n_x},\ \frac{n_{x\{a\}} + n_{x\{a,b\}}}{n_x}\right], \quad \hat{q}_{\{a,b\}|xy} \in \left[0,\ \frac{n_{x\{a,b\}}}{n_{x\{y\}} + n_{x\{a,b\}}}\right], \tag{7}$$

with $x \in \{0,\ 1\}$ and $y \in \{a,\ b\}$. Points in these intervals are constrained by the relationships in Φ. These estimators and the corresponding intervals may be refined by including assumptions about the coarsening justified from the application standpoint (in the spirit of [16]). Very strict assumptions may induce point-identified parameters, as estimation under CAR or SI in the categorical case shows.[3]

4.2. Basic argument of studying the identifiability

Discussing identifiability, we turn to the general case with $k = |\Omega_X|$ and $m = |\Omega_Y|$, using the setting of the example only for reasons of illustration. In Section 4.3 and 4.4, we briefly study the cases in which CAR/gCAR and SI/gSI can be point-identifying. The mapping Φ is definitely not injective, if $dim(\Theta_{obs}) < dim(\Theta_{lat})$. In this way, we need the degree of freedom under the assumption in focus (here generally noted as *aspt*), i.e.

$$df^{aspt} = dim(\Theta_{obs}) - dim(\Theta_{lat}^{aspt}), \tag{8}$$

to be non-negative, in order to be able to make Φ injective and thus to receive point-valued estimators under *aspt* at all. Including an assumption into the estimation problem has an impact on $dim(\Theta_{lat})$ only, while $dim(\Theta_{obs})$ stays to be equal to $k \cdot (2^m - 2)$ independently of the fact if the assumption of CAR/gCAR or SI/gSI is included.[4]

4.3. Identifiability and estimation under CAR/gCAR

Thus, we study the possibility of achieving point-valued estimators under CAR by checking whether $df^{CAR} \geq 0$ is satisfied (cf. (8)). Within each subgroup, every coarse category requires one coarsening parameter only, wherefore additionally to the $k \cdot (m-1)$ parameters representing the latent variable distribution, $k \cdot (2^m - 1 - m)$ coarsening parameters are estimated. In this way,

$$df^{CAR} = k \cdot (2^m - 2) - [k \cdot (m-1) + k \cdot (2^m - 1 - m)] = 0$$

is obtained, pointing to the well-known result that CAR is generally point-identifying.
By assuming CAR, i.e. restricting the set of possible coarsening mechanisms to $q_{\{a,b\}|xa} = q_{\{a,b\}|xb}$ with $x \in \{0,\ 1\}$, we receive the point-valued estimators

$$\hat{\pi}_{xa}^{CAR} = \frac{n_{x\{a\}}}{n_{x\{a\}} + n_{x\{b\}}}, \quad \hat{q}_{\{a,b\}|xa}^{CAR} = \hat{q}_{\{a,b\}|xb}^{CAR} = \frac{n_{x\{a,b\}}}{n_x}. \tag{9}$$

Interpreting these results, under this type of coarsening, $\hat{\pi}_{xa}$ corresponds to the proportion of $\{a\}$-observations in subgroup x ignoring all coarse values and

[3] Identifiability may not only be obtained by assumptions on the coarsening: e.g. for discrete graphical models with one hidden node, conditions based on the associated concentration graph are used in [26].

[4] For every subgroup, $|\Omega_Y| - 1$ parameters of the observed world have to be estimated (cf. Section 2).

Table 2: Minimum number of subgroups k for a given m

m	2	3, 4, 5	6, 7	8, 9
minimum k	2	3	4	5

$\hat{q}_{\{a,b\}|xa} = \hat{q}_{\{a,b\}|xb}$ is the proportion of observed $\{a, b\}$ in subgroup x.
Since the dimension of the parameter space under gCAR always corresponds to $dim(\Theta_{lat}^{CAR})$, we may receive point-valued estimators for the general version as well. For fixed values of $R_{0,a,b,\{a,b\}}$ and $R_{1,a,b,\{a,b\}}$, the parameter of main interest π_{xa} is point-identified, wherefore the ratios may be regarded as sensitivity parameters in the sense of Kenward, Goetghebeur and Molenberghs [13] (cf. [21]). Partial assumptions, as e.g. $R_{0,a,b,\{a,b\}},\ R_{1,a,b,\{a,b\}} \in [0, 1[$, can be included into the estimation by taking the collection of all point-valued results obtained by the estimation under fixed ratios that are compatible with this assumption (cf. [21]).

4.4. Identifiability and estimation under SI/gSI

If SI is incorporated into the estimation, $df^{SI} = dim(\Theta_{obs}) - dim(\Theta_{lat}^{SI})$ is not necessarily non-negative. Since the value of the subgroup does not play any role for the coarsening under SI, the number of coarsening parameters corresponds to the one in the homogeneous case, i.e. $m \cdot (2^{m-1} - 1)$, thus receiving $dim(\Theta_{lat}^{SI}) = k \cdot (m - 1) + m \cdot (2^{m-1} - 1)$. Solving

$$df^{SI} = k \cdot (2^m - 2) - [k \cdot (m - 1) + m \cdot (2^{m-1} - 1)] \geq 0$$

for k, we obtain the condition

$$k \geq \frac{m \cdot (2^{m-1} - 1)}{2^m - m - 1} \,. \tag{10}$$

that has to be satisfied to concede point-valued estimators at all. A first impression about the minimum number of necessary subgroups can be achieved by considering Table 2.

In this paper we focus on the setting where $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$ with all categories observable, but there are data situation where only specific coarse categories, i.e. a strict subset of $\mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$ can be observed, i.e. we are in fact considering a space $\tilde{\Omega}_{\mathcal{Y}} \subsetneq \Omega_{\mathcal{Y}}$. In these cases, the number $v = |\tilde{\Omega}_{\mathcal{Y}}|$, instead of $|\Omega_{\mathcal{Y}}| = 2^m - 1$, has to be included into df^{SI}, such that the minimum number of subgroups generally can no longer be expressed in terms of m exclusively. Although the prominent missing data case, which is of high practical relevance, is classified as a case of that kind, we do not need v to elaborate the subgroup criterion in this situation. We are concerned with m precise categories and one missing category, wherefore $|\tilde{\Omega}_{\mathcal{Y}}| = m+1$. The number of subgroups k has to be greater

or equal to m, since in this case

$$\begin{aligned} dim(\Theta_{obs}) &= k \cdot (m + 1 - 1) = k \cdot m \\ dim(\Theta_{lat}^{SI}) &= k \cdot (m - 1) + m, \quad \text{and thus} \\ df^{SI} &= k \cdot m - (k \cdot (m - 1) + m) \geq 0 \Leftrightarrow k \geq m \ . \end{aligned}$$

In the setting of our example, there are two subgroups available, which corresponds to the lower bound in (10), such that the respective condition is satisfied. This is in line with the result that under rather weak regularity conditions, namely $\pi_{0a} \neq \pi_{1a}$,[5] $\pi_{0a} \notin \{0, 1\}$, and $\pi_{1a} \notin \{0, 1\}$ for $x \in \{0, 1\}$, under SI the mapping Φ becomes injective (a proof is given in [21, p. 17, 20]). Hence, we obtain point-valued estimators

$$\begin{aligned} \hat{\pi}_{xa}^{SI} &= \frac{n_{x\{a\}}}{n_x} \, \frac{n_0 \, n_{1\{b\}} - n_{0\{b\}} \, n_1}{n_{0\{a\}} \, n_{1\{b\}} - n_{0\{b\}} \, n_{1\{a\}}}, \\ \hat{q}_{\{a,b\}|xa}^{SI} &= \frac{n_{0\{a,b\}} \, n_{1\{b\}} - n_{0\{b\}} \, n_{1\{a,b\}}}{n_0 \, n_{1\{b\}} - n_{0\{b\}} \, n_1}, \\ \hat{q}_{\{a,b\}|xb}^{SI} &= \frac{n_{0\{a,b\}} \, n_{1\{a\}} - n_{0\{a\}} \, n_{1\{a,b\}}}{n_0 \, n_{1\{a\}} - n_{0\{a\}} \, n_1}, \end{aligned} \tag{11}$$

provided they are well-defined and inside $[0, \ 1]$.
Turning to gSI again, all findings concerning the identifiability under SI are equally applicable to gSI, since $dim(\Theta_{lat}^{gSI})$ corresponds to $dim(\Theta_{lat}^{SI})$. By including partial knowledge about the ratios in (2), the estimators in (6) can again be refined substantially.

5. On the testability of CAR and SI

Due to the potentially substantial bias of $\hat{\pi}_{xy}$ if CAR or SI are wrongly assumed (cf. e.g. [21, p. 15, 18]), testing these assumptions would be of particular interest. Although it is already established that without additional information it is not possible to test whether the CAR condition holds (e.g. [16, p. 29]), it may be insightful, in particular in the light of Section 5.2, to address this impossibility in the context of the example.

5.1. Testability of CAR and gCAR

A closer consideration of (9) already indicates that CAR can never be rejected without including additional assumptions about the coarsening. This point is illustrated in Fig. 1 by showing the interaction between points in the intervals in (7). Spoken for the situation of the example: The coarsening scenario, where respondents from the low income category and respondents from

[5] The case of $\pi_{0a} = \pi_{1a}$ represents the homogeneous case, where multiple solutions result [20].

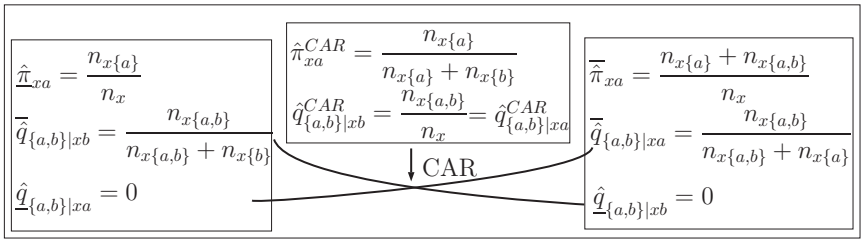

Figure 1: Since the relationships expressed via Φ in (5) have to be met, only specific points from the estimators in (7) are combinable, ranging from $(\underline{\hat{\pi}}_{xa},\ \underline{\hat{q}}_{\{a,b\}|xa},\ \overline{\hat{q}}_{\{a,b\}|xb})$ to $(\overline{\hat{\pi}}_{xa},\ \overline{\hat{q}}_{\{a,b\}|xa},\ \underline{\hat{q}}_{\{a,b\}|xb})$ with the CAR case always included.

the high income category tend to give coarse answers in the same way, can generally not be excluded. The in this sense uninformative coarsening, which here just ignores all coarse values, is always a possible scenario included in the estimators in (6).

For the example, under CAR we obtain

$$\hat{\pi}^{CAR}_{0a} = 0.09, \quad \hat{\pi}^{CAR}_{1a} = 0.46, \quad \hat{q}^{CAR}_{\{a,b\}|0y} = 0.18, \quad \hat{q}^{CAR}_{\{a,b\}|1y} = 0.10, \quad y \in \{a,b\},$$

which may not be excluded from the set-valued estimators, and also the intervals

$$\begin{aligned} \hat{\pi}_{0a} &\in [0.073,\ 0.26], \quad & \hat{q}_{\{a,b\}|0a} &\in [0,\ 0.71], \quad & \hat{q}_{\{a,b\}|0b} &\in [0,\ 0.20], \\ \hat{\pi}_{1a} &\in [0.41,\ 0.52], \quad & \hat{q}_{\{a,b\}|1a} &\in [0,\ 0.20], \quad & \hat{q}_{\{a,b\}|1b} &\in [0,\ 0.18], \end{aligned}$$

unless further assumptions as e.g. "respondents from the high income group tend to give coarse answers more likely" are justified. In the same way, specific dependencies of the coarsening process on the true underlying value in the sense of gCAR are generally not excludable, and thus the generalization neither can be tested. Nevertheless, there are several approaches that show how testability of CAR is achieved by distributional assumptions (e.g. [11]), like the naive Bayes assumption (cf. [12]), or by the inclusion of instrumental variables (cf. [1]).

5.2. Testability of SI and gSI

Our considerations concerning the testability of SI are mainly based on two findings from Section 4.4. There, we firstly elaborated the condition in (10) as a necessary condition to be able to obtain point-valued estimators at all. In this way, we could not generally conclude point-valued estimators as it works in the case of CAR. Similarly, this applies in context of studying the testability of SI, where two cases have to be distinguished: The case of $df^{SI} < 0$, where SI can not be tested in the sense that the "test statistic" is completely degenerate, and $df^{SI} \geq 0$, where we can test it indeed. Secondly, the (unconstrained)[6]

[6]probability restrictions are not explicitly included

estimators in (11) already indicated that – depending on the data situation – results partly outside the interval [0, 1] are conceivable. In order to illustrate this point, we apply the estimators in (11) to the example. We obtain the unconstrained estimates

$$\hat{\pi}_{0a}^{SI} = 0.070, \quad \hat{\pi}_{1a}^{SI} = 0.40, \quad \hat{q}_{\{a,b\}|xa}^{SI} = -0.04, \quad \hat{q}_{\{a,b\}|xb}^{SI} = 0.20, \quad x \in \{0,1\},$$

revealing that there are data situations that might hint to (partial) incompatibility with SI. Informally spoken, the reason for this indication of incompatibility can be explained as follows: The subgroup specific coarse observations have to be produced by the compatible, precise values within the considered subgroup. This might be prevented under the assumption of SI, representing a too strict coarsening rule in certain observed data situations, wherefore SI might be testable.
Thus, if restricted to the case with sufficiently many subgroups, two situations have to be distinguished, already suggesting the two possible test decisions: Either the likelihood optimized under SI achieves the computational maximum obtained by $\Phi^{-1}(\hat{\theta}_{obs})$, where Φ^{-1} is the inverse of Φ, or not. In the second case the optimization under SI induces a lower value of the likelihood compared to the case of refraining from strict coarsening assumptions and using those mentioned in Section 2 only. Thus, spoken for our example, the unconstrained estimators under SI, which are the unique (Φ is injective) inverse image of the MLE's $\hat{p}_{x\{a\}}$ and $\hat{p}_{x\{b\}}$, are partly outside the interval [0, 1]. In context of the determination of the test statistic in Section 6.2, we come back to this point. If the criterion given in (10) is satisfied, gSI is testable as well, where we devote ourselves to this question in Section 7.

6. Likelihood-ratio test for SI

6.1. The hypotheses

If sufficient subgroups are available in the sense that the condition in (10) is met, a statistical test for the following hypotheses can be constructed in the categorical case:

$$\begin{aligned} H_0 &: q_{\mathfrak{y}|xy} = q_{\mathfrak{y}|x'y} \text{ for all } \mathfrak{y} \in \Omega_{\mathcal{Y}},\ x, x' \in \Omega_X,\ y \in \Omega_Y, \\ H_1 &: q_{\mathfrak{y}|xy} \neq q_{\mathfrak{y}|x'y} \text{ for some } \mathfrak{y} \in \Omega_{\mathcal{Y}},\ x, x' \in \Omega_X,\ y \in \Omega_Y. \end{aligned} \tag{12}$$

In order to derive the distribution of the test statistic in Section 6.3, a restatement of the hypotheses in terms of the parameters of the observed world proves to be beneficial, which we here consider for the setting of the example:

$$\begin{aligned} H_0^* &: (p_{0\{a\}} \cdot p_{1\{a,b\}} - p_{1\{a\}} \cdot p_{0\{a,b\}}) \cdot (p_{0\{b\}} \cdot p_{1\{a,b\}} - p_{1\{b\}} \cdot p_{0\{a,b\}}) \leq 0 \\ H_1^* &: (p_{0\{a\}} \cdot p_{1\{a,b\}} - p_{1\{a\}} \cdot p_{0\{a,b\}}) \cdot (p_{0\{b\}} \cdot p_{1\{a,b\}} - p_{1\{b\}} \cdot p_{0\{a,b\}}) > 0. \end{aligned}$$

To explain the conditions therein, Figure 2 shows informally the subgroup

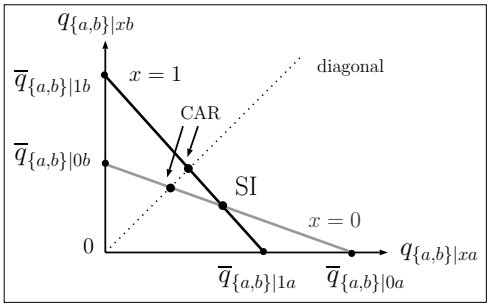

Figure 2: The gray and black solid lines symbolize all coarsening parameters within Γ (cf. (6)) for subgroup $x = 0$ and $x = 1$, respectively. While the CAR case is represented by the intersection points with the diagonal, the SI assumption is satisfied at the intersection point of both lines.

specific coarsening parameters $q_{\{a,b\}|xa}$ and $q_{\{a,b\}|xb}$ ranging from 0 to

$$\overline{q}_{\{a,b\}|xa} = \frac{p_{x\{a,b\}}}{p_{x\{a,b\}} + p_{x\{a\}}}\,, \qquad \overline{q}_{\{a,b\}|xb} = \frac{p_{x\{a,b\}}}{p_{x\{a,b\}} + p_{x\{b\}}} \tag{13}$$

respectively, $x \in \{0,\ 1\}$. The assumption of SI is only achievable, if both lines intersect, i.e.

$$\overline{q}_{\{a,b\}|1b} - \overline{q}_{\{a,b\}|0b} \geq 0 \quad \text{and} \quad \overline{q}_{\{a,b\}|1a} - \overline{q}_{\{a,b\}|0a} \leq 0\,, \tag{14}$$

or the other way round. After replacing the upper bounds for the coarsening parameters in (14) by (13) and making some little rearrangements, it turns out that an intersection requires

$$\begin{aligned} &p_{0\{b\}} \cdot p_{1\{a,b\}} - p_{1\{b\}} \cdot p_{0\{a,b\}} \geq 0 \qquad \text{and} \\ &p_{0\{a\}} \cdot p_{1\{a,b\}} - p_{1\{a\}} \cdot p_{0\{a,b\}} \leq 0 \quad , \end{aligned}$$

or the other way round, which corresponds to the null hypothesis H_0^*. To receive a first impression of the situations that are in accordance with H_0^*, Figure 6 in appendix A might be helpful, depicting over a grid of parameters $p_{0\{a\}}$, $p_{1\{a\}}$, $p_{0\{a,b\}}$ and $p_{1\{a,b\}}$, whether the condition in H_0^* is satisfied or not. Thereby, we also differentiate between a boundary and a non-boundary fulfillment, since this will be of importance in Section 6.3.

6.2. The test statistic

Since we here consider a likelihood-based approach directly based on the realizations in the observed level, applying a corresponding likelihood-ratio test is natural. Thus, our test for the general hypotheses H_0 and H_1 in (12) can be based on the classical test statistic (e.g. [30])

$$T = -2 \cdot \ln(\Lambda(\mathfrak{y}_1, \ldots, \mathfrak{y}_n, x_1, \ldots, x_n)) \tag{15}$$

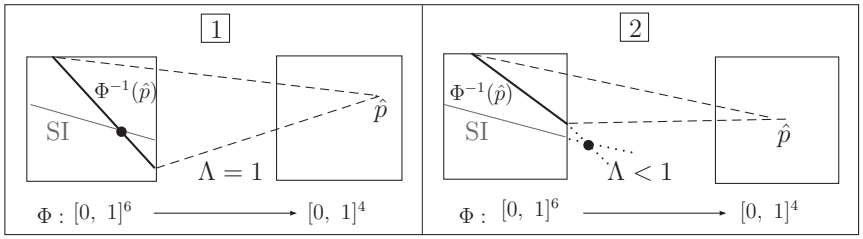

Figure 3: The impact on Λ of two substantially differing data situations is illustrated.

of this test with likelihood ratio

$$\Lambda(\mathfrak{y}_1, \ldots, \mathfrak{y}_n, x_1, \ldots, x_n) = \frac{\sup_{H_0} L(\theta_{lat} || \mathfrak{y}_1, \ldots, \mathfrak{y}_n, x_1, \ldots, x_n)}{\sup_{H_0 \cup H_1} L(\theta_{lat} || \mathfrak{y}_1, \ldots, \mathfrak{y}_n, x_1, \ldots, x_n)}, \tag{16}$$

(cf. (4)).[7] In fact, simulation studies corroborate the decrease of Λ with deviation from SI (cf. [21, p. 19]). The sensitivity of Λ with regard to the test considered here is also illustrated informally in Fig. 3 by depicting Φ in (3) for two data situations, where only the second one gives evidence against SI. The gray line symbolizes all arguments satisfying SI, while the bold line represents all arguments maximizing the likelihood if only the assumptions mentioned in Section 2 are imposed (i.e. all values in (7) compatible with each other). The intersection of both lines represents the values in (11), and if it is included in the domain of Φ (cf. first case of Fig. 3), the same maximal value of the likelihood is obtained regardless of including SI or not, resulting in $\Lambda = 1$, and thus $T = 0$. An intersection outside the domain (cf. second case of Fig. 3) induces a lower value of the likelihood under SI, also reflected in $\Lambda < 1$, causing $T > 0$. For the example one obtains $\Lambda \approx 0.93$ and $T \approx 0.14$, indicating a slight evidence against SI based on a direct interpretation of the test statistic, while the determination of a general decision rule depending on significance level α is the goal of the next section.

6.3. Test decision

In case of the likelihood-ratio test, the asymptotic distribution of the test statistic under the null hypothesis is typically given by a χ^2-distribution with degrees of freedom df, providing the basis for the critical value, namely its $(1-\alpha)$-quantile, that is used for the test decision (cf. [30]). Here, it turns out that the degrees of freedom df^{SI}, considered in Section 4.4, crucially determine the type of the asymptotic distribution. We have to differentiate between the situation $df^{SI} = 0$ and $df^{SI} > 0$, whereas subgroup independence is not testable under $df^{SI} < 0$ (cf. Section 5.2). While the quantile $\chi^2_{df,1-\alpha}$, with $df = df^{SI}$,

[7] While the denominator of Λ can be obtained using any values in (7) compatible with each other, the numerator must in general be calculated by numerical optimization. Alternatives to this statistic would include the construction of uncertainty regions, in the spirit of [29].

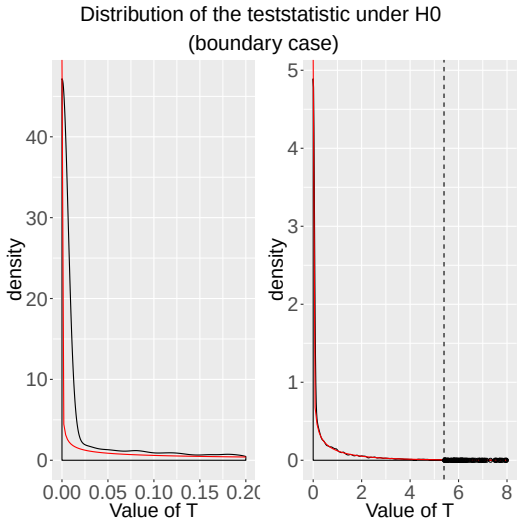

Figure 4: For an exemplary boundary case, the empirical distribution of the test statistic T under H_0 (black line) is compared to the theoretical asymptotic distribution (red line).

gives the critical value in case of $df^{SI} > 0$, the critical value is calculated based on a specific asymptotic distribution in case of $df^{SI} = 0$.
We now address this special case by focusing on the setting of the example, also revealing $df^{SI} = 0$. Here, two fundamentally differing situations, the boundary and the non-boundary case, have to be distinguished. By referring to the hypothesis H_0^* again, one can note that the boundary case is attained if either $p_{0\{a\}} \cdot p_{1\{a,b\}} = p_{1\{a\}} \cdot p_{0\{a,b\}}$ or $p_{0\{b\}} \cdot p_{1\{a,b\}} = p_{1\{b\}} \cdot p_{0\{a,b\}}$ (but not both, which would correspond to the case where both solid lines in Figure 2 completely overlap). In the non-boundary case, the value of the test statistic is asymptotically degenerate at $T = 0$ (as implied by the consistency of $\hat{\theta}_{obs}$), inducing that the null hypothesis generally cannot be rejected. Against this, according to Chernoff ([3]), in the boundary case

$$T \underset{H_0}{\overset{a}{\sim}} 0.5 \cdot \delta_0 + 0.5 \cdot \chi_1^2 \,, \tag{17}$$

is obtained, where δ_0 is the Dirac distribution at zero. In words, the asymptotic distribution of T in the boundary case is that of a random variable which is zero half of the time and has a χ^2-distribution with one degree of freedom the other half of the time.
Since we do not know, whether we are in the boundary case or not, we always go for the worst case scenario in case of $df^{SI} = 0$ and take the critical value of the boundary case, thus generally referring to the distribution in (17). Taking the $(1-\beta)$-quantile of the χ_1^2-distribution as critical value, the probability of wrongly rejecting H_0 is $0.5 \cdot \beta$, since one does not reject H_0 for sure in the δ_0 part of the mixture distribution. Therefore, in the boundary case β has to be

Table 3: Distribution of T under H_0 in dependence of k and m

$m = 2$	$m = 3$	$m \geq 4$
$k = 1: \ \delta_0$	$k \leq 2: \ \delta_0$	$k \leq \lfloor \frac{m}{2} \rfloor: \ \delta_0$
$k = 2: \ 0.5 \cdot \delta_0 + 0.5 \cdot \chi^2_1$	$k \geq 3: \chi^2_{df^{SI}}$	$k \geq \lceil \frac{m+1}{2} \rceil: \ \chi^2_{df^{SI}}$
$k \geq 3: \ \chi^2_{df^{SI}}$		

chosen as $2 \cdot \alpha$, thus obtaining the critical value $\chi^2_{1,1-2\cdot\alpha}$.[8]
Applying the decision rule to the data of the example, H_0 cannot be rejected at significance level $\alpha = 0.01$, since the value of the test statistic $T \approx 0.14$ falls below the critical value 5.4, i.e. the $(1 - 2 \cdot \alpha)$-quantile of the χ^2_1-distribution.

To quickly illustrate the finite sample distribution of the test, we calculated the test statistic T for $M = 10\,000$ simulation runs referring to the exemplary boundary case with $p_{0\{a\}} = 0.1$, $p_{0\{b\}} = 0.7$, $p_{0\{a,b\}} = 0.2$, $p_{1\{a\}} = 0.2$, $p_{1\{b\}} = 0.4$ and $p_{1\{a,b\}} = 0.4$. The zoomed extract in the left part of Figure 4 shows the theoretical asymptotic distribution in (17) as well as the empirical distribution of the obtained values for the test statistic, where both lines are quite close indeed. The vertical line in the right part marks the critical value determined by the $\chi^2_{1,1-2\cdot\alpha}$-quantile (here 5.4), where we choose $\alpha = 0.01$. By calculating the percentage of values exceeding this threshold (illustrated as points in the right part of Figure 4), we obtain the estimated type I error of ≈ 0.0110, basically complying with the level α.
To allow a first check of testability in more general cases, we make a convenient upper assessment for the minimum number of subgroups in (10), receiving

$$k \geq \frac{m \cdot (2^{m-1} - 1)}{2^m - m - 1} = \frac{m}{2} \frac{1 - \frac{1}{2}^{m-1}}{1 - (m-1)(\frac{1}{2})^m} > \frac{m}{2}.$$

Table 3 shows the distribution of the test statistic under the null hypothesis for a given number of subgroups and categories of the variable of interest.

7. Generalized version of the test

By using the ratios $R_{x,x',y,\mathfrak{y}}$ in (2), the hypothesis test for SI may be generalized straightforwardly for gSI. For this purpose, we establish hypotheses

$$\begin{aligned} H_0 &: \ q_{\mathfrak{y}|xy} = R_{x,x',y,\mathfrak{y}} \cdot q_{\mathfrak{y}|x'y}, \ \text{for all } \mathfrak{y} \in \Omega_{\mathcal{Y}},\ x, x' \in \Omega_X,\ y \in \Omega_Y, \\ H_1 &: \ q_{\mathfrak{y}|xy} \neq R_{x,x',y,\mathfrak{y}} \cdot q_{\mathfrak{y}|x'y}, \ \text{for some } \mathfrak{y} \in \Omega_{\mathcal{Y}},\ x, x' \in \Omega_X,\ y \in \Omega_Y. \end{aligned} \quad (18)$$

[8] Notice that this is similar to the one-sided t-test; in fact, the t-tests are likelihood-ratio tests: the two-sided ones have the standard asymptotic distribution χ^2_1 (since the t-distribution tends to the normal one), while the one-sided t-test have the (worst-case) asymptotic distribution given in (17).

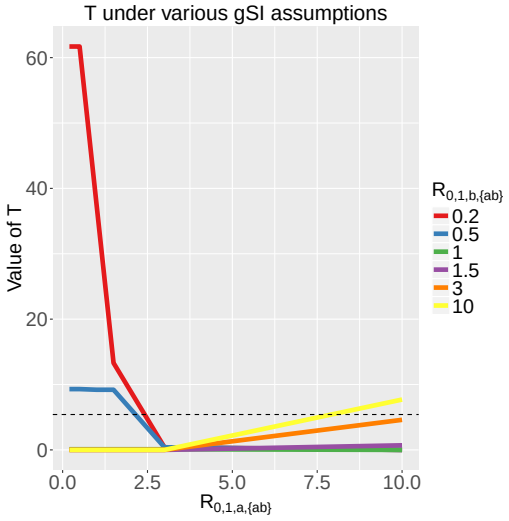

Figure 5: The figure gives some indication of the test decision for a selection of coarsening scenarios. Each solid line represents the value of the test statistic in dependence of $R_{0,1,a,\{a,b\}}$ for a given value of $R_{0,1,b,\{a,b\}}$, where only the points on the chosen grid are directly interpretable, the other values on the lines give rough information about the actual value of T only.

As a test statistic we again utilize T in (15), where the numerator of the likelihood ratio Λ in (16) is the only component that changes: Instead of optimizing the likelihood under SI, we refer to a specific coarsening scenario expressed by assuming certain values for the ratios $R_{x,x',y,\mathfrak{y}}$.

To illustrate this test, we refer to the PASS data example and the ratios in (1). Thus, we focus on the hypotheses

$$H_0: \; q_{\{a,b\}|0a} = R_{0,1,a,\{a,b\}} \cdot q_{\{a,b\}|1a} \text{ and } q_{\{a,b\}|0b} = R_{0,1,b,\{a,b\}} \cdot q_{\{a,b\}|1b}$$
$$H_1: \; q_{\{a,b\}|0a} \neq R_{0,1,a,\{a,b\}} \cdot q_{\{a,b\}|1a} \text{ or } q_{\{a,b\}|0b} \neq R_{0,1,b,\{a,b\}} \cdot q_{\{a,b\}|1b} \text{ or both}$$

and exemplarily assume $R_{0,1,a,\{a,b\}} = 1.2$ and $R_{0,1,b,\{a,b\}} = 0.5$. By maximizing the likelihood for this coarsening situation and determining the value of the test statistic, we receive $T = 9.2$, exceeding the obtained critical value of ≈ 5.4 ($(1 - 2 \cdot \alpha)$-quantile of the χ^2_1-distribution, with $\alpha = 0.01$), such that H_0 can be rejected.
Figure 5 gives an overview of the test decision for various tests on gSI in our data situation, including different specifications of $R_{0,1,a,\{a,b\}}$ and $R_{0,1,b,\{a,b\}}$ varying on a grid with values 0.2, 0.5, 1, 1.5, 3, 10, respectively. Coarsening scenarios expressed by values of $R_{0,1,a,\{a,b\}}$ and $R_{0,1,b,\{a,b\}}$ above the horizontal dashed line, which indicates the critical value, are rejected by the likelihood-ratio test based on $\alpha = 0.01$. Thus, subgroup independence (with $R_{0,1,a,\{a,b\}} = R_{0,1,b,\{a,b\}} = 1$) is represented by a point falling below the border, such that

the null hypothesis cannot be rejected. Against this, the point representing gSI with $R_{0,1,a,\{a,b\}} = 1.2$ and $R_{0,1,b,\{a,b\}} = 0.5$ considered here, exceeds the line, resulting in a rejection of H_0. Interpreting the dependencies depicted in Figure 5 as a whole, the null hypothesis is rejected if both ratios are jointly either very small or very large, i.e. either $q_{\{a,b\}|0a} << q_{\{a,b\}|1a}$ and $q_{\{a,b\}|0b} << q_{\{a,b\}|1b}$, or $q_{\{a,b\}|0a} >> q_{\{a,b\}|1a}$ and $q_{\{a,b\}|0b} >> q_{\{a,b\}|1b}$. This is reasonable, since the number of coarse observations for a given subgroup, here e.g. $n_{0\{a,b\}}$, has to be produced by the precise categories that are compatible with the observation, which is not the case in the rejection scenarios.

The construction as likelihood-ratio test, which relies on a test statistic including the ratio of suprema of likelihoods under different specifications of parameters, allows testing on partial knowledge as a substantial extension. While a test on partial assumptions including some ratios leading to values of T above and some ratios leading to values below the critical value cannot be rejected, there are also partial assumptions that can be rejected, in the example, e.g. $R_{0,1,a,\{a,b\}} \in [0.2,\ 2]$ and $R_{0,1,b,\{a,b\}} \in [0.2,\ 0.5]$ (cf. Figure 5).[9]

8. Conclusion

We studied the (non-)testability of the dual assumptions CAR and SI, as well as the extended assumptions gCAR and gSI. By calculating the number of degrees of freedom of the respective estimation problem under these assumptions, we could confirm the already well-known result that CAR, and equally gCAR, is generally point-identifying and elaborate the criterion of the minimum number of subgroups required to be able to obtain also point-valued estimators in the case of SI and gSI at all. The estimates of the running example illustrated the result that SI/gSI – in contrast to CAR/gCAR – is indeed testable in case of sufficiently many subgroups, wherefore the likelihood-ratio test for SI was presented. While the setting of the example is a particular case, where the calculation of the critical value is based on a mixture distribution, relying on the common χ^2-distribution with the number of degrees of freedom achieved in the estimation problem under SI is appropriate in most cases. Straightforwardly transferring this test to gSI and the facility of expressing partial knowledge about the coarsening process, substantially increase the relevance of this test, enabling the user to test for specific dependencies of the coarsening process on the value of categorical covariates.
Although both strict assumptions are in a certain manner uninformative in the sense that specific underlying values do not play any role for the coarsening

[9] This idea of testing on partial assumptions reminds on the hypothesis test by Nordheim [19], who formalized hypotheses about the latent variable (not the coarsening parameters) distribution and included $R_{x,y,y'\mathfrak{Y}} = \frac{q_{\mathfrak{Y}|xy}}{q_{\mathfrak{Y}|xy'}}$ (not $R_{x,x',y,\mathfrak{Y}}$) into the respective test statistic.

process, we could detect a substantial difference with regard to the testability, summed up as follows: CAR is characterized by the absence of information within the coarsening process itself, making the true underlying value irrelevant, which can not be refuted from observations. Against this, under SI the value of the covariate is negligible for the coarsening, and not the value of the variable of interest. As elaborated in this paper, this kind of assumption is incompatible with some data situations since SI may require too strong coarsening rules for each given subgroup.
Finally, we should take note of a general issue of applying statistical procedures in the presence of coarse data: Generally, two kinds of uncertainties should be distinguished – uncertainty due to a finite sample only and uncertainty arising from the incompleteness in the data. While a hypothesis test reacts to an increasing sample size reducing the first kind of uncertainty, the projections of the set-valued estimators do not respond sensitively, indicating the general impossibility to test the second kind of uncertainty. Thus, although the proposed test does test on the coarsening process directly, it does not – and should not – reduce the second kind of uncertainty in the sense of gathering extra information about the hidden coarsening process that goes beyond the information gained by the estimators in (6).

Acknowledgements

We are grateful to the Research Data Center at the Institute for Employment Research, Nuremberg, especially Mark Trappmann and Anja Wurdack, for the access to the PASS data and their support in practical matters. Moreover, we highly appreciate the very helpful remarks of the two anonymous reviewers of the SMPS submission [22] underlying this paper. The first author thanks the LMUMentoring program, providing financial support for young, female researchers.

References

[1] C. Breunig, Testing missing at random using instrumental variables, Tech. rep., Humboldt University, Collaborative Research Center 649, `https://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2015-016.pdf` (2016).

[2] M. Cattaneo, A. Wiencierz, Likelihood-based imprecise regression, International Journal of Approximate Reasoning 53 (2012) 1137–1154.

[3] H. Chernoff, On the distribution of the likelihood ratio, The Annals of Mathematical Statistics 25 (1954) 573–578.

[4] I. Couso, D. Dubois, Statistical reasoning with set-valued information: Ontic vs. epistemic views, International Journal of Approximate Reasoning 55 (2014) 1502–1518.

[5] I. Couso, D. Dubois, Maximum likelihood under incomplete information: toward a comparison of criteria, in: M. F. Ferraro, B. Giordani, P. Vantaggi, M. Gagolewski, M. A. Gil, P. Grzegorzewski, O. Hryniewicz (eds.), Soft Methods for Data Science (SMPS 2016), Intelligent Systems and Computing Series, Springer, 2016, pp. 141–148.

[6] T. Denœux, Likelihood-based belief function: justification and some extensions to low-quality data, International Journal of Approximate Reasoning 55 (2014) 1535–1547.

[7] R. D. Gill, P. D. Grünwald, An algorithmic and a geometric characterization of coarsening at random, The Annals of Statistics 36 (2008) 2409–2422.

[8] R. Guillaume, D. Dubois, Robust parameter estimation of density functions under fuzzy interval observations., in: T. Augustin, S. Doria, E. Miranda, E. Quaeghebeur (eds.), ISIPTA '15, Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications, SIPTA, 2015, pp. 147–156.

[9] D. F. Heitjan, D. B. Rubin, Ignorability and coarse data, The Annals of Statistics 19 (1991) 2244–2253.

[10] E. Hüllermeier, Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization., International Journal of Aproximate Reasoning 55 (2014) 1519–1534.

[11] M. Jaeger, Ignorability for categorical data, The Annals of Statistics 33 (2005) 1964–1981.

[12] M. Jaeger, On testing the missing at random assumption, in: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (eds.), ECML '06, Proceedings of the 17th European Conference on Machine Learning, Lecture Notes in Artificial Intelligence, Springer, 2006, pp. 671–678.

[13] M. G. Kenward, E. J. Goetghebeur, G. Molenberghs, Sensitivity analysis for incomplete categorical data, Statistical Modelling 1 (2001) 31–48.

[14] A. Korinek, J. Mistiaen, M. Ravallion, Survey nonresponse and the distribution of income, The Journal of Economic Inequality 4 (2006) 33–55.

[15] R. J. Little, D. B. Rubin, Statistical Analysis with Missing Data, 2nd edition, Wiley, 2014.

[16] C. F. Manski, Partial Identification of Probability Distributions, Springer Science & Business Media, 2003.

[17] F. Molinari, Partial identification of probability distributions with misclassified data, Journal of Econometrics 144 (2008) 81–117.

[18] H. T. Nguyen, An Introduction to Random Sets, CRC, 2006.

[19] E. Nordheim, Inference from nonrandomly missing categorical data: An example from a genetic study on Turners syndrome, Journal of the American Statistical Association 79 (1984) 772–780.

[20] J. Plass, T. Augustin, M. Cattaneo, G. Schollmeyer, Statistical modelling under epistemic data imprecision: Some results on estimating multinomial distributions and logistic regression for coarse categorical data., in: T. Augustin, S. Doria, E. Miranda, E. Quaeghebeur (eds.), ISIPTA '15, Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications, SIPTA, 2015, pp. 247–256.

[21] J. Plass, T. Augustin, M. Cattaneo, G. Schollmeyer, Statistical modelling under epistemic data imprecision, Tech. rep., LMU Munich, `http://jplass.userweb.mwn.de/forschung.html` (2016).

[22] J. Plass, M. Cattaneo, G. Schollmeyer, T. Augustin, Testing of coarsening mechanisms: Coarsening at random versus subgroup independence, in: M. F. Ferraro, B. Giordani, P. Vantaggi, M. Gagolewski, M. A. Gil, P. Grzegorzewski, O. Hryniewicz (eds.), Soft Methods for Data Science (SMPS 2016), Intelligent Systems and Computing Series, Springer, 2016, pp. 415–422.

[23] C. R. Rao, Maximum likelihood estimation for the multinomial distribution, Sankhya: The Indian Journal of Statistics 18 (1957) 139–148.

[24] D. B. Rubin, Inference and missing data, Biometrika 63 (1976) 581–592.

[25] G. Schollmeyer, T. Augustin, Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data, International Journal of Approximate Reasoning 56 (2015) 224–248.

[26] E. Stanghellini, B. Vantaggi, Identification of discrete concentration graph models with one hidden binary variable, Bernoulli 19 (2013) 1920–1937.

[27] R. Tourangeau, T. Yan, Sensitive questions in surveys, Psychological Bulletin 133 (2007) 859–883.

[28] M. Trappmann, S. Gundert, C. Wenzig, D. Gebhardt, PASS: A household panel survey for research on unemployment and poverty, Schmollers Jahrbuch 130 (2010) 609–623.

[29] S. Vansteelandt, E. J. Goetghebeur, M. G. Kenward, G. Molenberghs, Ignorance and uncertainty regions as inferential tools in a sensitivity analysis, Statistica Sinica 16 (2006) 953–979.

[30] S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, The Annals of Statistics 9 (1938) 60–62.

[31] M. Zaffalon, E. Miranda, Conservative inference rule for uncertain reasoning under incompleteness, Journal of Artificial Intelligence Research 34 (2009) 757–821.

[32] Z. Zhang, Profile likelihood and incomplete data, International Statistical Review 78 (2010) 102–116.

Appendices

A. Visual depiction of H_0^* over a grid of parameters (cf. Section 6.1)

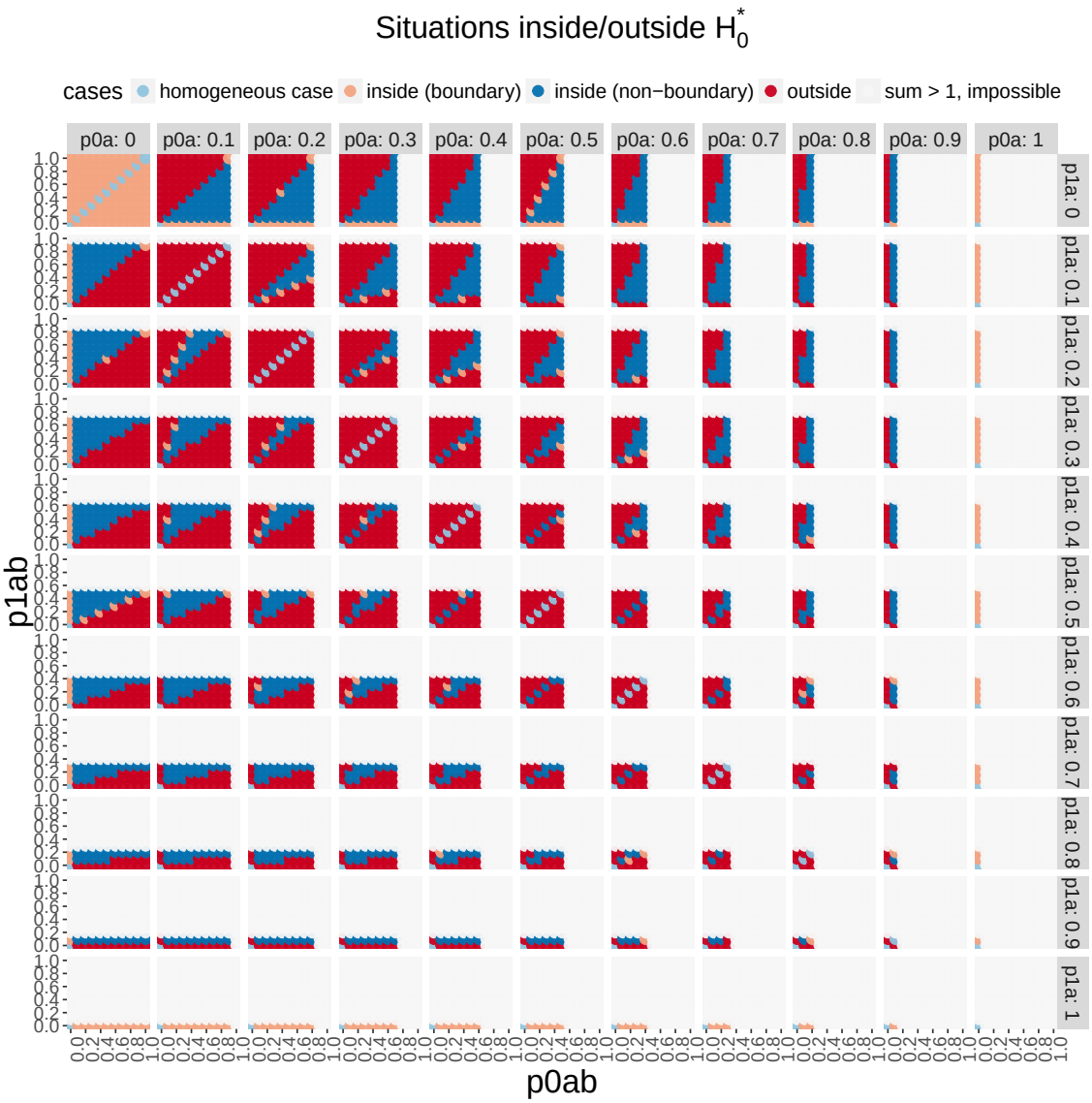

Figure 6: On a grid of values for the observed variable distribution different cases are distinguished: While the boundary case contains all combinations with either $p_{0\{a\}} \cdot p_{1\{a,b\}} = p_{1\{a\}} \cdot p_{0\{a,b\}}$ or $p_{0\{b\}} \cdot p_{1\{a,b\}} = p_{1\{b\}} \cdot p_{0\{a,b\}}$, joint equality is attained in the i.i.d. case. Moreover, it is differentiated between combinations that are (non-boundary) inside and outside H_0^*.