Towards Statistical Modelling under Epistemic Data Imprecision: Some Results on Estimating Multinomial Distributions and Logistic Regression for Coarse Categorical Data*

Julia Plass
Department of Statistics, LMU Munich
julia.plass@stat.uni-muenchen.de

Thomas Augustin
Department of Statistics, LMU Munich
augustin@stat.uni-muenchen.de

Marco E. G. V. Cattaneo
Department of Mathematics,
University of Hull
m.cattaneo@hull.ac.uk

Georg Schollmeyer
Department of Statistics, LMU Munich
georg.schollmeyer@stat.uni-muenchen.de

Abstract

The present paper is concerned with parameter estimation for categorical data under epistemic data imprecision, where for a part of the data only coarse(ned) versions of the true values are observable. For different observation models formalizing the information available on the coarsening process, we derive the (typically set-valued) maximum likelihood estimators of the underlying distributions. We discuss the homogeneous case of independent and identically distributed variables and present, by investigating logistic regression under a categorical covariate, some first steps towards statistical modelling of heterogenous multivariate data in this context. We start with the imprecise point estimator under an observation model describing the coarsening process without any further assumptions. Then we determine several sensitivity parameters that allow the refinement of the estimators in the presence of auxiliary information.

Keywords. Coarse data, missing data, epistemic data imprecision, sensitivity analysis, partial identification, categorical data, multinomial logit model, coarsening at random (CAR), likelihood, multinomial distribution

*Preliminary version of a technical report, supplementing a paper under review for *ISIPTA '15 (9th International Symposium on Imprecise Probabilities: Theories and Applications, Pescara, Italy, 2015)*.

1 The problem and its background

A frequent challenge in statistical modelling is *data imprecision*, where not all data are observed in the resolution intended in the subject matter context. In this paper, we utilize imprecise probability methodology for proper handling of *epistemic* data imprecision, also called *coarse(ned) data*. For categorical data as considered throughout here, this means that there exists a true precise value y of a generic variable Y of material interest taking values in a finite sample space $\Omega_Y = \{1, \ldots, K\}$, but we may only observe a non-singleton set $\mathfrak{y}$ containing y.[1] Missing data are included in this setting as the prominent special case where the whole sample space is observed.

Coarse categorical data emerge most naturally in a huge variety of applications. Missing data, for instance, arise directly by design in observational studies on treatment effects,[2] and unit non-response is quite frequent in surveys, in particular as refusals to answer sensitive questions. Typical examples of not missing but still coarse data include the numerous data sets where coarsening is deliberately applied as an anonymization technique (see, e.g., [7]), forecasts from opinion polls with respondents still undecided between some alternatives,[3] matched data sets with not completely identical categories, secondary data where the original coding produced categories that turn out to be not fine enough and, to give last but not least also a technical example, reliability analysis of a system whose components are tested separately prior to assembly [23].

Trapped in the framework of precise probabilities, traditional statistical methods are forced to neglect data imprecision or to impose quite strong, empirically untestable assumptions on the underlying coarsening process. Thus, except the very rare cases where the external information on the subject matter problem is rich enough to justify such an extent of precision of the modelling of the coarsening process, the price of the (seemingly) precise result is a substantial debilitation of the reliability of the conclusions drawn.

Against this background, set-valued approaches, aiming at a proper reflection of the available information, have been gathering momentum, also becoming a popular topic at the ISIPTA symposia ([4, 19, 12, 25, 26], to name just a few). In different areas of application concepts of cautious data completion emerged, where a classical procedure is extended by considering the set of all virtual precise observations in accordance with the coarse data (see, e.g., the exposition in [2], and the references

[1] Epistemic data imprecision has to be distinguished from an ontic view of sets, where the set is understood as holistic entity (see, in particular, [6], and [17] for an application in a regression setting.)

[2] See, for instance, the instructive case study [20].

[3] This problem is addressed e.g. in [17], where the here considered epistemic view is resolved in the ontic view.

therein). General investigations of coarse data from an imprecise probability-based Bayesian point of view include [5], [28]. Linear regression under metrical coarse data (interval data) is vividly discussed in the partial identification literature in the spirit of [14] (see also, e.g, [19], and the references therein). Mainly focusing on missing data, [27] suggest a framework for a systematic sensitivity analysis for statistical modelling under epistemic data imprecision. [4] introduces a profile likelihood approach for coarse data (for missing data see also [29]) and derive from it an uniform framework for robust regression analysis with imprecise data.
This paper will develop another likelihood-based approach.[4] It is strongly influenced by the methodology of partial identification, dealing with the tradeoff between information and credibility by first using the empirical evidence only, i.e. using information implied by the data and including only those assumptions about which there exists a common consensus concerning their validity (e.g., [14, 21, 15]). Sensitivity analysis pursues the same goal as partial identification, but the direction of proceeding differs. While partial identification starts from total uncertainty and gradually adds further assumptions, in the framework of sensitivity analysis the collection of all plausible point identified results from successively relaxed assumptions is considered. Thereby, the analysis is framed by a sensitivity parameter, a parameter that is not identified but given this parameter the parameter of interest is [27].

Our paper is structured as follows. In the next section we fix the notation used and formulate the problem setting more exactly for the cases considered in this paper: independent and identically distributed (i.i.d.) variables and the logistic regression model with a categorical covariate. The crucial technical argument underlying our paper to introduce an observation model and utilize invariance properties of the likelihood is developed in general terms in Section 3. In Section 4 we derive and discuss the set-valued estimators arising from a fully non-committal observation model, and we then turn to settings where this interval is narrowed when we benefit from the presence of additional auxiliary information. For technically handling this by sensitivity parameters, it is helpful to go to the other extreme, investigating point identifying additional assumptions in some special cases. For the homogeneous situ-

[4] In statistics, maximum likelihood estimation is a general procedure to derive estimators for the parameters of a statistical model. The likelihood function reinterprets the probability of observations in dependence on a parameter as describing the plausibility of the parameters given the data, and thus the maximum likelihood estimator selects that parameter value (or in the more general setting those parameter values) maximizing the likelihood function and in this way providing the most plausible explanation for the data (e.g., [3, § 6.3, 7.2.2]). The methodology is strictly observation-based (, i.e. without the need of specifying any prior distribution), conditional with many appealing frequentist properties (including asymptotic efficiency of the estimators) and generally applicable. (For regression models, like the logistic regression model considered here, it can be shown that, under the regularity condition that the marginal distribution of the covariates does not depend on the parameters of interest, it is sufficient to build the likelihood analysis on the conditional distribution of the outcome given the covariates.)

ation, after studying known coarsening in Section 5.1, we focus on the coarsening at random (CAR) assumption and illustrate the disastrous behaviour of the resulting point estimator when CAR is inappropriate (Section 5.2). Then in Section 5.3 we consider an extension of CAR and determine the corresponding ratio of coarsening probabilities as a sensitivity parameter. For the logistic regression case in Section 5.4 we work out that there is, as an alternative to CAR and its extensions, a further point identifying modelling assumption, which we call subgroup independent coarsening. Its generalization again can serve as a sensitivity parameter (Section 5.5). These sensitivity parameters frame a systematic sensitivity analysis, resulting in imprecise point estimators reflecting justifiable auxiliary information.

2 The basic setting

Let $Y_1, \ldots, Y_n$ be a random sample of a categorical response variable of interest Y with realizations $y_1, \ldots, y_n$ in sample space $\Omega_Y = \{1, \ldots, j, \ldots, K\}$. Problematically, some of those realizations are not known in a precise form, such that only realizations[5] $\mathfrak{y}_1, \ldots, \mathfrak{y}_n$ of a sample $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ of a random variable $\mathcal{Y}$ within sample space $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \emptyset$ can be observed, where $\mathcal{P}$ denotes the power set. All possible categories of $\mathcal{Y}$ represent singletons in $(\Omega_{\mathcal{Y}},\ \mathcal{P}(\Omega_{\mathcal{Y}}))$ with corresponding probability mass functions $p_{\mathfrak{y}_i} = P(\mathcal{Y}_i = \mathfrak{y}_i)$ $(i = 1, \ldots, n)$. But as we are interested in the random variables $Y_1, \ldots, Y_n$, our basic goal consists of gathering information about the individual probabilities $\pi_{i1} = P(Y_i = 1), \ldots, \pi_{iK} = P(Y_i = K)$.
We discuss the homogeneous case (i.i.d. case), in biometrical terms *prevalence* estimation, as well as situations with one precise categorical covariate X, in biometrical terms called *treatment*, with sample space Ω_X, being available. Here we confine ourselves to the case of one categorical covariate only, as this is technically equivalent to any finite set of categorical covariates. While in the i.i.d. case probabilities $\pi_{i1} = \pi_1, \ldots,\ \pi_{iK} = \pi_K$ are assumed to be independent of individual i, in the case with one covariate the probabilities $\pi_{i1} = P(Y_i = 1|X_i = x_i) = \pi_{x_i 1}, \ldots,\ \pi_{iK} = P(Y_i = K|X_i = x_i) = \pi_{x_i K}$ are influenced by individual i through the corresponding values of the covariate X_i. One of most generally applied models is the *multinomial logit model*. It describes the dependence of a categorical dependent variable Y of nominal scale on covariates X by

$$\pi_{ij} = P(Y_i = j|\mathbf{x}_i) = \frac{\exp(\beta_{j0} + \mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)} \tag{1}$$

[5] We assume throughout the paper that the coarsening process is error-free, in the sense that $\mathfrak{y}_i \supseteq y_i$, $i = 1, \ldots, n$.

		$\mathcal{Y}$			
		A	B	AB	
X	0	n_{0A}	n_{0B}	n_{0AB}	n_0
	1	n_{1A}	n_{1B}	n_{1AB}	n_1
		n_A	n_B	n_{AB}	n

Table 1: Contingency table that introduces used notation.

$i = 1, \ldots, n$ for categories $j = 1, \ldots, K-1$ and by

$$\pi_{iK} = 1 - \pi_{i1} - \ldots - \pi_{iK-1} = \frac{1}{1 + \sum_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)} \tag{2}$$

with category specific regression coefficients, that is $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm})^T$ referring to m covariates and intercept β_{j0}. As we here address the case of one categorical covariate $X_i \in \{1, \ldots, c\}$, dummy coded variables $X_{i1}, \ldots, X_{im}$ with $m = c - 1$ are included into the model. Equivalently, the multinomial logit model can be written as

$$\log \frac{\pi_{ij}}{\pi_{iK}} = \beta_{j0} + \mathbf{x}_i^T \boldsymbol{\beta}_j, \tag{3}$$

where log denotes the natural logarithm.
It is common to summarize categorical data in contingency tables by reporting the counts for possible outcomes, where the covariates X are supposed to be in the rows (e.g., [24]). Thus, in our case the contingency table in Table 1 will be addressed. The number of observations with $\mathcal{Y} = \mathfrak{y}$ and treatment group $X = x$ is denoted by $n_{x\mathfrak{y}}$, where $n_0 = n_{0A} + n_{0B} + n_{0AB}$, $n_1 = n_{1A} + n_{1B} + n_{1AB}$, $n_A = n_{0A} + n_{1A}$, $n_B = n_{0B} + n_{1B}$ and $n_{AB} = n_{0AB} + n_{1AB}$.

Example: In order to illustrate our results we consider the contingency table of Table 2 as a running example. It shows data from the German panel study "Labor Market and Social Security" (PASS, here wave 1, 2006/2007, [22]), where partial income knowledge (variable HEK0700; "na" denotes that no suitable answer has been reported) as well as the receipt of the so-called Unemployment Benefit II (variable alg2abez; here denoted by UBII) are collected.

3 Sketch of the basic argument

This paper, similar to [4, 29], relies on the likelihood as the fundamental concept to derive parameter estimators under epistemic data imprecision, but looks at it from a different angle. In order to support the appropriate incorporation of the available

		income			
		< 1000€	≥ 1000€	na	
UBII	yes	130	114	75	319
	no	108	721	263	1092
		238	835	338	1411

Table 2: Contingency table to illustrate some results by means of the PASS data.

information provided by the data and the background knowledge, we explicitly formulate, and utilize, an *observation model* relating the observable level and the ideal level. The observation model is a set $\mathfrak{Q}$ of (precise) coarsening probabilities,[6] and thus the medium to specify carefully and flexibly the available information about the coarsening process. By virtue of the theorem of total probability, the elements of $\mathfrak{Q}$ relate the probability distribution of the imprecise observation $\mathcal{Y}$ to the distribution of the underlying latent variable Y (and, if present, certain covariates).
Parametrizing the distributions, again possibly after splitting with respect to certain covariate values, let ϑ (the various p's in the following sections) and η (the various π's below) be the parameters determining the distribution of $\mathcal{Y}$ and Y, respectively, and let ζ be the parameter characterising the elements of $\mathfrak{Q}$ (the various q's, possibly constrained by the specified constraints: $\left(q_{\mathfrak{y}|y} := P(\mathcal{Y} = \mathfrak{y}|Y = y)\right)_{(\mathfrak{y}\in\Omega_{\mathcal{Y}}, y\in\Omega_Y)}$ in the i.i.d. case, while in the regression context the coarsening mechanisms generally also depend on the values of X_i, i.e., $(q_{\mathfrak{y}|xy} = P(\mathcal{Y} = \mathfrak{y}|X = x, Y = y)_{(\mathfrak{y}\in\Omega_{\mathcal{Y}}, y\in\Omega_Y, x\in\Omega_X)}$ has to be considered.
Then we can describe the relationship between $\gamma := (\eta^T, \zeta^T)^T$ (with domain Γ) and ϑ (with domain Θ) via the mapping

$$\begin{array}{rccc} \Phi(\cdot): & \Gamma & \to & \Theta \\ & \gamma & \mapsto & \vartheta\,. \end{array}$$

Figure 1 shows this mapping $\Phi(\cdot)$ and all parameters included.
Most important in our context is the invariance of the likelihood under parameter transformations; evaluating the likelihood in terms of γ and in terms of $\vartheta = \Phi(\gamma)$ is equivalent in the situations considered here. Our random set modelling will allow us to determine the ML-estimator $\hat{\vartheta}$ of ϑ, which moreover, apart from trivial extreme cases, can be shown to be single-valued. Then the possibly set-valued maximum-likelihood estimator for γ is obtained as

$$\hat{\Gamma} = \left\{\gamma \,\middle|\, \Phi(\gamma) = \hat{\vartheta}\right\}\,. \tag{4}$$

[6] More precisely, $\mathfrak{Q}$ is a generalized transition kernel, consisting of credal sets indexed by the values of Y.

Thus, adapting the concept of maximum likelihood estimators to a persistent set-based perspective and to random set-based situations, we achieve a general and powerful framework for handling coarse categorical data via the mapping $\Phi(\cdot)$. If $\Phi(\cdot)$ is injective, then $\hat{\Gamma}$ is a singleton as well, and γ so-to-say empirically point identified; otherwise $\hat{\Gamma}$ is set-valued in the literal sense and γ empirically partially identified.

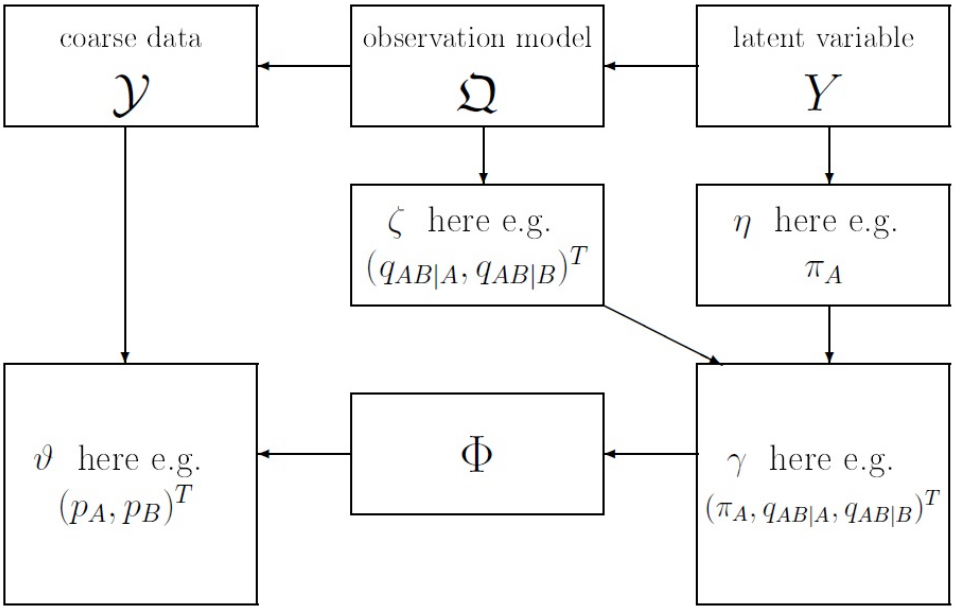

Figure 1: Observable and latent variable and the corresponding parameters.

The dimension of the parameter vectors η and ζ increases substantially with the cardinality of Ω_Y and Ω_X. In the i.i.d. case $m = (\sum_{z=1}^{|\Omega_Y|} \binom{|\Omega_Y|}{z} \cdot z) - 1$ or equivalently $m = K \cdot 2^{K-1} - 1$ parameters have to be estimated, where in the case with one covariate this number even increases to $|\Omega_X| \cdot m$. In this way, in the i.i.d. case with $\Omega_Y = \{1,\ 2,\ 3\}$ and corresponding $\Omega_{\mathcal{Y}} = \{1,\ 2,\ 3,\ 12,\ 13,\ 23,\ 123\}$, where for instance "12" denotes "either catgory 1 or category 2", already eleven parameters, that is $\pi_1,\ \pi_2,\ q_{12|1},\ q_{12|2},\ q_{13|1},\ q_{13|3},\ q_{23|2},\ q_{23|3},\ q_{123|1},\ q_{123|2}$ and $q_{123|3}$, have to be estimated.
Thus, for reasons of conciseness of the presentation, we mainly confine detailed explanations and derivations on the special, yet still representative cases of a binary response variable Y with sample space $\Omega_Y = \{A,\ B\}$ and observations within $\Omega_{\mathcal{Y}} =$

$\{A,\ B,\ AB\}$,[7] as well as a binary precise categorical covariate X with values 0 and 1.[8] In this case, the underlying model expressed in Equation (1) and (2) is called *logit model*. Nevertheless, the main results not only will be shown for this situation, where coarsening corresponds to missingness, but also in a general way.

4 Maximum likelihood estimation for coarse categorical data without additional information

In this section we derive the maximum likelihood estimators for the case where no additional information on the coarsening process is available, i.e. there are no constraints on the elements of $\mathfrak{Q}$. A crucial step is to rely on the random set view that treats data imprecision as a change of the sample space with corresponding random variables $\mathcal{Y}_i$, $i = 1, \ldots, n$, which then lead to multinomially distributed variables with parameter ϑ for the counts based on the new sample space. According to the argumentation in Section 3, the resulting likelihood in ϑ, and the estimator derived from maximizing it, will then be related to the parameters of the distribution of the latent variable (and the observation model). As discussed at the end of the previous section, we explain the construction in some detail for the representative special cases with $\Omega_Y = \{A,\ B\}$ (and $\Omega_X = \{0, 1\}$) and then report the general results.

4.1 Estimation in the i.i.d. case

Considering categorical i.i.d. random variables $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ with realizations $\mathfrak{y}_1, \ldots, \mathfrak{y}_n$ in the sample space $\Omega_{\mathcal{Y}} = \{A,\ B,\ AB\}$, we obtain the following likelihood function for the parameter $\vartheta = (p_A, p_B)^T$ given the data, summarized by the counts n_A, n_B and n_{AB} (with $p_{AB} = 1 - p_A - p_B$):[9]

$$\begin{aligned} L(\vartheta) &= L(p_A,\ p_B) = L(p_A,\ p_B || \mathfrak{y}_1, \ldots, \mathfrak{y}_n) = P(\mathfrak{y}_1, \ldots, \mathfrak{y}_n || p_A,\ p_B) \\ &\propto p_A^{n_A} \cdot p_B^{n_B} \cdot p_{AB}^{n_{AB}} \,. \end{aligned} \tag{5}$$

For $n = n_A + n_B + n_{AB} > 0$ this likelihood is uniquely maximized by the corresponding relative frequencies (see [18]),

$$\hat{p}_A^{(MLE)} = \frac{n_A}{n}, \qquad \hat{p}_B^{(MLE)} = \frac{n_B}{n}\,, \tag{6}$$

[7] For ease of presentation we denote in the binary case the different categories by A, B, and AB instead of numbers and sets of numbers.

[8] All results can be transferred straightforward to cases of covariates with more than two categories by including more dummy variables of that kind.

[9] In the following, we will use the abbreviated notation of the likelihood without referring to the data.

and thus $\hat{p}_{AB}^{(MLE)} = 1 - \hat{p}_A^{(MLE)} - \hat{p}_B^{(MLE)} = \frac{n_{AB}}{n}$.
Essentially, we are interested in the parameter $\eta = \pi_A$ determining the probabilities of the true, but unobserved variable Y being equal to particular categories and the associated maximum likelihood estimator. Those probabilities of interest, in our case π_A and $\pi_B = 1 - \pi_A$, can be related with probabilities p_A, p_B and p_{AB} corresponding to the observable variables by

$$\begin{aligned} p_A &= (1 - q_{AB|A}) \cdot \pi_A \,, \\ p_B &= (1 - q_{AB|B}) \cdot (1 - \pi_A) \,, \end{aligned} \tag{7}$$

where $p_{AB} = q_{AB|A} \cdot \pi_A + q_{AB|B} \cdot (1 - \pi_A)$ results from the law of total probability. This means that the likelihood in terms of ϑ in Equation (5) and in terms of $\gamma = (\pi_A, q_{AB|A}, q_{AB|B})^T$, i.e.

$$\begin{aligned} L(\gamma) = L(\pi_A,\ q_{AB|A},\ q_{AB|B}) \propto [(1 - q_{AB|A}) \cdot \pi_A]^{n_A} \cdot [(1 - q_{AB|B}) \cdot (1 - \pi_A)]^{n_B} \\ \cdot [q_{AB|A} \cdot \pi_A + q_{AB|B} \cdot (1 - \pi_A)]^{n_{AB}}, \end{aligned} \tag{8}$$

coincide, indeed.
By the invariance of the likelihood under parameter transformations, Equations (6) and (7) can be combined, resulting in the following system of equations:

$$\begin{aligned} (1 - \hat{q}_{AB|A}) \cdot \hat{\pi}_A &= \frac{n_A}{n} = \hat{p}_A^{(MLE)}, \\ (1 - \hat{q}_{AB|B}) \cdot (1 - \hat{\pi}_A) &= \frac{n_B}{n} = \hat{p}_B^{(MLE)}, \\ \hat{q}_{AB|A} \cdot \hat{\pi}_A + \hat{q}_{AB|B} \cdot (1 - \hat{\pi}_A) &= \frac{n_{AB}}{n} = \hat{p}_{AB}^{(MLE)}. \end{aligned} \tag{9}$$

For reasons of redundancy we can leave the third equation out of consideration. As there typically are multiple triples $\hat{\gamma} = (\hat{\pi}_A,\ \hat{q}_{AB|A},\ \hat{q}_{AB|B})^T$ that lead to the same values of $\hat{\vartheta} = (\hat{p}_A^{(MLE)},\ \hat{p}_B^{(MLE)})^T$ (cf. Figure 2), the mapping $\Phi :\ [0,\ 1]^3 \to [0,\ 1]^2$ with

$$\Phi \begin{pmatrix} \pi_A \\ q_{AB|A} \\ q_{AB|B} \end{pmatrix} = \begin{pmatrix} \pi_A \cdot (1 - q_{AB|A}) \\ (1 - \pi_A) \cdot (1 - q_{AB|B}) \end{pmatrix} = \begin{pmatrix} p_A \\ p_B \end{pmatrix} \tag{10}$$

(cf. Figure 1) connecting both parametrizations in general is not injective. Thus

$$\hat{\Gamma} = \left\{ \gamma \middle| \Phi(\gamma) = \hat{\vartheta} \right\} \tag{11}$$

is set-valued in the literal sense. Points in this set are constrained through the relationships in (9), and thus $\hat{\Gamma}$ is not a cuboid in $[0, 1]^3$. Building the one dimensional projections, set-valued estimators of the single components of γ are obtained via

$$\begin{aligned} \hat{\pi}_A &\in \left[\frac{n_A}{n},\ \frac{n_A + n_{AB}}{n} \right], \quad \hat{q}_{AB|A} \in \left[0,\ \frac{n_{AB}}{n_A + n_{AB}} \right] \quad \text{and} \\ \hat{q}_{AB|B} &\in \left[0,\ \frac{n_{AB}}{n_B + n_{AB}} \right]. \end{aligned} \tag{12}$$

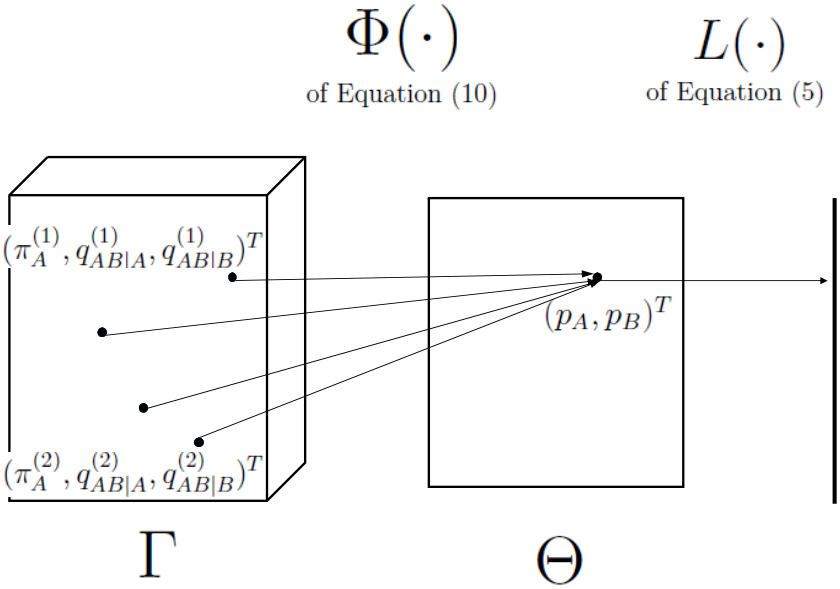

Figure 2: Illustration of the estimation problem in the i.i.d. case.

Extending the discussion here to the general case of $\Omega_Y = \{1, \ldots, K\}$ and the corresponding $\Omega_{\mathcal{Y}}$, the estimators in Equations (12) generalize to

$$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n}, \ \frac{\sum_{\mathfrak{y} \ni y} n_{\mathfrak{y}}}{n} \right], \quad \hat{q}_{\mathfrak{y}|y} \in \left[0, \ \frac{n_{\mathfrak{y}}}{n_{\{y\}} + n_{\mathfrak{y}}} \right], \tag{13}$$

where $y \in \Omega_y = \{1, \ldots, K\}$ and $\mathfrak{y} \in \Omega_{\mathcal{Y}}$.[10]

4.2 Logistic regression with a categorical covariate X

Now we consider the heterogenous situation expressed by a discrete covariate X, which also has been depicted in Table 1. Again we can derive set-valued estimators of the parameters of interest $\eta = (\pi_{0A}, \pi_{1A})^T$ (and the auxiliary parameter ζ characterizing the coarsening mechanisms) by taking the random set perspective, setting up the corresponding likelihood function and applying the appropriate parameter transformations. Proceeding in this way, for fixed treatment group x the cell counts

[10] The estimators of the probability components of the distribution of Y_i prove to be the same as arising from a belief functions like construction of empirical probabilities and also coincide with the estimator obtained from cautious data completion, plugging in all potential precise sample outcome compatible with the observations $\mathfrak{y}_1, \ldots, \mathfrak{y}_n$ (see, e.g., [2])

$(n_{xA}, n_{xB},\ n_{xAB})$ follow a multinomial distribution, i.e.

$$(n_{xA}, n_{xB}, n_{xAB}) \sim M(n_x, (p_{xA}, p_{xB}, p_{xAB}))$$

with conditional probabilities

$$p_{x\mathfrak{y}} = P(\mathcal{Y} = \mathfrak{y}|X = x)$$

(see [24, 1]).[11] Therefore, the corresponding likelihood function for parameter $\vartheta = (p_{0A}, p_{1A}, p_{0B}, p_{1B})^T$ (with $p_{0AB} = 1 - p_{0A} - p_{0B}$ and $p_{1AB} = 1 - p_{1A} - p_{1B}$) is given by

$$L(\vartheta) = L(p_{0A},\ p_{1A},\ p_{0B},\ p_{1B}) \propto p_{0A}^{n_{0A}} \cdot p_{0B}^{n_{0B}} \cdot p_{0AB}^{n_{0AB}} \cdot p_{1A}^{n_{1A}} \cdot p_{1B}^{n_{1B}} \cdot p_{1AB}^{n_{1AB}}, \quad (14)$$

For $n_x > 0$ the maximum likelihood estimators for the parameters are unique and given by (see [18])

$$\hat{p}_{x\mathfrak{y}}^{(MLE)} = \frac{n_{x\mathfrak{y}}}{n_x},\ \text{for } x \in \{0, 1\}.$$

Reparametrizing the likelihood function of Equation (14) in terms of the parameters of interest and the parameters of the observation model, i.e. $\gamma = (\pi_{0A}, \pi_{1A}, q_{AB|xA}, q_{AB|xB})^T$, we obtain the likelihood function

$$\begin{aligned} & L(\gamma) = L(\pi_{0A},\ \pi_{1A},\ q_{AB|0A},\ q_{AB|1A},\ q_{AB|0B},\ q_{AB|1B}) \quad (15) \\ \propto\ & [(1 - q_{AB|0A}) \cdot \pi_{0A}]^{n_{0A}} \cdot [(1 - q_{AB|1A}) \cdot \pi_{1A}]^{n_{1A}} \\ & \cdot [(1 - q_{AB|0B}) \cdot (1 - \pi_{0A})]^{n_{0B}} \cdot [(1 - q_{AB|1B}) \cdot (1 - \pi_{1A})]^{n_{1B}} \\ & \cdot [q_{AB|0A} \cdot \pi_{0A} + q_{AB|0B} \cdot (1 - \pi_{0A})]^{n_{0AB}} \\ & \cdot [q_{AB|1A} \cdot \pi_{1A} + q_{AB|1B} \cdot (1 - \pi_{1A})]^{n_{1AB}}. \end{aligned}$$

Analogously to Section 4.1, we consider the mapping which connects both parametrizations, $\Phi :\ [0,\ 1]^6 \to [0,\ 1]^4$ with

$$\Phi \begin{pmatrix} \pi_{0A} \\ \pi_{1A} \\ q_{AB|0A} \\ q_{AB|1A} \\ q_{AB|0B} \\ q_{AB|1B} \end{pmatrix} = \begin{pmatrix} \pi_{0A} \cdot (1 - q_{AB|0A}) \\ \pi_{1A} \cdot (1 - q_{AB|1A}) \\ (1 - \pi_{0A}) \cdot (1 - q_{AB|0B}) \\ (1 - \pi_{1A}) \cdot (1 - q_{AB|1B}) \end{pmatrix} = \begin{pmatrix} p_{0A} \\ p_{1A} \\ p_{0B} \\ p_{1B} \end{pmatrix} \quad (16)$$

and observe that in this case it is also not injective and thus $\hat{\Gamma}$, constructed along the line of (4), is strictly set-valued, too. Illustrating $\hat{\Gamma}$ again by the corresponding

[11] This corresponds to a product-multinomial sampling scheme (e.g. [24, 1]).

projections along the axes, we obtain for given value $x \in \{0, 1\}$ in the general case with more than two categories in Y, i.e. $y \in \Omega_Y = \{1, \ldots, K\}$ and $\mathfrak{y} \in \Omega_{\mathcal{Y}}$,

$$\hat{\pi}_{xy} \in \left[\frac{n_{xy}}{n_x},\ \frac{\sum_{\mathfrak{y} \ni y} n_{x\mathfrak{y}}}{n_x} \right],\ \hat{q}_{\mathfrak{y}|xy} \in \left[0,\ \frac{n_{x\mathfrak{y}}}{n_{x\mathfrak{y}} + n_{xy}} \right]. \tag{17}$$

Example (continued): Applying Equation (17) to our example, one obtains

$$\begin{aligned} \hat{\pi}_{0A} &\in \left[\frac{130}{319},\ \frac{130 + 75}{319} \right] &&= [0.41,\ 0.64]\,, \\ \hat{\pi}_{1A} &\in \left[\frac{108}{1092},\ \frac{108 + 263}{1092} \right] &&= [0.10,\ 0.34]\,. \end{aligned}$$

By recurring on the relation defined in Equation (1) and (2), and utilizing the injectivity of the logistic function, the likelihood function considered here can also be uniquely expressed in terms of the regression coefficients. In this way, instead of the estimators $\hat{\pi}_{0A}$ and $\hat{\pi}_{1A}$ of Equation (17), equivalently one can consider the estimators

$$\begin{aligned} \hat{\beta}_{A0} &\in \left[\log\left(\frac{n_{0A}}{n_0 - n_{0A}} \right), \log\left(\frac{n_{0A} + n_{0AB}}{n_{0B}} \right) \right] \\ \hat{\beta}_{A1} &\in \left[\log\left(\frac{n_{1A} \cdot (n_0 - n_{0A})}{(n_1 - n_{1A}) \cdot n_{0A}} \right),\ \log\left(\frac{(n_{1A} + n_{1AB}) \cdot n_{0B}}{n_{1B} \cdot (n_{0A} + n_{0AB})} \right) \right], \end{aligned} \tag{18}$$

assuming all expressions to be well defined. Equations (18) can directly be obtained via solving the relation of the logistic function for the regression coefficients and incorporating the results of Equation (17).

Example (continued): In terms of the regression coefficients, we obtain the estimates $\hat{\beta}_{0A} \in [-0.37,\ 0.59]$ and $\hat{\beta}_{1A} \in [-1.83,\ -1.25]$.

Reminiscing about the derivation given here, we see that the categorical covariate case for the logistic model – in strict contrast to the continuous case (see Section 6) – in essence consists of a subgroup-specific consideration of the i.i.d. case.

5 Reliable incorporation of auxiliary information: sensitivity parameters and partial identification

The set-valued estimators from Equation (12) (and analogously from Equation (17)) are a typical application of the methodology of partial identification, emphasizing that only justified assumptions should be made which do not have to induce point

identified parameters, but at least identify the parameter of interest in parts compared to the set of parameters that seemed to be possible in the beginning of the analysis (e.g., [14]). In this way, the trivial bounds $[0,\ 1]$ have been refined substantially.
In the spirit of partial identification and sensitivity analysis we can further refine the set-valued estimators from Equation (12) and Equation (17) if, and also only if, auxiliary information beyond the empirical evidence is available.[12] To handle this technically, we start with distinguishing and investigating point identifying additional assumptions, in order to utilize them as a technical mean to derive sensitivity parameters, governing the incorporation of additional information.
Due to the fact that the imprecise point estimators in Equation (17) directly result from considering Equation (12) in a subgroup specific way, in Section 5.1 to Section 5.3 the detailed presentation is confined on the i.i.d. case. In Section 5.4, considering explicitly the regression model, another point-identifying assumption is suggested, where again the corresponding generalization may be used as a sensitivity parameter which allows the inclusion of partial knowledge.

5.1 Known coarsening

If one or both coarsening parameters $q_{AB|A}$ and $q_{AB|B}$ are known (and different from 1), one can conclude directly that the corresponding mapping $\Phi(\cdot)$ from (10) is injective as in this case the parameter π_A can be uniquely related to the parameter p_A. Therefore, the set-valued estimator for π_A specified in Equation (12) can be shrunk to a single-valued estimator.
The exact value of the coarsening parameters is most often unknown, but in case there is material information available that allows to bound it in a nontrivial interval, the consideration here gives a first way to perform a systematic sensitivity analysis. In most situations however such direct bounds will not be available. Therefore we look for alternative ways to introduce auxiliary knowledge.

5.2 Coarsening at random (CAR)

If the coarsening is nonstochastic, the underlying degree of coarsening is predetermined and known. For instance, if respondents are requested to give their answer in

[12] Vansteelandt et al. [27] suggest to determine a sensitivity parameter $\delta \in \Delta$ under which the problem is identified and then to calculate the parameter of interest η for different values of the sensitivity parameter, where the whole region of the resulting parameters of interest is called Ignorance Region $ir(\eta, \Delta)$ and the corresponding region of estimates Honestly Estimated Ignorance Region (HEIR) $\hat{ir}_n(\eta, \Delta)$. In order to account for statistical uncertainty due to finite sample size as well, in context of sensitivity analysis uncertainty regions are addressed that either can be constructed as covering the parameter of interest or the whole ignorance region with a probability of at least $(1 - \alpha)$ [10, 27].

a grouped way and we assume that all respondents answer correctly, then the coarsening is predefined in the sense that there is an unique coarsened outcome for every true answer. In the context of distinguishing between nonstochastic and stochastic coarsening mechanisms, Heitjan and Rubin [9] investigated under which properties the corresponding likelihood can be simplified to the so-called grouped likelihood and introduced the concept of *coarsening at random (CAR)*. This is a simplifying property requesting that the probability $q_{\mathfrak{y}|y}$ is constant, no matter which true value y is underlying as long as it fits to the observed value $\mathfrak{y}$. Thus, in the addressed i.i.d. data situation, probability $q_{AB|y}$ takes the same value for all true values that correspond with the observed data, namely true value "$y = A$" and "$y = B$", i.e., under CAR, $q_{AB|A} = q_{AB|B}$ is assumed. Illustrated by the running example, CAR postulates that the probability of giving no suitable answer should not depend on the true income category. In the dichotomous situation of this example, we are then actually concerned with the assumption of missing at random (MAR) [13], which can be regarded as a special case of CAR.
Focusing again on the i.i.d. case, incorporating the CAR assumption of $q_{AB|A} = q_{AB|B}$ into the likelihood in Equation (8) and in the observation model specifying $\Phi(\cdot)$, the situation simplifies substantially. Indeed, Φ is (almost) injective now, and we get the empirically point identified estimators, corresponding to having simply ignored the units with coarse values:

$$\begin{aligned} \hat{\pi}_A &= \frac{n_A}{n_A + n_B}, \\ \hat{q}_{AB|A} &= \hat{q}_{AB|B} = \frac{n_{AB}}{n_A + n_B + n_{AB}}. \end{aligned}$$

There are several ideal-type situations in which CAR can be justified indeed.[13] Nevertheless, this assumption must be treated with greatest care. Deviating from such an ideal-type situation and wrongly assuming CAR can lead to a bias of an extent that for sure destroys the practical relevance of the analysis, as is also illustrated in Figure 3. There the estimation of π_A under obstinately assumed CAR but varying coarsening probabilities is evaluated by the median relative empirical bias $\frac{\hat{\pi}_A - \pi_{A,\text{true}}}{|\pi_{A,\text{true}}|}$ based on 100 simulated datasets (here with $\pi_A = 0.6$).[14] The relative median bias increases the more one deviates from the case of CAR, indeed, up to a median relative bias of almost 80%. It can be noted that one does not face a symmetric

[13]For instance, rounding, type I censoring, which is present if the censoring times are fixed, and progressive type II censoring, which investigates censoring after the fixed d-th failure, in their pure form are CAR [11, 8].

[14]Thereby, in all addressed situations characterized by different true underlying coarsening mechanisms ($q_{AB|A}$ and $q_{AB|B}$ varying between 0.1 and 0.9 in equidistant breaks of 0.1, respectively), the assumption of CAR is involved into the estimation by plugging $q_{AB|A} = q_{AB|B}$ into the likelihood of Equation (8) that is maximized.

problem. This can be explained by the fact that in the simulated data the number of true values of A exceeds the number of true values of B and consequently there are in total more coarsened values if $q_{AB|A}$ is larger compared to $q_{AB|B}$ which increases uncertainty.

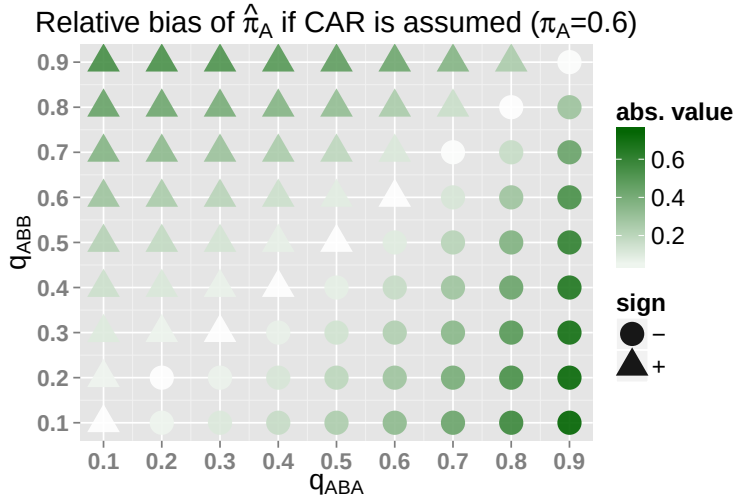

Figure 3: Consequences for the median relative bias of $\hat{\pi}_A$ if there is a deviation from assumed CAR.

5.3 Ratio of coarsening parameters

In our context the paper by Nordheim [16] obtains new importance. He considers the ratio between different mechanisms in the context of nonrandomly missing and misclassified data. By fixing the ratio between the coarsening probabilities the corresponding maximum likelihood problem leads to quadratic equations, where one solution is contained in the interval of $\hat{\pi}_A$ from Equation (12), while the other solution lies outside of $[0,\ 1]$ (cf. [16, p. 774]). Here we set $R = \frac{q_{B|B}}{q_{A|A}} = \frac{1-q_{AB|B}}{1-q_{AB|A}}$ slightly modifying the ratio of Nordheim by referring to the probabilities of the complementary events. Treating this ratio between the probabilities of precise observation fixed and including it into the likelihood problem in Section 4.1, unique, empirically point identified estimators

$$\begin{aligned} \hat{\pi}_A &= \frac{R \cdot n_A}{n_B + n_A \cdot R}, \\ \hat{q}_{AB|A} &= \frac{n_B \cdot (R-1) + n_{AB}}{n \cdot R}, \end{aligned} \tag{19}$$

are obtained. The construction of R and the estimators from Equation (19) directly show that the case of assuming $R = 1$ corresponds to the assumption of CAR. In this way, the incorporation of an assumed value R generalizes the CAR assumption. As in the case of CAR, the impact of assuming a wrong value of R has been investigated (results are available on request, see also [16]), where again a substantial bias can occur.[15]

Because of the fact that the parameter of interest π_A is identified given the typically unknown value of R, the ratio R can be used as a sensitivity parameter. In many cases it might be difficult to gain information about the exact value of R, but it seems quite realistic that a rough evaluation of R can be derived from contentual considerations, former studies or experiments.

Thus, it is interesting to investigate the gain of information resulting from implying a factor R that is roughly known only compared to the situation without any additional assumptions. Assessing a precise observation for category A as at least as probable as for category B, we can impose the assumption $R \leq 1$. Turning for instance to the second group of the running example (i.e. $n_A = 108,\ n_B = 721,\ n_{AB} = 263$) the corresponding HEIR[16] $\hat{ir}_n(\pi_A, \mathbb{R}_0^+) = \left[\frac{n_A}{n},\ \frac{n_A+n_{AB}}{n}\right] = [0.10,\ 0.34]$ can be shrunk to $ir(\pi_A, [0,1]) = \left[\frac{n_A}{n},\ \frac{n_A}{n_A+n_B}\right] = [0.10,\ 0.13]$. Thus, even by implying a vague assumption about the relation R valuable information about the parameter of interest can be gained provided this assumption is satisfied indeed.

In more general cases of $|\Omega_Y| > 2$, the relations between the precise observation probabilities are not sufficient and relations concerning different coarsening mechanisms have to be known in order to obtain point identified estimators. For instance in case of $\Omega_Y = \{A, B, C\}$, assumptions about relations $R_1 = \frac{q_{AB|B}}{q_{AB|A}}$, $R_2 = \frac{q_{AC|C}}{q_{AC|A}}$, $R_3 = \frac{q_{BC|C}}{q_{BC|B}}$, $R_4 = \frac{q_{ABC|C}}{q_{ABC|B}}$ and $R_5 = \frac{q_{ABC|B}}{q_{ABC|A}}$ have to be imposed [16]. In this context, we have shown (results are available upon request) that the relation that concerns for instance category AC (e.g. $R = \frac{q_{AC|C}}{q_{AC|A}}$) does only marginally influence the estimation of the coarsening mechanism of other categories as for instance $q_{AB|B}$. Apart from this, similar trends as in the case with two categories seem to result.

5.4 Subgroup independent coarsening mechanism

In the situation with covariates, there is apart from CAR, i.e. $\hat{q}_{AB|xA} = \hat{q}_{AB|xB}$, an alternative kind of uninformative coarsening, namely the independence of the un-

[15] The fact that there a similar variance of the estimators is obtained independently of the amount of deviation from the true value of R shows drastically that such deviations do not increase statistical uncertainty in the traditional sense and thus cannot be discovered by a traditional statistical analysis.

[16] Considering the ratio R as a sensitivity parameter leads to the HEIR (see footnote 12 and, e.g., [10, 27]).

derlying covariate value. We will establish injectivity of the corresponding mapping $\Phi(\cdot)$ under an intuitive regularity condition and then, analogously to the procedure in Sections 5.2 and 5.3, this idea will be generalized in Section 5.5 by again considering the corresponding fraction as a sensitivity parameter.
Imposing such *subgroup independent coarsening mechanisms*[17] with

$$\begin{aligned} q_{AB|0A} &= q_{AB|1A} =: q_{AB|A} \\ q_{AB|0B} &= q_{AB|1B} =: q_{AB|B}, \end{aligned} \tag{20}$$

in the estimation problem of Section 4.2, the mapping of Equation (16) changes to

$$\begin{aligned} &\Phi: \ [0,\ 1]^4 \to [0,\ 1]^4 \qquad \text{with} \\ &\Phi \begin{pmatrix} \pi_{0A} \\ \pi_{1A} \\ q_{AB|A} \\ q_{AB|B} \end{pmatrix} = \begin{pmatrix} \pi_{0A} \cdot (1 - q_{AB|A}) \\ \pi_{1A} \cdot (1 - q_{AB|A}) \\ (1 - \pi_{0A}) \cdot (1 - q_{AB|B}) \\ 1 - \pi_{1A}) \cdot (1 - q_{AB|B}) \end{pmatrix} = \begin{pmatrix} p_{0A} \\ p_{1A} \\ p_{0B} \\ p_{1B} \end{pmatrix}. \end{aligned} \tag{21}$$

Note that $\Phi(\cdot)$ is injective if

$$\pi_{0A} \neq \pi_{1A} \quad \underline{\text{and}} \quad \pi_{0A} \notin \{0, 1\} \quad \underline{\text{and}} \quad \pi_{1A} \notin \{0, 1\}, \tag{22}$$

where the case $\pi_{0A} = \pi_{1A}$ reproduces the i.i.d. case.

To prove this, note that injectiveness is violated whenever there exist two distinct vectors $\gamma^{(1)}$ and $\gamma^{(2)}$, namely $(\pi_{0A}^{(1)},\ \pi_{1A}^{(1)},\ q_{AB|A}^{(1)},\ q_{AB|B}^{(1)})^T$ and $(\pi_{0A}^{(2)},\ \pi_{1A}^{(2)},\ q_{AB|A}^{(2)},\ q_{AB|B}^{(2)})^T$, leading to the same ϑ, namely $(p_{0A},\ p_{1A},\ p_{0B},\ p_{1B})$, i.e.

$$\begin{aligned} \pi_{0A}^{(1)} \cdot (1 - q_{AB|A}^{(1)}) &= \pi_{0A}^{(2)} \cdot (1 - q_{AB|A}^{(2)}) \\ \pi_{1A}^{(1)} \cdot (1 - q_{AB|A}^{(1)}) &= \pi_{1A}^{(2)} \cdot (1 - q_{AB|A}^{(2)}) \\ (1 - \pi_{0A}^{(1)}) \cdot (1 - q_{AB|B}^{(1)}) &= (1 - \pi_{0A}^{(2)}) \cdot (1 - q_{AB|B}^{(2)}) \\ (1 - \pi_{1A}^{(1)}) \cdot (1 - q_{AB|B}^{(1)}) &= (1 - \pi_{1A}^{(2)}) \cdot (1 - q_{AB|B}^{(2)}). \end{aligned} \tag{23}$$

Assuming $\pi_{1A} \neq 0$ as well as $1 - \pi_{1A} \neq 0$ (i.e. $\pi_{1A} \neq 1$), we rearrange this system of equations in Equation (23) by dividing the first equation by the second one and the third by the fourth and obtain

$$\frac{\pi_{0A}^{(1)}}{\pi_{1A}^{(1)}} = \frac{\pi_{0A}^{(2)}}{\pi_{1A}^{(2)}} \qquad \underline{\text{and}} \qquad \frac{1 - \pi_{0A}^{(1)}}{1 - \pi_{1A}^{(1)}} = \frac{1 - \pi_{0A}^{(2)}}{1 - \pi_{1A}^{(2)}}. \tag{24}$$

As for all cases in Equation (22) there is only one solution that accounts for both conditions from Equation (24) at the same time, the mapping $\Phi()$ in Equation (21) is injective in these situations.

[17] Illustrating this assumption by means of the example of Table 2, subgroup independent coarsening mechanism means that answering in a coarsened form, i.e., giving no suitable answer, does not depend on the receipt of the unemployment benefit.

In the i.i.d. case there are multiple solutions indeed (cf. Section 4.1), since both conditions from Equation (24) are always valid.

Because of the injectiveness of the mapping in Equation (21), for all cases in Equation (22) the system of equations

$$\begin{aligned}
(1-\hat{q}_{AB|A})\cdot\hat{\pi}_{0A} &= \frac{n_{0A}}{n_0} \quad (=\hat{p}_{0A}^{(MLE)}) \\
(1-\hat{q}_{AB|A})\cdot\hat{\pi}_{1A} &= \frac{n_{1A}}{n_1} \quad (=\hat{p}_{1A}^{(MLE)}) \\
(1-\hat{q}_{AB|B})\cdot(1-\hat{\pi}_{0A}) &= \frac{n_{0B}}{n_0} \quad (=\hat{p}_{0B}^{(MLE)}) \\
(1-\hat{q}_{AB|B})\cdot(1-\hat{\pi}_{1A}) &= \frac{n_{1B}}{n_1} \quad (=\hat{p}_{1B}^{(MLE)})
\end{aligned} \tag{25}$$

can be solved uniquely and one obtains the estimators

$$\begin{aligned}
\hat{\pi}_{0A} &= \frac{n_{0A}}{n_0}\frac{n_{1B}n_0 - n_1 n_{0B}}{n_{0A}n_{1B} - n_{0B}n_{1A}} \\
\hat{\pi}_{1A} &= \frac{n_{1A}}{n_1}\frac{n_{1B}n_0 - n_1 n_{0B}}{n_{0A}n_{1B} - n_{0B}n_{1A}} \\
\hat{q}_{AB|A} &= 1 - \frac{n_{0A}n_{B1} - n_{0B}n_{1A}}{n_{1B}n_0 - n_1 n_{0B}} \\
\hat{q}_{AB|B} &= 1 - \frac{n_{0A}n_{1B} - n_{0B}n_{1A}}{n_{0A}n_1 - n_{1A}n_0},
\end{aligned} \tag{26}$$

when these are inside the interval $[0,\ 1]$. Otherwise the maximum likelihood estimation is more challenging, but it can be shown that asymptotically ($n \to \infty$) the estimators of Equation (26) typically for all cases in Equation (22) will be in the interval $[0,\ 1]$.

It has to be emphasized that in practical applications one must carefully reflect the plausibility of the subgroup independent coarsening assumption of Equation (20). In addition, the restrictions

$$p_{0A} \leq \frac{P(X=0)\cdot p_{1B} - p_{0B}\cdot P(X=1)}{p_{1B} - p_{0B}\cdot\frac{p_{1A}}{p_{0A}}} \leq 1 - p_{0B}$$

offer, at least under large sample sizes, via the condition

$$n_{0A} \leq \frac{n_0\cdot n_{1B} - n_{0B}\cdot n_1}{n_{1B} - n_{0B}\cdot\frac{n_{1A}}{n_{0A}}} \leq n_{0A} + n_{0AB},$$

a possibility to check whether the subgroup independent coarsening is appropriate at all.

5.5 Generalization of subgroup independent coarsening mechanism

There are situations in which one might have an idea about the magnitude of the probabilities of precise observation in both subgroups. For instance, knowledge from former studies could be available concerning the question whether respondents who do receive Unemployment Benefit II rather report their income class in a precise or a coarse way compared to the respondents that do not receive this benefit.
Analogously to the generalization of CAR in Section 5.3, we now generalize the assumption of subgroup independent coarsening by considering the ratio between the subgroup specific probabilities of precise observation, i.e., $R_1 = \frac{q_{A|1A}}{q_{A|0A}}$ and $R_2 = \frac{q_{B|1B}}{q_{B|0B}}$, where the case of $R_1 = R_2 = 1$ corresponds to assuming subgroup independent coarsening.
If the estimation problem of Section 4.2 is adapted to this assumption, the corresponding mapping $\Phi(\cdot)$ in Equation (16) becomes injective for all cases in Equation (22). This can be shown by considering the equations

$$\begin{aligned}
\pi_{0A}^{(1)} \cdot q_{A|0A}^{(1)} &= \pi_{0A}^{(2)} \cdot q_{A|0A}^{(2)} \\
\pi_{1A}^{(1)} \cdot R_1 \cdot q_{A|0A}^{(1)} &= \pi_{1A}^{(2)} \cdot R_1 \cdot q_{A|0A}^{(2)} \\
(1 - \pi_{0A}^{(1)}) \cdot q_{B|0B}^{(1)} &= (1 - \pi_{0A}^{(2)}) \cdot q_{B|0B}^{(2)} \\
(1 - \pi_{1A}^{(1)}) \cdot R_2 \cdot q_{B|0B}^{(1)} &= (1 - \pi_{1A}^{(2)}) \cdot R_2 \cdot q_{B|0B}^{(2)} ,
\end{aligned}$$

where from dividing the first by the second equation and the third by the fourth equation, the conditions of Equation (24) follow. Thus, for all cases in Equation (22) there is only one solution and unique estimators

$$\begin{aligned}
\hat{\pi}_{0A} &= \frac{R_1 \cdot n_{0A}}{n_0} \cdot \frac{n_{1B} \cdot n_0 - R_2 \cdot n_1 \cdot n_{0B}}{R_1 \cdot n_{0A} \cdot n_{1B} - R_2 \cdot n_{0B} \cdot n_{1A}}, \\
\hat{\pi}_{1A} &= \frac{n_{1A}}{n_1} \cdot \frac{n_{1B} \cdot n_0 - R_2 \cdot n_1 \cdot n_{0B}}{R_1 \cdot n_{0A} \cdot n_{1B} - R_2 \cdot n_{0B} \cdot n_{1A}}, \\
\hat{q}_{AB|A} &= 1 - \frac{R_1 \cdot n_{0A} \cdot n_{1B} - R_2 \cdot n_{0B} \cdot n_{1A}}{R_1 \cdot (n_{1B} \cdot n_0 - R_2 \cdot n_1 \cdot n_{0B})}, \\
\text{and } \hat{q}_{AB|B} &= \frac{\text{numerator}}{\text{denominator}},
\end{aligned}$$

with numerator $= n_{0B} \cdot (R_1 \cdot n_{0A} \cdot n_{1B} - R_2 \cdot n_{0B} \cdot n_{1A})$ and denominator $= n_0 \cdot (R_1 \dot{n}_{0A} \cdot n_{1B} - R_2 \cdot n_{0B} \cdot n_{1A}) - R_1 \cdot n_{0A} \cdot (n_{1B} \cdot n_0 - R_2 \cdot n_1 \cdot n_{0B})$ are obtained, when they are in the interval $[0,\ 1]$. One can note that for $R_1 = R_2 = 1$ the estimators of Equation (26) result. The problem that the estimators are not within $[0,\ 1]$ already discussed in Section 5.4 as well as the considerations concerning generalizations to

non-binary response variables Y sketched in Section 5.3 are applicable in this context as well.
Again, inclusion of partial knowledge is possible by regarding R_1 and R_2 as sensitivity parameters and considering all estimators resulting from incorporating a region of plausible values R.

6 Concluding remarks

We presented a maximum likelihood analysis of categorical data under epistemic data imprecision. Our approach working with possibly set-valued maximum likelihood estimators overcomes the dilemma of the precise probability based approaches, often damned to debilitate contentual conclusions by the need to incorporate unjustified formal assumptions to ensure identifiability of parameters. The explicit reliance on an observation model specifying the coarsening process allows us to incorporate properly auxiliary information whenever it is present, in order to refine appropriately estimates derived from the empirical evidence alone.
The crucial arguments were developed, mutatis mutandis, for the i.i.d. case as well as a logistic regression based on one (or more) categorical covariates. From the applied point of view, an extension to metrical covariates is highly desirable. Although then a subgroup specific investigation is not possible any more, appropriate generalizations seem achievable in further work, especially for situations where sensitivity parameters can be determined. However, to allow estimation of the underlying distribution from the data and to maintain the metric character, (partially) parametric modelling is needed. This implicitly restricts the set of distributions considered and in particular raises further issues in the understanding of statistical models as discussed, e.g., in Section 3.1 of [19] for linear regression modelling.
In addition to this, the invariance property of the likelihood under different parametrizations, which is the technical basis of our results, offers two further directions of generalization. Further work may utilize these relationships beyond maximum likelihood estimation, in order to derive likelihood-based regions taking finite sample variability into account explicitly. These estimators also should be compared to confidence intervals derived along the lines of [27] in those situations where an appropriate sensitivity parameter could be determined. Another area of further research is the consideration of other "deficiency" processes, most notably misclassification, which can be formalized in a very similar way. Our methodology therefore offers an alternative to, and a generalization to logistic regression of, recent work on misclassification from a partial identification perspective [15, 12].

References

[1] A. Agresti. *Categorical Data Analysis*. Wiley, 2013[3].

[2] T. Augustin, G. Walter, F. Coolen. Statistical inference. In: T. Augustin, F. Coolen, G. de Cooman, M. Troffaes (eds.): *Introduction to Imprecise Probabilities*, Wiley, 2014, pp. 135–189.

[3] G. Cassella, R. Berger. *Statistical Inference.* Duxbury Pacific Grove, CA, 2002.

[4] M. Cattaneo, A. Wiencierz. Likelihood-based imprecise regression. *Int. J. Approx. Reasoning*, 53:1137–1154, 2012. [based on an ISIPTA '11 paper]

[5] G. de Cooman, M. Zaffalon. Updating beliefs with incomplete observations. *Artif. Intell.*, 159:75–125, 2004.

[6] I. Couso, D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *Int. J. Approx. Reasoning*, 55, 1502–1518, 2014.

[7] A. Dobra, S. Fienberg. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *P. Natl. Acad. Sci. USA*, 97: 11885–11892, 2000.

[8] D. Heitjan. Ignorability and coarse data: Some biomedical examples. *Biometrics*, 49:1099–1109, 1993.

[9] D. Heitjan, D. Rubin. Ignorability and coarse data. *Ann. Stat.*, 19:2244–2253, 1991.

[10] G. Imbens, C. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72:1845–1857, 2004.

[11] J. Kalbfleisch, R. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, 2002[2].

[12] H. Küchenhoff, T. Augustin, A. Kunz. Partially identified prevalence estimation under misclassification using the kappa coefficient. *Int. J. Approx. Reasoning* 53:1168–1182, 2012. [based on an ISIPTA '11 paper]

[13] R. Little, D. Rubin, *Statistical Analysis with Missing Data*. Wiley, 2002[2].

[14] C. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

[15] F. Molinari. Partial identification of probability distributions with misclassified data. *J. Econom.*, 144:81–117, 2008.

[16] E. Nordheim. Inference from nonrandomly missing categorical data: An example from a genetic study on Turner's syndrome. *J. Am. Stat. Assoc.*, 79:772–780.

[17] J. Plass, P. Fink, N. Schöning, T. Augustin. Statistical modelling in surveys without neglecting "the undecided": Multinomial logistic regression models and imprecise classification trees under ontic data imprecision. *under review for ISIPTA 2015*. See also: Technical Report, Number 179, Dep. Statistics, LMU Munich, 2015 (url: www.epub.ub.uni-muenchen.de/23816).

[18] C. Rao. Maximum likelihood estimation for the multinomial distribution. *Indian J. Stat.*, 18:139–148, 1957.

[19] G. Schollmeyer, T. Augustin. Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *Int. J. Approx. Reasoning*, 56:224–248, 2015. [based on an ISIPTA '13 paper]

[20] J. Stoye. Partial identification and robust treatment choice: An application to young offenders. *J. Statistical Theory and Practice*, 3:239–254, 2009.

[21] E. Tamer. Partial identification in econometrics. *Annu. Rev. Econ.*, 2:167–195, 2010.

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

[23] M. Troffaes, F. Coolen. Applying the imprecise Dirichlet model in cases with partial observations and dependencies in failure data. *Int. J. Approx. Reasoning*, 50:257–268, 2009.

[24] G. Tutz. *Regression for Categorical Data.* Cambridge University Press, 2011.

[25] L. Utkin, T. Augustin. Decision making under imperfect measurement using the imprecise Dirichlet model. *Int. J. Approx. Reasoning*, 44: 322–338, 2007. [based on an ISIPTA '05 paper]

[26] L. Utkin, F. Coolen. Interval-valued regression and classification models in the framework of machine learning. In: F. Coolen, G. de Cooman, T. Fetz, M. Oberguggenberger (eds.), *ISIPTA'11*, pp. 371–380.

[27] S. Vansteelandt, E. Goetghebeur, M. Kenward, G. Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Stat. Sin.*, 16:953–979, 2006.

[28] M. Zaffalon, E. Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *J. Artif. Intell. Res.*, 34:757–821, 2009.

[29] Z. Zhang. Profile likelihood and incomplete data. *Int. Stat. Rev.*, 78:102–116, 2010.