Statistical modelling of categorical data under ontic and epistemic imprecision:

Contributions to power set based analyses, cautious likelihood inference and (non-)testability of coarsening mechanisms

Julia Irina Plaß

München 2018

Statistical modelling of categorical data under ontic and epistemic imprecision:

Contributions to power set based analyses, cautious likelihood inference and (non-)testability of coarsening mechanisms

Julia Irina Plaß

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität München

vorgelegt von
Julia Irina Plaß
aus Nürnberg

München, den 31. Januar 2018

Erstgutachter: Prof. Thomas Augustin
Zweitgutachterin: Prof. Inés Couso
Drittgutachterin: Prof. Frauke Kreuter
Tag der Einreichung: 31. Januar 2018
Tag der Disputation: 08. Mai 2018

Acknowledgement

I would like to express my sincere gratitude to everyone who contributed to this dissertation! A special thanks goes to ...

... *Thomas Augustin for the pleasant collaboration, motivation, trust, support and all the very helpful advice, or in short, for the excellent supervision. Thank you for always taking time and encouraging me, especially when the feared ISIPTA deadline nears!*

... *Inés Couso and Frauke Kreuter for their willingness to fill the role of the external reviewers and Helmut Küchenhoff and Volker Schmid for their willingness to be part of the examination committee.*

... *my coauthors Thomas Augustin, Marco Cattaneo, Paul Fink, Christian Heumann, Aziz Omar, Norbert Schöning and Georg Schollmeyer for the great collaboration.*

... *Marco Cattaneo for the kind invitation to the School of Mathematics and Physical Sciences at the University of Hull. Thank you, Marco, for very helpful discussions, significantly pushing forward our project about reliable regression analysis. Moreover, I appreciate the hospitality of Andrea Wiencierz and Marco Cattaneo planning plenty of activities to make my time in England perfect!*

... *Francesca Biagini, my mentor in the LMUMentoring program. Thank you also for the financial support that gave me the opportunity to present our work at several conferences. Meeting researchers from similar topics definitely inspired further research.*

... *Eyke Hüllermeier and Sabine Zinn for the kind invitation to present our work in the Intelligent Systems group at the University of Paderborn and the Leibniz Institute for Educational Trajectories, respectively. Thank you also for the fruitful discussions.*

... *the Research Data Center at the Institute for Employment Research, Nuremberg, for the access to the PASS data. Moreover, I thank all organizers of the "mouse movement" study, in particular Malte Schierholz, for including our question(s).*

... *all former and current members of my working group: Thank you Andrea, Aziz, Christoph, Eva, Georg, Johanna, Marco, Patrick, Paul and Thomas for the great collaboration, stimulating discussions at the research seminar, your unique indecision in the choice of restaurants and an unforgettable time!*

... *all remaining former and current colleagues at the Department of Statistics for the excellent general atmosphere. Thank you, Jona and Sarah for the great breaks in the Eisbach, thank you for the ice-cream, Margret!*

... *"knorke" office 248! Thank you for all the singing, drumming, chocolate and the jumps in the air. I would like to have you in my office forever!*

... *Ari, Basti, Carola, Christoph, Georg, Eva, Oscar and Silke for proofreading parts of this thesis, especially Malte and Matthias for reading through the sections about survey statistics.*

... *the best family and the best boyfriend who always support me, understand me and are there for me.*

Zusammenfassung

Grobe Daten, d.h. Daten, die nicht in der ursprünglich gewünschten Genauigkeit beobachtet werden, entstehen aus den verschiedensten Gründen. Beispielsweise kann es sein, dass Befragte sich nicht zwischen Antwortmöglichkeiten (wie z.B. Parteien) entscheiden können oder, dass sie – insbesondere bei sensitiven Fragen – nicht bereit sind, genauere Informationen preiszugeben. Während die Variable im ersten Beispiel von Natur aus impräzise Werte aufweist, liegt in der zweiten Situation ein impräziser Beobachtungsprozess eines präzisen Wertes zugrunde. Somit weisen die Fälle auf zwei sich grundsätzlich unterscheidende Interpretationen von groben Daten hin, die in der Literatur mit ontischer und epistemischer Datenimpräzision bezeichnet werden.

Diese kumulative Dissertation verfolgt das Ziel einer vertrauenswürdigen statistischen Modellierung grober Daten, welche die gesamte verfügbare Information – und nur diese – ausnutzt. Dabei werden eine grobe, kategoriale Responsevariable und präzise, kategoriale Kovariablen betrachtet. Die Arbeit motiviert das Thema, indem das Erheben von groben kategorialen Daten als mögliche Strategie aufgezeigt wird, verschiedene bei der Beantwortung von Surveyfragen auftretende Fehler zu minimieren. Nach einer Einordnung der Arbeit in die allgemeine Literatur werden die eingehenden Beiträge zusammengefasst, Querverbindungen herausgearbeitet und Ideen für die weitere Forschung skizziert. Abschließend wird nochmals Bezug zur Surveyforschung genommen, wobei sich zeigt, dass die Vorschläge dieser Arbeit auch von aktuellen Entwicklungen, wie der Erhebung von Paradaten, profitieren. In die Dissertation geht ein Beitrag (*Beitrag 1*) ein, der sich mit ontischer und vier Beiträge (*Beitrag 2* bis *5*), die sich mit epistemischer Datenimpräzision befassen.

Durch Zulassen von Mehrfachantworten und die Betrachtung derselben als eigene Entitäten wird in ***Beitrag 1*** ein neuer Weg für den Umgang mit Antworten von Unentschlossenen aufgezeigt. Dieser Wechsel des Zustandsraumes zur Potenzmenge des ursprünglichen Zustandsraumes wird für das multinomiale Logitmodell und Klassifikationsbäume ausgearbeitet und durch die Daten der German Longitudinal Election Study (GLES) illustriert.

Ein wesentlicher Bestandteil der Arbeit, der den verbleibenden Beiträgen gemein ist, ist durch die „vertrauenswürdige Maximum Likelihood Schätzung“ gegeben. Diese nutzt ein Beobachtungsmodell, um den Einbezug von inhaltlichen Informationen über den Vergröberungsprozess zu steuern. In ***Beitrag 2*** wird dieser Ansatz für den i.i.d. Fall und die logistische Regression entwickelt, wobei – wie in *Beitrag 3* und *4* – Illustrationen anhand der Daten der Panelstudie "Arbeitsmarkt und soziale Sicherung" (PASS) des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) erfolgen.

Beitrag 3 erweitert *Beitrag 2* um die Verwendung nicht-saturierter Modelle. Es werden zwei Herangehensweisen vorgestellt und der Einfluss der parametrischen Annahme an das Regressionsmodell auf die geschätzten Vergröberungsparameter wird untersucht.

In ***Beitrag 4*** wird die coarsening at random (CAR) Annahme mit der sogenannten subgroup independence (SI) bezüglich Identifizierbarkeit und Testbarkeit verglichen. Es werden Fälle charakterisiert, in welchen SI nicht nur punktidentifizierend wie CAR, sondern anders als CAR auch testbar ist, woraufhin der Likelihood Ratio Test für SI ausgearbeitet wird.

Da alle gängigen Ansätze für fehlende Daten in Small Area Estimation (SAE) starke Annahmen an den Fehlendmechanismus stellen, werden in ***Beitrag 5*** vorsichtige Versionen für bekannte SAE Schätzer entwickelt. Die Ergebnisse werden durch die Daten des ALLBUS illustriert.

Summary

There are different reasons why coarse data, i.e. data that cannot be observed in the originally required resolution, may arise. These include the inability to give a precise answer due to indecision between several categories (such as political parties) and lacking willingness to disclose more detailed information especially in case of sensitive questions. In the first example the variable shows values that are imprecise by nature in the sense that indecisive respondents are not able to choose a single category. Against this, an imprecise observation process of a precise value is underlying in the second situation. Both cases mentioned already point to the two fundamentally differing interpretations of coarse data, in literature referred to as ontic and epistemic data imprecision.

This cumulative PhD thesis aims at a reliable statistical modelling of coarse data that fully exploits – and at the same time restricts to – all available information, throughout focusing on a coarse categorical response variable and precisely observed categorical covariates. The present work starts with a motivation presenting the collection of coarse categorical data as a possible strategy to minimize some errors arising when answering survey questions. After embedding this work into the general literature, the involved contributions are summarized, links are elaborated and ideas for further research are discussed. Finally, it is referred to a survey context again, where the benefit of our proposals from recent developments, such as the collection of paradata, becomes directly apparent. This dissertation includes one contribution (*Contribution 1*) dealing with ontic and four contributions (*Contribution 2* to *5*) with epistemic data imprecision:

By allowing for multiple answers in questionnaires and regarding them as entities of their own, ***Contribution 1*** motivates a new way to deal with the answers of indecisive respondents. This change of the state space to the power set of the original state space is worked out for the multinomial logit model and classification trees and illustrated by the data of the German Longitudinal Election Study (GLES).

A crucial component of this work, framing the remaining contributions, is given by the reliable maximum likelihood estimation under coarse data. It uses an observation model to guide the procedure of including subject-driven auxiliary information about the coarsening. ***Contribution 2*** develops this approach in an i.i.d. and a logistic regression setting, where – as in *Contribution 3* and *4* – illustrations are based on the data of the German Panel Study "Labour Market and Social Security" (PASS) provided by the Institute for Employment Research (IAB).

Contribution 3 extends *Contribution 2* by now admitting non-saturated models. Two approaches are presented, and it is investigated how the parametric assumption on the regression model may affect the estimated coarsening parameters.

In ***Contribution 4*** the coarsening at random (CAR) assumption is compared to the so-called subgroup independence (SI) with regard to identifiability and testability. Situations are characterized where SI is not only point-identifying like CAR, but also testable unlike CAR and the likelihood ratio test for SI is elaborated.

Since all usual approaches for nonresponse in Small Area Estimation (SAE) make strong missingness assumptions, in ***Contribution 5*** cautious versions of prominent estimators from SAE are developed by exploiting the results of the reliable maximum likelihood approach. Results are illustrated by the data of the German General Social Survey (GGSS).

Contents

Contributions of the thesis

This PhD project is composed of the following five publications, referred to as *Contribution 1* to *Contribution 5*:

1. Plass, J. Fink, P. Schöning, N. Augustin, T. Statistical modelling in surveys without neglecting the undecided: Multinomial logistic regression models and imprecise classification trees under ontic data imprecision. In Augustin, T. Doria, S. Miranda, E. and Quaeghebeur, E. editors, *Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications*, pages 257–266, Rome, 2015. Aracne.

2. Plass, J. Augustin, T. Cattaneo, M. Schollmeyer, G. Statistical modelling under epistemic data imprecision: Some results on estimating multinomial distributions and logistic regression for coarse categorical data. In Augustin, T. Doria, S. Miranda, E. and Quaeghebeur, E. editors, *Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications*, pages 247–256, Rome, 2015. Aracne.

3. Plass, J. Cattaneo, M. Augustin, T. Schollmeyer, G. Heumann, C. Towards a reliable categorical regression analysis for non-randomly coarsened observations: An analysis with German labour market data. *Technical Report 206*, Department of Statistics, LMU Munich, 2017, accessible from `https://epub.ub.uni-muenchen.de/view/subjects/160102.html` *and currently under review for the Journal of the Royal Statistical Society: Series A (Statistics in Society).*

4. Plass, J. Cattaneo, M. Schollmeyer, G. Augustin, T. On the testability of coarsening assumptions: A hypothesis test for subgroup independence. *International Journal of Approximate Reasoning*, 90: 292–306, 2017.

5. Plass, J. Omar, A. Augustin, T. Towards a cautious modelling of missing data in small area estimation. In Antonucci, A. Corani, G. Couso, I. and Destercke, S. editors, *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications, Volume 62 of Proceedings of Machine Learning Research*, pages 253–264, PMLR, 2017.

Declaration of the author's specific contributions

All contributing papers are the result of a fruitful collaboration with several co-authors. By separately referring to each of the papers, in the following the own contribution of the author is clarified:

- *Contribution 1:* The suggested power-set based analysis and the forecasting part were developed by Thomas Augustin, Paul Fink and the author in joint discussions, where findings in Couso et al. (2014) and Plass (2013) served as a basis. While the elaboration of the power-set based idea for classification trees is supplied by Paul Fink, the author devoted to the multinomial regression part. The author constructed the data application including the "ontic" variable in consultation with Norbert Schöning and computed all results concerning the general theory and multinomial regression.
  Except for the Section 2.3 and 5.3 (drafted by Paul Fink), the draft of the paper was written by the author. In joint discussions the paper was improved. All authors contributed to revising the paper.

- *Contribution 2:* Thomas Augustin and the author jointly worked out the basic frame of this contribution. After presenting the current work in the working group's research seminar and at the WPMSIIP '14 (7th workshop on principles and methods of statistical inference with interval probability; cf. `http://www.sipta.org/blog/?p=342`), all authors developed the idea to exploit the connection between the latent and the observed world in the reliable maximum likelihood estimation. The introductory and concluding words as well as the part about the basic setting and the sketch of the basic argument were drafted by Thomas Augustin, where the draft of the remaining sections was written by the author. The data example was provided by the author. All authors contributed to revising the paper.

- *Contribution 3:* The project was set-up by Thomas Augustin and the author, who both continued working on *Contribution 2* by studying general categorical regression models for coarse dependent variables. The author elaborated a draft of the paper, which was revised in connection with the author's stay abroad at Marco Cattaneo's home University (University of Hull): Marco Cattaneo suggested considering the (relative) profile log-likelihood, and he and the author jointly derived some further results on how the parametric assumption on the regression model can affect the coarsening assumption. The draft including these new findings, the data example as well as all computations were provided by the author. All authors contributed with further remarks, where in particular Christian Heumann shared his knowledge on the latest state of research in missing data based on classical methods.

- *Contribution 4:* A preliminary version of this paper was published as a conference paper (cf., Plass et al., 2016), drafted by the author and revised by remarks of all the other authors. The work was then presented at the SMPS '16 conference (8th International Conference on Soft Methods in Probability and Statistics;

cf. `http://www.sbai.uniroma1.it/smps2016/index.php`), where it received the "IJAR Best paper award". Following the invitation to the special issue on "Soft methods in probability and statistics" of the International Journal of Approximate Reasoning, the conference version was significantly extended. A main new contribution was the study of the distribution of the test statistic under the null hypothesis, in big parts worked out by Marco Cattaneo and the author, who also conducted some simulation studies. The draft of this version was again written by the author and discussed by all the authors. All authors contributed to revising the paper.

- *Contribution 5:* The aim of this project was to bring together the research areas the authors are working in to be able to deal with nonresponse in Small Area Estimation (SAE) without making untenable assumptions about the missingness process. Aziz Omar shared his knowledge about Small Area Estimation (SAE) and pointed to prominent estimators that we decided to consider. The author developed an approach to include the reliable maximum likelihood approach into these SAE estimators. While the parts referring to SAE were drafted by Aziz Omar, the draft of the remaining parts was supplied by the author. In several meetings with all three authors participating, the basic concept was improved and rewritten. Specifically, Thomas Augustin indicated the problem of finding the bounds of the cautious LGREG-estimator and suggested a proper representation for the case without auxiliary information about the missingness process. While being supported by joint discussions with Aziz Omar, the author mainly developed the data example and computed the results. Aziz Omar solved the problem of the inclusion of design weights. All authors contributed to revising the paper.

1 Motivation: Why to collect coarse categorical data in surveys?

How do respondents typically act in surveys, when they are unable to report one of the provided (categorical) options? In case that "Don't know" (DK) is admitted as an answer, this option may represent the category of their choice. However, some respondents who could actually give a substantive response might also decide for DK, e.g. to protect private information or to minimize the efforts associated with the memory and decision process, and hence allowing for DK in surveys is debatable. Moreover, there are many cases where respondents are able to exclude several categories and hence their knowledge or attitude is too strong to be best expressed by the DK category. Here, we firstly turn to the principal discussion whether to provide the DK category or not. Afterwards, the idea of offering coarse categories, i.e. options that are not collected in the resolution originally intended in the subject matter context, is embedded as a main proposal of this work in order to maximize the information about the respondents' opinion.

While most authors (cf., e.g., Gilljam and Granberg, 1993; Krosnick and Presser, 2010; Poe et al., 1988) rather suggest to drop the DK option, there are also some holding the view that the DK option is needed to filter out respondents without real opinion or knowledge on a topic (cf., e.g., Vaillancourt, 1973). Also in more recent research it seems still inadequate to form a clear, general recommendation, since the respondents' answering behavior, and hence the different drawbacks arising with and without the DK option, may depend on several factors: Examples are the type of the question (like e.g. factual (cf., e.g., Poe et al., 1988) or interpretable question), the topic of the questionnaire (like e.g. sensitive topic or not), the interviewer and the mode of data collection (cf., e.g., Kreuter et al., 2008a).

Nevertheless, the problem can be approached by understanding the respondents' cognitive process: There is a wide consensus that respondents have to interpret the question first (step 1), then they have to retrieve all relevant, available knowledge (step 2) and form a judgement (step 3), which finally has to be converted into a response by comparing the supplied options with the own decision (step 4) (cf., e.g., Krosnick and Presser, 2010; Tourangeau et al., 2000, p. 8). In all steps cognitive efforts are needed, where so-called optimizing respondents invest some time and energy to conclude an answer, while so-called satisficing respondents show a low level of motivation (cf., e.g., Krosnick and Alwin, 1987). In this way, especially in case of non-trivial questions satisficing respondents might favor the DK option and perceive it as an "easy out". But also for optimizing respondents there are motives to choose DK in every cognitive step (cf., e.g., Krosnick and Presser, 2010), namely difficulties to understand the meaning of the question (step 1), poor knowledge/experience in the topic of the question (step 2 and 3) and problems with matching

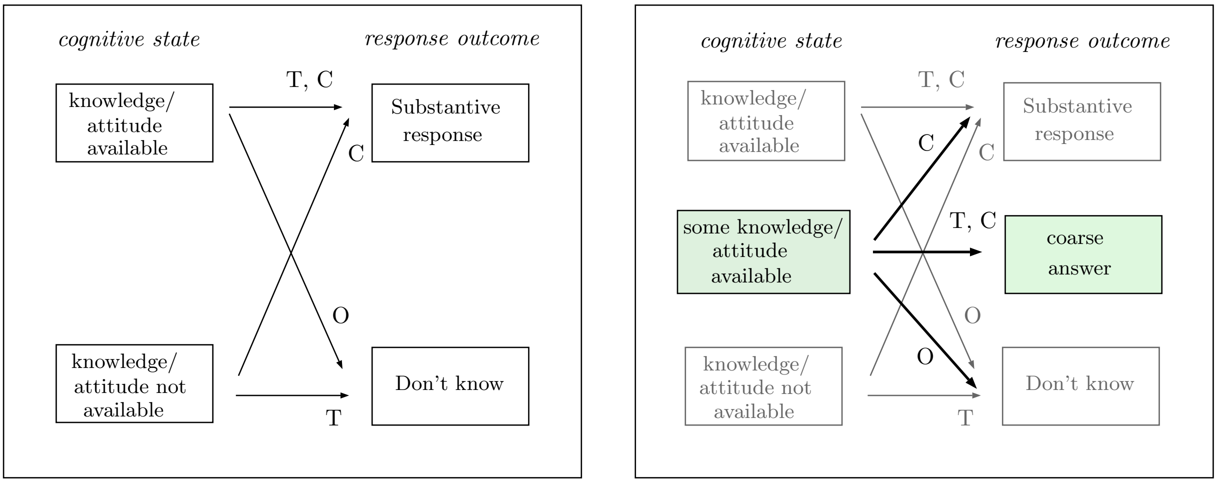

cognitive state
response outcome
knowledge/ attitude available
T, C
Substantive response
C
O
knowledge/ attitude not available
T
Don't know
cognitive state
response outcome
knowledge/ attitude available
T, C
C
C
Substantive response
some knowledge/ attitude available
T, C
coarse answer
O
O
knowledge/ attitude not available
T
Don't know

Figure 1.1: State-response mapping; left: classical model with two states (cf., Beatty and Herrmann, 1995); right: adapted model with the new parts colored.

the judgement with one of the supplied answers (step 4). Furthermore, in case of sensitive questions social desirability (cf., e.g., DeMaio, 1984) may encourage respondents to go for the DK option. Most respondents with the above listed motives for DK are actually able to report a meaningful answer, i.e. an answer that might not be in the required accuracy, but at least bears an increased information compared to DK. For that reason, Krosnick and Presser (2010, p. 285) finally recommend to refrain from an explicit DK category, but to ask follow-up questions that aim at the strength of previously reported attitudes.

An alternative consequence of investigating these reasons – promoted in the present work – is to meet the respondents' needs by offering different kinds of coarse options, i.e. respondents do no longer have to commit to one single option, but answers as "option a or option b" or scale point "1-3" are acceptable as well. Hence, respondents who have insufficient knowledge to form a judgment in the required accuracy (step 3) or cannot decide between some of the provided answers (step 4) are given the possibility to adequately express their answer. Moreover, respondents refusing to disclose their answer might be prompted to give at least some coarse information instead of DK or a wrong answer.

A further motivation for allowing for coarse answers is given by considering the state-response mapping for two cognitive states by Beatty and Herrmann (1995), illustrated in the left part of Figure 1.1. The two cognitive states (knowledge available, no knowledge available) are connected with the response outcome (substantive response, DK) by different response paths, representing either a "truthful response (T)" or an "error of commission (C)" or an "error of omission (O)". Thus, it is distinguished between a deliberate wrong answer (C) and withholding an actually available answer (O). We now extend this model by an intermediate cognitive state between "full" and "no knowledge", additionally considering the state of "some knowledge". This state resembles the second and third state of the mapping for four cognitive states by Beatty et al. (1998)'s, that for reasons of clarity is not considered here. But unlike in the four-state model, we make a consequent adaption in the response outcome by additionally admitting coarse answers (cf. right part of Figure 1.1), which enables respondents with some knowledge to make a truthful statement (T). Some errors remain, when respondents do not confess that they cannot give a substantive response (C), decide for a wrong (coarse) answer (C) or when there is still no willingness to disclose any information (O). These groups of respondents already existed before, while the related error will be reduced due to the respondents (truthfully) answering with a coarse answer. However, respondents driven by their intuition might give a (correct) substantive report instead of a weak DK in the left model, but decide for a (too) coarse answer in the right model, now where this option is available. To counter this behavior, the choice of a questioning technique that first asks about precise answers, providing coarse answers only in case of a previous nonresponse is recommendable.

In general, a two-step questioning technique of that kind is employed comparably rarely. However, in the context of the sensitive income question, one sometimes relies on this procedure (cf., e.g., Kennickell, 1996): For example, in the German General Social Survey (cf., GESIS Leibniz Institute for the Social Sciences, 2016) and in the German Panel Study "Labour Market and Social Security" (PASS, cf., Trappmann et al., 2010) income classes are directed to respondents refusing to disclose their precise income. In the latter one, several follow-up questions generate coarse categorical answers with different levels of accuracies. The gain of the explicit collection of coarse answers in the second step becomes directly apparent from Kuha et al. (2017), where initial nonresponders are simply encouraged to give a substantial (precise) response, decreasing (item-)nonresponse bias, but increasing measurement errors. By allowing initial nonresponders to state their answer in their individually required accuracy, measurement errors are expected to be reduced.

To conclude this motivation, we add some general points on coarse data in surveys, also embedding the considerations above. We addressed the quite common situation that the respondents of some survey have to choose between several (categorical) alternatives. In fact, the DK category represents a coarse answer, namely the one where there is no information on specific categories at all. Additionally, the considerations above motivate a more extensive collection of coarse categorical data, recommending that one should not only restrict to the DK option, but also gather coarse answers with different levels of accuracies. Further stressing the gain of information resulting from collecting coarse data by a questioning technique as accomplished by the PASS study and a proper statistical modelling of these data is part of this work.

Moreover, we discussed a variety of reasons for reporting DK, which points to the ambiguous meaning of this category from uncertainty about the interpretation of the question, via lacking knowledge about the topic or willingness to give a (precise) answer through to indecision about the answer. Consequently, a differentiated use of this category is strongly advisable (cf., e.g., Sanchez and Morchio, 1992), where one should make an explicit distinction between "Don't know (because of lack of knowledge in the topic)", "Undecided" and "Prefer not to say / Refused". The claim of this work to distinguish between the so-called ontic and epistemic data imprecision (cf., e.g., Couso and Dubois, 2014) is closely linked to this point. We start by explaining this differentiation in the next section.

2 Current state of research and aim of this work

2.1 General literature review and gaps that are filled by this work

Every analysis of coarse (categorical) data should be preceded by a careful distinction between ontic and epistemic data imprecision (cf., e.g., Couso and Dubois, 2014), because the aim of the analysis and the way to proceed are strongly reliant upon the type of coarse data at hand. The origin of this differentiation can be found in the two interpretations of sets (cf., e.g., Dubois and Prade, 2009):

- On the one hand, a set can be regarded as a collection of elements forming an entity of its own. In this **ontic** (or conjunctive) view, coarse values, such as $\{a, b\}$ composed of categories a and b, are the precise observation of something imprecise and we refer to data under ontic imprecision in this case. Multiple responses such as the languages one is able to speak (cf., e.g., Couso et al., 2014) or the opinion of partially undecided respondents can be mentioned as examples. In this work, the party affiliation of undecided respondents is studied. In pre-election studies several respondents might be indifferent between some parties in the sense that they themselves do not know which of those parties to elect. Hence, we address imprecise opinions, which can be collected in a precise way.

- On the other hand, a set can reflect incomplete information about a specific actually precise value. When observing $\{a, b\}$, in the **epistemic** (or disjunctive) view, one of those elements represents the true one. Thus, under epistemic data imprecision actually precise values are partly observed in a coarse way due to an underlying coarsening process (cf., e.g., Couso et al., 2014). Missing data represent a special case reflecting complete lack of knowledge about the observation. While surveys explicitly collect coarse categorical data beyond the missing case to a limited extent in surveys so far (cf. Section 1), this is different for continuous data: In this way, grouped answers, such as income classes, are frequently provided and interval data are indirectly produced by the rounding (or generally heaping, cf., e.g., Zinn and Würbach 2016) or the censoring problem (cf., e.g., Wang et al., 2001). An example where coarse categorical data implicitly occur is given by incomplete rankings (cf., e.g., Couso and Hüllermeier, 2018; Fahandar et al., 2017).

While under ontic data imprecision it is of main interest to find a way how the naturally coarse values can be incorporated in the analysis, under epistemic data imprecision one tries to understand the underlying coarsening structure to be able to estimate quantities referring to the latent variable. Due to this implied differentness of the purpose, we now separately review for each type of data imprecision established methods to deal with the respective challenge. Yet, there are a few publications jointly accounting for both types, i.e. uncertain ontic information (cf., e.g., Denœux et al., 2010, by relying on the formalism of belief functions). We set a focus on coarse categorical data and their handling in a survey context and show how our contributions fit into the existing literature.

2.1.1 Ontic data imprecision

The theory of random sets gives a proper framework for the formal representation of the considered kinds of data imprecision, where the respective interpretation of a random set determines the underlying view (cf., e.g., Couso et al., 2014). Although random sets where already indirectly addressed by Kolmogoroff (1933, p. 46) speaking of "a measurable region of the plane whose shape depends on chance", interest in this topic only increased when Matheron (1975) introduced random closed sets (cf., e.g., Stoyan, 1998) and plenty of applications followed (such as in image analysis, cf., e.g., Molchanov 2004, or in econometrics, cf., e.g., Molchanov and Molinari 2014).

Since we restrict to coarse categorical data, studying finite random sets is sufficient, where it is directly visible that random sets can be regarded as generalized random variables, representing a measurable mapping on the power set $\mathcal{P}(S)$ instead of the state space S itself (cf., e.g., Nguyen, 2006). In this way, a finite random set Y is given by

$$Y : \Omega \to \mathcal{P}(S), \tag{2.1}$$

where for the inverse image of each $A \subseteq S$ it has to be valid that $Y^{-1}(\{A\}) = \{\omega \in \Omega : Y(\omega) = A\} \in \mathcal{A}$ with $(\Omega, \mathcal{A})$ denoting the underlying measurable space (cf., e.g., Nguyen, 2006, p. 35). Understanding a random set as a multiple-valued random variable taking values in $\mathcal{P}(S)$ directly complies with the ontic view (cf., Couso and Dubois, 2014), where this formalization serves as a basis in our contribution. Alternatively, data under ontic imprecision can be formally regarded as functional data (cf., e.g., Guillaume and Dubois, 2015, p. 147). Although this is elaborated by González-Rodríguez et al. (2012) for fuzzy data, all results are directly applicable for coarse data, which represent the special case where all subcomponents of our precisely observed imprecise entity are identified with corresponding indicator functions.

Coarse categorical data under ontic imprecision are collected in surveys, whenever respondents are given the opportunity to tick more than one of the provided options. In this context, one refers to multiple response data. As illustration we consider the question

> Which mediums do you use to be informed about the events of the day? *(choose all that apply)* □ newspaper □ TV □ smartphone app □ internet □ radio

All elements within the power set of the state space $S = \{\text{newspaper,TV, smartphone app, internet, radio}\}$ (one may explicitly exclude the empty set) can then be selected as possible answers, where the answer "{TV, radio}" for example is interpreted as own, precisely observed answer expressing that one utilizes TV and radio as channel of information. A consequent analysis should be based on the power set.

While in multi-label classification this is exactly how one deals with multiple responses (cf., e.g., Tsoumakas and Katakis, 2006), in most remaining methods a concrete procedure that gains general acceptance is still missing. Agresti and Liu (1999) (p. 936) recommend to take the contingency table referring to all combinations of answers as a basis, which is equivalent to treating multiple responses as ontic sets, but in many cases the evaluation of multiple response data is still restricted to item specific frequencies (cf., e.g., Santos, 2000). When taking the ontic interpretation of multiple responses seriously, one should calculate frequencies for each combination of items (without order) instead, hence summing up the frequencies of all supersets containing the considered combination (cf., Couso and Dubois, 2014, p. 1505). The frequent ignorance of the nature of multiple responses is even more incomprehensible, when bearing in mind that most statistical software, such as R, SPSS, STATA or SAS, is already able to treat each combination as own category. In this way, statistical software either represents multiple answers as (ontic) sets and/or understands each item as yes (1) / no (0) question (cf., e.g., Koziol and Bilder, 2014). It is obvious that the number of "ontic" categories increases in the cardinality of the state space S and thus can become very large. But specifically in survey questions requiring to tick a specific number of boxes, such as "choose the three most important reasons", this problem is kept within a limit.

Due to the lack of clear rules how to deal with multiple responses in statistical analyses, in many cases one refrains from providing the "choose all that apply" supplement. One application where this is especially noticeable is given by the question about the voting intention in pre-election studies – at least we did not find any example, where multiple answers were allowed in this context. In this way, this question is frequently asked as follows:

> Which of the following parties do you favor?
> □ party A □ party B □ party C □ other party □ Don't know

Beyond the problem that there are different reasons prompting respondents to tick the DK category (cf. Section 1), there is a loss of information induced by undecided respondents who are able to exclude specific parties. For that reason, we recommend to add the "choose all that apply" instruction in questions of that kind and then to draw on the formal framework of Couso et al. (2014), interpreting the answers of different groups of "the Undecided" as ontic sets. In our work we also study how to incorporate these multiple responses induced by indecision into commonly used statistical procedures. In doing so, we throughout refer to the application of election studies. For that reason, an overview about common practices in this context is given next.

In fact, political analysts are frequently more interested in the "Undecided" than in respondents who are already firmly convinced about their voting intention, since undecided voters are able to swing an election (cf., e.g., Gawronski and Galdi, 2011). Thus, it is all the more surprising that characterizing the undecided respondents and investigating how they come to their decision is an understudied topic (cf., e.g., Orriols and Martínez, 2014). In our work, we aim at this goal and try to explain the (coarse) voting intention represented as ontic sets by means of several demographic variables, but also variables related to measures of election campaigns. In this way, we mainly refer to the ontic view, taking the current preferences of the undecided respondents seriously.

However, most political analysts address the epistemic view, studying the final decision when precise voting decisions are made or forced. Since they refrain from explicitly collecting the voting intention of undecided voters in most cases, they either base their voting prediction on the decided respondents only or allocate all DK responders to parties in a specific manner (cf., e.g., Bon et al., 2017). Usual ways of allocations are given by even assignments between the major parties or proportional assignments reflecting the voting intention of the decided respondents (cf., e.g., Martin et al., 2005, referred to as "missing (completely) at random" in the missing data literature). Since the answers of the undecided and the decided respondents may substantially differ, a substantial bias is expected for the voting prediction. More sophisticated, but rarely applied allocations of "Undecided" include imputation-based (cf., Fenwick et al., 1982) and Bayesian assignments both making use of information about other variables, while the latter one additionally exploits prior information about voting for each specific party (cf., Press and Yang, 1974). Even though these approaches account for some available information about the undecided respondents, an adequate understanding of uncertainty in prediction models is still missing (cf., e.g., Rothschild, 2015). However, in some cases the uncertain behavior of voters is at least captured in the data collection process. In this way, verbal statements, such as "Lean towards party A", or probabilistic votes , quantifying the degree of certitude in the party preference, are offered (cf., Burden, 1997; Delavande and Manski, 2010).

Although *Contribution 1*, which addresses this topic, mainly refers to ontic data imprecision, it also suggests an interval-valued prediction reflecting the underlying uncertainty. This is in accordance with the conception of epistemic data imprecision used in the remaining contributions. A corresponding literature review is given next.

2.1.2 Epistemic data imprecision

Coming back to the definition of a random set given in Equation (2.1), apart from its ontic interpretation applied in Section 2.1.1 there is also an epistemic one: Instead of considering a random set as a multiple-valued random variable on $\mathcal{P}(S)$, we can also understand it as a multiple-valued mapping $Y_{epist} : \Omega \to \mathcal{P}(S)$ representing the disjunctive set of precise random variables $Y_{precise} : \Omega \to S$ (cf., Couso and Dubois, 2014) that are compatible with the (incomplete) realizations of Y_{epist} and are often called selections. Hence, when taking the epistemic view, we interpret the random set as

$$\{Y_{precise}(\omega) \in Y_{epist}(\omega), \forall \omega \in \Omega\} \ . \qquad (2.2)$$

In our contributions concerning epistemic data imprecision (i.e. in *Contribution 2* to *Contribution 5*), the random variable Y, which refers to the true underlying construct, has a key role. In this way, Y denotes a specific selection $Y_{precise}$, when regarding the epistemic interpretation of a random set. Statistical inference about the distribution of Y represents our main interest, where maximum likelihood estimation is used as an estimation technique.

Since Y_{epist} is regarded as the collection of several precise models that can be inferred from the incomplete knowledge, point-identification of the distribution of Y is only guaranteed in special cases. Point-identification is a general property meaning that different values of parameters have to induce different probability distributions of the considered random variables (cf., e.g., Lehmann and Casella, 2006, p. 24). To ensure point-identification, classical statistics mostly breaks down the problem by including technical restrictions that are strong enough to point-identify the parameters of interest. In this way, in the context of coarse data strict assumptions on the coarsening process are incorporated. The origin of these assumptions and also of most approaches dealing with coarse data lies in the area of missing data. For that reason, commonly used approaches for missing data are presented first, then proceeding to the situation of coarse data beyond the missing case.

Current methods for dealing with missing data

In the missing data literature[1] the differentiation between various **types of missingness mechanisms**, i.e. "missing completely at random" (MCAR), "missing at random" (MAR) and "missing not at random" (MNAR), is essential. While the missingness is independent of the observed and the missing values under MCAR, under MAR it is dependent on the observed values and under MNAR even on the actual missing values (cf., e.g., Little and Rubin, 2014). In the context of likelihood inference, MCAR and MAR (plus parameter distinctness, cf., Little and Rubin 2014, p. 119) are especially desirable, since under these assumptions the complete-data likelihood is proportional to the likelihood ignoring the missingness mechanism (cf., Little and Rubin, 2014, p. 119), whose parameters are point-identified. However, it is important to be aware of the fact that the correct assumption about the true underlying mechanism is a necessary prerequisite to receive unbiased estimators. This point turns out to be a major difficulty, since testing of missingness assumptions is generally impossible (cf., e.g., Manski, 2003, p. 26). Hence, it is problematic that widespread techniques dealing with missing data, such as imputation or the EM-algorithm, are based upon the MAR assumption (cf., e.g., Jaeger, 2006). Moreover, complete-case and available-case analyses completely ignoring the missingness are still quite common (cf., e.g., Geva et al., 2013; Rombach et al., 2016).

Imputation methods aim at valid statistical inferences by replacing missing values by plausible ones. For this purpose, the imputed values are derived from a predictive distribution described by the observed values, where the way of drawing from this distribution varies for the different imputation methods, such as mean imputation, regression

[1] We throughout refer to missing data produced by item-nonresponse, cf., e.g., Kreuter (2013) to get an overview of commonly used methods addressing unit-nonresponse.

imputation or hot-deck imputation (cf., e.g., Weisberg, 2009). Recent examples recommending multiple imputation in a survey context are given in Pampaka et al. (2016), who refer to a case study with educational data, as well as Frick and Grabka (2014) and Spieß (2009) discussing this method for the data of the German Socio-Economic Panel (SOEP). The **EM algorithm** (cf., Dempster and Laird, 1977) is likelihood-based iterative procedure that relies on starting values for the parameters of interest and then determines the expectation of the joint distribution (complete-data likelihood) given the observation. This expectation provides the basis for re-estimating the parameters of interest, while this procedure then continues until some stability is achieved (cf., e.g., Schafer, 1997). Both, methodological reviews about dealing with missing data (cf., e.g., Dong and Peng, 2013) and practical applications (cf., e.g., Kariuki et al., 2015) include the EM-algorithm as a standard method.

Although one mostly sticks to the already mentioned procedures, there are also a few approaches that explicitly refrain from the MAR assumption and intend to **model the underlying missingness process** instead. In this way, selection models split the joint distribution[2] of the missingness variable[3] M and the variable Y into one part referring to Y (outcome model) and one part referring to M given the variable Y (generally called selection model, but here missingness model to distinguish it from the general term) (cf., e.g., Toutenburg et al., 2004). Choosing the missingness model represents a crucial difficulty of this procedure. A specific, popular variant of the selection model is the Heckman selection model (cf., Heckman, 1979), combining an outcome model and a missingness model as follows: Based on the assumption of a multivariate normal distribution, the correlation between the error terms of the two model equations is explicitly included. The selection bias is then fully traced back to this correlation and used to correct the estimators obtained under MAR (cf., e.g., Amemiya, 1985, for more details). Beyond the problem of imposed distributional assumptions, – as in the general selection model– finding variables that appropriately explain the missingness process definitely stays a challenging and mostly impossible task.

For that reason, a **systematic sensitivity analysis** is performed in some cases, regarding different missingness models that impose various, from a practical viewpoint conceivable missingness assumptions. Each (specific) precise missingness model point-identifies the distribution of Y, wherefore the parameter specifying the missingess process can be regarded as a kind of nuisance parameter, also called sensitivity parameter in this context (cf., e.g., Kenward et al., 2001). Considering the whole range of results inferred from different, reasonable missingness models, then gives a set-valued estimator of the distribution of Y. In this way, the idea of a systematic sensitivity analysis differs from a conventional sensitivity analysis (as e.g. performed in Goldsmith, 2005) simply investigating the impact of a deviation from the imposed assumptions without understanding it as a part of the result. Examples of a systematic sensitivity analysis in a categorical setting are given in

[2] This joint distribution corresponds to the complete-data likelihood.

[3] The missingness variable is an indicator variable with value 1, whenever the value of Y is missing, and 0 otherwise.

Baker et al. (1992), Kenward et al. (2001) Molenberghs et al. (1999), Molenberghs et al. (2001) and Nordheim (1984), where in the context of regression this topic is raised by Baker and Laird (1988) and Moreno-Betancur et al. (2015). Strongly related to selection models are pattern-mixture models. In pattern-mixture models one factorizes the joint distribution of Y and M in the opposite way, hence regarding the marginal distribution of M and the distribution of Y given M (cf., e.g., Little, 1993), where again systematic sensitivity analyses show to be beneficial (cf., e.g., Daniels and Hogan, 2000). However, in most contributions and also throughout this work, the representation of a selection model is used.

The idea of the methodology of **partial identification** (cf., e.g., Manski, 2003) resembles the one of systematic sensitivity analyses, except for proceeding in the opposite direction: While the collection of all results obtained under plausible assumptions about the missingness process is taken if a systematic sensitivity analysis is performed, partial identification starts by making no assumptions on the missingness process at all, but then successively includes all available, (weak) auxiliary information that is frequently not strong enough to ensure point-identification. This careful inclusion of auxiliary information about the missingness gradually refines the result in such a manner that imprecision is reduced, but does not disappear, except sufficient information was available (cf., e.g., Manski, 2005a). Thus, by admitting the possibility that parameters are partially identified, identification does no longer have to be taken as a binary concept in the sense that parameters are either identified or not (cf., e.g., Tamer, 2010). Practical examples for the inclusion of auxiliary information are for instance given in Jiang and Ding (2016).

Manski's original application of partial identification addresses the selection problem arising when the estimation of treatment effects on outcomes is the main goal (cf., e.g., Manski, 1989, 1990, 2003, 2005b; Stoye, 2009), such as the influence of the family structure on children's outcomes (cf., Manski, 1999). In this context, only tenable assumptions on the so-called counterfactuals ("what-if probabilities") are imposed (cf., e.g., Morgan and Winship, 2014, for more details about causal inference in general). Furthermore, areas forcing point-identification by strong, and sometimes questionable assumptions may profit from the underlying idea. Examples are given in Di Zio and Vantaggi (2017), Molinari (2008), Küchenhoff et al. (2012) and Tamer (2010), who exploit partial identification in statistical matching, misclassification and more general in econometrics.

Recently, Manski stressed the importance of the methodology of partial identification in official statistics. He postulates that the communication of the uncertainty attributable to the nonresponse error should be part of every dissemination of results (cf., Manski, 2015, 2016). Also referring to survey/official statistics, **in our work** we tie on the latter publications, but study the more general situation of coarse data (under epistemic imprecision). For that reason, a brief overview of commonly used methods for dealing with coarse data is given now. Since most methods for coarse data just draw on the approaches for missing data recalled in this section, the classification of missing data as a special case of coarse data becomes directly apparent.

Current methods for dealing with coarse data

Analogously as in the missing data situation, considering the likelihood under coarse data may lead us to the following question: "Which conditions allow a simplification of the complete-data likelihood in the sense that the coarsening can be ignored?" This question is answered by Heitjan and Rubin (1991) requiring parameter distinctness and **"coarsening at random" (CAR)**. In their definition of CAR they postulate that the probability of each fixed (coarse) observation does not depend on the true underlying value, as long as this value is consistent with the observation. In Jaeger (2005b) this assumption is called distributional CAR (d-car) and is distinguished from the **G-car** variant by Heitjan (1997), which gives a more direct extension of MAR by relying on a representation in terms of the coarsening variable G. This variable G is a generalization of M, not only differentiating between observed and missing values, but between several degrees of coarseness. Whenever asymmetric, mixture or probabilistic rounding/heaping is present (cf., e.g., Schneeweiß et al., 2010), the formalization of a coarsening variable G is widely spread. In these cases, one either models the rounding by the exceedance of certain thresholds of G (cf., e.g. Drechsler et al., 2015; Heitjan and Rubin, 1991, Example 2) or by a so-called rounding profile function (cf., e.g. Torelli and Trivellato, 1993; Schneeweiß and Komlos, 2009), e.g. in dependence of the interval width. Taking the rounding process into account in this way and relying on certain smoothness conditions (cf., Kendall, 1938), the estimation of the variable of interest's mean is nearly unbiased, while the variance can be adjusted by the Sheppard's correction (cf., Sheppard, 1897) extended in Schneeweiß and Komlos (2009) for cases beyond simple rounding. However, although one explicitly decides to model the rounding process, in most cases the rounding is assumed to be independent on the true underlying value, in the sense that one relies on the G-car assumption. Like most of the papers, this work also refers to the CAR variant as formulated by Heitjan and Rubin (1991), which is less restrictive compared to G-car (cf., Jaeger, 2005b). Nevertheless, the MCAR analogue for coarse data, i.e. "coarsening completely at random" (CCAR), is only reasonable when presented in terms of a variable G. Like G-car, CCAR requests the distribution of G to be independent of the true value, but now one refrains from requiring that this true value has to be consistent with the observation (cf. Heitjan 1994).

Another assumption that expresses a kind of lack of information about the observation process is the superset assumption (cf., e.g., Couso and Dubois, 2018; Hüllermeier and Cheng, 2015). It resembles the original CAR assumption, but switches the values that are held fixed. Hence, the probability of an observation given a fixed, true, underlying value is assumed to be constant for each observation that is compatible with this true value (cf., Couso and Dubois, 2018). In the context of updating probability distributions (cf., e.g., Grünwald and Halpern, 2003), the CAR assumption[4] naturally arises in a reverse[5] form (also called RCAR) (cf., Theorem 10 in van Ommen et al., 2016). In this context, the RCAR condition is not an assumption, but is rather obtained as a result whenever one pursues a minimax strategy (cf., Section 2.1 in van Ommen et al., 2016).

[4] more exactly strong CAR as defined in Jaeger (2005a)

[5] in the sense of switching the event and the condition

In this work, the CAR assumption is considered in a regression context and hence it is additionally conditioned on the values of the covariates. This assumption corresponds to the conventional assumption made in survival analysis that conditional on some covariates the censoring time is independent of the survival time. This is often referred to as "independent censoring". Since the estimators of survival rates may be biased whenever independent censoring is wrongly assumed (cf., e.g., Zheng and Klein, 1995), frameworks were developed that rely on dependent censoring, such as the copula-based approaches by Huang and Zhang (2008) and Emura and Chen (2016). Moreover, in *Contribution 2* a new coarsening assumption called subgroup independence (SI) is introduced: While under CAR the coarsening is independent of the values of Y, but dependent on the covariate values, the (in)dependence structure under SI is the other way around. We do not only study maximum likelihood estimation under these assumptions, but also investigate the (im)possibility of testing these assumptions. Although CAR is generally impossible to test, some hypothesis tests have been suggested in the literature, all relying on strong assumptions: For instance, testability of CAR can be achieved under the availability of instrumental variables and bounded completeness (cf., Breunig, 2017) or when distributional constraints on the structure of a network are incorporated (cf., Jaeger, 2006). Generally, the challenge remains to distinguish between situations where CAR is justifiably rejected or not rejected and situations where the test decision is meaningless, since the included additional assumptions were wrongly made.

Commonly used approaches for missing data have been extended for coarse data, such as imputation (cf., e.g., Heitjan and Rubin, 1990; Kim and Hong, 2012) and the EM-algorithm. While both imputation and the EM-algorithm are always reliant upon the quite restrictive CAR assumption (cf., e.g., Jaeger, 2006), the latter one was additionally investigated to be reasonable only in situations where the coarse observation of each true value is predetermined, as e.g. satisfied in case of grouped data (cf., Couso and Dubois, 2016). Since the EM-algorithm is likelihood-based as our approach, a more detailed comparison of the respective results is of interest. We postpone this purpose to Section 3.2.4, where all necessary notations already have been clarified. In this connection we will also contrast our idea to that of other likelihood-based ideas, relying on different optimization strategies, such as the minimax (cf., e.g., Guillaume and Dubois, 2015) or the maximax (cf., e.g., Hüllermeier 2014) approach. Jaeger (2016) presents an algorithm whose underlying idea resembles the maximax strategy: His AI & M (adjusting imputation and maximization) procedure explicitly refrains from concrete coarsening assumptions, but relies on the view that given the data some mechanisms appear to be more likely than others.

Imposing the CAR assumption or making use of specific optimization strategies to force point-identified parameters is not always justified. For that reason, there are some approaches that only include **weak or even no assumptions** about the coarsening process. Some likelihood-based approaches of that kind are given in Cattaneo and Wiencierz (2012) and Zhang (2010), where set-valued results are obtained by considering the maxima of the respective profile likelihood function, and in Denœux (2014) relying on belief functions. Our contributions on epistemic data imprecision get in line with these approaches, but mostly regard the problem in a regression context with a coarse categorical response variable and categorical precisely observed covariates. A general framework for learning from coarse data is given in Couso and Sánchez (2016) and Sánchez and Couso (2018).

There are also **Bayesian approaches**, describing information about the coarsening process via corresponding prior assumptions. An example is given in Zaffalon and Miranda (2009): Just as our approach, their conservative inference rule (CIR) aims at refining the result under total ignorance of the missingness/coarsening assumption, where it practically coincides with the conservative updating rule presented in De Cooman and Zaffalon (2004). By applying CIR, a compromise between a too optimistic and too pessimistic knowledge about the missingness/coarsening process is given, assuming CAR in the context of some variables and total ignorance about the coarsening process of other variables. Our approach does not force us to decide for one extreme case, i.e. either no or complete knowledge about the coarsening, but allows us to incorporate (arbitrary) weak coarsening assumptions in a careful way. For instance, the coarsening probability can be assumed to be higher for specific groups of respondents compared to others. The inclusion of such auxiliary information is an important part of our contribution (cf. Section 3.2.1 and the summary of *Contribution 2* in Section 3.2.3).

2.2 Aim of this work

This section is closed by briefly outlining the highlights of this work. This PhD thesis aims at providing ways for a natural inclusion of all tenable information about the coarse structure of the data into commonly used statistical models. In particular, we promote ...

- ... the explicit differentiation between epistemic and ontic data imprecision in data collection as well as in data analysis.
- ... the explicit collection of coarse categorical data: We allow "the Undecided" to report multiple answers and give respondents with insufficient knowledge in the topic of the question or poor willingness to disclose their answer the opportunity to give coarse answers. This can also be regarded as a possible strategy to minimize the errors arising in the four cognitive states of respondents when answering survey questions (cf. Section 1).
- ... a medium to regulate the inclusion of auxiliary information, which allows to incorporate subject-driven coarsening assumptions instead of strong, untestable ones, such as CAR. In this way, auxiliary information about the coarsening process that is not sufficient to point-identify the parameters of interest can now explicitly be exploited, while it would have to be left out of consideration under traditional approaches.
- ... a new coarsening assumption called subgroup independence, which is indeed testable in specific settings. A hypothesis test for SI is proposed.

Throughout, the setting is restricted to a coarse categorical response variable and categorical covariates that are precisely observed. A focus is set on epistemic data imprecision.

3 About the contributing material: Relations, summaries and outlooks

In this chapter we take a closer look at the papers contributing to this thesis. All the contributions address methods for carefully handling coarse data exploiting all available information about the coarse structure of the data. When referring to ontic data imprecision, i.e. in *Contribution 1*, this goal is reflected by explicitly collecting coarse data and taking its coarse nature seriously in the analysis, while under epistemic data imprecision, i.e. in *Contribution 2* to *5*, only tenable assumptions about the coarsening process are contemplated. To some extent, *Contribution 1* addresses epistemic data imprecision as well: In this case, interval-valued forecasts are suggested by relying on Dempster's lower and upper probability (cf., Dempster, 1967). When no assumptions about the "decision process" are included, these intervals are in line with the ones obtained from the framework used in the contributions referring to epistemic imprecision. Under the availability of some auxiliary information, such as "undecided respondents rather vote for the SPD compared to the Green party", the interval-valued forecasts may profit from the estimation framework developed in *Contribution 2*.

This chapter is again structured by the respective type of data imprecision (cf. Section 3.1 for the ontic and Section 3.2 for the epistemic case). Both parts give an overview of the corresponding contributions followed by some remarks and ideas for future research. Since the contributions referring to epistemic data imprecision are built upon a joint basis, it is reasonable to start the associated part more generally by presenting this common ground as well as interrelations between the included contributions.

3.1 Ontic data imprecision: Contribution 1

3.1.1 Summary

As already discussed in Section 2.1.1, in most surveys there is no clearly favored procedure to deal with "The Undecided" yet. The problem already starts in the data collection process, where undecided respondents do not have the opportunity to express their current opinion adequately; they are either forced to give a precise answer or may choose an additional category "Don't know" (cf. the debate around the DK category, also addressed in Section 1). By proceeding in the latter way, different types of undecided respondents cannot be distinguished within the analysis. This induces a remarkable loss of information, since from answers as "a or b" one can at least conclude that answers as "c" or "d" are ruled out, when a state space $S = \{a,\ b,\ c,\ d\}$ is underlying. As if this wasn't enough, frequently the analysis is based on the decided respondents only, which – due to a potential systematic difference between decided and undecided respondents – may provoke biased results (cf., e.g., Bon et al., 2017).

Allowing for multiple answers and regarding them as categories of their own gives a way out of the explained dilemma: One not only respects the heterogeneity in the group of undecided respondents, but even reflects the opinion of the respondents in the most informative way. More formally, this corresponds to reinterpreting a random conjunctive set (cf., Couso and Dubois, 2014), defined as a measurable mapping into the power set, as precise random object (cf. Section 2.1.1, in particular Equation (2.1)). Hence, one simply needs to extend the original state space S of our variable of interest to the power set of S (without the empty set) to account for ontic data imprecision. Relying on this new state space $S^* = \mathcal{P}(S) \setminus \{\emptyset\}$ is the only thing that changes, while the statistical methods stay the same. We stress this point by elaborating the idea for the multinomial logit model and classification trees. In case of the multinomial logit model with ontic data imprecision in the response variable, the power-set based analysis not only gives us category specific regression coefficients for each precise, but also for each coarse category, perfectly representing the conception of understanding different types of undecided respondents as groups of their own. In the context of classification (trees), relying on a class variable with values within S^* instead of S already represents a comparatively well-established procedure, then referred to as multi-label classification (cf., e.g., Tsoumakas and Katakis, 2006).

Illustrating the ontic approach faces us with new challanges: As far as we know, there is not any pre-election study that allows undecided respondents to express their voting intention by multiple answers. The "German Longitudinal Election Study (GLES) 2013" (cf., Rattinger et al., 2014) at least collects some information on the certainty of the voting intention as well as the assessment of several parties, along with the current, precise (and thus partly forced) voting intention. This gives us the opportunity to construct a new variable "ontic" (cf. Table 1 of *Contribution 1* and corresponding explanations), partly consisting of multiple answers, which reflect the respondent's indecision. In this way, we can compare the obtained results based on this new variable with the ones from a traditional analysis, only including answers of decided respondents. The regression estimates from the multinomial logit model in both analyses indicate remarkable differences (cf. Table 4 of *Contribution 1*), partly even connected with a change in sign. Due to the underrepresentation of undecided respondents induced by the underlying sampling design, the results are even expected to differ to a higher extent. Moreover, the general reduction of the sample size in the course of the construction of the ontic variable might cause to vanish the significance of some estimators.

The attractiveness of the power-set based analysis is especially given by the generality of the idea in the sense that all statistical methods and their refinements (such as penalization in regression analysis) can account for ontic data imprecision. As now appropriate statistical methodology has been proven to be available, we strongly recommend allowing for multiple answers directly within questionnaires. In particular, in election studies this is getting more and more important, because an increasing number of voters make their vote decision shortly before the election day (cf., e.g., Dassonneville, 2016). The transfer to S^* may substantially increase the computational complexity, but by restricting to the most important groups, e.g. determined via substance matter reasons (indecision between certain parties is more likely compared to others) or regularization techniques, one can cope with this difficulty.

Furthermore, we calculate interval-valued forecasts, which demands to change the underlying perspective in the sense that we make an epistemic reinterpretation of the data (also underlying *Contribution 2* to *5*, summarized in Section 3.2.3): In the election example, we assume that the election day has come, forcing the respondents to make a decision. Using the GLES'13 data to forecast the proportion of respondents electing a specific party, we now understand coarse answers no longer as entity of their own, but as incomplete knowledge. We formally anchor this idea by relying on the notion of ill-known random variables (cf. Equation (2.2)), hence considering several precise models that – due to this incomplete knowledge – cannot be distinguished (cf., Couso and Dubois, 2014), which then gives us proper interval-valued forecasts (cf. p. 265 of *Contribution 1*).

3.1.2 Comments and perspectives

The framework presented in *Contribution 1* can be adapted for several modifications of the setting addressed there. Not only extensions to various categorical regression models besides the multinomial logit model, but also to coarse ordinal response variables are straightforward. In this section, apart from a brief motivation and some considerations with regard to these points, also first results from a recent study providing multiple responses are given. In this way, one is able to evaluate the suggested new formulation of the question without being forced to rely on a artificially constructed variable "ontic".

More general discrete choice models

In our application example we used the multinomial logit model, which only represents one specification of a variety of discrete choice models (cf., e.g., Train, 2009, to get an overview). While the multinomial logit model is restricted to describe the voting intention by voter characteristics, the conditional logit model introduced by McFadden (1973) shows a linear predictor that is limited to account for the party attributes as perceived by the voter, mostly relying on the ideological distance between the party and the voter (concerning different issues, such as environment or taxes etc.) (cf., e.g., Alvarez and Nagler, 1998). In current practices of voting research more flexible modelling approaches are favored that combine both model specifications, thus considering the linear predictor (cf., e.g., Dow and Endersby, 2004)

$$\eta_{ij} = x_i \cdot \beta_j + z_{ij} \cdot \gamma \;, \tag{3.1}$$

where the index i refers to the voter and the index j to the party. The first component of the sum in (3.1) corresponds to the linear predictor in the multinomial logit model, where party specific coefficients β_j are included, while the covariates x_i are constant across party choices. The idea of the conditional logit model is expressed by the second component of the sum in (3.1), where a global regression coefficient γ and voter characteristics z_{ij} varying through party choices are incorporated.

Due to the general character of our idea to deal with "the Undecided", we can directly apply the ontic approach to the combined model with the linear predictor in (3.1). Apart from the adaption of the state space S to S^*, nothing has to be changed. Just like in the multinomial logit model, party-specific regression coefficients β_j for all groups of undecided respondents are a direct consequence of this adaption. Since the coefficients γ from the conditional logit model component in (3.1) are global, no additional modifications are needed here.

Coarse ordinal data

So far, we addressed a coarse categorical response variable of nominal scale. However, the phenomenon of indecision may also arise in connection of variables showing an ordinal scale of measurement. Considering rating scales, there are different possibilities to deal with indecision. Some surveys provide a DK or "Undecided" category additionally to the rating scale; consequences of this procedure have already been discussed in Section 1 and Section 2.1.1. Whenever an additional category of that kind is omitted, the behavior of "the Undecided" is guided by the number of response categories: Offering an even number of response categories forces undecided respondents to a decision, while an odd number prompts them to choose the midpoint of the scale. One generally detects a "tendency to the middle" (cf., e.g., Friedman and Amoo, 1999), which is undesirable, since the meaning of the middle category as "neutral response" might be distorted by comprising e.g. undecided and satisficing respondents as well as respondents avoiding social embarrassment (cf., e.g., Sturgis et al., 2014).

Against this background, Iannario and Piccolo (2011) and others consider the CUB model[1]. The core of this approach is the distinction between the attractiveness towards the item (also called feeling) and fuzzyness around the final choice (also referred to as uncertainty). While the feeling is traced back to several individual characteristics such as age or previous experience, the uncertainty component is determined by circumstances as tiredness, time devoted to the question or the incentive to satisfice. The final CUB model is then composed as a mixture of the feeling and the uncertainty component. In this way, the CUB model deals with the "tendency to the middle" problem by an explicit inclusion of a random variable U describing confounding factors, called "uncertainty". The original model represents the special case where any "tendency to the middle" is neglected, just utilizing $U \sim \mathcal{U}(1, k)$ (with $\mathcal{U}$ representing the uniform distribution) and hence attributing a constant probability to each of the alternatives $1, \ldots, k$. A proposal to explicitly reflect the "tendency to the middle" is given in Tutz and Schneider (2017): There a beta-binomial distribution is assumed for the uncertainty, i.e. $U \sim BetaBinom(k, \alpha, \beta)$ with $\alpha,\ \beta > 0$,

[1] for Combination of discrete Uniform and shifted Binomial random variables

where $\alpha = \beta$ is throughout assumed. The parameter α determines the concentration of the distribution in the middle and is explicitly modelled by the covariates used in the model for the feeling component.

By allowing for multiple answers and then relying on the ontic view, the problem of a proper description of the uncertainty component could be avoided. Again, an extension of the state space represents the crucial adaption. However, this may bear the challenge of analyzing partially ordered (cf., e.g., Schollmeyer (2017a) for descriptive analyses or Schollmeyer et al. (2017) with regard to stochastic dominance) coarse data: Considering different levels of undecided respondents, such as $\{3, 4\}$ and $\{3, 4, 5\}$ when referring to a rating scale, expresses a clear ordering between some categories (e.g., $\{3, 4\} < \{3, 4, 5\}$), while leaving it open for others (e.g. $\{3, 4, 5\} \lesseqgtr \{4\}$). By referring to regression analysis, a first idea for dealing with this problem is given: A commonly applied model for ordinal responses is the cumulative logit model (cf., e.g. Fahrmeir et al., 2013, p. 334–337). It relies on the proportional odds assumption, which requires global regression coefficients, and hence no own regression estimators have to be estimated for each coarse category (just as in the conditional logit model, see above). The model reflects the (strict) ordinal structure by additional restrictions on the category-specific intercept β_{0r}, $r \in \{1, \ldots, k\}$, assuming

$$\beta_{01} < \beta_{02} < \ldots < \beta_{0k}$$

for the k ordered categories (cf., e.g. Fahrmeir et al., 2013, p. 336). As a direct consequence of the partially ordered structure implied by the “ontic” variable, restrictions between incomparable categories have to be omitted. This should correspond to a procedure that repeatedly estimates regression coefficients based on a cumulative logit model, but varies between all conceivable orders of the category-specific intercepts. Those regression estimators that achieve the maximum value of the likelihood are then finally taken. Further research should be devoted to this, here only briefly discussed, problem, also exploiting findings from already existing literature about categorical regression with partially ordered response variables (cf., e.g., Zhang and Ip, 2012). Considering party preferences on a left-right continuum and hence applying ordinal models as suggested above in this context as well, gives us results about the placement of several coarse party preferences on the left-right continuum as a by-product.

Gain of information induced by the new voting question

To the best of our knowledge there is not a single pre-election study allowing respondents to report their voting intention in terms of multiple answers. In this way, in *Contribution 2* we were forced to the artificial construction of the variable “ontic”, intended to reflect the indecision of the respondents. As a consequence, the evaluation of the gain of information by means of comparing the results from an ontic and a traditional analysis is somehow limited. To allow a more expressive statement about the effectiveness from allowing for multiple responses, the author submitted appropriate voting questions to the mouse movement web survey with the topic “challenges at the German labour market”

AfD	CD	CD-AfD	CD-FDP	CD-SPD	FDP
40	90	10	12	21	26
Green	Left	SPD	SPD-Green	DK	$\leq$ 10 supporters
99	38	78	21	91	84

Table 3.1: Group 2: Absolute frequencies of party affiliations. The delimiter "-" separates the parties between which respondents are undecided.

jointly conducted by the University of Mannheim and the Institute for Employment Research, Nuremberg (cf., Horwitz et al., 2017). A split ballot experiment was used: One half of the respondents was asked to choose a single party (group 1), the other half was given a multiple choice option (group 2). Since the respondents were randomly assigned to the type of question and both groups show large sample sizes ($n = 611$ respondents could choose one party only, while $n = 610$ were allowed to decide for multiple parties), it is justified to assume that similar true voting intentions are underlying in both groups.

The reduction of the DK proportion induced by the multiple-choice option is (notably) visible: While 18.8% chose DK within group 1, a respective proportion between 14.9% (when restricting to DKs as a single answer) and 16.1% (when additionally considering all supersets) is determined based on the data of group 2. Due to the non-neglectable amount of respondents who decided for multiple parties in group 2 (133 of 610 respondents), it additionally appears that several respondents in group 1 felt forced to give a single option.

In a first illustrative study, we compare the regression estimators obtained from a multinomial logit model separately applied for both groups (similarly as in *Contribution 1*). The party affiliation is taken as response variable, in group 1 with single responses only and in group 2 with multiple responses as given in Table 3.1. To avoid that the regression estimators are calculated based on a small number of respondents only, we restrict to the party affiliations that were reported by at least ten respondents and summarize the remaining ones in a joint category (called "less than 10 supporters")[2]. The variables "university entrance" (abbreviated by university, with values "no" (=reference category (ref)) and "yes"), sex (with values "male" (ref) and "female") and the dichotomized personal assessment of the general economic situation (abbreviated by economy, with values "very good or good" (ref) as well as "fairly, bad or very bad") are used as covariates. In both analyses, we choose "CDU/CSU" (abbreviated by CD for Christian Democrats) as a reference category, which is – according to corresponding[3] pre-election studies (cf., e.g., infratest dimap, 2016) – the most popular party.

In Table 3.2 some results are presented. Most regression estimators are remarkably different in both analyses, in some cases even connected with changes in sign (cf., e.g., estimator for Green/economy or FDP/economy). Moreover, some significant regression coefficients

[2] Referring to the data of group 1, party "NPD" and party "Pirate" were comprised in this joint category, while in group 2 additionally several rare combinations of multiple party choices enter this category.

[3] the data collection period was in summer 2016

	university		economy		sex	
	single	multiple	single	multiple	single	multiple
AfD	-0.49	-0.75*	1.05***	1.12***	-0.27	-1.10***
CD-AfD	–	-1.06	–	1.29*	–	-1.94**
CD-FDP	–	-0.25	–	-0.19	–	1.41*
CD-SPD	–	0.47	–	-0.15	–	-0.29
Left	0.32	0.35	1.04***	0.71*	-0.75**	-0.44
FDP	0.11	-0.35	0.65	-0.57	-0.07	-0.33
Green	0.87***	0.50*	-0.26	0.37	0.51*	0.31
SPD	-0.01	-0.12	0.21	0.14	-0.08	-0.40
SPD-Green	–	1.20**	–	0.25	–	0.66
DK	-0.30	-0.48	1.02***	1.17***	0.12	0.48
$\leq$ 10 supp.	0.14	0.23	1.04**	0.78**	-0.97*	0.06

Table 3.2: Comparison of regression estimators for the two groups (single, multiple). The stars refer to the significances (on level 0.1 (*), 0.05 (**) and 0.01 (***)). For ease of conciseness, the estimated (category-specific) intercepts are not shown.

are received for some multiple response categories. Exemplary, it is referred to the interpretation of the estimated coefficient for CD-AfD/economy: For respondents assessing the general economy as fair or even (very) bad, the probability of having party affiliation "CD-AfD" instead of "CD" is increased by the multiplicative factor $\exp(1.29) = 3.63$, compared to respondents who consider the general economy as (very) good.

The comparably high absolute frequency of category "$\leq$ 10 supporters" in the data of group 2 (cf. Table 3.1) definitely represents a drawback, since several different groups of respondents are mixed up. This problem might be weakened when referring to studies with larger sample sizes; here due to the split ballot experiment the sample size was halved. Moreover, one could think about asking twice, firstly refraining from the multiple response option, then explicitly allowing it. In this way, at least some forced answers could be used from undecided voters showing unusual party combinations. Due to the resulting additional burden respondents are exposed to, many general surveys are expected to decide against this procedure. However, some pre-election studies, mainly intended to collect current party affiliations, might take this loss.

Also turning to epistemic data imprecision and comparing the forecast of specific parties, such as "AfD", illustrates the gain of allowing for multiple responses: While the forecast within group 1 is calculated as 6.7% (41/611), in group 2 an interval-valued forecast of [6.5% (= 40/610), 9.3% (= 57/610)] is received. Only the result in group 2 points to the high proportion of respondents who can potentially imagine to support party "AfD". In the federal elections 2017, in fact there was a big surprise about the high proportion of AfD voters. Asking for multiple responses and calculating interval-valued forecasts[4] might be

[4] Political studies proceed to calculate the proportion of potential voters of a party and hence start to be

able to catch this error. Further comparisons based on these data are planned, specifically devoting to models that are commonly used in political science, such as the conditional logit model presented previously.

3.2 Epistemic data imprecision: Contribution 2 to 5

3.2.1 The joint basis

Contributions 3 to *5* are all based on the reliable (cautious) estimation technique for coarse data proposed in *Contribution 2*. Generally, the labels "reliable" and "cautious" should be used carefully, especially when claiming that the obtained estimators fulfill these characterizations (cf., Schollmeyer, 2017b, p. 30). While "cautious" could be (exclusively) understood as the most conservative situation, where no coarsening assumptions are imposed at all, the term "reliable" may pretend that every kind of uncertainty[5] is caught and that model assumptions are right (cf., Schollmeyer, 2017b, p. 30). In this work, we use both labels (synonymously), but keep in mind that the obtained "reliability" can be ascribed to the communicated uncertainty associated with the coarse data problem only and refrain from referring it to other aspects. We include and restrict to all available knowledge about the coarsening process, hence embedding the most conservative situation as a special case.

In this section we summarize the main aspects of the underlying idea, before we elaborate how the contributions draw on these findings to solve related problems only then in Section 3.2.2. Moreover, a short overview of the illustration example based on the PASS data is given that runs like a common thread through three of the four contributions.

Reliable likelihood inference under coarse data

Let $(x_{11}, \dots, x_{1p},\ y_1), \dots, (x_{n1},\ \dots, x_{np},\ y_n)$ be a sample of n independent realizations of categorical random variables $(X_1, \dots,\ X_p,\ Y)$. We lay a special focus on the variable Y with values in Ω_Y,[6] where our main goal is the estimation of

$$\pi_{\mathbf{x}y} = P(Y = y|\mathbf{X} = \mathbf{x}),\ y \in \Omega_Y, \mathbf{x} = (x_1, \dots, x_p) \in \Omega_X\ .$$

Due to the epistemic data imprecision, some values of Y cannot be observed precisely. In this way, we only observe a sample $(x_{11},\ \mathfrak{y}_1), \dots, (x_{n1},\ \mathfrak{y}_{n1})$ of n independent realizations of $(X_1, \dots, X_p,\ \mathcal{Y})$, where $\mathcal{Y}$ is a random object mapping into $\Omega_{\mathcal{Y}} \subseteq \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$.

interested in the upper bound of this interval-valued forecast.

[5] Continuing with approaches that aim at the total survey error (cf., e.g. Weisberg, 2009) is recommendable, which encompasses further problems, such as sampling uncertainty and measurement errors.

[6] To stay consistent with the notation in *Contribution 2* to *Contribution 5*, we denote the image space of Y in this way, but it corresponds to S from Section 2.1 and *Contribution 1*. However, it has to be clarified that *Contribution 1* to *Contribution 5* (wrongly) use the term "sample space" when referring to Ω_Y, Ω_X, and $\Omega_{\mathcal{Y}}$; actually it is meant the image of the respective random object.

An observation model that relates Y of the latent world to $\mathcal{Y}$ of the observed world and is governed by the coarsening parameters

$$q_{\mathfrak{y}|xy} = P(\mathcal{Y} = \mathfrak{y}|\mathbf{X} = \mathbf{x}, Y = y),\ \mathfrak{y} \in \Omega_{\mathcal{Y}},\ \mathbf{x} \in \Omega_X,\ y \in \Omega_Y$$

constitutes the core of our approach. Since we aim at a reliable estimation of $\pi_{\mathbf{x}y}$, in the spirit of partial identification (cf., e.g., Manski, 2003) we start by refraining from any assumptions on the coarsening parameters, before successively including those assumptions about the coarsening parameters that are truly justified from an application standpoint.

The connection between $\theta_{lat} = \big((\pi_{xy})_{\mathbf{x}\in\Omega_x, y\in\Omega_Y},\ (q_{\mathfrak{y}|xy})_{\mathfrak{y}\in\Omega_Y, \mathbf{x}\in\Omega_X, y\in\Omega_Y}\big)^T \in \Theta_{lat}$ and $\theta_{obs} = (p_{\mathbf{x}\mathfrak{y}})_{\mathbf{x}\in\Omega_X, \mathfrak{y}\in\Omega_{\mathcal{Y}}} \in \Theta_{obs}$, i.e. the latent variable distribution and the coarsening parameters with parameter space Θ_{lat} as well as the observed variable distribution $p_{\mathbf{x}\mathfrak{y}} = P(\mathcal{Y} = \mathfrak{y}|\mathbf{X} = \mathbf{x})$ with corresponding parameter space Θ_{obs}, is established via a function

$$\Phi : \Theta_{lat} \to \Theta_{obs},\ \theta_{lat} \mapsto \theta_{obs}$$

(cf. Figure 3.1), basically relying on the law of the total probability, i.e.

$$p_{\mathbf{x}\mathfrak{y}} = \sum_{y\in\mathfrak{y}} \pi_{\mathbf{x}y} \cdot q_{\mathfrak{y}|\mathbf{x}y}\ , \tag{3.2}$$

for all $\mathbf{x} \in \Omega_X$, $\mathfrak{y} \in \Omega_{\mathcal{Y}}$ (also cf. Equation (8) in *Contribution 2* for the binary case). Utilizing the invariance properties of the likelihood (cf., e.g., Casella and Berger, 2002, p. 299), we can determine the maximum likelihood estimator of θ_{lat} as the inverse image under Φ of the maximum likelihood estimator $\hat{\theta}_{obs}$, simply determined by the respective relative frequencies. Due to the non-injectivity of function Φ, the set-valued estimator

$$\hat{\Gamma} = \{\hat{\theta}_{lat} \mid \Phi(\hat{\theta}_{lat}) = \hat{\theta}_{obs}\} \tag{3.3}$$

is obtained, which we illustrate by building its one-dimensional projections represented as the intervals

$$\hat{\pi}_{\mathbf{x}y} \in \left[\frac{n_{\mathbf{x}\{y\}}}{n_{\mathbf{x}}}, \frac{\sum_{\mathfrak{y}\ni y} n_{\mathbf{x}\mathfrak{y}}}{n_{\mathbf{x}}}\right], \quad \hat{q}_{\mathfrak{y}|\mathbf{x}y} \in \left[0, \frac{n_{\mathbf{x}\mathfrak{y}}}{n_{\mathbf{x}\{y\}} + n_{\mathbf{x}\mathfrak{y}}}\right]\ , \tag{3.4}$$

where $n_{\mathbf{x}\mathfrak{y}}, n_{\mathbf{x}\{y\}}$ and $n_{\mathbf{x}}$ denote the respective cell counts in the underlying contingency table (cf. Equation (10) of *Contribution 2*). Points in these intervals are constrained by the relationships in Φ. The result in (3.4) corresponds to the one obtained from cautious data completion, plugging in all potential precise sample outcomes compatible with the observations (cf. Augustin et al., 2014, §7.8).

Whenever we benefit from some auxiliary information about the coarsening mechanism, the results in (3.3) and hence (3.4) can be refined. In order to handle this technically, we generalize CAR by imposing assumptions about coarsening ratios (cf., Nordheim, 1984, who considers missingness ratios)

$$R_{\mathbf{x},y,y',\mathfrak{y}} = \frac{q_{\mathfrak{y}|\mathbf{x}y}}{q_{\mathfrak{y}|\mathbf{x}y'}}, \quad \mathfrak{y} \in \Omega_{\mathcal{Y}},\ y,\ y' \in \mathfrak{y},\ \mathbf{x} \in \Omega_X\ , \tag{3.5}$$

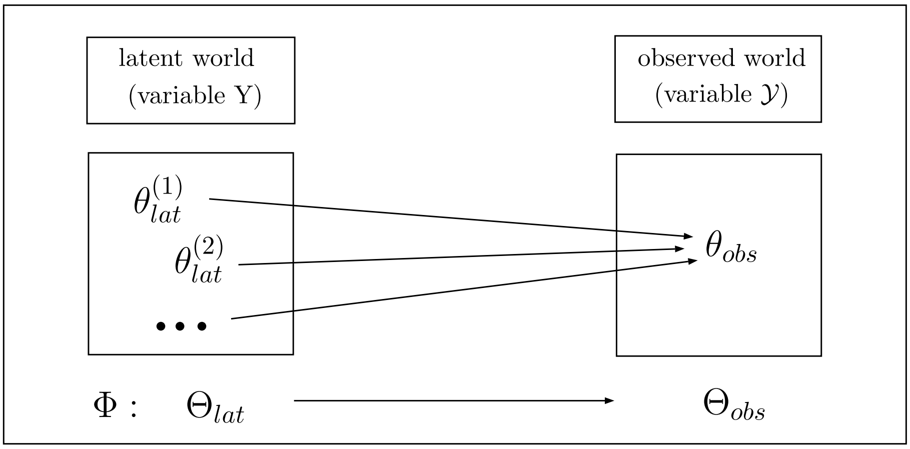


Figure 3.1: Connecting the latent and the observed world

with $R_{\mathbf{x},y,y',\mathfrak{y}} \in \mathcal{R} \subseteq \mathbb{R}_0^+$ and $\mathcal{R}$ as the set of coarsening ratios reflecting our assumptions. We define coarsening ratios as given in (3.5) for all pairs of directly successive categories y and y'. The special case of CAR is expressed by setting all these ratios equal to 1.

While specific values of R force point-identification, partially identified parameters are achieved by allowing for a whole range of values. In this way, we can incorporate frequently available, subject-driven weak assumptions as for instance "rich respondents rather tend to give a coarse answer compared to poor respondents", where in a traditional, precise approach there would be no opportunity to exploit such information. The obtained estimators under the respective assumptions are denoted by $\hat{\pi}_{\mathbf{x}y}^{\mathcal{R}}$ and $\hat{q}_{\mathfrak{y}|\mathbf{x}y}^{\mathcal{R}}$, $\mathbf{x} \in \Omega_X, y \in \Omega_Y, \mathfrak{y} \in \Omega_{\mathcal{Y}}$. In a similar way, another strict assumption called subgroup independence (SI) is generalized, which we introduce as a so-to-say dual assumption to CAR in the sense that the coarsening is not independent of the true underlying value, but the values of the covariate. This further extends our possibilities to express weak coarsening assumptions.

The illustration example (PASS data)

Unlike most surveys, the German Panel study "Labour Market and Social Security" (conducted by the Institute for Employment Research, Nuremberg, cf., Trappmann et al., 2010) makes use of nonresponse follow-up questions, to compensate the high number of missing data in the context of the sensitive income question. In this way, apart from precise and missing values, coarse answers showing different levels of accuracy are explicitly collected. For ease of presentation, we refer to the data situation given in Figure 3.2, while the PASS study even provides some finer categories. Whenever one refuses to disclose the precise income, one is directed to a question asking about the income by comparably coarse income categories, i.e. "$<$ 1000€" and "$\geq$ 1000€", where all responders are asked again in terms of more precise categories. We regard the most precise categories that can be received by this technique (here: $<$ 500€, $<$ 750€, $\geq$ 750€, $<$ 1500€, $\geq$ 1500, $<$ 3000€, $\geq$ 3000€) as the true categorical answer and represent the reported answers as coarse categorical data.

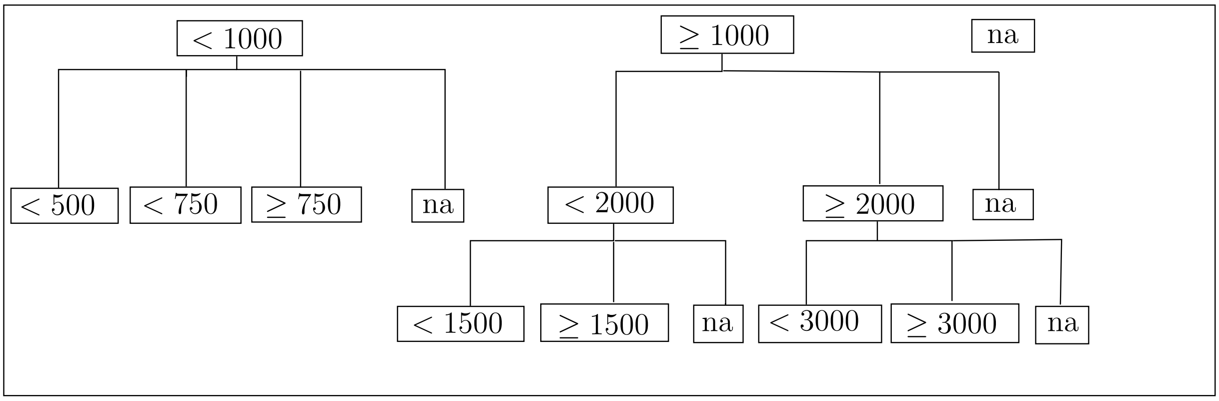


Figure 3.2: (Simplified) questioning technique used in the PASS study to collect the respondents' income (in euro).

3.2.2 Interrelations between the contributions

The literature review in Section 2.1.2 pointed towards the problem that most approaches dealing with coarse data are based on the CAR assumption. The biased estimators resulting from wrongly assuming CAR and the non-testability of this assumption make this procedure debatable in many situations. The contributions of this thesis aim at dealing with these difficulties: Hence, a reliable maximum likelihood approach for coarse data that only includes available knowledge about the coarsening process is elaborated (*Contribution 2*) and the issue of (non-)testability of coarsening assumptions is studied (*Contribution 4*). Furthermore, we stress the applicability of the basic framework of the reliable approach in more specific problems as regression analysis (mainly *Contribution 3*) and research areas as Small Area Estimation (*Contribution 5*).

The main link between the contributions is given by the reliable likelihood inference under coarse data (cf. Section 3.2.1). Figure 3.3 shows how *Contribution 3* to *5* adjoin this estimation technique developed in *Contribution 2*, where the respective extension of each contribution is marked by a different color. In *Contribution 3* the set-valued estimator $\hat{\Gamma}$ in (3.3) obtained by the reliable likelihood approach is taken as a basis to study reliable regression estimators in case of coarse data, again considering the situation of no assumptions on the coarsening process first, then gradually incorporating some subject-driven auxiliary information. Assuming a saturated model, the reliable regression estimators follow from a direct transformation of the bounds of the estimated latent variable distribution in (3.4), exploiting the connection given by the chosen link function g, which is bijective in this case. Whenever a non-saturated model is regarded, the relation is not that direct; however, the log-likelihood for the regression coefficients accounting for the parametric assumption on the regression model can be considered: While inference about $\theta_{lat} = \big((\pi_{xy})_{\mathbf{x}\in\Omega_x, y\in\Omega_Y},\ (q_{\mathbf{y}|xy})_{\mathbf{y}\in\Omega_Y, \mathbf{x}\in\Omega_X, y\in\Omega_Y}\big)^T$ is based on the log-likelihood $l(\Phi^{-1}(\theta_{obs}))$, inference about the regression coefficients β_{0y} and $\boldsymbol{\beta}_y$ takes the very same log-likelihood as a basis, then replacing π_{xy}, $\mathbf{x} \in \Omega_X, y \in \Omega_Y$, by $g^{-1}(\eta_{\mathbf{x}y})$, where $\eta_{\mathbf{x}y} = \beta_{0y} + d(\mathbf{x})^T\boldsymbol{\beta}_y$ is the linear predictor expressing the parametric assumption on the regression model and d fills the role of transferring the covariates into appropriate dummy-coded ones. Considering the maxima of the profile log-likelihood of each regression coefficient then gives us the reliable regression estimators. Studying the relation between θ_{lat} and β_y under the impact of the parametric assumption on the regression model represents a main part of *Contribution 3*.

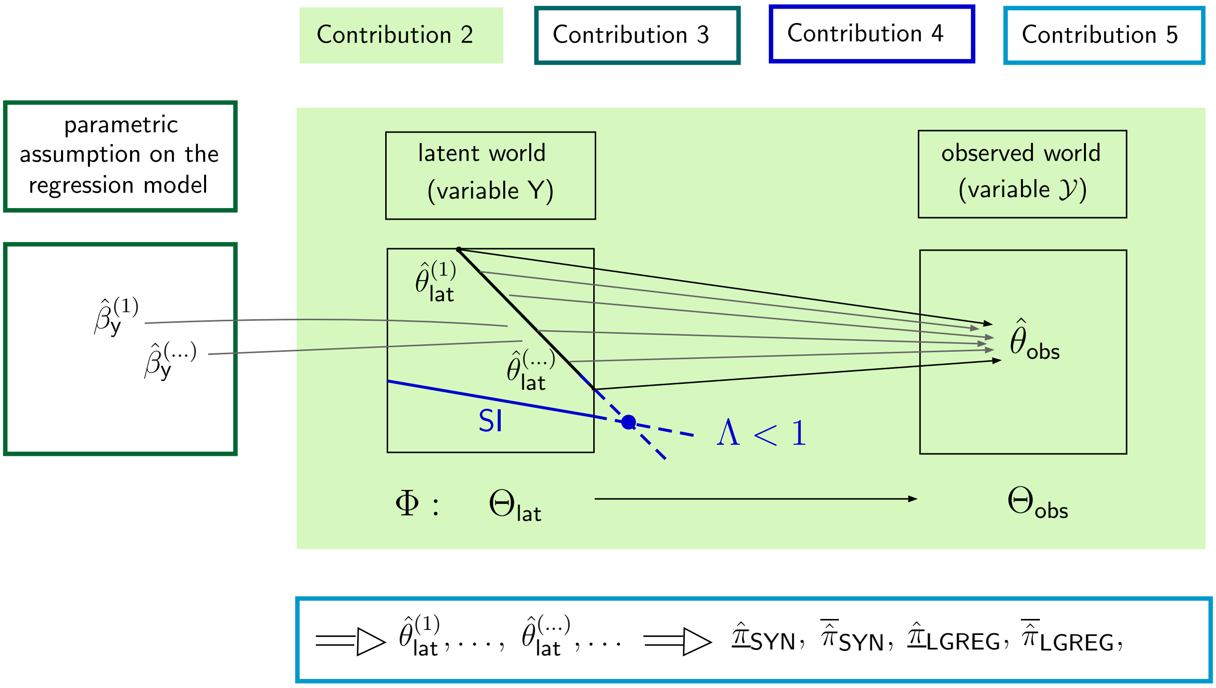

Contribution 2
Contribution 3
Contribution 4
Contribution 5
parametric assumption on the regression model
latent world (variable Y)
observed world (variable $\mathcal{Y}$)
$\hat{\beta}_{\mathrm{y}}^{(1)}$
$\hat{\beta}_{\mathrm{y}}^{(\dots)}$
$\hat{\theta}_{\mathrm{lat}}^{(1)}$
$\hat{\theta}_{\mathrm{lat}}^{(\dots)}$
$\hat{\theta}_{\mathrm{obs}}$
SI
$\Lambda < 1$
$\Phi : \Theta_{\mathrm{lat}} \longrightarrow \Theta_{\mathrm{obs}}$
$\Longrightarrow \hat{\theta}_{\mathrm{lat}}^{(1)}, \dots, \hat{\theta}_{\mathrm{lat}}^{(\dots)}, \dots \Longrightarrow \hat{\underline{\pi}}_{\mathrm{SYN}}, \hat{\overline{\pi}}_{\mathrm{SYN}}, \hat{\underline{\pi}}_{\mathrm{LGREG}}, \hat{\overline{\pi}}_{\mathrm{LGREG}},$

Figure 3.3: Interrelation between *Contribution 2* and *Contribution 3, 4,* and *5.*

Also the idea of the hypothesis test for SI developed in *Contribution 4* is based upon the reliable likelihood approach: Here the estimation technique is applied twice, firstly refraining from any assumptions about the coarsening process, then imposing SI. Comparing the maximal value of the likelihood in both cases, already points to the two possible test decisions. In Figure 3.3 a situation is sketched, where a lower maximal value of the likelihood is achieved under SI.[7] This induces a likelihood ratio Λ smaller than one and hence a rejection of SI if the reduction of the likelihood value under SI is large enough in the light of the significance level. Whenever the same maximal value of the likelihood is obtained with and without SI, the null hypothesis of SI can not be rejected.

In *Contribution 5* common estimators from Small Area Estimation, the synthetic estimator $\hat{\pi}_{SYN}$ (cf., e.g., Rao, 2015, p. 36) and the logistic generalized regression estimator $\hat{\pi}_{LGREG}$ (cf., e.g., Lehtonen and Veijanen, 1998), are expressed in terms of the estima-

[7] The intersection point of the blue line (informally representing all arguments satisfying SI) and the black line (symbolizing all arguments maximizing the likelihood under no assumptions) lies outside Θ_{lat}. Hence, only the unconstrained likelihood under SI achieves the maximal value of the likelihood under no assumptions. Further explanations are given in *Contribution 4*, using a similar illustration.

tors from the reliable likelihood approach. By directly exploiting the results in (3.4), very cautious variants are obtained, while the inclusion of auxiliary information about the missingness process is guided by relying on $\hat{\pi}^{\mathcal{R}}_{\mathbf{x}y}$ and $\hat{q}^{\mathcal{R}}_{\mathfrak{y}|\mathbf{x}y}$, $\mathbf{x} \in \Omega_X, y \in \Omega_Y, \mathfrak{y} \in \Omega_{\mathcal{Y}}$.

3.2.3 Summaries

Contribution 2

In this contribution, the framework of the reliable likelihood inference, already presented in Section 3.2.1, is developed. While the first part of the contribution is devoted to the case without any auxiliary information about the coarsening process, later on the inclusion of some weak assumptions is investigated. After having explained the basic idea of the observation model in a general way (Section 2 and 3 of *Contribution 2*), the approach is explicitly written down by referring to the situation of a binary response variable and a binary covariate with values in $\Omega_Y = \{A, B\}$ and $\Omega_X = \{0,\ 1\}$, respectively. It is started by considering the setting of the homogeneous case, where the link between the parameters of the latent and the observed world, i.e. $p_A = P(\mathcal{Y} = A)$ and $p_B = P(\mathcal{Y} = B)$ as well as $\pi_A = P(Y = A)$, $q_{AB|A} = P(\mathcal{Y} = AB|Y = A)$ and $q_{AB|B} = P(\mathcal{Y} = AB|Y = B)$, is established via the mapping $\Phi :\ [0,\ 1]^3 \to [0,\ 1]^2$ with

$$\Phi\begin{pmatrix}\pi_A \\ q_{AB|A} \\ q_{AB|B}\end{pmatrix} = \begin{pmatrix}\pi_A \cdot (1 - q_{AB|A}) \\ (1 - \pi_A) \cdot (1 - q_{AB|B})\end{pmatrix} = \begin{pmatrix}p_A \\ p_B\end{pmatrix} \tag{3.6}$$

(cf. Equation (8) of *Contribution 2*). Studying the regression case with a coarse response variable turns out to be parallel, one just has to make all considerations in a subgroup specific way, i.e. condition on the respective value of the covariate. The regression estimators of a saturated regression model are then obtained by exploiting the relation given by the link function of the (multinomial) logit model (cf. Equation (14) of *Contribution 2*). In both settings by means of the relation defined by a function as presented in (3.6), estimators as given in (3.3) and (3.4) are obtained.

Moreover, a subject-driven inclusion of auxiliary information about the coarsening process is investigated. We elaborate the estimators under known coarsening parameters, the CAR assumption and a new coarsening assumption introduced here, called subgroup independence (SI). In the binary case, SI requests

$$q_{AB|0A} = q_{AB|1A},\ q_{AB|0B} = q_{AB|1B} \tag{3.7}$$

(cf. Equation (16) of *Contribution 2*). Generalizations of CAR and SI by means of coarsening ratios enable the user to include auxiliary information about the coarsening process in a powerful and flexible way (cf. Section 3.2.1, in particular (3.5)).

Contribution 3

While we already started to study reliable categorical regression analysis in *Contribution 2* restricting to saturated models, here we look at the problem more generally, also accounting for parametric assumptions of the regression model in the sense that certain interactions are set equal to zero. For this purpose, we mainly refer to the most cautious situation where no assumptions about the coarsening are imposed, proposing a way to include auxiliary information in the end. Apart from determining reliable (maximum likelihood) regression estimators, we aim at elaborating how the parametric assumption on the regression model can affect the compatible coarsening assumptions by comparing the procedure and results under saturated and non-saturated models.

In saturated models, the bijective response function allows us to make use of a two step procedure that firstly estimates the bounds of the latent variable distribution (cf. Equation (3.4) in Section 3.2.1), which are then simply transformed to obtain the bounds of the regression coefficients (cf. Equation (8) of *Contribution 3*). Due to the reduction of the dimension of the parameter space, this is no longer possible in non-saturated models, yet we could present a method that demands to estimate the latent variable distribution $\pi_{\mathbf{x}y}$, $\mathbf{x} \in \Omega_x,\ y \in \Omega_Y$, first. We started by applying this two-step method in a binary setting, reducing to the missing data situation. Studying logistic regression, we could detect that the parametric assumption on the regression model can have a different impact on the estimated coarsening parameters, from no effect, via tighter bounds, through to point-identification. In specific situations – that we characterized by giving a proper criterion (cf. Equation (11) of *Contribution 3*) – the two-step procedure is not useful in the sense that the underlying optimization problem is not solvable. For that reason, we turned to a more natural approach – here called the direct method – where considerations are based on the maximization of the (relative) profile log-likelihood of the regression coefficients to determine reliable regression estimators.

To contrast both methods, we addressed a more general setting, also considering an illustrative study with coarse data in the strict sense. Since the categories show an ordinal structure in this case, we decide to refer to the cumulative logit model. While the direct method is always applicable, it may lead to technical difficulties. The proposed two-step method can be very useful in specific situations, where it may simplify the calculation (as in the binary setting), however, it is not always worthwhile (see above). Comparing the obtained reliable regression estimates with the ones from the usually applied procedure relying on CAR shows that although the latter are always included in the reliable results, they may even suggest specific signs of the effect in situations where the direction would be actually unclear if no assumptions about the coarsening were imposed (cf. Table 4 of *Contribution 3*). Nevertheless, there might be cases where some auxiliary information about the coarsening process is tenable: To incorporate this frequently only weak knowledge, we seize on the procedure developed in *Contribution 2* and practically add the assumptions on the coarsening parameters by incorporating constraints accordingly in the maximization of the (relative) profile log-likelihood. To additionally account for sampling uncertainty, we also study confidence intervals obtained by relying on the (relative) profile log-likelihood.

Contribution 4

One of the major challenges in the analysis of coarse data is the impossibility to test most coarsening assumptions that provide the basis for commonly used approaches. In this contribution, we compare the prominent CAR assumption to subgroup independence (SI) introduced in *Contribution 2*, where the probability of giving a coarse answer is independent of the values of the covariates. Both assumptions are uninformative in the sense that specific underlying values do not play any role for the coarsening. Nevertheless, we can elaborate substantial differences with regard to identifiability as well as testability: It is already well-known (and re-illustrated here) that CAR is generally point-identifying and not testable. Against this, in the context of SI we can demonstrate that both aspects strictly depend on the number of the covariate and response variable values, and we elaborate a proper criterion (cf. Equation (11) of *Contribution 4*). Our argumentation is mainly based on comparing the dimensions of the parameter spaces underlying the mapping Φ (cf. Section 3.2.1). In this way, we calculate the number of the degrees of freedom

$$df^{aspt.} = dim(\Theta_{obs}) - dim(\Theta_{lat}^{aspt})$$

under the assumption (aspt.) in focus (cf. Equation (9) of *Contribution 4*). Analogous conclusions can be drawn by considering the generalized version of CAR and SI (gCAR/gSI) obtained by making assumptions about coarsening ratios.
Furthermore, we elaborate the likelihood-ratio test for SI. For this purpose, we illustrate the reaction of the test statistic to the deviation from the null hypothesis and study the asymptotic distribution of the test statistic under the null hypothesis to obtain a decision rule in dependence of the significance level. Having a closer look at the binary setting of the PASS data example, we identify this situation as a special case where the calculation of the critical value has to be based on the mixture distribution

$$0.5 \cdot \delta_0 + 0.5 \cdot \chi_1^2 \ ,$$

where δ_0 is the Dirac distribution at zero (cf. Equation (19) of *Contribution 4*). In a small simulation study we compare the finite sample distribution to the theoretical one, corroborating this result. In all other cases, we can use the common χ^2-distribution with the number of degrees of freedom df^{SI}. By directly transferring the likelihood ratio test to gSI, we can enable the user to test for specific dependencies of the coarsening process on the covariate values. Beyond that, the facility of expressing partial knowledge about the coarsening process substantially increases the relevance of this test.

Contribution 5

Considering samples from sub-populations (areas) that are too small to permit a satisfying precision represents a popular topic of official statistics. Small Area Estimation (SAE) provides a variety of ways to deal with this problem, mainly concentrating on the estimation of the area-specific mean. A further relevant issue is given by the missing data problem. Nonresponse may not only dramatically reduce the already small sample size in SAE, but also leads to a substantial bias, whenever wrong assumptions about the missingness are imposed. To our best knowledge, already existing approaches for nonresponse in SAE are all based on strong assumptions, such as missing at random or missing not at random plus strict distributional assumptions. In this contribution, we propose cautious versions of prominent estimators from SAE, refraining from these strong and frequently untenable assumptions about the missingness. For this purpose, we exploit the results from the reliable likelihood inference under coarse data developed in *Contribution 2*.

We mainly focus on two prominent design-based estimators, the synthetic estimator and the LGREG-synthetic estimator, and start by looking at the special case where no assumptions about the missingness are imposed, before we then turn to more general situations. Relying on the bounds of the maximum likelihood estimators in (3.4), where either all nonresponders are attributed to the category of interest or none of them, naturally gives us cautious variants of the small area estimators under consideration (cf. Equation (5) and (6) of *Contribution 5*). To practically frame the inclusion of some auxilairy information about the coarsening process, we restate the small area estimators in terms of the estimators from the reliable likelihood approach, such as $\hat{\pi}^{\mathcal{R}}_{\mathbf{x}y}$ and $\hat{q}^{\mathcal{R}}_{\boldsymbol{\mathfrak{y}}|\mathbf{x}y}$ (cf. notation introduced in Section 3.2.1). In a next step, we incorporate those estimations that minimize/maximize the restated small area estimator to find its lower and upper bounds.

In the restated cautious LGREG-estimator, in fact two connected estimators for the latent variable distribution appear that are obtained by separate likelihood optimizations: One borrows strength by referring to all areas, the other is based on the area of interest only (cf. Equation (8) and (9) of *Contribution 5*). For that reason, we discuss how to account for important interrelations afterwards, also motivating an approach that is based on one overall likelihood. All results are illustrated by an application example giving cautious estimations of the poverty rate based on the data of the German General Social Survey 2014, where area-specific auxiliary information is taken from the German Federal Statistical Office's data report. Furthermore, we discuss why our approach cannot be directly extended to prominent model-based estimators, which are built upon mixed models. We then perform a first sensitivity analysis studying the estimation of the regression coefficients and the random effect under different missingness processes.

3.2.4 Comments and perspectives

Although the contributions dealing with epistemic imprecision were already embedded into the current literature in Section 2.1.2, now – where the most important notations have been introduced and the contributions have been summarized – a more detailed discussion becomes possible. The reliable likelihood inference presented here is placed into some recent contributions on maximum likelihood estimation under coarse data, while our results on the impact of the parametric assumption on the regression model are opposed to those from some publications on statistical learning under coarse data. Furthermore, it is investigated how the idea of the observation model can be exploited in some problems from survey statistics. In this context, the focus lies on dealing with measurement errors by means of a latent class analysis and statistical matching.

Maximum likelihood estimation under coarse data

The topic of maximum likelihood estimation under missing/coarse data received increased attention recently. An overview was given by Couso and Dubois (2018), where several approaches are structured by differentiating between various types of likelihoods, optimization strategies and coarsening assumptions. Depending on the kind of data we consider, i.e. the observed sample $\mathfrak{y}_1, \dots, \mathfrak{y}_n$, the precise, latent sample $y_1, \dots, y_n$, or the joint sample $(\mathfrak{y}_1, y_1), \dots, (\mathfrak{y}_n, y_n)$, different types of likelihoods arise, i.e. the visible, the latent and the total likelihood, respectively.[8]

We start by having a closer look at the latent and the total likelihood, both referring to the ill-known realizations, where the latter is helpful when some knowledge about the measurement process is available (cf., e.g., Couso and Dubois, 2018). Since there are several possible data completions, the likelihood is considered to be set-valued. One usually argues to maximize either its lower or its upper bound, depending on the chosen optimization strategy (cf., e.g., Couso et al., 2017): The minimax strategy (cf., e.g., Guillaume and Dubois, 2015) aims at the arg max of the lower bound of the set-valued likelihood, which results in the optimal (latent/joint) sample that induces the empirical distribution with the maximal entropy (cf. Proposition 15 in Couso et al. 2017). Consequently, this strategy is regarded as a robust procedure, especially appropriate whenever the data generating process about Y is non-deterministic. Against this, the maximax strategy (cf., e.g., Hüllermeier, 2014) takes the upper bound of the set-valued likelihood as a basis, favoring entropy minimizing estimated distributions in such a manner that frequently occurring precisely observed categories are imputed more likely and model assumptions about the data generating process of Y become more important (cf., e.g. Guillaume et al., 2017; Couso et al., 2017, Proposition 13). The spirit underlying our contributions (*Contribution 2* to *5*) explicitly refrains from choosing a specific strategy, thus taking all possible data completions into consideration and ending up with a set-valued maximum likelihood estimator. Although this is often condemned as too cautious (cf., e.g., Hüllermeier, 2014, p. 1521), we motivate this procedure for data examples from survey statistics, where one may frequently benefit from weak auxiliary information about the coarsening process refining our results.

In doing so, we regard the visible likelihood, thus studying the probability referring to the observed sample, but parametrized by the latent variable distribution and the coarsening parameters (cf., e.g., Example 6 in Couso and Dubois, 2018, where the visible likelihood under no assumptions on the measurement process is discussed). For $\Omega_Y = \{a, b\}$ and $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$, this likelihood $L(\pi_a, q_{\{a,b\}|a}, q_{\{a,b,\}|b})$ is given as

$$\underbrace{\left(\pi_a(1 - q_{\{a,b\}|a})\right)}_{p_{\{a\}}}{}^{n_{\{a\}}} \underbrace{\left((1 - \pi_a)(1 - q_{\{a,b\}|b})\right)}_{p_{\{b\}}}{}^{n_{\{b\}}} \underbrace{\left(q_{\{a,b\}|a}\pi_a + q_{\{a,b\}|b}(1 - \pi_a)\right)}_{(1 - p_{\{a\}} - p_{\{b\}})}{}^{n_{\{a,b\}}} \tag{3.8}$$

with $n_{\mathfrak{y}}$ as the counts in the respective cells.[9] The parametrization allows us to directly

[8] Additionally, the face likelihood is investigated, which equals the visible likelihood when grouped data are considered and both likelihoods are proportional to each other under the CAR assumption (cf., e.g., Heitjan and Rubin, 1991)

[9] In our contributions we additionally condition on the values of the covariates.

refer to the parameters of interest, while guaranteeing that the estimators maximizing the likelihood are in agreement with the observation. Exploiting the relation between the parameters of the observed and the latent world is a major point of our work.

The EM-algorithm relies on the visible likelihood like our approach, but refrains from the parametrization as used in (3.8). Since a set-valued maximum likelihood estimator is generally[10] obtained, the EM-algorithm typically ends after the first iteration giving a solution that is consistent with the observed data, but is strongly dependent on the starting values (cf., e.g., Couso and Dubois, 2016, p. 10). In this way, the assumption about the initial, virtual values has a severe impact, while we perceive the inclusion of (weak) auxiliary information via assumptions on the coarsening process as the more natural procedure to include some knowledge. The restricted utilization of the EM-algorithm represents a further drawback: Whenever coarse observations overlap, problems arise due to a non-careful definition of the likelihood represents a further drawback (cf., e.g., Couso and Dubois, 2016, p. 9 to 11). The inclusion of a parametrization in terms of coarsening parameters and the latent variable distribution again represents a way out of the problem.

The role of the parametric assumption on the regression model

The impact of model assumptions within the disambiguation process[11] is an interesting topic. Now our conception with regard to this question, mainly presented in *Contribution 3*, is compared to the view in some literature about statistical learning under coarse data.

An approach where the model assumption is taken that seriously that it even drives the disambiguation process is given in Hüllermeier (2014). Due to the availability of model assumptions some virtual values are perceived as more plausible than others, which goes along with the maximax strategy as described in the previous section. A totally different view is taken up by the "conservative" approach refraining from the use of model assumptions in the disambiguation process, hence learning a separate parametric model (e.g. a linear model) for each possible combination of virtual values (cf., e.g., Ramoni and Sebastiani, 2001), then summarizing all estimated parameters in intervals and finally taking their midpoints. A third procedure given by the possibilistic risk function is less conservative, admittedly taking all combinations of virtual values into account, but now referring to a single model (e.g. a concrete linear model with pre-determined regression coefficients) for all completions. According to this model, a separate loss for each virtual value is calculated, where the thereby obtained interval-valued losses are then compressed by taking the midpoint (cf., e.g., Wójtowicz et al., 2016), serving as a basis for the determination of the risk function that is used to choose the best model (cf., e.g., Sánchez and Couso, 2018).

Breaking down intervals to midpoints is not imperative: In this way, Sánchez and Couso (2018) promote to refrain from the aggregation of losses in the calculation of possibilistic risk functions, hence comparing vectors of (fuzzy) losses instead. The Marrow Region

[10] except strong, point-identifying constraints are imposed on the coarsening parameters

[11] The disambiguation process, sometimes referred to when dealing with coarse data in a machine learning context, is akin to the coarsening process. However, it takes up the opposite view, describing the mechanism that converts coarse values into precise virtual values.

and the Collection Region studied in Schollmeyer and Augustin (2015), are more in line with leaving out the aggregation in the conservative approach. A fundamental part of this work is the elaboration of the principally differing impact of the two interpretations of a (linear) model when considering partially identified models: The structural view (in the sense of Freedman (1987)) is characterized by the assumption of a truly underlying linear relationship. Under this view, models are considered based on the values that are not only compatible with the coarse observation, but also comply with the linear relationship. The estimated Marrow Region is obtained by then taking the set of regression estimators received by these models. Against this, the descriptive view (also cf., Freedman, 1987) regards the linear relationship only as a rough approximation. In this spirit, a linear model for all possible combinations of virtual values is estimated. All regression estimators are unified in the estimated Collection Region.

Just like the approaches investigated in Schollmeyer and Augustin (2015), in our contribution we also avoid aggregating intervals and are closest to the conservative strategy. But instead of considering the linear model, we focus on the (cumulative) logit model. In particular, we study the impact of the parametric assumption on the regression model in the sense that specific interactions are set equal to zero. In this regard, we investigate that depending on these parametric assumptions and the number of precisely observed categories the disambiguation process is affected, potentially leading to refined – and sometimes even to precise – regression estimates. Our approach accounting for the parametric assumption on the regression model takes the model assumption seriously and appears to go more in the direction of the structural view, where the Marrow Region is of interest. Dropping this parametric assumptions in our categorical setting means that the same information is represented by the estimated latent variable distribution $\hat{\pi}_{\mathbf{x}y}$ and the regression coefficients. In this way, all combinations of virtual values fit to the model assumption and one comes closer to the Collection Region.

Despite this influence of model assumptions on the disambiguation process, we decide explicitly against a procedure completely mixing up the problem of model identification and data disambiguation as aimed at in Hüllermeier (2014) and Hüllermeier and Cheng (2015). Especially in situations where the assumption of CAR/MAR is rather doubtful, validating model assumptions from the precisely observed data has to be treated with caution. Then, the imputation model should not exclusively be built upon the model assumption, but a separate model including all available information about the coarsening process should guide the disambiguation process. In this way, by referring to survey data applications, we propose the observation model as a powerful, practical possibility to incorporate frequently available, subject-driven, rough statements about the coarsening that could not be exploited in traditional approaches that are damned to give precise results.

Relaxation of assumptions

Turning away from the coarse data problem, the question is raised how strict assumptions that are commonly used in other survey problems can be weakened by relying on the reliable likelihood inference. In particular, the local independence and the conditional independence assumption are relaxed. These strong assumptions are frequently applied in latent class analysis (e.g. used to deal with measurement errors) and statistical matching, respectively.

Latent class analysis for measurement errors

In order to evaluate the magnitude of measurement errors, gold standard measurements, such as administrative data, are widely used. However, in many cases there are no gold standard measurements available or their justification of error-freeness is rather doubtful. For the case of categorical variables and the availability of repeated measurements of the variable of interest, latent class analysis (LCA) has been presented as a proper method (cf., e.g., Biemer, 2011).

Referring to an example with three measurements, we consider the three indicator variables $M^{[1]}$, $M^{[2]}$ and $M^{[3]}$. Moreover, we assume both, the indicator variables and the latent (true) variable of interest T to be binary, thus considering their values $m^{[1]}$, $m^{[2]}$ and $m^{[3]}$ as well as t to be in $\{0, 1\}$. The basic representation of the LCA (cf., e.g., McCutcheon, 1987) is then given by

$$P(M^{[1]} = m^{[1]}, M^{[2]} = m^{[2]}, M^{[3]} = m^{[3]}, T = t) = P(T = t) \cdot \prod_{j=1}^{3} P(M^{[j]} = m^{[j]} | T = t), \quad (3.9)$$

where the conditional probabilities $P(Y^{[j]} = y^{[j]} | T = t)$, $j = 1, \dots 3$ are the main focus, when the evaluation of measurement errors is studied. These are referred to as misclassification probability, whenever $m^{[j]} \neq t$. From (3.9) it is directly inferable that the indicator variables are assumed to be dependent on each other through the variable of interest only. This so-called local independence is the central assumption of the LCA. In the setting considered here with a binary (latent) variable and three binary measurements the inclusion of the local independence assumption is sufficient to guarantee identifiability of the parameters of interest, i.e. $P(T = t)$ and $P(M^{[j]} = m^{[j]} | T = t)$, $j = 1, \dots, 3$ (cf., e.g., McCutcheon, 1987, p. 25). However, in more general cases one always has to check that a positive number of degrees of freedom is considered. If it is negative, one includes the local independence assumption and/or other strong assumptions on the relation between indicator variables and the (latent) variable of interest, where the inclusion of grouping variables turns out to be helpful (cf., e.g., Biemer, 2011). The justification of local independence or other strict assumptions of that kind is rather questionable in many cases.

Several practical studies investigated that the LCA may be beneficial whenever the identification of inappropriate survey questions is of interest, but the estimation of misclassification probabilities performs rather poorly (cf., Kreuter et al., 2008b; Yan et al., 2012). In large parts, this may be attributable to the violation of the local independence assumption or other strong assumptions. Possible reasons for local dependence are given by bivocality, behaviorally correlated error, and latent heterogeneity, explained in detail in Berzofsky et al. (2014). Despite the awareness that assumptions like the local independence are frequently untenable, one mostly keeps relying on them, simply for reasons of identifiability. Nevertheless, there are some first suggestions in literature controlling for local independence by the introduction of a method factor accounting for the individual strategy to answer survey questions. Details are given in Oberski et al. (2015) and Oberski (2017).

Whenever weak assumptions about the conditional probabilities $P(M^{[j]} = m^{[j]}|T = t)$ are justified only, allowing for partially identified parameters in the spirit of Manski (2003) may be promising. In this way, one could relax the strong local independence assumption for instance by requiring

$$P(M^{[2]} = m^{[j]}|T = t, M^{[1]} = 0) = R \cdot P(M^{[2]} = m^{[j]}|T = t, M^{[1]} = 1) \qquad (3.10)$$

with $R \in [\underline{R}, \overline{R}]$. In this way, at least some dependence on other indicator variables is admitted, where $\underline{R}$ and $\overline{R}$ should be determined by subject-driven considerations (cf. inclusion of auxiliary information summarized in Section 3.2.1). The local independence assumption is a special case, given whenever $R = 1$. In (3.10) it is assumed that the second indicator is only dependent on the first, but not the third indicator variable. This is an assumption that comes from the modified path analysis and is frequently motivated by memory effects that only occur with regard to previously asked survey questions (cf., e.g., Goodman, 1973).

Also relaxing other restrictions, such as for instance an equal misclassification probability for female and male respondents in the sense of $P(M^{[1]} = m^{[1]}|T = t, G = 'female') = P(M^{[1]} = m^{[1]}|T = t, G = 'male')$ with G denoting the grouping variable sex, could be of interest. For this purpose one could e.g. rely on $P(M^{[1]} = m^{[1]}|T = t, G = 'female') = R \cdot P(M^{[1]} = m^{[1]}|T = t, G = 'male')$ with $R \in [0, 1[$ if a higher misclassification probability is wished to be assumed for male respondents. Incorporating these weak assumptions into the estimation should be straightforward: The likelihood of the traditional LCA (cf., e.g., Biemer, 2011, p. 130) can be taken as a basis[12], but now constraints on the parameters expressing the weak assumptions should be included.

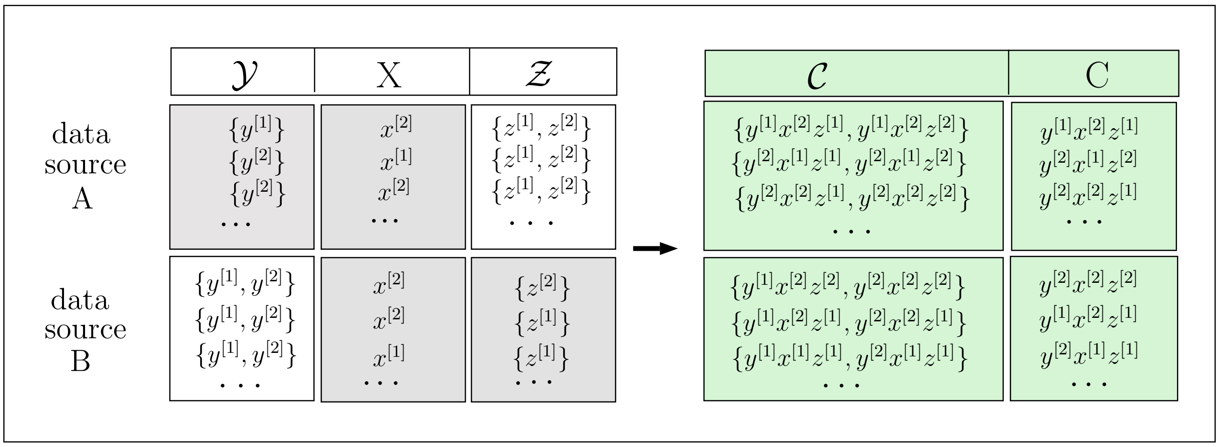
Statistical matching

The main goal of statistical matching consists in revealing joint information on variables from different data sources containing different respondents (cf., e.g., Rässler, 2012). Many approaches are based on the conditional independence assumption[13], although this procedure induces misleading estimates, whenever this untestable assumption is not justified (cf., e.g., Barry, 1988; Rodgers, 1984). For that reason, during the last decades one started to strive for cautious approaches that aim at all joint distributions that are consistent with the available information (cf., e.g., Moriarity and Scheuren (2001) and Kadane (2001) for the continuous case and D' Orazio et al. (2006) and Endres et al. (2018) for the categorical

[12] The joint cell counts from the contingency table of the variables T, $M^{[1]}$, $M^{[2]}$ and $M^{[3]}$ follow a multinomial distribution. In this way, the likelihood considered in LCA resembles the one we considered in this work in the context of coarse data, where the role of the coarsening parameters is now replaced by the misclassification probabilities.

[13] This means that the variables available only in one data source are conditionally independent given the joint variable(s).

case). By interpreting the statistical matching problem as a coarse data situation[14], we motivate how one could enqueue in these cautious approaches taking the reliable likelihood approach of this work as a starting point.



	$\mathcal{Y}$	X	$\mathcal{Z}$	$\mathcal{C}$	C
data source A	$\{y^{[1]}\}$	$x^{[2]}$	$\{z^{[1]}, z^{[2]}\}$	$\{y^{[1]}x^{[2]}z^{[1]}, y^{[1]}x^{[2]}z^{[2]}\}$	$y^{[1]}x^{[2]}z^{[1]}$
	$\{y^{[2]}\}$	$x^{[1]}$	$\{z^{[1]}, z^{[2]}\}$	$\{y^{[2]}x^{[1]}z^{[1]}, y^{[2]}x^{[1]}z^{[2]}\}$	$y^{[2]}x^{[1]}z^{[2]}$
	$\{y^{[2]}\}$	$x^{[2]}$	$\{z^{[1]}, z^{[2]}\}$	$\{y^{[2]}x^{[2]}z^{[1]}, y^{[2]}x^{[2]}z^{[2]}\}$	$y^{[2]}x^{[2]}z^{[1]}$
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
data source B	$\{y^{[1]}, y^{[2]}\}$	$x^{[2]}$	$\{z^{[2]}\}$	$\{y^{[1]}x^{[2]}z^{[2]}, y^{[2]}x^{[2]}z^{[2]}\}$	$y^{[2]}x^{[2]}z^{[2]}$
	$\{y^{[1]}, y^{[2]}\}$	$x^{[2]}$	$\{z^{[1]}\}$	$\{y^{[1]}x^{[2]}z^{[1]}, y^{[2]}x^{[2]}z^{[1]}\}$	$y^{[1]}x^{[2]}z^{[1]}$
	$\{y^{[1]}, y^{[2]}\}$	$x^{[1]}$	$\{z^{[1]}\}$	$\{y^{[1]}x^{[1]}z^{[1]}, y^{[2]}x^{[1]}z^{[1]}\}$	$y^{[2]}x^{[1]}z^{[1]}$
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

Figure 3.4: Statistical matching interpreted as coarse data situation (the left part is based on D' Orazio et al. (2006)). The gray parts mark the precisely observed values, while the white areas exclusively show missing values.

Let us consider the simple example, where the variable Y is only observed in data source A, variable Z only in data source B, but variable X is observed in both data sources (cf. Figure 3.4). In this way, we study a missing data problem with two variables $\mathcal{Y}$ and $\mathcal{Z}$ partly showing missing values, i.e. values in $\{y^{[1]}, y^{[2]}, \{y^{[1]}, y^{[2]}\}\}$ and $\{z^{[1]}, z^{[2]}, \{z^{[1]}, z^{[2]}\}\}$, respectively, and one variable X with precise values in $\{x^{[1]}, x^{[2]}\}$. By summarizing the three variables into one, we represent this situation by a variable $\mathcal{C}$ with coarse values in the strict sense. The construction of this new variable is illustrated by the green part in Figure 3.4, where also a possible latent variable C is indicated.

Although reliable likelihood inference for coarse data as presented in this work could be technically applied, the parametrization in terms of $q_{\mathfrak{c}|c} = P(\mathcal{C} = \mathfrak{c}|C = c)$, is not appropriate in the context of statistical matching: In both data sources all values of a specific variable are missing by design. In this way, – depending on the combination of $\mathfrak{c}$ and c in a given data source – the design determines $\hat{q}_{\mathfrak{c}|c}$ to be equal to either one or zero, which makes the inclusion of assumptions on coarsening parameters meaningless. Illustrated by the data example presented in Figure 3.4, for instance

$$\hat{P}(\mathcal{C} = \{y^{[1]}x^{[2]}z^{[1]}, y^{[1]}x^{[2]}z^{[2]}\}|C = y^{[1]}x^{[2]}z^{[1]})$$

is predefined to be one in data source A and zero in data source B. Relying on the parametrization of the pattern mixture model and then including assumptions about $q_{c|\mathfrak{c}} = P(C = c|\mathcal{C} = \mathfrak{c})$ could be part of further research.

[14] Also confer Endres et al. (2018), who achieve coarse data by a cautious hot deck imputation.

4 Concluding remarks

This work developed statistical approaches for coarse categorical data under the two different kinds of imprecision: For ontic data imprecision, a power-set based analysis was suggested as a possible way to incorporate the answers of "The Undecided" (cf. *Contribution 1*). To ensure reliable statistical inference under epistemic data imprecision, a likelihood-based approach was elaborated that uses an observation model to include all available, frequently very weak, but tenable knowledge about the coarsening process (cf. *Contribution 2*). The corresponding framework was also used in the context of other problems (categorical regression analysis, cf. *Contribution 3*; testing of coarsening assumptions, cf. *Contribution 4*) and fields of application (such as nonresponse in Small Area Estimation, cf. *Contribution 5*). The transfer of already existing likelihood-based approaches for precise data to the situation of coarse data represents a fruitful field of study, since the connection established by means of the observation model can be directly exploited.

Furthermore, we promoted the explicit collection of coarse categorical data (cf. Section 1). In this way, some contribution to a variety of problems typically arising in surveys might be conceivable, such as "*how to ...*

- *... decrease item-nonresponse without increasing measurement errors?"* (cf., e.g., Kuha et al., 2017, and the motivation in Section 1)
- *... deal with respondents' different knowledge with regard to survey questions, e.g. induced by varying familiarity with the topic or ability to memorize some events?"* (cf., e.g., Auriat, 1991; Tourangeau et al., 2000)
- *... gather maximal achievable information in sensitive questions?"* (cf., e.g., Tourangeau and Yan, 2007)
- *... deal with indecision between several options to answer?"* (cf., e.g., literature review in Section 2.1.1)

Beyond the listed aspects, satisficing respondents (cf., e.g., Barge, 2012, and Section 1 of this work) are expected to give more useful answers: Reporting coarse answers demands a comparably small cognitive effort, especially in case of complex questions. In this way, some respondents with a low motivation might be moved to report a coarse answer that is in accordance with the true precise answer, instead of satisficing.

The concrete design of an adequate questioning technique, which collects coarse answers in a proper level of accuracy, should be part of further research. Yet, some ideas have already been presented in this work, which are specified now: Whenever a notable number of undecided respondents is expected (and thus ontic data imprecision arises), it might be reasonable to provide precise categories plus a category "Undecided" first. A further question should then be directed to all undecided respondents requesting them to state their answer in terms of multiple responses. The aim of the first question is to prevent decided respondents from choosing multiple answers. In the context of sensitive topics or questions demanding a comparably high knowledge or a difficult memory process (i.e. if epistemic data imprecision is presumed to underlie), we recommend to offer questions with coarse options to previous nonresponders. These coarse answers should then gradually be refined by further questions to gather as much information as possible. The PASS study gives an example that exactly proceeds in this way (cf., Trappmann et al., 2010), used as illustration in some contributions of this work[1] (cf. Figure 3.2). The promoted questioning technique would not only guarantee respondents' privacy as individually required, but could also account for the personal states of knowledge. In questions requiring high memory efforts some respondents might perceive this procedure as facilitation, since it guides them through the retrieval process in stages.

Although the cognitive burden is diminished when asked in a coarse way, some further burden is induced by the additional (filter) questions. Satisficing respondents have some incentive to give motivated misreports already in the first question, thus circumventing the filter(s), while spending minimal efforts (cf., e.g., Eckman et al., 2014; Tourangeau et al., 2015). The inconvenience induced by a number of filter questions could be prevented by individually adjusted (computer-assisted) questionnaires that exploit some information collected in earlier questions to provide proper (coarse) options already from the beginning. Whenever previous questions indicate that respondents are rather unfamiliar with a topic or have a high tendency to protect their privacy, categories showing a higher degree of coarseness could be offered. The choice of coarse categories could be further supported by some information obtained by pretesting procedures: The "think-aloud" method explicitly instructs respondents to reveal their cognitive process and tell their thoughts during answering a question (cf., e.g., Collins, 2003; Willis, 2004). In this way, not only possible misinterpretations of questions, but also aspects of social desirability and lack of knowledge in different groups of respondents (for example a lower expertise of older respondents when asking about new technologies) can be detected, inspiring the selection of appropriate coarse options. Similarly, recent achievements by means of new methods, such as eye-tracking (cf., e.g., Galesic et al., 2008) or mouse movement studies (cf., e.g., Horwitz et al., 2017), may give some insights about the cognitive process, which can then guide the creation of coarse answering categories. Since providing individually adjusted coarse options contradicts the principle of standardized interviews[2], this idea demands further discussion in future research.

Survey statistics may not only profit from the explicit collection of coarse data, but also

[1] Although the example in the PASS study refers to the actually continuous variable income, this technique can also be analogously applied for categorical variables, e.g. asking about rather coarse and finer job categories.

[2] claiming that each respondents should be asked exactly the same questions in the same order (cf., e.g., Fowler and Mangione, 1990)

from the opportunity to relax strong assumptions commonly included in survey methods. In this work, first ideas have been presented for weakening the local independence assumption used in latent class analysis and conditional independence frequently representing the central assumption in statistical matching.

So far, the most important (practical) benefits of this work have been outlined. The major limitation is given by the restricted setting: Throughout this work, we confined ourselves to coarse categorical data from small state spaces consisting of few categories only. While in categorical cases showing large state spaces one can focus on the most important coarse categories to avoid the explosion of the number of coarse categories, the problem has to be approached totally differently under continuous variables (cf., e.g., Schollmeyer and Augustin, 2015). Since most questionnaires involve questions showing a manageable number of categorical answers (cf., e.g., the German General Social Survey, GESIS Leibniz Institute for the Social Sciences (2016), or the European Social Survey, Norwegian Centre for Research Data (2016)), there is a variety of situations where our proposals can be employed.

The change from traditional methods relying on CAR to the presented reliable likelihood approach is associated with a loss of information of the obtained results, which might frequently be perceived as a further drawback. However, a possibly small content of information should not be regarded as a weakness of the reliable approach, but associated to sparse additional knowledge about the coarsening process. Generally, the analysis should be driven by the available information about the coarsening process, instead of – maybe unfoundedly chosen – optimization criteria or point-identifying coarsening assumptions. Although assumptions are part of nearly every statistical analysis, the inclusion of strict missingness/coarsening assumptions is especially disastrous: The CAR assumption is not testable and – to make matters worse – it leads to a substantial bias whenever wrongly assumed. Our contributions showed that traditional approaches might even lead to specific signs about regression coefficients that are not justified when restricting to the weak, available information about the coarsening process.

New developments might positively affect the popularity of our approach. In this way, paradata are expected to bring additional (weak) auxiliary information. They are a by-product of the regular data collection and refer to the data collection process itself (cf., e.g., Durrant and Kreuter, 2013). Examples are interviewer observations about the housing situation of (unit-)nonresponders to account for nonresponse bias. Information of that kind, such as an interviewer assessment of the respondent's embarrassment with regard to certain questions or response times (cf., e.g., Couper and Kreuter, 2013), might enrich knowledge about the coarsening process. In many cases, this knowledge is expected to be of a rather weak nature, such as "respondents with a long response time rather tend to give a coarse answer compared to respondents answering quickly": While traditional approaches have to leave this information unconsidered, our approach is able to incorporate this weak knowledge properly.

Further references

Agresti, A. and Liu, I. (1999). Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics*, 55:936–943.

Alvarez, R. and Nagler, J. (1998). When politics and models collide: Estimating models of multiparty elections. *American Journal of Political Science*, 42:55–96.

Amemiya, T. (1985). *Advanced econometrics.* Harvard University Press.

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

Auriat, N. (1991). Who forgets? An analysis of memory effects in a retrospective survey on migration history. *European Journal of Population*, 7:311–342.

Baker, S. and Laird, N. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83:62–69.

Baker, S., Rosenberger, W., and Dersimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, 11:643–657.

Barge, S.and Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education*, 53:182–200.

Barry, J. (1988). An investigation of statistical matching. *Journal of Applied Statistics*, 15:275–283.

Beatty, P. and Herrmann, D. (1995). Framework for evaluating "don't know" responses in surveys. In *Proceedings of the Section on Survey Research Methods*, pages 1005–1010. Alexandria, VA: American Statistical Association. No editors mentioned; accessible at `https://ww2.amstat.org/sections/srms/Proceedings/papers/1995_175.pdf`, last access: 01/30/18.

Beatty, P., Herrmann, D., Puskar, C., and Kerwin, J. (1998). "Don't know" responses in surveys: Is what I know what you want to know and do I want you to know it? *Memory*, 6:407–426.

Berzofsky, M., Biemer, P., and Kalsbeek, W. (2014). Local dependence in latent class analysis of rare and sensitive events. *Sociological Methods & Research*, 43:137–170.

Biemer, P. P. (2011). *Latent class analysis of survey error*. Wiley.

Bon, J., Ballard, T., and Baffour, B. (2017). Biased polls and the psychology of voter indecisiveness. *arXiv:1703.09430*. Last access: 01/30/18.

Breunig, C. (2017). Testing missing at random using instrumental variables. *Journal of Business & Economic Statistics*, (accepted). Discussion Paper accessible at `https://edoc.hu-berlin.de/bitstream/handle/18452/19401/2017-007.pdf?sequence=1`, last access: 01/30/18.

Burden, B. (1997). Deterministic and probabilistic voting models. *American Journal of Political Science*, 41:1150–1169.

Casella, G. and Berger, R. (2002). *Statistical inference*. Duxbury Pacific Grove.

Cattaneo, M. and Wiencierz, A. (2012). Likelihood-based imprecise regression. *International Journal of Approximate Reasoning*, 53:1137–1154.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, 12:229–238.

Couper, M. and Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176:271–286.

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

Couso, I. and Dubois, D. (2016). Belief revision and the EM algorithm. In Carvalho, J., Lesot, M., Kaymak, U., Vieira, S., Bouchon-Meunier, B., and Yager, R., editors, *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 279–290. Springer.

Couso, I. and Dubois, D. (2018). A general framework for maximizing likelihood under incomplete data. *International Journal of Approximate Reasoning*, 93:238–260.

Couso, I., Dubois, D., and Hüllermeier, E. (2017). Maximum likelihood estimation and coarse data. In Moral, S., Pivert, O., Sánchez, D., and Marín, N., editors, *11th International Conference on Scalable Uncertainty Management*, pages 3–16. Springer.

Couso, I., Dubois, D., and Sánchez, L. (2014). Random sets and random fuzzy sets as ill-perceived random variables. In Kacprzyk, J., editor, *Springer Briefs in Applied Sciences and Technology, Subseries: Springer Briefs in Computational Intelligence*. Springer. doi: 10.1007/978-3-319-08611-8.

Couso, I. and Hüllermeier, E. (2018). Statistical inference for incomplete ranking data: A comparison of two likelihood-based estimators. In Mostaghim, S., Nürnberger, A., and Borgelt, C., editors, *Frontiers in Computational Intelligence*, pages 31–46. Springer.

Couso, I. and Sánchez, L. (2016). Machine learning models, epistemic set-valued data and generalized loss functions: An encompassing approach. *Information Sciences*, 358:129–150.

D' Orazio, M., Di Zio, M., and Scanu, M. (2006). Statistical matching for categorical data: Displaying uncertainty and using logical constraints. *Journal of Official Statistics*, 22:137–157.

Daniels, M. and Hogan, J. (2000). Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics*, 56:1241–1248.

Dassonneville, R. (2016). Volatile voters, short-term choices? An analysis of the vote choice determinants of stable and volatile voters in Great Britain. *Journal of Elections, Public Opinion and Parties*, 26:273–292.

De Cooman, G. and Zaffalon, M. (2004). Updating beliefs with incomplete observations. *Artificial Intelligence*, 159:75–125.

Delavande, A. and Manski, C. (2010). Probabilistic polling and voting in the 2008 presidential election: Evidence from the American life panel. *Public Opinion Quarterly*, 74:433–459.

DeMaio, T. (1984). Social desirability and survey measurement: A review. In Turner, C. and Martin, E., editors, *Surveying Subjective Phenomena*, pages 257–282. Sage.

Dempster, A. (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38:325–339.

Dempster, A. and Laird, N.and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 39:1–38.

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

Denœux, T., Younes, Z., and Abdallah, F. (2010). Representing uncertainty on set-valued variables using belief functions. *Artificial Intelligence*, 174:479 – 499.

Di Zio, M. and Vantaggi, B. (2017). Partial identification in statistical matching with misclassification. *International Journal of Approximate Reasoning*, 82:227–241.

Dong, Y. and Peng, C. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2:222.

Dow, J. and Endersby, J. (2004). Multinomial probit and multinomial logit: A comparison of choice models for voting research. *Electoral Studies*, 23:107–122.

Drechsler, J., Kiesl, H., and Speidel, M. (2015). MI double feature: Multiple imputation to address nonresponse and rounding errors in income questions. *Austrian Journal of Statistics*, 44:59–71.

Dubois, D. and Prade, H. (2009). Gradualness, uncertainty and bipolarity: Making sense of fuzzy sets. *Fuzzy Sets Systems*, 192:3–24.

Durrant, G. and Kreuter, F. (2013). The use of paradata in social survey research. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176:1–3.

Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Tourangeau, R., and Presser, S. (2014). Assessing the mechanisms of misreporting to filter questions in surveys. *Public Opinion Quarterly*, 78:721–733.

Emura, T. and Chen, Y. (2016). Gene selection for survival data under dependent censoring: A copula-based approach. *Statistical Methods in Medical Research*, 25:2840–2857.

Endres, E., Fink, P., and Augustin, T. (2018). Imprecise imputation: A nonparametric micro approach considering the natural uncertainty of statistical matching with categorical data. Manuscript in preparation for submission.

Fahandar, M., Hüllermeier, E., and Couso, I. (2017). Statistical inference for incomplete ranking data: The case of rank-dependent coarsening. *arXiv:1712.01158*. Last access: 01/30/18.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, methods and applications*. Springer.

Fenwick, I., Wiseman, F., Becker, J., and Heiman, J. (1982). Dealing with indecision-should we... or not? In Mitchell, A., editor, *Advances in Consumer Research*, pages 247–250. Association for Consumer Research.

Fowler, F. and Mangione, T. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Sage.

Freedman, D. (1987). A rejoinder on models, metaphors, and fables. *Journal of Educational Statistics*, 12:206–223.

Frick, J. and Grabka, M. (2014). Missing income data in the German SOEP: Incidence, imputation and its impact on the income distribution. Discussion Paper 376, German Institute for Economic Research, accessible at `https://www.diw.de/documents/publikationen/73/diw_01.c.40900.de/dp376.pdf`, last access: 01/30/18.

Friedman, H. and Amoo, T. (1999). Rating the rating scales. *Journal of Marketing Management*, 9:114–123.

Galesic, M., Tourangeau, R., Couper, M., and Conrad, F. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, 72:892–913.

Gawronski, B. and Galdi, S. (2011). Using implicit measures to read the minds of undecided voters. In Cadinu, M. and Maass, A., editors, *Social Perception, Cognition, and Language in Honour of Arcuri*, pages 203–216. CLEUP.

GESIS Leibniz Institute for the Social Sciences (2016). German General Social Survey – ALLBUS 2014. GESIS Data Archive, Cologne. ZA5242 Data file Version 1.0.0, doi: 10.4232/1.12437, last access: 01/30/18.

Geva, D., Shahar, D., Harris, T., Tepper, S., Molenberghs, G., and Friger, M. (2013). Snapshot of statistical methods used in geriatric cohort studies: How we treat missing data in publications? *International Journal of Statistics in Medical Research*, 2:289–296.

Gilljam, M. and Granberg, D. (1993). Should we take don't know for an answer? *Public Opinion Quarterly*, 57:348–357.

Goldsmith, C. (2005). Sensitivity analysis. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*. Wiley.

González-Rodríguez, G., Colubi, A., and Gil, M. (2012). Fuzzy data treated as functional data: A one-way ANOVA test approach. *Computational Statistics & Data Analysis*, 56:943–955.

Goodman, L. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika*, 60:179–192.

Grünwald, P. and Halpern, J. (2003). Updating probabilities. *Journal of Artificial Intelligence Research*, 19:243–278.

Guillaume, R., Couso, I., and Dubois, D. (2017). Maximum likelihood and robust optimisation on coarse data. In Antonucci, A., Corani, G., Couso, I., and Destercke, S., editors, *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, volume 62 of *Proceedings of Machine Learning Research*, pages 169–180. PMLR.

Guillaume, R. and Dubois, D. (2015). Robust parameter estimation of density functions under fuzzy interval observations. In Augustin, T., Doria, S., Miranda, E., and Quaeghebeur, E., editors, *Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications*, pages 147–156. Aracne.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Heitjan, D. (1994). Ignorability in general incomplete-data models. *Biometrika*, 81:701–708.

Heitjan, D. (1997). Ignorability, sufficiency and ancillarity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59:375–381.

Heitjan, D. and Rubin, D. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, 85:304–314.

Heitjan, D. and Rubin, D. (1991). Ignorability and coarse data. *The Annals of Statistics*, 19:2244–2253.

Horwitz, R., Brockhaus, S., Henninger, F., Kieslich, P., Schierholz, M., Keusch, F., and Kreuter, F. (2017). Learning from mouse movements: Improving questionnaire and respondents' user experience through passive data collection. Discussion Paper 34, Institute for Employment Research, accessible at `http://doku.iab.de/discussionpapers/2017/dp3417.pdf`, last access: 01/30/18.

Huang, X. and Zhang, N. (2008). Regression survival analysis with an assumed copula for dependent censoring: A sensitivity analysis approach. *Biometrics*, 64:1090–1099.

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

Hüllermeier, E. and Cheng, W. (2015). Superset learning based on generalized loss minimization. In A., A., Rodrigues, P., Santos, C., Gama, J., Jorge, A., and Soares, C., editors, *Machine Learning and Knowledge Discovery in Databases, ECML 2015*, volume 9285 of *Lecture Notes in Computer Science*, pages 260–275. Springer.

Iannario, M. and Piccolo, D. (2011). CUB models: Statistical methods and empirical evidence. In Kenett, R. and Salini, S., editors, *Modern Analysis of Customer Surveys*, pages 231–258. Wiley.

infratest dimap (2016). Vote intention (nationwide). Accessible at `https://www.infratest-dimap.de/en/analyses-results/nationwide/vote-intention/`, last access: 01/30/18.

Jaeger, M. (2005a). Ignorability for categorical data. *Annals of Statistics*, 33:1964–1981.

Jaeger, M. (2005b). Ignorability in statistical and probabilistic inference. *Journal of Artificial Intelligence Research*, 24:889–917.

Jaeger, M. (2006). On testing the missing at random assumption. In Fürnkranz, J., Scheffer, T., and Spiliopoulou, M., editors, *Machine Learning, ECML 2006*, volume 4212 of *Lecture Notes in Artificial Intelligence*, pages 671–678. Springer.

Jaeger, M. (2016). The AI&M procedure for learning from incomplete data. In Dechter, R. and Richardson, T., editors, *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 225–232. AUAI Press.

Jiang, Z. and Ding, P. (2016). Using the information of missing types and monotonicity to improve partial identification with binary outcomes missing not at random. *arXiv:1610.01198*. Last access: 01/30/18.

Kadane, J. (2001). Some statistical problems in merging data files. *Journal of Official Statistics*, 17:423–433.

Kariuki, S., Gichuhi, A., and Wanjoya, A. (2015). Comparison of methods of handling missing data: A case study of KDHS 2010 data. *American Journal of Theoretical and Applied Statistics*, 4:192–200.

Kendall, M. (1938). The conditions under which Sheppard's corrections are valid. *Journal of the Royal Statistical Society*, 101:592–605.

Kennickell, A. (1996). Using range techniques with CAPI in the 1995 survey of consumer finances. No editors mentioned; accessible at `http://ww2.amstat.org/sections/srms/Proceedings/papers/1996_073.pdf`, last access: 01/30/18.

Kenward, M., Goetghebeur, E., and Molenberghs, G. (2001). Sensitivity analysis for incomplete categorical data. *Statistical Modelling*, 1:31–48.

Kim, J. and Hong, M. (2012). Imputation for statistical inference with coarse data. *Canadian Journal of Statistics*, 40:604–618.

Kolmogoroff, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer.

Koziol, N. and Bilder, C. (2014). MRCV: A package for analyzing categorical variables with multiple response options. *The R journal*, 6:144–150.

Kreuter, F. (2013). Facing the nonresponse challenge. *The Annals of the American Academy of Political and Social Science*, 645:23–35.

Kreuter, F., Presser, S., and Tourangeau, R. (2008a). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, 72:847–865.

Kreuter, F., Yan, T., and Tourangeau, R. (2008b). Good item or bad—can latent class analysis tell?: The utility of latent class analysis for the evaluation of survey questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171:723–738.

Krosnick, J. and Alwin, D. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51:201–219.

Krosnick, J. and Presser, S. (2010). Question and questionnaire design. In Marsden, P. and Wright, J., editors, *Handbook of Survey Research*, pages 264–313. Emerald Group.

Küchenhoff, H., Augustin, T., and Kunz, A. (2012). Partially identified prevalence estimation under misclassification using the kappa coefficient. *International Journal of Approximate Reasoning*, 53:1168–1182.

Kuha, J., Butt, S., Katsikatsou, M., and Skinner, C. (2017). The effect of probing "don't know" responses on measurement quality and nonresponse in surveys. *Journal of the American Statistical Association*. accepted, doi: 10.1080/01621459.2017.1323640.

Lehmann, E. and Casella, G. (2006). *Theory of point estimation*. Springer.

Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24:51–56.

Little, R. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125–134.

Little, R. and Rubin, D. (2014). *Statistical analysis with missing data*. Wiley. 2nd edition.

Manski, C. (1989). Anatomy of the selection problem. *Journal of Human Resources*, 24:343–360.

Manski, C. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80:319–323.

Manski, C. (1999). *Identification problems in the social sciences*. Harvard University Press.

Manski, C. (2003). *Partial identification of probability distributions*. Springer.

Manski, C. (2005a). Partial identification with missing data: concepts and findings. *International Journal of Approximate Reasoning*, 39:151–165.

Manski, C. (2005b). *Social choice with partial knowledge of treatment response*. Princeton University Press.

Manski, C. (2015). Communicating uncertainty in official economic statistics: An appraisal fifty years after Morgenstern. *Journal of Economic Literature*, 53:631–653.

Manski, C. (2016). Credible interval estimates for official statistics with survey nonresponse. *Journal of Econometrics*, 191:293–301.

Martin, E., Traugott, M., and Kennedy, C. (2005). A review and proposal for a new measure of poll accuracy. *Public Opinion Quarterly*, 69:342–369.

Matheron, G. (1975). *Random sets and integral geometry*. Wiley.

McCutcheon, A. (1987). *Latent class analysis.* Sage.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press.

Molchanov, I. (2004). Applications of random sets in image analysis. How to average a cat and a dog? In Lopez-Diaz, M., Angeles Gil, M., Grzegorzewski, P., Hryniewicz, O., and Lawry, J., editors, *Soft Methodology and Random Information Systems*, pages 8–18. Springer.

Molchanov, I. and Molinari, F. (2014). Applications of random set theory in econometrics. *Annual Review of Economics*, 6:229–251.

Molenberghs, G., Goetghebeur, E., and Lipsitz, S.and Kenward, M. (1999). Nonrandom missingness in categorical data: Strengths and limitations. *The American Statistician*, 53:110–118.

Molenberghs, G., Kenward, M., and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: The Slovenian plebiscite case. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50:15–29.

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

Moreno-Betancur, M., Rey, G., and Latouche, A. (2015). Direct likelihood inference and sensitivity analysis for competing risks regression with missing causes of failure. *Biometrics*, 71:498–507.

Morgan, S. and Winship, C. (2014). *Counterfactuals and causal inference.* Cambridge University Press.

Moriarity, C. and Scheuren, F. (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, 17:407–422.

Nguyen, H. (2006). *An introduction to random sets.* CRC Press.

Nordheim, E. (1984). Inference from nonrandomly missing categorical data: An example from a genetic study on Turner's syndrome. *Journal of the American Statistical Association*, 79:772–780.

Norwegian Centre for Research Data (2016). ESS Round 8: European Social Survey Round 8 Data (2016). Data Archive and distributor of ESS data for ESS ERIC. Data file edition 1.0. NSD, last access: 01/30/18.

Oberski, D. (2017). Estimating error rates in an administrative register and survey questions using a latent class model. In Biemer, P., Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L., Tucker, N., and West, B., editors, *Total Survey Error in Practice*, pages 339–358. Wiley.

Oberski, D., Hagenaars, J., and Saris, W. (2015). The latent class multitrait-multimethod model. *Psychological methods*, 20:422–443.

Orriols, L. and Martínez, Á. (2014). The role of the political context in voting indecision. *Electoral Studies*, 35:12–23.

Pampaka, M., Hutcheson, G., and Williams, J. (2016). Handling missing data: Analysis of a challenging data set using multiple imputation. *International Journal of Research & Method in Education*, 39:19–37.

Plass, J. (2013). Coarse categorical data under epistemic and ontologic uncertainty: Comparison and extension of some approaches. Master's thesis, Department of Statistics, accessible at `http://jplass.userweb.mwn.de/forschung.html`, last access: 01/30/18.

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

Poe, G., Seeman, I., McLaughlin, J., Mehl, E., and Dietz, M. (1988). "Don't know" boxes in factual questions in a mail questionnaire: Effects on level and quality of response. *Public Opinion Quarterly*, 52:212–222.

Press, S. and Yang, C. (1974). A Bayesian approach to second guessing "undecided" respondents. *Journal of the American Statistical Association*, 69:58–67.

Ramoni, M. and Sebastiani, P. (2001). Robust learning with missing data. *Machine Learning*, 45:147–170.

Rao, J. (2015). *Small-Area Estimation.* Wiley.

Rässler, S. (2012). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches.* Springer.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., and Wolf, C. (2014). Vorwahl-Querschnitt (GLES 2013). GESIS Datenarchiv, Köln. ZA5700 Datenfile Version 2.0.0, accessible at `https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5700&tab=3&ll=10&notabs=&af=&nf=1&search=gles&search2=&db=e`, last access: 01/30/18.

Rodgers, W. (1984). An evaluation of statistical matching. *Journal of Business & Economic Statistics*, 2:91–102.

Rombach, I., Rivero-Arias, O., Gray, A., Jenkinson, C., and Burke, O. (2016). The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: A review of the current literature. *Quality of Life Research*, 25:1613–1623.

Rothschild, D. (2015). Combining forecasts for elections: Accurate, relevant, and timely. *International Journal of Forecasting*, 31:952–964.

Sánchez, L. and Couso, I. (2018). A framework for learning fuzzy rule-based models with epistemic set-valued data and generalized loss functions. *International Journal of Approximate Reasoning*, 92:321–339.

Sanchez, M. and Morchio, G. (1992). Probing "dont know" answers: Effects on survey estimates and variable relationships. *Public Opinion Quarterly*, 56:454–474.

Santos, J. (2000). Getting the most out of multiple response questions. *ACE*, 2:1–4.

Schafer, J. (1997). *Analysis of incomplete multivariate data.* CRC press.

Schneeweiß, H. and Komlos, J. (2009). Probabilistic rounding and Sheppard's correction. *Statistical Methodology*, 6:577–593.

Schneeweiß, H., Komlos, J., and Ahmad, A. (2010). Symmetric and asymmetric rounding: A review and some new results. *Advances in Statistical Analysis*, 94:247–271.

Schollmeyer, G. (2017a). Application of lower quantiles for complete lattices to ranking data: Analyzing outlyingness of preference orderings. Technical Report 208, Department of Statistics, accessible at `https://epub.ub.uni-muenchen.de/40452/`, last access: 01/30/18.

Schollmeyer, G. (2017b). *Reliable statistical modeling of weakly structured information.* PhD thesis, LMU Department of Statistics. Accessible at `http://nbn-resolving.de/urn:nbn:de:bvb:19-213670`, last access: 01/30/18.

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

Schollmeyer, G., Jansen, C., and Augustin, T. (2017). Detecting stochastic dominance for poset-valued random variables as an example of linear programming on closure systems. Technical Report 209, Department of Statistics, accessible at `https://epub.ub.uni-muenchen.de/40416/`, last access: 01/30/18.

Sheppard, W. (1897). On the calculation of the most probable values of frequency-constants, for data arranged according to equidistant division of a scale. *Proceedings of the London Mathematical Society*, 1:353–380.

Spieß, M. (2009). Compensating for missing data in the SOEP. Data Documentation 41, German Institute for Economic Research, accessible at `http://hdl.handle.net/10419/129252`, last access: 01/30/18.

Stoyan, D. (1998). Random sets: Models and statistics. *International Statistical Review*, 66:1–27.

Stoye, J. (2009). Partial identification and robust treatment choice: An application to young offenders. *Journal of Statistical Theory and Practice*, 3:239–254.

Sturgis, P., Roberts, C., and Smith, P. (2014). Middle alternatives revisited: How the neither/nor response acts as a way of saying "I don't know"? *Sociological Methods & Research*, 43:15–38.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, 2:167–195.

Torelli, N. and Trivellato, U. (1993). Modelling inaccuracies in job-search duration data. *Journal of Econometrics*, 59:187–211.

Tourangeau, R., Kreuter, F., and Eckman, S. (2015). Motivated misreporting: Shaping answers to reduce survey burden. In Engel, U., editor, *Survey Measurements. Techniques, Data Quality and Sources of Error*, pages 24–41. Campus.

Tourangeau, R., Rips, L., and Rasinski, K. (2000). *The psychology of survey response.* Cambridge University Press.

Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133:859–883.

Toutenburg, H., Heumann, C., and Nittner, T. (2004). Statistische Methoden bei unvollständigen Daten. Discussion Paper 386, Department of Statistics, accessible at `https://epub.ub.uni-muenchen.de/1750/1/paper_380.pdf`, last access: 01/30/18.

Train, K. (2009). *Discrete choice methods with simulation.* Cambridge University Press.

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

Tsoumakas, G. and Katakis, I. (2006). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3:1–13.

Tutz, G. and Schneider, M. (2017). Mixture models for ordinal responses with a flexible uncertainty component. Technical Report 203, Department of Statistics, accessible at `https://epub.ub.uni-muenchen.de/34783/`, last access: 01/30/18.

Vaillancourt, P. (1973). Stability of children's survey responses. *Public Opinion Quarterly*, 37:373–387.

van Ommen, T., Koolen, W., Feenstra, T., and Grünwald, P. (2016). Robust probability updating. *International Journal of Approximate Reasoning*, 74:30–57.

Wang, M., Qin, J., and Chiang, C. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, 96:1057–1065.

Weisberg, H. (2009). *The total survey error approach: A guide to the new science of survey research*. University of Chicago Press.

Willis, G. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Sage.

Wójtowicz, A., Żywica, P., Stachowiak, A., and Dyczkowski, K. (2016). Solving the problem of incomplete data in medical diagnosis via interval modeling. *Applied Soft Computing*, 47:424–437.

Yan, T., Kreuter, F., and Tourangeau, R. (2012). Evaluating survey questions: A comparison of methods. *Journal of Official Statistics*, 28:503–529.

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

Zhang, Q. and Ip, E. (2012). Generalized linear model for partially ordered data. *Statistics in Medicine*, 31:56–68.

Zhang, Z. (2010). Profile likelihood and incomplete data. *International Statistical Review*, 78:102–116.

Zheng, M. and Klein, J. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82:127–138.

Zinn, S. and Würbach, A. (2016). A statistical approach to address the problem of heaping in self-reported income data. *Journal of Applied Statistics*, 43:682–703.

Attached contributions

Contribution 1: p. 57–65;
accessible at `http://www.sipta.org/isipta15/data/paper/19.pdf`

Contribution 2: p. 67–75;
accessible at `http://www.sipta.org/isipta15/data/paper/20.pdf`

Contribution 3: p. 77–102;
accessible at `https://epub.ub.uni-muenchen.de/41600/`

Contribution 4: p. 105–117;
accessible at `https://doi.org/10.1016/j.ijar.2017.07.014`

Contribution 5: p. 121–129;
accessible at `http://proceedings.mlr.press/v62/plass17a/plass17a.pdf`.

9th International Symposium on Imprecise Probability: Theories and Applications, Pescara, Italy, 2015

Statistical Modelling in Surveys without Neglecting *The Undecided*: Multinomial Logistic Regression Models and Imprecise Classification Trees under Ontic Data Imprecision

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

Paul Fink
Department of Statistics, LMU Munich
paul.fink@stat.uni-muenchen.de

Norbert Schöning
Geschwister Scholl Institute of
Political Science, LMU Munich
norbert.schoening@gsi.uni-muenchen.de

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

Abstract

In surveys, and most notably in election polls, undecided participants frequently constitute subgroups of their own with specific individual characteristics. While traditional survey methods and corresponding statistical models are inherently damned to neglect this valuable information, an ontic random set view provides us with the full power of the whole statistical modelling framework. We elaborate this idea for a multinomial logistic regression model (which can be derived as a discrete choice model for voting behaviour) and an imprecise classification tree, and apply them as a prototypic illustration to the German Longitudinal Election Study 2013. Our results corroborate the importance of a sophisticated, random set-based modelling. Furthermore, by reinterpreting the undecided respondents' answers as disjunctive random sets, general forecasts based on interval-valued point estimators are calculated.

Keywords. Ontic data imprecision, survey methodology, election polls, multinomial logistic models, discrete choice models, imprecise classification trees, conjunctive random sets, disjunctive random sets, epistemic prediction, German Longitudinal Election Study 2013 (GLES 2013)

1 Introduction

Although pondering between several options is characteristic for human beings, indecisiveness of respondents is not reflected in most surveys. Instead it is common to force a precise answer, and at best to provide an additional category "Don't know" for those that are not decided. Frequently, in the framework of the analysis respondents reporting this "Don't know" category are no longer taken into consideration as those answers are understood as unusable. In many cases indecisive respondents are able to definitely exclude some options, which is not expressed by category "Don't know", and additionally characteristics of indecisive and decisive respondents may systematically differ. Consequently, the common proceeding leads to a substantial loss of information in data collection and biased results in the analysis of data.

In order to deal with this problem, it is necessary that questionnaire designers allow for multiple answers as "option A or option B" or at least provide ways to construct them. Hence, the preferences of the indecisive respondents are reflected in the most informative way and we are able to distinguish between different types of indecisive respondents. In this sense, we explicitly account for the heterogeneity within the group of indecisive respondents.

In order to embed this idea into a proper statistical modelling framework, we mainly will rely on the notion of *ontic sets* in the sense of Dubois and Prade ([15, 16]) as well as Dubois and Couso ([11]). They stressed the importance of differentiating between two views of a set, one representing precise collections of elements (*ontic view*) and the other reflecting incomplete knowledge about a particular precise value (*epistemic view*) ([12]). As answers of indecisive respondents are interpreted as ontic sets, we will call data that are coarse induced by indecision like "A or B" *data under ontic imprecision*.

Our paper is structured as follows. In Section 2 we will recapitulate some notions mainly based on random set theory ([19]) that have already been investigated in the framework of ontic sets ([11, 12]). In this context, we will emphasize the applicability of ontic sets to the general analysis in the presence of answers of indecisive respondents, where the focus will be on incorporating the idea of the ontic view into multinomial logistic regression analysis and classification trees in order to model heterogeneity of respondents by their covariates. By briefly digressing into the epistemic view, in Section 3 interval-valued forecasts will be constructed. The aforementioned techniques are used in an illustrative analysis based on the German Longitudinal Election Study that is briefly presented in Section 4. Corresponding results are shown and compared to those obtained from classical statistical analyses in Section 5.

For sake of simplicity, we focus on categorical data of nominal scale, yet adaptation to ordinal scale for other applications may be derived only with little additional effort. Moreover, an extension to coarse categorical covariates under ontic data imprecision may be achieved with similar arguments.

2 Data under Ontic Imprecision: Basic Idea and Extending some Statistical Approaches

As argued in the introduction, it is crucial to distinguish between the ontic and epistemic view and thus between *random conjunctive sets* and *ill-known random variables* ([11, 12]). In this section we focus on *random conjunctive sets*, underlying the ontic view.

2.1 General Analysis

As we regard the case of categorical data with a finite state space, it is sufficient to focus on the definition of *finite random sets*, which can be considered as a simplification of the more general definition of random closed sets. A finite random set is a mapping $Z^* : \Omega \to \mathcal{P}(S)$ such that for any $A \subseteq S$ holds: $Z^{*-1}(\{A\}) = \{\omega \in \Omega : Z^*(\omega) = A\} \in \mathcal{A}$, where S denotes the state space, $\mathcal{P}$ the power set and $(\Omega, \mathcal{A})$ the underlying measurable space, equipped later with a probability measure P (e.g. [20]). In other words, a finite random set is characterized by a measurable mapping on the power set. Couso and Dubois call this notion *random conjunctive set* or *(ontic) set* ([11, 12]).

The important characteristic of an ontic set is that it represents a precise collection of elements in the sense that there is no true element of S underlying, but the set itself constitutes an entity of its own ([11]). Answers like “A or B” may be regarded as an ontic set $\{A,\ B\}$ as there is no unique preference. Therefore, the nature of coarse data under ontic imprecision is well represented by the ontic view. Consequently, this leads to a power set based view, meaning an extension of the classical precise state space S to $S^* = \mathcal{P}(S) \setminus \emptyset$, with the asterisk stressing ontic imprecision. Thus, basing the analysis on S^*, and therefore regarding coarse categories as own entities, provides the main idea of dealing with ontic imprecision. The one and only difference compared to the classical case is the adapted state space S^*.

Hence, by reinterpreting the random conjunctive set as precise random variable, classical probability theory and all statistical methods based on it are applicable. In other words, the idea of the adapted state space is independent of the statistical method and exploiting this idea further for formulating regression models and classification trees in the next sections should be regarded as an example.

A short example shall be given already here. It consists of calculating the probability of respondents, who are at least indecisive between particular options C_0, by the probability of the family of corresponding supersets $\mathcal{C} = \{T \subseteq S : C_0 \subseteq T\}$ to

$$P_{Z^*}(\mathcal{C}) = \sum_{C \in \mathcal{C}} P_{Z^*}(C)\ , \tag{1}$$

which is essentially a summation over singletons of the space S^* (cf. [11, p. 8]).

2.2 Regression Analysis

Generally, the main goal of regression analysis consists of modelling the relation between several covariates X and a dependent variable Y, without claiming to describe necessarily the causal impact of variables. In our case the dependent variable is assumed to be coarse under ontic imprecision, whereas we address precise covariates. As we restrict ourselves to a coarse categorical variable of nominal scale, a multinomial logit model is an appropriate statistical model.

2.2.1 Multinomial Logit Model

In this section it is mainly referred to [17, pp. 329-331]. A more thorough treatment of discrete choice models can be found for instance in [29]. We denote by $Y_i \in S = \{1, \dots, c\}$ the random variable describing the response of individual $i = 1, \dots, n$. Assuming a multinomial logit model, the probability of occurrence of category $s \in \{1, \dots, c-1\}$ for i with given covariate values $\mathbf{x}_i$ is set to be

$$P(Y_i = s \,|\, \mathbf{x}_i) \;=\; \pi_{is} \;=\; \frac{\exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s)}{1 + \sum_{r=1}^{c-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r)}\ , \tag{2}$$

with $\tilde{\mathbf{x}}_i^T = (1, \mathbf{x}_i^T)$ and category specific regression coefficients $\boldsymbol{\beta}_s = (\beta_{s0}, \beta_{s1}, \dots, \beta_{sp})^T$ referring to p covariates. Because of the redundancy resulting from the fact that all probabilities add up to one, the corresponding probability for the so-called reference category c can be determined by

$$P(Y_i = c \,|\, \mathbf{x}_i) \;=\; \pi_{ic} \;=\; 1 - \pi_{i1} - \ldots - \pi_{ic-1}$$
$$\;=\; \left(1 + \textstyle\sum_{r=1}^{c-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r)\right)^{-1} .$$

This corresponds to the side constraint that the regression coefficients of category c are set to zero.[1]

Expressing Equation (2) in terms of the linear predictor $\eta_{is} = \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s$, one obtains the logarithmised chances and the relative risks of category $s \in \{1, \ldots, c-1\}$ and reference category c by

$$\log\left(\frac{\pi_{is}}{\pi_{ic}}\right) = \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s \quad \text{and} \quad \frac{\pi_{is}}{\pi_{ic}} = \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s) \; . \quad (3)$$

Accordingly, the exponential of β_{sj} $(j = 1, \ldots, p)$ expresses how the chance for category s compared to the reference category c changes if the value of a certain covariate x_j is increased by one unit in the case of metric covariates or if x_j is taken instead of reference category x_J in the case of categorical covariates.

2.2.2 A Multinomial Logit Model Based Approach under Ontic Imprecision

The redefinition of the original precise state space $S = \{1, \ldots, c\}$ of Y to the state space $S^* = \mathcal{P}(S) \setminus \emptyset$ of Y^* is crucial for adapting the multinomial logit model to account for ontic imprecision, treating answers of indecisive respondents as own categories, as already pointed out in Section 2.1.

Consequently, the number of categories of the dependent variable Y^* amounts to the cardinality of the new state space S^* $(m = |S^*| = |\mathcal{P}(S) \setminus \emptyset| = 2^{|S|} - 1)$. It formalizes the idea that no longer for each $Y \in \{1, \ldots, c\}$ but for each $Y_i^* \subseteq \{1, \ldots, c\}$ probabilities $\pi_{i1}^*, \ldots, \pi_{im}^*$ are modeled and coefficients $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_{m-1}^*$ are estimated. Hence, the probability of occurrence of category $s \in \{1, \ldots, m-1\}$ for i with given covariate values $\mathbf{x}_i$ is determined by

$$P^*(Y_i^* = s \,|\, \mathbf{x}_i) \;=\; \pi_{is}^* \;=\; \frac{\exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s^*)}{1 + \sum_{r=1}^{m-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r^*)}$$

and for reference category m by

$$P^*(Y_i^* = m \,|\, \mathbf{x}_i) \;=\; \pi_{im}^* \;=\; 1 - \pi_{i1}^* - \ldots - \pi_{im-1}^*$$
$$\;=\; \left(1 + \textstyle\sum_{r=1}^{m-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r^*)\right)^{-1} .$$

In this way, one obtains own regression coefficients for each coarse category, which exactly reflects the underlying idea that different types of indecisive respondents are regarded as own group.

In summary, one can account for ontic imprecision within categorical variable Y of nominal scale by incorporating coarse answers as own categories into a multinomial logit model. Apart from the up to exponential increase in the number of categories nothing changes: All statistical methods refining and extending the classical multinomial logit model, like penalization approaches, flexible covariate modelling or random effects under repeated measurements (e.g. [30]), and their fundamental statistical properties, like consistency and asymptotic normality of estimators, can be transferred. In this way, the here addressed adaptation of the multinomial logit model serves as an example for incorporating the power set based idea into categorical regression models.

2.3 Classification Trees

Whereas in regression we are mainly interested in the estimation of the regression coefficients, which provide a structural interpretation of the data, in the framework of classification trees one major goal is to predict the value(s) of a dependent variable (called *class variable* Y later on) of a future observation, based on values of some independent, so-called feature, variables. Learning a classification tree involves recursively partitioning the full data space as it is available in the beginning, into disjoint subspaces by splitting with respect to some (in-)homogeneity criterion. A most favourable property of a single classification tree from a statistical modelling point of view is that it still allows a structural interpretation, while such is lacking in the even more prediction orientated ensemble of trees, so-called bags or forests.

In the framework of classification trees there are numerous algorithms available that are able to deal both with nominal and numerical variables, some even account for missingness at random, for instance Quinlan's ID3 [23] and Breiman's CART [9] and their successors. They share the concept of selecting splitting feature variables performing the partitioning by a similarity measure, in our context the entropy. For sake of simplicity we confine ourselves to class and feature variables of nominal scale.

In order to calculate the entropy and decide on a splitting feature variable, it is required to estimate the class' probabilities, classically achieved by the corresponding relative frequency. Abellan and Moral [4] introduced *imprecise classification trees* by changing the estimation to involve imprecise probability mod-

[1] In order to ensure identifiability it is important to include a side constraint for the regression coefficients into the basic model. Alternatively, any other category may be chosen as reference category or a symmetric type of constraint like $\sum_{r=1}^{c} \boldsymbol{\beta}_r^T = (0, \ldots, 0)^T$ can be applied (e.g. [30]).

els. As a split criterion they favoured a maximum entropy approach and presented in [4] an adaptation of Quinlan's ID3 algorithm, both of which for sake of simplicity we employ.

Yet there are more general approaches, where for instance the full entropy range is taken into account, as in [18] or [13], the latter naturally growing a forest. Further improvements of the initial imprecise algorithm also include the concept of bagging [2, 3].

In our analyses in Section 5.3 we grow classification trees accordingly to [4] but relying on a Nonparametric Predictive Inference (NPI) model for estimation of the class probability distribution within a node instead, yet an Imprecise Dirichlet Model would have been also applicable; see [10] for a more detailed introduction to NPI for categorical data and [4] or [18] for a description on how an imprecise classification tree based on it is actually constructed. Yet, we briefly recall the estimation with NPI within a tree's node.

Each node of the tree consists of a collection of observations. They are assigned to nodes in such a way that they form the aforementioned disjoint subspaces in an optimal way with respect to the splitting criterion. In the context of an entropy based splitting criterion the probability distribution of the class variable is required. In [4] the assumption of a precise probability distribution is relaxed to a credal set leading to a maximum entropy split criterion approach. According to NPI the predictive probability that for a virtual next observation the class variable attains a value y_i of its state space is within the following interval

$$P(Y=y_i) \in \left[\max\left(0,\frac{n_i-1}{n}\right), \min\left(\frac{n_i+1}{n},1\right)\right], \quad (4)$$

with n_i the number of observations having a class value of y_i and n the overall number of observations, both with respect to the node under consideration.

In the situation where the class variable is only observable under ontic imprecision, we embed ontic sets into the framework of classification trees properly by a redefinition of the class variable as a finite random set, thus basing the analysis on the power set of the class variable space, similarly to the regression analysis. This is a direct implementation of the crucial idea, allowing us to reinterpret the ontic sets as a new precise class variable, i.e. an answer "A or B" is interpreted now as the precise class "AB". Therefore, any classification tree technique might be applied that is able to deal with a precise classification variable, regardless of the underlying probability model(s). This power set based technique is frequently applied in the framework of multi-label classification (e.g. MODEL–n in [8]). Due to the increased number of classes the concept of entropy correction ([27]) becomes more important, besides substituting Y by Y^* in (4).

Furthermore, basically any classification technique may be applied, after the state space of the variables under ontic uncertainty is substituted by its power set. The classification trees serve as a feasible example.

3 Interval-valued Forecast

We consider the same data situation, but change our perspective and the aim of our analyses. Instead of modelling the underlying structure of voting (in)decisions, we now turn to forecasts based on an epistemic reinterpretation of our data.

Let's assume that our main interest lies now in forecasting certain events by enforcing a final decision expressed by a variable Y_{final}. In the context of voting behaviour such a situation arises when a forecast on the election result is required. Under the assumption that the final decision is precise and consistent with the data collected now, this means a precise true value is underlying the set-valued response.

In this way, set-valued elements A^* of S^* are no longer interpreted as own entities, but are regarded as incomplete knowledge, which for every event B from the space $(S, \mathcal{P}(S))$ is given by (cf. [7, p.185])

$$P(Y_{\text{final}} \in B \,|\, Y^* = A^*) \in \begin{cases} \{0\}, & \text{if } B \cap A^* = \emptyset \\ \{1\}, & \text{if } B \supseteq A^*\} \\ [0,1], & \text{otherwise} \end{cases} ,$$

postulating that the final answer is compatible with the initial information from the ontic view.

This corresponds to an epistemic view of modelling[2]. However, models should be cautiously interpreted as the data were originally obtained under ontic imprecision, yet it may be justified for modelling purpose.

In the context of the epistemic view Couso and Dubois ([11]) consider *ill-known random variables* Y_{epist} with precise, but incomplete realizations y_{epist}. An *ill-known random variable* Y_{epist} is a multiple-valued mapping $Y_{\text{epist}} : \Omega \to \mathcal{P}(S)$ described by the disjunctive set of mappings

$$\left\{Y_{\text{precise}} : \; Y_{\text{precise}}(\omega) \in Y_{\text{epist}}(\omega) \;,\; \forall\, \omega \in \Omega\right\} ,$$

where $Y_{\text{precise}} : \Omega \to S$ is a precise random variable. Thus, Y_{epist} is interpreted as the collection of several precise models that can be deduced from incomplete knowledge.

[2] First steps towards statistical modelling under epistemic data imprecision can be found in ([21]).

Taking the reinterpretation as disjunctive sets seriously, the range covering the true probability of a certain event of interest E can be expressed by Dempster's lower and upper probabilities ([14]) that are

$$\begin{aligned} \underline{P}_{Y_{\text{epist}}}(E) &= \sum_{Y_{\text{epist}}(\omega)\subseteq E} p(\omega) \ , \\ \overline{P}_{Y_{\text{epist}}}(E) &= \sum_{Y_{\text{epist}}(\omega)\cap E\neq\emptyset} p(\omega) \ , \end{aligned}$$

where p is the probability mass function of P ([11]).

Thus, the proportion of an option E can be forecasted by the sample counterparts $\widehat{I}(E)$ of the interval

$$I(E) \ = \ \Big[\ \underline{P}_{Y_{\text{epist}}}(E) \ , \ \ \overline{P}_{Y_{\text{epist}}}(E) \ \Big] \ . \qquad (5)$$

As the difference between the values of the lower and the upper probability represents the lack of knowledge induced by indecisive answers, it is apparent that the length of this interval can be interpreted as the extent of the underlying epistemic imprecision.

In order to account additionally for statistical uncertainty due to finite sampling, confidence intervals for $I(E)$ may be calculated. This leads to so-called uncertainty regions aiming to cover both: imprecision due to incompleteness and statistical uncertainty ([31]).

4 Data

Until now the German Longitudinal Election Study (GLES) ([25]) is the most elaborated German electoral poll and currently focuses on three federal elections (2009, 2013, 2017). The sampling method of the initial data set of the *GLES* 2013 is a (3-step) random sample, which is treated here in our illustrative analysis as a simple random sample. As voting intentions before the election day are of main interest, we consider the preliminary study of *GLES 2013*, which is a face-to-face interview two months prior to the election day.

To our present knowledge there is not any pre-election study allowing indecisive respondents to express their voting intention by multiple answers. The main advantage of *GLES 2013* is that respondents are also explicitly required to report their voting intention's certainty ("certainty")[3] along with the assessments of several parties (*q21a-q21h*[4]). Those and the respondent's current voting intention[5], collected in a precise answer, allow us the construction of a variable "ontic", reflecting the respondent's indecision by multiple answers. The procedure for our construction of the variable "ontic" is as follows: While for all "very certain" respondents the reported party of the variable "vote" is taken, the party or parties with maximal assessment are chosen for the respondents that are "fairly certain" explicitly allowing by construction indecision between the corresponding parties. For the respondents that decide for "neither/nor" or "not certain at all" parties with maximal and second highest assessments are taken.The chosen way of construction of the variable "ontic" is to some extent arbitrary, but at least it accounts reasonably for ontic imprecision. In the following we focus on the second vote, as similar steps and explanations hold for the first vote as well.

	case 13	case 126	case 1515
certainty	very certain	fairly certain	neither/ nor
vote	GREEN	SPD	CD
assessCD	-1	-1	+3
assessSPD	+2	+1	+3
assessFDP	-4	0	0
assessLEFT	-4	+1	-5
assessGREEN	+4	-3	+2
	⇓	⇓	⇓
ontic	GREEN	LEFT:SPD	CD:GREEN:SPD

Table 1: Construction of variable "ontic" (example)

The examples in Table 1 illustrate the way of construction by means of three randomly chosen respondents.[6]As our goal consists of demonstrating the difference in results from an analysis including ontic imprecision and a classical analysis, such a constructed variable is required.

Partly due to the construction of variable "ontic" several respondents had to be excluded[7]. All conducted filtering steps (e.g. excluding voters of smaller parties or non-voters) that reduced the sample of initially 2003 to 1196 respondents can be found in [22]. The associated loss of information caused by the reduced

[3] *q13* with categories "very certain", "fairly certain", "neither/nor" and "not certain at all"

[4] Each measured on a scale from "-5" ("a very negative view of this political party") to "+5" ("a very positive view of this political party")

[5] The German election system mixes elements of election by proportionality and by majority. The voters have two votes (*q11ab*: second vote, *q11aa*: first vote). The second vote is generally considered as more important, because the proportion of seats in the German Bundestag mainly is allocated according to the second vote. The first vote determines the direct representative of an election district in the Bundestag.

[6] Translations of German abbreviations of political parties are used here. Considered parties are: *Christlich Demokratische Union Deutschlands* (CDU) and *Christlich-Soziale Union in Bayern* (CSU) representing throughout Germany one option only (here denoted by CD), *Sozialdemokratische Partei Deutschlands* (SPD), *Die Linke* (LEFT), *Bündnis 90/Die Grünen* (GREEN), *Freie Demokratische Partei* (FDP).

[7] In voting studies sample loss is rather common. Usually empirical analyses are reduced to those parties, who entered the German Bundestag finally (e.g. [28]).

CD	SPD	GREEN
495	271	125
LEFT	FDP	GREEN:SPD
106	39	36
CD:SPD	CD:FDP	GREEN:LEFT
35	18	15
LEFT:SPD	CD:GREEN:SPD	GREEN:LEFT:SPD
14	17	13
CD:FDP:SPD		
12		

Table 2: Absolute frequencies of constructed variable "ontic" (second vote)

sample size is undesirable, but unavoidable for an ontic analysis illustrated by this data set. Because of the underrepresentation of indecisive persons induced by the current design of the questionnaire, which implicitly excludes indecisive respondents by the preceding filtering of the "certainty" item (cf. [22]), we expect less marked differences between an ontic and a classical analysis, described in the following sections.

The resulting illustrative data set containing variable "ontic", whose absolute frequencies are given in Table 2, forms the basis of the following analysis.[8]

5 Data Analysis

The principal goal consists of comparing the results obtained by an analysis using the constructed variable "ontic" (cf. Section 4 and [22]) to a classical analysis excluding all uncertain respondents. This issue will be considered in this section with regard to the findings from Section 2. Hereby, we focus on the second vote, only where mentioned explicitly the first vote is considered. All analyses are based on complete cases, dependent on the variables effectively under consideration. We performed our analyses with the open-source statistical software R [24]. The code is available on request from the authors.

5.1 General Analysis

The analysis incorporating ontic imprecision is based on $S^* = \mathcal{P}(S) \setminus \emptyset$, where

$$S = \{\mathrm{CD}, \mathrm{SPD}, \mathrm{GREEN}, \mathrm{LEFT}, \mathrm{FDP}\}$$

is the state space. Since only 13 elements of S^* are attained in the addressed data set, we adapted S^* to cover those values of variable "ontic" only (see Table 2).

If for instance the probability of respondents is of interest that are (at least) indecisive between party "SPD" and "GREEN", according to Equation (1) all probabilities referring to respondents that are (at least) indecisive between both parties have to be summed up, which can be estimated by associated relative frequencies to

$$\begin{aligned} &\widehat{P}_{Z^*}\big(Z^* \supseteq \{\text{GREEN, SPD}\}\big) \\ &= \widehat{P}\Big(\big\{\omega : Z^*(\omega) = \{\text{GREEN, SPD}\}\big\}\Big) \\ &\quad + \widehat{P}\Big(\big\{\omega : Z^*(\omega) = \{\text{CD, GREEN, SPD}\}\big\}\Big) \\ &\quad + \widehat{P}\Big(\big\{\omega : Z^*(\omega) = \{\text{GREEN, LEFT, SPD}\}\big\}\Big) \\ &= \frac{36}{1196} + \frac{17}{1196} + \frac{13}{1196} \approx 0.06\,. \end{aligned}$$

The estimated proportion of indecisive respondents is 0.13, calculated analogously. Consequently, if just decisive respondents are considered an amount of 13% of respondents are not taken into account. As respondents are excluded because of the value of the variable of interest itself, we are concerned with a *not missing at random* situation and thus ignoring the indecisive respondents may lead to biased results. This is particularly fatal for a theoretical understanding of voting decisions as well as from a practical campaigners' view, because this percentage covers those respondents that are of particular interest.

5.2 Regression Analysis

In order to analyse the heterogeneity within the coarse dependent variable Y under ontic data imprecision, the models presented in Section 2.2 are applied. The multinomial logit model has a longstanding tradition in the context of modelling voting behaviour[9].

In our analysis the variable "ontic" represents the coarse dependent variable, where "SPD" is chosen as reference category. Generally, it is important to choose all reference categories in such a way that interpretations enable answering the question of interest. For our illustrative purpose we use a very simple voting model with only two covariates[10], namely socio-demographical variable "religious denomination" (*q228*) as well as variable "most important source of information" (*q97*). In both variables certain categories were aggregated. Thus, variable "religious denomination" here only takes values "Christian" and "non-Christian", where the categories of "most important

[8] Absolute frequencies of singletons differ from those of variable "vote" due to the construction of variable "ontic".

[9] Actually, the multinomial logit model is the simplest model of the discrete choice family. Although it has several disadvantages for the modelling of voting behaviour as discussed by [6], for the sake of our illustrative application yet the multinomial logit model is appropriate, because it shows the basic concept in handling data under ontic imprecision, which can be extended analogously to more tailored models.

[10] Recent models of voting behaviour use policy distance, party identification and socio-demographical variables and yield a remarkable fit and prognostic validity (cf. [5])

Coefficient	ontic		classical
	CD	G:S	CD
intercept	0.37	-1.47 ***	0.13
rel.christ	0.32 *	-0.05	0.49 ***
info.tv	0.01	-0.29	-0.01
info.np	-0.05	-1.67 **	-0.01

Table 3: Comparison of results (second vote).

source of information" are translated to "television", "newspaper" and "other source", the latter also covering "radio", "internet" and "talking to other people". Every reclassification is subject to avoid categories with only few observations in order to decrease statistical uncertainty. By including "most important source of information" as a covariate into the model, we assume that the way how voters inform themselves of the federal election influences their voting intention. Nevertheless, one cannot exclude an opposite (causal) direction as respondents who vote for particular parties potentially avoid or prefer certain information sources because of the way this party is represented in it. This needs to be kept in mind when interpreting the model's results.

For reasons of conciseness estimated regression coefficients are shown just for category "CD" and "GREEN:SPD" (G:S) here.[11] With $n_{\text{CD}} = 508$ and $n_{\text{G:S}} = 36$ they form the largest groups of decisive and indecisive respondents, respectively, such that the interpretation of corresponding regression coefficients is comparably trustworthy. Especially in the context of estimators for indecisive groups, we remark that some of the regression coefficients' calculations are based on few observations, and thus corresponding interpretations have to be treated cautiously.

Furthermore, in context of interpretation one should check by taking the statistical significance[12] into account whether the regression coefficients vary just randomly. The small sample size within several groups of variable "ontic" may be responsible for non-significant estimators. Thus, from an increase in sample size statistical uncertainty is reduced and potentially significant results can be obtained.

Considering the results of the second vote analysis presented in Table 3 (ontic)[13], for Christian respondents the probability of electing "CD" instead of "SPD" is increased by the multiplicative factor $\exp(0.32) = 1.38$ compared to non-Christian respondents under the ceteris paribus assumption of unchanged other covariates.[14] Furthermore, regression coefficients closely to zero indicate that no influence of covariate "most important information source" on the probability of electing "CD" in comparison to the reference category "SPD" may be verified.

The crucial property of the multinomial regression under ontic imprecision consists of estimating own coefficients for the different indecisive groups. For instance, for respondents reporting "newspaper" as their most important information source in comparison to those naming another information source the probability of being indecisive between the two parties "GREEN" and "SPD" instead of voting for "SPD" is decreased by the factor $\exp(-1.67) = 0.19$ on the ceteris paribus premise. Likewise investigations are important for election campaigners to adjust their strategies adequately, as they show how potential voters differ from the core voters of a party (as here "SPD") in the choice of their favourable information source.

Results from a classical analysis that chooses variable "vote" as response variable and takes only those respondents into consideration that are "very certain" or "certain" may be found in Table 3 as well, again just displaying coefficients for "CD".

Comparing results from both analyses, estimators of similar magnitude are obtained throughout. In this way, the classical and the generalized approach reflecting ontic imprecision do not contradict each other.

The importance of our ontic set based modelling is corroborated even stronger when we consider the first vote instead. Now the analyses reveal remarkable differences partly associated with a change in sign. Thus, some covariates have an amplifying effect on the dependent variable in one analysis, while in the other analysis a weakening effect is underlying (cf. Table 4), yet those are not statistically significant.

Although the complete case analysis and the carried out filtering steps mainly induced by the questionnaire design led to a further decrease in the number of indecisive respondents, this illustrative analysis already shows striking differences between both analyses. Because of the here provided proof of concept for an ontic analysis, it is strongly suggested to include the option of reporting multiple answers such that those can be

[11] Estimated regression coefficients for the other categories may be found in [22]

[12] "***", "**" and "*" denotes statistical significance of level $\alpha = 0.01$, $\alpha = 0.05$ or $\alpha = 0.1$, respectively.

[13] Covariates "religious denomination" and "most important information source" are dummy coded with "non-Christian" and "other source" as reference category, respectively. The estimates quantify the difference between the group under consideration and the reference category (rel.christ: "religious denomination" is "Christian"; info.tv, info.np: "most important information source" is television, newspaper, respectively).

[14] Despite the name "CD" and the above results indicating a strong Christian relation, nowadays the "CD" parties understand themselves as a general conservative party with members and supporters regardless their religious affiliation.

Coefficient	ontic		classical
	CD	G:S	CD
intercept	0.33	-1.41^{**}	-0.12
rel.christ	0.37^{**}	-0.25	0.52^{***}
info.tv	-0.02	-0.32	0.25
info.np	-0.12	-1.69^{**}	0.13

Table 4: Comparison of results (first vote).

	ontic	vote	classical
Scenario 1	0.407 (0.040)	0.425 (0.050)	0.446 (0.041)
Scenario 2	0.704 (0.026)	0.796 (0.031)	0.817 (0.042)

Table 5: Correct classification rate (standard deviation) for second vote based on 10-fold cross-validation

included into the analysis in an appropriate way. In cases of large data sets with numerous indecisive respondents, we even expect increased differences in the estimation of regression coefficients.

5.3 Classification Trees Analysis

In a first scenario the settings are the same as we explored in the regression analysis, thus considering "ontic" coarse class variable and "religious denomination" and "most important source of information" as split feature variables, in the same scaling as previously in section 5.2 (Scenario 1). We are considering this setting to retain direct comparability with the regression analysis, yet we are aware that a classification tree's ability lies in reducing the sample space by discovering few favourable independent variables out of a potentially huge number of candidates. Therefore, we are not expecting an outstanding performance in this scenario. As discussed above we decided in favour of a Nonparametric Predictive Inference model as underlying (imprecise) model of the classification tree. We choose the most frequent class as prediction rule in the leaves, thus enforcing a precise result. Furthermore, we grew imprecise classification trees on the data set neglecting the undecided, but in this case we chose "vote" as the dependent variable as a counter part to the classical regression analysis. In order to assess the predictive ability of the trees a 10-fold cross-validation each was performed.

The results are to be found in the first row of Table 5, with respect to the second vote. For a fair comparison we measure the accuracy for both data situations by the correct classification rate (columns *ontic* and *classical*), and furthermore in case of the ontic data sets we checked the prediction result of "ontic" against "vote" (column *vote*). Any value of "vote" which was contained in the predicted coarse category was considered correctly classified. Furthermore the standard deviation is reported.

As it is clearly visible the predictive ability of the imprecise trees is unsurprisingly poor, and an inspection of the underlying trees reveals the culprits. The selection of the independent variables only allows growing of 13 different trees, which only in case of a strong dependency between the independent and depend variables leads to reasonable accuracy results. Furthermore when looking at the relative class frequencies in the root nodes, the category of "CD" is with over 40% by far the most observed one. While the construction of most trees involved at least one split, category "CD" is still predicted in a vast majority of the tree's leaves, in few cases even in all.

In further analyses, we incorporated more independent variables, allowing a higher variation in potential trees (Scenario 2). Further splitting candidate variables were the party identification (*q119*), the person's social stratum (*q192*), the sex (*q1*), general political interest (*q3*) and the personal economic situation (*q17*). With those and the previous variables the same analysing steps were repeated, but now with the accuracy nearly doubling in either scenario as the second row of Table 5 indicates. Especially the party identification has a high influence.

Similar prediction results as above are obtained when considering the first vote, instead of the second, displayed in [22]. Quite interestingly, the correct classification rate is lower when we are predicting the "ontic" variable than in the case when predicting "vote". In the second scenario there is a notable gap of around 10%, which is mainly caused by an ontic coarse class prediction, whereas vote is (naturally) precise.

In both scenarios the classical procedure of omitting the undecided persons leads to better results, when just considering the predictive ability, yet with the help of our ontic view we are able to identify hard to classify respondents.

A major reason for the small differences between the classical and ontic analyses is the comparably little percentage of undecided persons (less than 10% within the data under consideration). As mentioned in the discussion in the regression analyses, this is partly due to the conducted complete case analysis and the construction of variable "ontic", but more gravely imposed by the design of the questionnaire. When allowing for multiple answers directly in variable "vote", we expect an increase in the accuracy of the ontic prediction, as the number of hard to precisely classify, indecisive persons raises.

5.4 Interval-valued Forecast

In Section 3 the epistemic view has been used in order to calculate interval-valued forecast $I(E)$, which will be illustrated in this section.

For instance, if one is interested in the forecasted proportion of respondents electing "CD", by referring to the absolute frequencies of variable "ontic" in Table 2 and to Equation (5), the interval-valued forecast

$$\widehat{I}(\{\text{CD}\}) \;=\; \left[\frac{495}{1196}\,,\;\frac{495+35+18+17+12}{1196}\right]$$

is obtained. All fractions that are included in the lower bound refer to respondents who vote for the "CD" party for sure while all fractions that are used within the calculation of the upper bound concern respondents who generally could imagine to vote for it. Political studies gradually proceed to calculate the fraction of "potential voters" which corresponds to the upper bound of interval $\widehat{I}(E)$ (cf. [1]).

Nevertheless, forecasts are commonly based on respondents that are characterized by a high degree of certainty concerning their voting intention only. In our data example there are $n = 1096$ respondents that are "very certain" or "fairly certain" according to their voting intention, where 490 of those intend to vote for "CD" and thus the naive estimated forecasting probability results in

$$\widehat{P}_{\text{naive}}(\{\text{CD}\}) \;=\; \frac{490}{1096}\,.$$

As indecisive voters may systematically differ from respondents that are sure of their voting intention, the proportion in terms of interval $\widehat{I}(E)$ contains valuable information that is not expressed by $\widehat{P}_{\text{naive}}(E)$. Because of the difference between these groups it is important to treat results ignoring indecisive respondents with caution.

In practice forecasting the proportion of a set containing more than one element is of considerable relevance: Frequently, for instance in Germany, the main interest is the voters' percentage not just for a particular single party, but for a coalition. In this context the interval-valued forecast $\widehat{I}(E)$ becomes of particular interest, as respondents that are indecisive between the parties contained in the coalition of interest E are incorporated for sure. Thus, these coarse observations constitute a precise vote for the coalition (e.g. [22]).

6 Concluding Remarks

While currently data under ontic imprecision are still neglected in statistical analysis, they could prove a valuable source of information. Especially in context of election studies incorporating the different types of "The Undecided" into statistical analyses becomes increasingly important as more and more voters decide shortly before the election day (cf., e.g. [26]). Once the practitioner changes the state space, the statistical methods remain the same, as we could demonstrate. Even as the group was comparably small and we were forced to assess indecisiveness indirectly by constructing an ontic variable, we corroborated in our data example that including the undecided respondents did make a difference. Therefore, as now appropriate statistical methodology has been proven to be available, we strongly recommend allowing for multiple answers directly within questionnaires.

As the underlying idea is somewhat generic, the in here presented analyses by a multinomial regression model and imprecise classification trees are just the tip of the iceberg. One may think of more complex methods to study the data set, mutatis mutandis. For simplicity we restricted ourselves to the case of a nominal scale of the variable under ontic imprecision, yet the adaptation to an ordinal scale is achievable with little additional effort as well. In further studies it is worth considering not only the dependent variable under ontic imprecision but also the covariates. In principle, this is achievable by involving the power-set based idea again.

Acknowledgements

We are grateful to two of three anonymous reviewers for their very helpful remarks, also stimulating further research.

References

[1] Großteil der Wähler würde sich noch umstimmen lassen. *Süddeutsche Zeitung*, 16 August 2013. Accessed 24 January 2015, http://www.sueddeutsche.de/politik/umfrage-zur-bundestagswahl-die-meisten-waehler-wuerden-sich-noch-umstimmen-lassen-1.1747539.

[2] J. Abellán and A. Masegosa. Bagging decision trees on data sets with classification noise. In S. Link and H. Prade, editors, *Foundations of Information and Knowledge Systems*, pages 248–265. Springer Berlin Heidelberg, 2010.

[3] J. Abellán and A. Masegosa. An ensemble method of using credal decision trees. *European Journal of Operations Research*, 205(1):218–226, 2010.

[4] J. Abellán and S. Moral. Building classification trees using the total uncertainty criterion. *International Journal of Intelligent Systems*, 18(12):1215–1225, 2003.

[5] J. Adams, S. Merrill, and B. Grofman. *A Unified Theory of Party Competition: A Cross-National Analysis Integrating Spatial and Behavioral Factors.* Cambridge University Press, Cambridge, 2005.

[6] R. Alvarez and J. Nagler. When politics and models collide: Estimating models of multiparty elections. *American Journal of Political Science*, 42(1):55–96, 1998.

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

[8] M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth Books, Monterey, CA, 1984.

[10] F. Coolen and T. Augustin. A nonparametric predictive alternative to the Imprecise Dirichlet Model: The case of a known number of categories. *International Journal of Approximate Reasoning*, 50(2):217–230, 2009.

[11] I. Couso and D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning*, 55(7):1502–1518, 2014.

[12] I. Couso, D. Dubois, and L. Sánchez. *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables.* Springer, Cham, 2014.

[13] R. Crossman, J. Abellán, T. Augustin, and F. Coolen. Building imprecise classification trees with entropy ranges. In F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger, editors, *ISIPTA '11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pages 129–138, Innsbruck, 2011. SIPTA.

[14] A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[15] D. Dubois and H. Prade. *Possibility Theory.* Plenum Press, New York, 1988.

[16] D. Dubois and H. Prade. Formal representations of uncertainty. In D. Bouyssou, D. Dubois, M. Pirlot, and H. Prade, editors, *Decision-Making Process: Concepts and Methods*, pages 85–156. ISTE & Wiley, London, 2009.

[17] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: Models, Methods and Applications.* Springer, Berlin, 2013.

[18] P. Fink and R. Crossman. Entropy based classification trees. In F. Cozman, T. Denœux, S. Destercke, and T. Seidenfeld, editors, *ISIPTA '13: Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications*, pages 139–147, Compiègne, 2013. SIPTA.

[19] G. Matheron. *Random Sets and Integral Geometry.* Wiley, New York, 1975.

[20] H. Nguyen. *An Introduction to Random Sets.* CRC Press, Boca Raton, Florida, 2006.

[21] J. Plass, T. Augustin, M. Cattaneo, and G. Schollmeyer. Towards statistical modelling under epistemic data imprecision: Some results on estimating multinomial distributions and logistic regression under coarse categorical data. Under revision for ISIPTA '15, preprint temporary available at `http://www.statistik.lmu.de/~jplass/forschung.html` (20.03.2015).

[22] J. Plass, P. Fink, N. Schöning, and T. Augustin. Statistical Modelling in Surveys without Neglecting "The Undecided": Multinomial Logistic Regression Models and Imprecise Classification Trees under Ontic Data Imprecision - extended version. Technical Report 179, University of Munich, Department of Statistics, 2015. `http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-23816-6`.

[23] J. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[24] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014.

[25] H. Rattinger, S. Roßteutscher, R. Schmitt-Beck, B. Weßels, and C. Wolf. Vorwahl-Querschnitt (GLES 2013), 2014. GESIS Datenarchiv, Köln. ZA5700 Datenfile Version 2.0.0, Accessible from `https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5700&tab=3&ll=10&notabs=&af=&nf=1&search=gles&search2=&db=e`.

[26] O. Schirg. Wahlforscher: Jeder Dritte ist noch unentschlossen. *Die Welt*, 10 August 2001. Accessed 22 January 2015, `http://www.welt.de/print-welt/article467015/Wahlforscher-Jeder-Dritte-ist-noch-unentschlossen.html`.

[27] C. Strobl. Variable selection in classification trees based on imprecise probabilities. In F. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, pages 339–348, Carnegie Mellon University, Pittsburgh, 2005. SIPTA.

[28] P. Thurner. The empirical application of the spatial theory of voting in multiparty systems with random utility models. *Electoral Studies*, 19(4):493–517, 2000.

[29] K. Train. *Discrete Choice Methods with Simulation.* Cambridge University Press, 2009.

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

[31] S. Vansteelandt, E. Goetghebeur, M. Kenward, and G. Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16(3):953–979, 2006.

9th International Symposium on Imprecise Probability: Theories and Applications, Pescara, Italy, 2015

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 paper deals 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 as well as logistic regression under a categorical covariate. 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.

1 The Problem and its Background

A frequent challenge in statistical modelling is *data imprecision*, where some data are *coarse*, i.e. they are not observed in the resolution originally intended in the subject matter context. Throughout this paper, we focus on the case where the coarse observations are data under *epistemic data imprecision.* For categorical data as considered here this means that there exists a true precise value y of a generic variable Y taking values in a finite sample space $\Omega_Y = \{1, \ldots, K\}$, but we may only observe a non-singleton set $\mathscr{Y}$ containing y. It is important to distinguish epistemic from *ontic* data imprecision, where data are coarse by nature and thus have to be interpreted as indivisible entities of their own (see, in particular, [7, 8]; [24] for an application in a multinomial logit model and classification.)

Epistemic data imprecision emerges most naturally in a huge variety of applications. Missing data, interpreted as the prominent special case where the whole sample space is observed only, arise, for instance directly by design in observational studies on treatment effects, see, e.g., [27], and unit non-response is quite frequent in surveys, in particular as refusals to answer sensitive questions. Typical instances of not missing but still coarse data include the numerous data sets where coarsening is deliberately applied as an anonymization technique (see, e.g., [10]), matched data sets with not completely identical categories, secondary data where the originally coded categories turn out to be not fine enough and, as a technical example, reliability analysis of a system whose components are tested separately prior to assembly [30].

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 ([5, 26, 17, 32, 33], to name just a few contributions). 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 therein). General investigations of coarse data from an imprecise-probability-based Bayesian point of view include [6, 36]; random set-based perspectives are developed for instance in

[8, 21]. Linear regression under metrical coarse data (interval data) is vividly discussed in the partial identification literature in the spirit of [19] (see also, e.g, [26], and the references therein). Mainly focusing on missing data, [34] suggests a framework for a systematic sensitivity analysis for statistical modelling under epistemic data imprecision. [5] introduces a profile likelihood approach for coarse data (for missing data see also [37]) and derive from it a uniform framework for robust regression analysis with imprecise data.

This paper will develop another likelihood-based (see, e.g., [4, § 6.3, 7.2.2] for a general introduction) approach and we will in addition briefly sketch Bayesian approaches in Section 3. Our work is strongly influenced by the methodology of partial identification, dealing with the trade-off 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., [19, 28, 20]). Sensitivity analysis pursues the same goal, but proceeds in a different direction. While partial identification starts from total uncertainty and gradually adds assumptions, in the framework of sensitivity analysis the collection of all precise results from successively relaxed assumptions is considered. Thereby, the analysis is framed by a sensitivity parameter, which is not identified but suffices to identify the parameter of interest, (e.g., [34]).

Our paper is structured as follows. In the next section we fix the notation and formulate the problem setting more exactly for the cases considered in this paper: independent and identically distributed (i.i.d.) variables and logistic regression with a categorical covariate. The crucial technical argument underlying our paper (developed in general terms in Section 3) is to introduce an observation model and utilize invariance properties of the likelihood. 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 situation, 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 assumption refining the initial set of estimators to a precise result. This assumption is called subgroup independent coarsening and 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 $Y_1, \dots, Y_n$ be a random sample of a categorical response variable of interest Y with realizations $y_1, \dots, y_n$ in sample space $\Omega_Y = \{1, \dots, j, \dots, K\}$. Problematically, some of those realizations are not known in a precise form, and thus only realizations $\mathscr{y}_1, \dots, \mathscr{y}_n$ of a sample $\mathcal{Y}_1, \dots, \mathcal{Y}_n$ 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. The possible categories of $\mathcal{Y}$ constitute the singletons of $(\Omega_{\mathcal{Y}},\ \mathcal{P}(\Omega_{\mathcal{Y}}))$, with corresponding probability mass functions $p_{\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i)$ $(i = 1, \dots, n)$. But as we are interested in the random variables $Y_1, \dots, Y_n$, our basic goal consists of gathering information about the individual probabilities $\pi_{i1} = P(Y_i = 1), \dots, \pi_{iK} = P(Y_i = K)$. Thereby, we assume throughout the paper that the coarsening process is error-free, in the sense that $\mathscr{y}_i \ni y_i$, $i = 1, \dots, n$.

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. Both situations will be illustrated by means of the following example.

Running Example: *We refer to the data from the German panel study “Labor Market and Social Security” (PASS, wave 1, 2006/2007, [29]). As asking for the income may be regarded as a sensitive question and thus the response rate is expected to be low, in this study non-responders are required to report their income in classes starting from rather large classes that are narrowed by following questions. By proceeding in this way, anonymization is guaranteed in the level that is requested by the respondents and answers of different degrees of coarseness are obtained. Keeping things simple, here we refer to the data from question “HEK0700”, where respondents are asked to report if their income Y is < 1000€ or ≥ 1000€ ($y_i \in \{<, \geq\}$; “$<$” and “$\geq$” abbreviating these classes, respectively) and our main goal is the estimation of $\pi_<$. As some respondents gave no suitable answer (“na”) and cannot be allocated to one of the classes, partly only coarsened values of the variable $\mathcal{Y}$ are observed ($\mathscr{y}_i \in \{<, \geq, na\}$).*

Example, version 1: In order to illustrate the i.i.d. case, we only consider the reported answers of the income question, where 238, 835 and 338 respondents reported "<", "≥" and "na", respectively ($n_< = 238,\ n_\geq = 835,\ n_{\text{na}} = 338$).

In the case with categorical covariates, we here confine ourselves to 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 $\pi_{i1} = P(Y_i = 1|X_i = x_i) = \pi_{x_i1}, \ldots,\ \pi_{iK} = P(Y_i = K|X_i = x_i) = \pi_{x_iK}$ are influenced by individual i through the corresponding value 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

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

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

$$\pi_{iK} = \left(1 + \sum\nolimits_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)\right)^{-1} \tag{2}$$

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.[1]

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., [31]). Thus, in our case the contingency table in Table 1 will be addressed. The number of observations with $\mathcal{Y} = \mathscr{y}$ and treatment group $X = x$ is denoted by $n_{x\mathscr{y}}$, 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, version 2: Illustrating the case with a categorical covariate, apart from the partial income knowledge, the receipt of the so-called Unemployment Benefit II (variable alg2abez; here denoted by UBII) is considered and serves in the model in Expressions (1) and (2) as covariate X_i, $i, \ldots, n$. The data are summarized in Table 2.

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

Table 1: Contingency table that introduces used notation.

		income			
		<	≥	na	total
UBII	yes (0)	130	114	75	319
	no (1)	108	721	263	1092
	total	238	835	338	1411

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

3 Sketch of the Basic Argument

This paper, similarly to [5, 37], 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 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 $\mathcal{Q}$ of (precise) coarsening probabilities,[2] 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 $\mathcal{Q}$ 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 $\mathcal{Q}$ (the various q's, possibly constrained by the specified constraints: $\big(q_{\mathscr{y}|y} := P(\mathcal{Y} = \mathscr{y}|Y = y)\big)_{(\mathscr{y} \in \Omega_\mathcal{Y}, y \in \Omega_Y)}$ 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., $(q_{\mathscr{y}|xy} := P(\mathcal{Y} = \mathscr{y}|X = x, Y = y)_{(\mathscr{y} \in \Omega_\mathcal{Y}, y \in \Omega_Y, x \in \Omega_X)}$ has to be considered).

Then we can describe the relationship between $\gamma := (\eta^T, \zeta^T)^T \in \Gamma$ and $\vartheta \in \Theta$ via the mapping $\Phi : \Gamma \to \Theta$, $\gamma \mapsto \vartheta$. Figure 1 and the running example illustrate

[1] Dummy variable X_{il} ($l = 1, \cdots, m$) attains value 1 if the l-th category is chosen by individual i, otherwise it is 0. In this way, reference category c is represented by all dummy variables being 0.

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

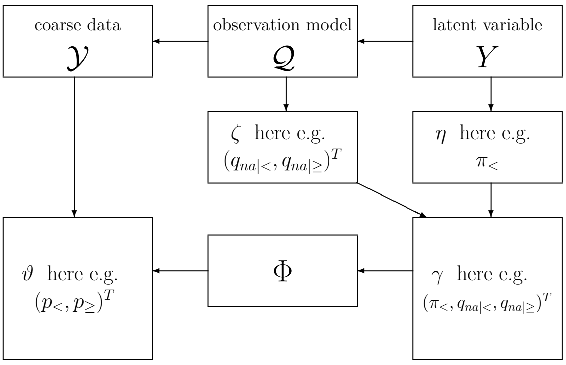


Figure 1: Observable and latent variable and the corresponding parameters.

this mapping $\Phi(\cdot)$ and all parameters involved.

Example, version 1 (cont.): The mapping $\Phi(\cdot)$ with arguments $\zeta = (q_{\text{na}|<}, q_{\text{na}|\geq})^T$ and $\eta = \pi_<$ establishes a connection to the parameters determining the probabilities of the observable income variable $\mathcal{Y}$, namely $\vartheta = (p_<, p_\geq)^T$.

In a first step (Section 4), we will only assume that the coarsening process is error-free and therefore take $\mathcal{Q}$ as the set of all coarsening mechanisms compatible with error-freeness. Then (Section 5), by using auxiliary information, we sharpen this set $\mathcal{Q}$. Note that we do neither assume anything about the plausibility of different elements ζ of $\mathcal{Q}$ nor do we treat different $y \in \mathscr{Y}$ as differently plausible. To derive the estimators, the invariance of the likelihood under parameter transformations is crucial: evaluating the likelihood in terms of γ and in terms of $\vartheta = \Phi(\gamma)$ is equivalent 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

$$\hat{\Gamma} = \left\{\gamma \,\middle|\, \Phi(\gamma) = \hat{\vartheta}\right\} \tag{3}$$

(see also [5, Section 2]). Thus, adapting the concept of maximum likelihood (ML) 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.

This compares to other approaches: A classical Bayesian analysis would put some prior on ζ and on η (cf., e.g., [23, 14]) while a generalized Bayesian analysis would replace one or both priors by a set of priors. This can be seen as imposing imprecise priors on ζ and on η. The non-committal analysis would start with a near-ignorance prior, for instance based on Dirichlet distributions adapting [35]'s imprecise Dirichlet model, and auxiliary information can be expressed by smaller credal sets; compare also the general Bayesian treatment of incomplete information in [6, 36]. Partially differently, in [3, Section 4.4.] an approach is presented that puts a precise prior on η and no prior on ζ and models the coarsening process with a multivalued mapping. This may be seen as imposing a vacuous imprecise probability on ζ. In another direction, one could impose some prior knowledge w.r.t. the imprecise data point $\mathscr{Y}$ by assuming different $y \in \mathscr{Y}$ as differently plausible. This can be done for example by imposing a possibility distribution on y (cf., e.g., [9, Section 3.2.]) or constructing observations directly by data augmentation (cf., e.g., [18]).

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 $|\Omega_X| \cdot m$. Thus, for reasons of conciseness of presentation, we 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 $\Omega_{\mathcal{Y}} = \{A,\ B,\ AB\}$, as well as a binary precise categorical covariate X with values 0 and 1. Then the underlying model expressed in Expression (1) and (2) is called *logit model*. As the inclusion of more than one dummy variable simply leads to an increase of the number of subgroups, all results can be transferred straightforwardly to more general cases, namely cases with more than one non-binary covariates. Furthermore, the main results not only will be shown for the situation of a binary Y, where coarsening corresponds to missingness, but also in a general way.

4 Maximum Likelihood Estimation 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 $\mathcal{Q}$. 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 just discussed, 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 $\mathcal{Y}_1, \dots, \mathcal{Y}_n$ with realizations $\mathscr{y}_1, \dots, \mathscr{y}_n$ 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 $p_{AB} = 1 - p_A - p_B$):[3]

$$\begin{aligned} L(\vartheta) &= L(p_A,\ p_B) = L(p_A,\ p_B || \mathscr{y}_1, \dots, \mathscr{y}_n) \qquad (4)\\ &= P(\mathscr{y}_1, \dots, \mathscr{y}_n || p_A,\ p_B) \propto p_A^{n_A} \cdot p_B^{n_B} \cdot p_{AB}^{n_{AB}} \,. \end{aligned}$$

For $n = n_A + n_B + n_{AB} > 0$ this likelihood is uniquely maximized by the relative frequencies (see [25]),

$$\hat{p}_A^{(MLE)} = \frac{n_A}{n}, \qquad \hat{p}_B^{(MLE)} = \frac{n_B}{n}\,, \qquad (5)$$

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

$$\begin{aligned} p_A &= (1 - q_{AB|A}) \cdot \pi_A\,, \qquad (6)\\ p_B &= (1 - q_{AB|B}) \cdot (1 - \pi_A)\,, \end{aligned}$$

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 $\vartheta = (p_A, p_B)^T$ in Expression (4) and in terms of $\gamma = (\pi_A, q_{AB|A}, q_{AB|B})^T$, coincide, indeed.

By the invariance of the likelihood under parameter transformations, Expressions (5) and (6) can be combined, resulting in the following system of equations:

$$\begin{aligned} (1 - \hat{q}_{AB|A}) \cdot \hat{\pi}_A &= \frac{n_A}{n} = \hat{p}_A^{(MLE)},\\ (1 - \hat{q}_{AB|B}) \cdot (1 - \hat{\pi}_A) &= \frac{n_B}{n} = \hat{p}_B^{(MLE)}, \quad (7)\\ \hat{q}_{AB|A} \cdot \hat{\pi}_A + \hat{q}_{AB|B} \cdot (1 - \hat{\pi}_A) &= \frac{n_{AB}}{n} = \hat{p}_{AB}^{(MLE)}. \end{aligned}$$

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 $\hat{\vartheta} = (\hat{p}_A^{(MLE)},\ \hat{p}_B^{(MLE)})^T$, the mapping $\Phi:\ [0,\ 1]^3 \to [0,\ 1]^2$ with

$$\Phi\begin{pmatrix} \pi_A \\ q_{AB|A} \\ q_{AB|B} \end{pmatrix} = \begin{pmatrix} \pi_A \cdot (1 - q_{AB|A}) \\ (1 - \pi_A) \cdot (1 - q_{AB|B}) \end{pmatrix} = \begin{pmatrix} p_A \\ p_B \end{pmatrix} \quad (8)$$

(cf. Figure 1 for the case of the running example) connecting both parametrizations in general is not injective. Thus the maximum likelihood estimate $\hat{\Gamma}$ from Expression (3) is set-valued in the literal sense. Points in this set are constrained through the relationships in (7), 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

$$\begin{aligned} \hat{\pi}_A \quad &\in \quad \left[\frac{n_A}{n},\ \frac{n_A + n_{AB}}{n}\right]\,, \qquad (9)\\ \hat{q}_{AB|A} \quad &\in \quad \left[0,\ \frac{n_{AB}}{n_A + n_{AB}}\right]\,, \end{aligned}$$

and analogously for $\hat{q}_{AB|B}$, where $\frac{0}{0} := 1$.

Extending the discussion here to the general case of $\Omega_Y = \{1, \dots, K\}$ and the corresponding $\Omega_{\mathcal{Y}}$, the estimators in Expression (9) generalize to

$$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n},\ \frac{\sum_{\mathscr{y} \ni y} n_{\mathscr{y}}}{n}\right] \quad \hat{q}_{\mathscr{y}|y} \in \left[0,\ \frac{n_{\mathscr{y}}}{n_{\{y\}} + n_{\mathscr{y}}}\right]\,, \qquad (10)$$

(where as above $\frac{0}{0} := 1$) for all $y \in \Omega_y = \{1, \dots, K\}$ and all $\mathscr{y} \in \Omega_{\mathcal{Y}}$ such that $\{y\} \subset \mathscr{y}$.[4]

Example, version 1 (cont.): Applying Expression (10) to our example, one obtains

$$\hat{\pi}_< \in \left[\frac{238}{1411},\ \frac{238 + 338}{1411}\right] = [0.17,\ 0.41]\ .$$

4.2 Logistic Regression with a Categorical Covariate

Now we consider the heterogeneous 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

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

[4] 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 $\mathscr{y}_1, \dots, \mathscr{y}_n$ (see, e.g., [2])

applying the appropriate parameter transformations. Proceeding in this way, for fixed treatment group x the cell counts $(n_{xA}, n_{xB},\ n_{xAB})$ 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 $p_{x\mathscr{Y}} = P(\mathcal{Y} = \mathscr{Y}|X = x)$ (see [31, 1]).[5] Therefore, the corresponding likelihood function is given by

$$\begin{aligned} L(\vartheta) &= L(p_{0A},\ p_{1A},\ p_{0B},\ p_{1B}) \qquad (11)\\ &\propto p_{0A}^{n_{0A}} \cdot p_{0B}^{n_{0B}} \cdot p_{0AB}^{n_{0AB}} \cdot p_{1A}^{n_{1A}} \cdot p_{1B}^{n_{1B}} \cdot p_{1AB}^{n_{1AB}} . \end{aligned}$$

For $n_x > 0$ the maximum likelihood estimators for the parameters are unique and given by (see [25])

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

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

$$\Phi \begin{pmatrix} \pi_{0A} \\ \pi_{1A} \\ q_{AB|0A} \\ q_{AB|1A} \\ q_{AB|0B} \\ q_{AB|1B} \end{pmatrix} = \begin{pmatrix} \pi_{0A} \cdot (1 - q_{AB|0A}) \\ \pi_{1A} \cdot (1 - q_{AB|1A}) \\ (1 - \pi_{0A}) \cdot (1 - q_{AB|0B}) \\ (1 - \pi_{1A}) \cdot (1 - q_{AB|1B}) \end{pmatrix} = \begin{pmatrix} p_{0A} \\ p_{1A} \\ p_{0B} \\ p_{1B} \end{pmatrix} \qquad (12)$$

(cf. Figure 1) and observe that in this case it is also not injective and thus $\hat{\Gamma}$, constructed along the line of (3), is strictly set-valued, too. Illustrating $\hat{\Gamma}$ again by the corresponding 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 $\mathscr{Y} \in \Omega_{\mathcal{Y}}$ with $\{y\} \subset \mathscr{Y}$,

$$\hat{\pi}_{xy} \in \left[\frac{n_{x\{y\}}}{n_x},\ \frac{\sum\limits_{\mathscr{Y} \ni y} n_{x\mathscr{Y}}}{n_x} \right],\ \hat{q}_{\mathscr{Y}|xy} \in \left[0,\ \frac{n_{x\mathscr{Y}}}{n_{x\{y\}} + n_{x\mathscr{Y}}} \right], \qquad (13)$$

where again $\frac{0}{0} := 1$.[6]

Example, version 2 (cont.): Applying Expression (13) to our example, one obtains

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

By recurring on the relation defined in Expression (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 $\hat{\pi}_{1A}$ determined by Expression (13), equivalently one can consider the estimators

$$\begin{aligned} \hat{\beta}_{A0} &\in \left[\log\left(\frac{n_{0A}}{n_{0B} + n_{0AB}} \right), \log\left(\frac{n_{0A} + n_{0AB}}{n_{0B}} \right) \right] \\ \hat{\beta}_{A} &\in \left[\log\left(\frac{n_{1A} \cdot (n_{0B} + n_{0AB})}{n_{0A} \cdot (n_{1B} + n_{1AB})} \right),\right. \qquad (14) \\ &\qquad \left. \log\left(\frac{n_{0B} \cdot (n_{1A} + n_{1AB})}{n_{1B} \cdot (n_{0A} + n_{0AB})} \right) \right], \end{aligned}$$

assuming all expressions to be well-defined.

Example, version 2 (cont.): In terms of the regression coefficients, we obtain the estimates $\hat{\beta}_{<0} \in [-0.37,\ 0.59]$ and $\hat{\beta}_{<} \in [-1.83,\ -1.25]$.

Interpreting the indeterminate sign of intercept $\beta_{<0}$, one notes that for the group of persons that receives UBII (i.e. $X = 0$) the chance of being in the lower income group ($<$ 1000€) in comparison to being in the higher income group ($\geq$ 1000€) varies between $\exp(-0.37) = 0.69$ and $\exp(0.59) = 1.89$. In this way, one cannot judge the impact of the UBII on the dependent variable income without implying further assumptions about the coarsening. Unjustifiably ignoring the coarsening (see Section 5.2) pretends a particular sign of the regression coefficients. This corroborates the importance of including all imaginable coarsening mechanisms for obtaining a trustworthy result, which will be discussed now more in detail.

5 Reliable Incorporation of Auxiliary Information: Sensitivity Parameters and Partial Identification

The set-valued estimators from Expression (9) (and analogously from Expression (13)) 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., [19]). In this way, the trivial bounds [0, 1] on the probabilities have been refined substantially. In the spirit of partial identification and sensitivity analysis we can further refine the analysis if, and also only if, auxiliary information beyond the empirical evidence is available. Vansteelandt et al. [34] suggests to determine a sensitivity parameter δ in some range Δ 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

[5] This corresponds to a product-multinomial sampling scheme (e.g. [31, 1]).

[6] 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.

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 $(1 - \alpha)$ [13, 34].

To handle the inclusion of reliable information technically, we start with distinguishing and investigating point identifying additional assumptions, in order to utilize them as a technical means to derive sensitivity parameters, governing the incorporation of additional information.

Due to the fact that the imprecise point estimators in Expression (13) directly result from considering Expression (9) 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 (8) 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 Expression (9) can be shrunk to a single-valued estimator. The exact values of the coarsening parameters are most often unknown, but in case there is material information available that allows to bound them in non-trivial intervals, 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 non-stochastic, the underlying degree of coarsening is predetermined and known. For instance, if respondents are requested to give their answer in a grouped way and we assume that all respondents answer correctly, then the coarsening is predefined in the sense that there is a unique coarsened outcome for every true answer. In the context of distinguishing between non-stochastic and stochastic coarsening mechanisms, Heitjan and Rubin [12] 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 $q_{\mathscr{Y}|y}$ is constant, no matter which true value y is underlying as long as it fits to the observed value $\mathscr{Y}$. Illustrated by the running example, CAR postulates that the probability of giving no suitable answer should not depend on the true income category, which contradicts practical experiences (e.g., [16]). In the dichotomous situation of this example we are then actually concerned with the assumption of missing at random (MAR) [18], 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 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 ideal-type situations in which CAR can be justified indeed.[7] 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 relevance of the analysis, as is also illustrated in Figure 2. There the estimation of π_A under obstinately assumed CAR but varying coarsening probabilities is evaluated by the median relative empirical bias $\frac{\hat{\pi}_A - \pi_A}{\pi_A}$ based on 100 simulated datasets (here with $\pi_A = 0.6$).[8] The absolute value of the relative median bias increases the more one deviates from the case of CAR, indeed, up to a median relative bias of almost 80%.

5.3 Ratio of Coarsening Parameters

In our context the paper by Nordheim [22] obtains new importance. He considers the ratio between different mechanisms in the context of non-randomly missing and misclassified data. By fixing the ratio between the coarsening probabilities the corresponding maximum likelihood problem leads to quadratic equations, where

[7] 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 [15, 11].

[8] 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 that is maximized.

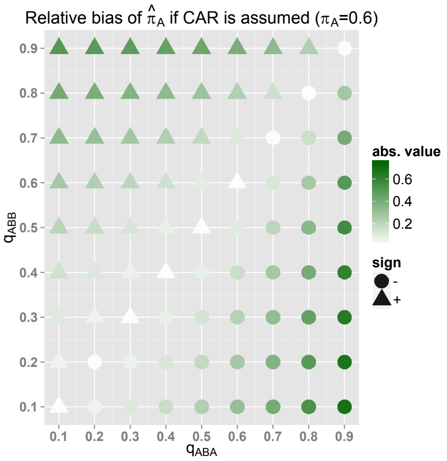


Figure 2: Consequences for the median relative bias of $\hat{\pi}_A$ if there is a deviation from assumed CAR.

one solution is contained in the interval of $\hat{\pi}_A$ from Expression (9), while the other solution lies outside of [0, 1] (cf. [22, 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 in Section 4.1, unique, empirically point identified estimators are obtained as

$$\begin{aligned} \hat{\pi}_A &= \frac{n_A \cdot R}{n_B + n_A \cdot R}, \qquad (15) \\ \hat{q}_{AB|A} &= \frac{n_B \cdot (R-1) + n_{AB} \cdot R}{n \cdot R} \end{aligned}$$

containing CAR as the special case $R = 1$. 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 [22]), where again a substantial bias can occur. 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.

Because 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 the magnitude of R can be derived from material 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.[9] Considering the ratio R as a sensitivity parameter leads to the HEIRs.[10]

5.4 Subgroup Independent Coarsening

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 underlying covariate value. Illustrated by the running example, imposing this kind of assumption means that answering in a coarse form, i.e., giving no suitable answer, does not depend on the receipt of unemployment benefit. As the receipt of unemployment benefit depends on the income, and the value of the income may influence the non-response to the income question (cf. Section 5.2), this assumption should be treated with particular caution here.

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*

$$\begin{aligned} q_{AB|0A} &= q_{AB|1A} =: q_{AB|A} \qquad (16) \\ q_{AB|0B} &= q_{AB|1B} =: q_{AB|B}, \end{aligned}$$

in the estimation problem of Section 4.2, the mapping $\Phi(\cdot)$ from Expression (12) is now injective[11] if restricted to the arguments $(\pi_{0A}, \pi_{1A}, q_{AB|A}, q_{AB|B})^T \in (0,1)^4$ such that

$$\pi_{0A} \notin \{0,1\},\ \pi_{1A} \notin \{0,1\}\ \underline{\text{and}}\ \pi_{0A} \neq \pi_{1A}. \qquad (17)$$

One obtains the following unique estimators

$$\begin{aligned} \hat{\pi}_{0A} &= \frac{n_{0A}}{n_0} \frac{n_{1B}n_0 - n_1 n_{0B}}{n_{0A}n_{1B} - n_{0B}n_{1A}}, \qquad (18) \\ \hat{\pi}_{1A} &= \frac{n_{1A}}{n_1} \frac{n_{1B}n_0 - n_1 n_{0B}}{n_{0A}n_{1B} - n_{0B}n_{1A}}, \\ \hat{q}_{AB|A} &= 1 - \frac{n_{0A}n_{1B} - n_{0B}n_{1A}}{n_{1B}n_0 - n_1 n_{0B}}, \\ \hat{q}_{AB|B} &= 1 - \frac{n_{0A}n_{1B} - n_{0B}n_{1A}}{n_{0A}n_1 - n_{1A}n_0}, \end{aligned}$$

[9] An example is given in the preliminary version of a technical report available at `http://www.statistik.lmu.de/~jplass/forschung.html`

[10] 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. More detailed information can be found in the preliminary version of a technical report cited in footnote 9.

[11] A proof of the injectivity of Φ in this situation is given in the preliminary version of a technical report cited in footnote 9. The case of $\pi_{0A} = \pi_{1A}$ reproduces the i.i.d. case, where there are multiple solutions.

when these are well-defined and 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 Expression (18) typically for all cases satisfying Expression (17) will be in [0, 1]. It has to be re-emphasized that in practical applications one must carefully reflect the plausibility of the subgroup independent coarsening assumption of Expression (16). 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, a possibility to check whether the subgroup independent coarsening is appropriate at all.

5.5 A Generalization of Subgroup Independent Coarsening

There are situations in which one might have an idea about the relative magnitude of the probabilities of precise observations 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. As in Section 5.4, the mapping $\Phi(\cdot)$ from Expression (12) is injective for all cases in Expression (17) and thus unique estimators result.[12] 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_1 and R_2.

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 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 when 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 [26, Sec. 3.1] 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 hypotheses tests and regions taking finite sample variability into account explicitly. These estimators also should be compared to confidence intervals derived along the lines of [34] when an appropriate sensitivity parameter could be determined.

Other areas of further research include a deeper investigation of the alternative generalized Bayesian (and possibilistic) approaches briefly mentioned in Section 3 as well as the consideration of other "deficiency" processes, most notably misclassification, which can be formalized in a very similar way. Our methodology thus also offers an alternative to, and a generalization to logistic regression of, recent work on misclassification from a partial identification perspective [20, 17].

Acknowledgements We are very grateful for the helpful remarks of three anonymous reviewers and appreciate their open sharing of stimulating ideas very much.

References

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

[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] A. Benavoli. Belief function and multivalued mapping robustness in statistical estimation. *Int. J. Approx. Reasoning*, 55:311–329, 2014.

[12] They are given in the preliminary version of the technical report cited in footnote 9.

[4] G. Casella, R. Berger. *Statistical Inference*. 2nd edn., Duxbury, 2002.

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

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

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

[8] I. Couso, D. Dubois, L. Sánchez. *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables*. Springer, Cham, 2014.

[9] T. Denoeux. Likelihood-based belief function: justification and some extensions to low-quality data. *Int. J. Approx. Reasoning*, 55:1535–1547, 2014.

[10] 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.

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

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

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

[14] T. Jiang, J. Dickey, Bayesian methods for categorical data under informative censoring, *Bayesian Anal.*, 3:541–553, 2008.

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

[16] A. Korinek, J. Mistiaen, M. Ravallion. Survey nonresponse and the distribution of income. *J. Econ. Inequal.*, 4:33–55, 2006.

[17] 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]

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

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

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

[21] H. Nguyen, B. Wu. Random and fuzzy sets in coarse data analysis. *Comput. Stat. Data. An.*, 51:70–85, 2006.

[22] 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, 1984.

[23] D. Paulino, C. De B. Pereira. Bayesian analysis of categorical data informatively censored. *Commun. Stat., Theory Methods*, 21:2689–2705, 1992.

[24] 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 revision for ISIPTA ’15*. See also: *Techn. Rep., 179, Dep. Statistics, LMU Munich,* 2015 (url: www.epub.ub.uni-muenchen.de/23816).

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

[26] 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]

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

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

[29] 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.

[30] 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.

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

[32] 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]

[33] 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, 2011.

[34] 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.

[35] P. Walley. Inferences from multinomial data: Learning about a bag of marbles (with discussion). *J. R. Stat. Soc. B*, 58:3–57, 1996.

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

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

LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Julia Plass, Marco Cattaneo, Thomas Augustin,
Georg Schollmeyer, Christian Heumann

Towards a reliable categorical regression analysis for non-randomly coarsened observations: An analysis with German labour market data

Technical Report Number 206, 2017
Department of Statistics
University of Munich

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

Towards a reliable categorical regression analysis for non-randomly coarsened observations: An analysis with German labour market data

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

Marco Cattaneo
School of Mathematics and Physical Sciences
University of Hull
m.cattaneo@hull.ac.uk

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

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

Christian Heumann
Department of Statistics, LMU Munich
christian.heumann@stat.uni-muenchen.de

Abstract

In most surveys, one is confronted with missing or, more generally, coarse data. Many methods dealing with these data make strong, untestable assumptions, e.g. coarsening at random. But due to the potentially resulting severe bias, interest increases in approaches that only include tenable knowledge about the coarsening process, leading to imprecise, but credible results. We elaborate such cautious methods for regression analysis with a coarse categorical dependent variable and precisely observed categorical covariates. Our cautious results from the German panel study "Labour market and social security" illustrate that traditional methods may even pretend specific signs of the regression estimates.

Keywords: coarse data, (cumulative) logit model, missing data, partial identification, PASS data, (profile) likelihood

1 Introduction: How to respect the (lack of) knowledge about incompleteness

In almost all surveys the problem of item-nonresponse occurs [e.g. 19, 40]. One of the principal challenges in the statistical analysis of missing data is the impossibility to test the associated missingness mechanism without adding strong assumptions [e.g. 20]. Despite the awareness of this problem, frequently untestable assumptions on the missingness process are still included in situations where the validity of these assumptions might actually be doubtful. Examples are the missing at random assumption [introduced by 34] or approaches relying on a specific pattern-mixture or selection model [e.g. developed by 12]. In this way, point-identifiability, i.e. uniqueness of parameters, is forced, which is an important prerequisite for the applicability of traditional statistical methods, as for instance the EM algorithm or imputation techniques [e.g. 22].

Especially due to the substantial bias induced by wrongly imposing such point-identifying assumptions, a proper reflection of the available information about the underlying missingness assumption is indispensable [e.g. 23]. To this end, one departs from insisting on point-identifying assumptions by turning to strategies that only include the achievable knowledge, typically ending up in set-valued estimators. In this way, approaches based on the methodology of partial identification start with no missingness assumptions at all, but then add successively assumptions compatible with the obtainable knowledge [e.g. 23]. A practical example is given in [1], where the worst-case bounds for the HIV rate resulting from an approach without any assumptions about the missingness are then refined by exploiting the longitudinal nature of the data. Similarly, sensitivity analyses for selection models take several different models of missing data processes into account [e.g. 14, 21, 45]. Recently, Manski [24] gave a new impetus to this topic by stressing the advantage of reliable, so to say interval-valued point estimates for official statistics with survey nonresponse.

Against this background, we rely on cautious likelihood-based strategies for incomplete data [similarly as in 6, 8, 21, 47], and pursue the goal of determining regression estimators reflecting the available information about the incompleteness in a careful way.[1] Motivated by the two considered examples regarding the income questions from the German panel study “Labour market and social security” [PASS, 42], the focus is set on the logit model for binary response data and the cumulative logit model for ordinal response data. In doing so, we not only restrict to the issue of nonresponse, but also look at the problem of missingness more generally: Apart from fully observed and fully unobserved values, we additionally consider partially observed values where subsets of the full sample space are observed, thus addressing the coarse data problem [e.g. 13]. Consequently, coarse data contain more information than missing data, wherefore we argue in favor of collecting coarse data in case of a preceding nonresponse. Throughout, we restrict to cases of coarse categorical response variables and precisely observed categorical covariates.

Although analysts might be aware of the consequences of traditional approaches mostly making simplified assumptions such as coarsening at random, they frequently prefer them to cautious approaches for pragmatic reasons. To face this dilemma, we provide an estimation technique that includes all available information about the coarsening in a very natural and flexible way. There are already several methods that try to exploit additional infor-

[1] While in Plass et al. [32] first considerations have already been presented for the special case of a multinomial logit model that included all interactions between the covariates, we here investigate general model specifications.

mation about the incompleteness, as e.g. knowledge about the number of failed-contact attempts in Wood et al. [46] or prior expert beliefs about the differences between responders and nonresponders in Jackson et al. [19]. But since these approaches are mostly restricted to either give a precise result or no answer at all, they are incapable to make use of potential available partial knowledge about the missingness that is not sufficient to point-identify the parameters of interest [e.g. 39]. Consequently, the users might conceive the explicit allowance of partially identified parameters as an advantage, since partial knowledge no longer has to be left out of consideration. In our data example we show how partial information about the coarsening such as "respondent with a high income rather tend to give a coarse answer compared to respondents with a low income" can refine the initial results without coarsening assumptions. Furthermore, we give the opportunity to consider "coarsening at random" instead of "exact coarsening at random" models improving the credibility of classical approaches.

The relevance of such a cautious approach, and hence the need of quantifying the underlying uncertainty due to incompleteness, is also apparent from the following latest practical example: Results on the job-seeking refugees in Germany without school-leaving qualification were published by the Federal Employment Agency and provoked a heated debate, mainly reasoned by a different dealing with item-nonresponse. While ignoring the 24.7% nonresponders leads to the result that 34.3% job-seeking refugees are without school-leaving qualification and assumes the refusals to be made randomly, the newspaper "Bild" disseminates an extreme interpretation of the Federal Institute for Vocational Education and Training's (BIBB) conjecture that job-seeking refugees without school-leaving qualification rather tend to disclose their answer and simply counted all nonresponders to this group, hence speaking of 59% in this context [cf., e.g., 15, 4]. A clear communication of the underlying uncertainty would have avoided the discussions and should generally be part of every trustworthy data analysis. As a reaction to the incident several statistical agencies pointed to the importance of reflecting about the reasons why the respondents refused their answers [cf., e.g., 5]. The cautious approach presented in this paper is able to express the underlying uncertainty attributed to nonresponse and could potentially derive weak, but tenable knowledge about the coarsening from the main reasons for nonresponse. In fact, we not only deal with the uncertainty associated to the incompleteness of the data leading to imprecise results, but also two further kinds of uncertainty: By constructing confidence intervals, we capture the uncertainty arising from the availability of a finite sample only. Studying regression models, we additionally address model uncertainty arising from the parametric assumptions implied by non-saturated regression models. The interaction between the different kinds of uncertainty will be a further aspect of investigation in this paper.

Our paper is structured as follows: In Section 2 we motivate the collection of coarse data, introduce the running example based on the PASS data, explain the way we look at the problem and briefly show the principal idea of the two methods to determine cautious regression estimates that we present and discuss in this paper. Both methods are firstly developed in context of a data example with a binary response variable reducing to the missing data problem in Section 3, where also a way to obtain respective likelihood-based confidence intervals is given. The synergy of the included parametric assumptions on the regression model and the observed data strongly determines the type of results, where three substantially differing cases are elaborated. Afterwards, the applicability of the previous major developments is discussed in the context of coarse data in the strict sense in Section 4. In Section 5 we turn to situations where we benefit from weak auxiliary information about the coarsening. Section 6 concludes by giving a summary and some remarks on further research.

2 Coarse categorical data

In most surveys, respondents can choose between several predetermined options to answer. Nevertheless, providing answers associated to a specific level of accuracy may be considered as problematic for different reasons: Firstly, respondents might be able to give a more precise answer, but there is no possibility to express it. Secondly, the other way round, respondents potentially may at most be able to decide for a set of categories, but not for the one category they actually belong to, since they are not acquainted enough with the topic of the question. Thirdly, respondents may deliberately refuse their precise answer for reasons of data privacy. While the consequence in the first situation is (only) loss of information, in the second and third situation non-ignorable nonresponse or measurement errors occur in a classical questionnaire design. All these problems could be attenuated by asking in different ways allowing the respondent to report in the required level of accuracy. An example for such an explicit collection of coarse categorical data is given in the following section by introducing the setting of the running example.

2.1 The running data example

Since the income question is known to be highly affected by nonresponse [e.g. 41], the German Panel study “Labour market and social security” [PASS study[2], 42] intends to mitigate this problem by using the following questioning technique illustrated in Figure 1: Respondents refusing to disclose their precise income (in the following called nonresponders) are asked to answer additional questions starting from providing rather large income classes (e.g. < 1000 € or not) that are successively narrowed (e.g. < 500 €).[3] In this way, answers with different levels of coarseness are received by simultaneously ensuring the individual degree of data privacy demanded by the respective respondent. This strategic questioning technique to increase response rates is sometimes referred to as non-response follow-up [e.g. 30, where this is distinguished from “follow-up attempts”, i.e. repeated efforts to contact respondents]. Depending on the research question, various ways to integrate the answers from the respondents reporting their precise, non-categorical income are conceivable, where we first point to some general options before we mention how we proceed here: To include all answers in the most precise level inferable from the data, a mixture model [e.g. 25] may be used differentiating between nonresponders and responders. In some situations, as e.g. in the context of poverty measurement, an answer on a certain ordinal level might be sufficient, hence the precise answers could be classified to the most precise income categories reported by the nonresponders, allowing a joint analysis. An alternative might be a joint likelihood approach accounting for responders, nonresponders and different groups of partial responders by distinct likelihood contributions [cf. 10, who use an imputation based technique and illustrate their results by the PASS data as well]. Restricting to the answers of the nonresponders, here we consider the most precise collectable categorical income as the true income category, ignoring that a (quasi-) continuous variable is underlying. In a second step a mixture model or a comparative analysis

[2] Here we rely on the data from wave 1 to 4

[3] For ease of presentation, we here restrict to the granularity of categories given in Figure 1. In fact, the PASS data partly provide even finer categories.

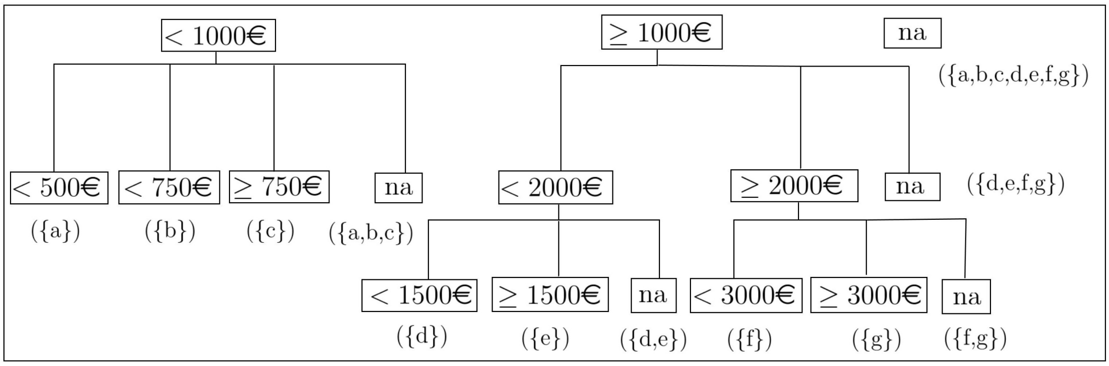


Figure 1: In the PASS study for nonresponders the income questions are individually adjusted, providing for instance categories abbreviated by "< 500 €", "< 750 €" (actually meaning < 750 € and ≥ 500 €) and "≥ 750 €" (≥ 750 € and < 1000 €) to original nonresponders who already reported to be in class < 1000 € in an earlier question. The notation in brackets refers to **Example 2**, introduced later on, where the cardinality of the sets gives some indication about the level of accuracy.

to the responders could follow.

Our main goal will be the investigation of some covariates' impact on a true categorical response variable partly observed in a coarse way. In the example, the true categorical income is used as a response variable distinguishing the following two settings, referred to as "Example 1" and "Example 2" later on:

Example 1: Binary response variable
Here we restrict the available income data to the answers obtained from the first question. Thus, categories "< 1000 €", "≥ 1000 €" and "no answer" (i.e. coarse answer "either < 1000 € or ≥ 1000 €") are observed, reducing the coarsening probem to the missing data problem. When we consider **Example 1**, the categories are abbreviated by "<", "≥" and "na" in the following.

Example 2: Ordinal response variable
Here we account for the whole ordinal structure inherent in the data, and the observed income variable includes different levels of coarseness. In the context of **Example 2**, the abbreviations given in brackets in Figure 1, i.e. categories "{a}" to "{a,b,c,d,e,f,g}", are utilized, where the latter one is interpreted as "either a or b or ... or g".

In this way, we constructed one data situation with a binary and one with an ordinal true response variable (with values "< 1000 €" and "≥ 1000 €" and values "{a}" to "{g}", respectively) in order to exemplify the results obtained by the two considered models. In Section 3, we use **Example 1** to illustrate the respective proposals, while in Section 4 the applicability of the previous ideas for coarse data, not reducing to the missing data case, is studied by referring to **Example 2**.

We use the highest school leaving certificate (first covariate) and age (second covariate) as covariates. Both variables are dichotomized, thus showing values "Abitur no (Abi 0)" and "Abitur yes (Abi 1)"[4] as well as "< 40 (0)" and "≥ 40 (1)", respectively. Since the categorical income questions are only directed to respondents refusing to disclose their precise

[4] The "Abitur" is the general qualification for university entrance in Germany.

Table 1: Contingency table for the data of **Example 1** (binary response variable).

	Observed income class		
Abi, age	<	≥	na
0, 0	97	63	102
0,1	69	115	131
1,0	33	50	41
1,1	38	79	59

Table 2: Contingency table for the data of **Example 2** (ordinal response variable).

	Observed income class											
Abi, age	{a}	{b}	{c}	{a,b,c}	{d}	{e}	{d,e}	{f}	{g}	{d–g}	{f,g}	{a–g}
0, 0	50	17	18	12	22	11	*	9	*	9	*	102
0, 1	24	18	21	6	23	18	6	16	9	33	10	131
1, 0	21	*	*	*	10	7	5	7	8	9	4	41
1 , 1	20	9	*	*	*	9	*	14	20	17	10	59

income, a group expected to be small in a study concerning the labour market, the number of individuals included in our analysis is comparably small. The contingency tables in Table 1 and Table 2 summarize the considered unweighted data including information of 877 individuals. To comply with our data access contract and the non-disclosure regulations of the Federal Employment Agency [cf. 3], we have to prohibit any back-calculations and delete all frequencies that are ≤ 3, here marking them by "*". In each line of Table 2 the sums of the frequencies referring to the categories $\{a\}$, $\{b\}, \{c\}$ and $\{a, b, c\}$ (group 1) as well as to $\{d\}$, $\{e\}$, $\{d, e\}$, $\{f\}$, $\{g\}$, $\{d - g\}$ and $\{f, g\}$ (group 2) can be inferred from Table 1. For that reason, we additionally hide the next smallest entry in each group showing deleted entries; to increase possibilities of potential replacements, one further entry is marked by "*", whenever the sum of the frequencies in the deleted entries is smaller than seven. All frequencies are > 0 (cf. assumptions in Section 2.2).

2.2 The general view of the problem

To frame the problem of coarse data technically, we distinguish between an observed and a latent world.

Let $(x_{11}, \ldots,\ x_{1p},\ y_1),\ \ldots, (x_{n1}, \ldots,\ x_{np},\ y_n)$ be a sample of n independent realizations of categorical random variables $(X_1, \ldots,\ X_p,\ Y)$. Unfavorably, some values y_i are not known precisely, hence the random variable Y refers to the latent world. Instead, we only observe a sample $(x_{11}, \ldots,\ x_{1p},\ \mathfrak{y}_1), \ldots, (x_{n1}, \ldots,\ x_{np},\ \mathfrak{y}_n)$ of n independent realizations of $(X_1, \ldots,\ X_p,\ \mathcal{Y})$, where the random set $\mathcal{Y}$ [e.g. 28] belongs to the observed world. We lay a special focus on the variable Y with sample space Ω_Y and the random set $\mathcal{Y}$ with sample space $\Omega_{\mathcal{Y}} \subset \mathcal{P}(\Omega_Y)$, where we assume the empty set to be generally excluded, but all precise categories $\{y\}$ to be included. Since we aim for a regression analysis here, we are interested in the estimation of the probabilities $\pi_{\mathbf{x}y} = P(Y = y|\mathbf{X} = \mathbf{x})$, $y \in \Omega_Y$, given the – assumed to be – precise values $\mathbf{x} = (x_1, \ldots, x_p)^T \in \Omega_X$ of categorical covariates $X_1, \ldots, X_p$. The associated dependence on the covariates is described by an appropriate

response function, $\pi_{\mathbf{x}y} = h(\eta_{\mathbf{x}y})$, with linear predictor $\eta_{\mathbf{x}y} = \beta_{0y} + d(\mathbf{x})^T\boldsymbol{\beta}_y$, where d fills the role of transferring the covariates into appropriate dummy-coded ones [cf., e.g. 11, p. 31]. Our main goal will be a cautious estimation of the regression coefficients β_{0y} and $\boldsymbol{\beta}_y$ that only includes the available information about the coarsening process.
By means of the law of total probability that includes coarsening parameters $q_{\mathfrak{y}|\mathbf{x}y} = P(\mathcal{Y} = \mathfrak{y}|\mathbf{X} = \mathbf{x}, Y = y)$ with $\mathbf{x} \in \Omega_X$, $y \in \Omega_Y$ and $\mathfrak{y} \in \Omega_{\mathcal{Y}}$ (cf. Section 3.1.1), we formalize the connection between both worlds, i.e. the latent world with parameters $\pi_{\mathbf{x}y}$, $y \in \Omega_Y$, $\mathbf{x} \in \Omega_{\mathbf{X}}$ and the observed world with parameters $p_{\mathbf{x}\mathfrak{y}} = P(\mathcal{Y} = \mathfrak{y}|\mathbf{X} = \mathbf{x})$, $\mathfrak{y} \in \Omega_{\mathcal{Y}}$, $\mathbf{x} \in \Omega_X$. Apart from requiring error-freeness in the sense that the true value is contained in the coarse value, $\mathfrak{y} \ni y$, and distinct parameters [cf. 34], we mainly refrain from making assumptions about the coarsening, only discussing in Section 5 how frequently available weak knowledge about the coarsening can be included in a powerful way. Considering the contingency table framework, $n_{\mathbf{x}\mathfrak{y}}$ and $n_{\mathbf{x}}$ represent the counts within the respective cells.

2.3 Two ways of approaching the problem

In this paper, we discuss two procedures to determine cautious maximum likelihood estimators for the regression coefficients:

- **Two-step method:** We firstly estimate the bounds of the latent variable distribution $\pi_{\mathbf{x}y} = P(Y = y|\mathbf{X} = \mathbf{x})$, $y \in \Omega_Y$, $\mathbf{x} \in \Omega_{\mathbf{x}}$, from which the cautious regression estimates are determined in a second step.
- **Direct method:** We rely on the (relative) profile log-likelihood for the regression coefficients of interest, where the set of maxima gives the cautious regression estimates.

Being interested in maximum likelihood estimators of the regression coefficients, maximizing the corresponding (profile log-) likelihood, i.e. the direct method, represents the natural procedure, which is always applicable. In specific situations – which we will characterize here – a two-step method will turn out as a useful alternative. Additionally, the way through the estimation of the latent variable distribution shows to be beneficial when we study how the parametric assumption on the regression model affects the estimated coarsening parameters, since we can implicitly control for the compatibility with the observed data. Nevertheless, it is important to point out that there are situations where only the direct method is worthwhile and hence the two methods cannot be regarded as at the same level.
Both ways aim at the cautious maximum likelihood estimators for each component of the vector of regression coefficients. Consequently, we gain an impression about the magnitude of each effect when no assumptions about the coarsening are imposed, but we cannot directly infer which one-dimensional regression estimates are combinable to achieve the maximum of the likelihood.

3 Cautious estimation of regression coefficients

An important contribution of this paper consists of elaborating how the presence of parametric assumptions on the regression model – in the sense that at least one effect or interaction of the saturated model is set equal to zero – can affect the assumptions about the coarsening process. By comparing the results from a two-step method (cf. Section 2.3) for the case with and without any parametric assumptions on the regression model, interesting insights with regard to this point can be gained. For that reason, we firstly devote ourselves to the case of a saturated model that includes all interactions between the covariates (cf. Section 3.1) and account for the uncertainty induced by parametric assumptions on the regression model only afterwards (cf. Section 3.2).

If a saturated model is chosen, a two-step method appears to be quite natural and hence we will restrict to this way here: Since there is no reduction of the parameter space and the latent variable distribution basically represents the same information as the regression estimators, we can determine the cautious regression estimators (cf. Section 3.1.2; also cf. 32, where the multinomial logit model is used in this context) by simply transforming the bounds of the latent variable distribution obtained in a first step (cf. Section 3.1.1).

Things become substantially different in the presence of parametric assumptions on the regression model, i.e. if a non-saturated model is specified. Now, due to the reduction of the parameter space a transformation as in the saturated model is no longer valid and the direct approach (cf. Section 2.3) is becoming more important. Nevertheless, basing considerations on a two-step method in some cases still may be useful and we formulate a constraint optimization problem that incorporates the bounds of the latent variable distribution (cf. Section 3.1.1).

3.1 The saturated model

3.1.1 Maximum likelihood estimation for the latent variable distribution

In order to estimate the latent variable distribution, we basically split the argumentation by completing three steps [32]: Firstly, we use the random set perspective interpreting all elements in $\Omega_{\mathcal{Y}}$ as categories of their own. Thus, in contrast to the situation in the latent world, knowledge about the "precise" values in the observed world is available, which allows to determine the maximum likelihood estimator (MLE) for the observed variable distribution $p_{\mathbf{x}\mathfrak{y}}$, $\mathbf{x} \in \Omega_X, \mathfrak{y} \in \Omega_{\mathcal{Y}}$ based on the $n = \sum_{\mathbf{x}\in\Omega_X} n_{\mathbf{x}}$ observations. Since for fixed covariate values $\mathbf{x} \in \Omega_X$, the cell counts $(n_{\mathbf{x}\mathfrak{y}})_{\mathfrak{y}\in\Omega_{\mathcal{Y}}}$ are multinomially distributed, the MLE for the observed variable distribution is uniquely obtained by the respective conditional relative frequency, i.e. $\hat{p}_{\mathbf{x}\mathfrak{y}} = \frac{n_{\mathbf{x}\mathfrak{y}}}{n_{\mathbf{x}}}$, $\mathbf{x} \in \Omega_X, \mathfrak{y} \in \Omega_{\mathcal{Y}}$, assuming that $n_{\mathbf{x}} > 0$.

Secondly, the information from the observation model relating the latent to the observed world is included. For this purpose, a mapping $\Phi : \gamma \mapsto \vartheta$, with $\gamma = (\pi_{\mathbf{x}y}, q_{\mathfrak{y}|\mathbf{x}y})_{\mathbf{x}\in\Omega_X,\mathfrak{y}\in\Omega_{\mathcal{Y}},y\in\Omega_Y}$ and $\vartheta = (p_{\mathbf{x}\mathfrak{y}})_{\mathbf{x}\in\Omega_X,\mathfrak{y}\in\Omega_{\mathcal{Y}}}$, is defined. This mapping describes the transfer between the parametrization in terms of the components of γ and the ones of ϑ by using the theorem of total probability. Consequently, the prescription of the reparametrization is given by

$$p_{\mathbf{x}\mathfrak{y}} = \sum_{y\in\mathfrak{y}} \pi_{\mathbf{x}y} \cdot q_{\mathfrak{y}|\mathbf{x}y} \ , \tag{1}$$

for all $\mathbf{x} \in \Omega_X, \ \mathfrak{y} \in \Omega_{\mathcal{Y}}$. Since we already calculated the MLE of ϑ and may express it as a function of the parameter of interest γ, i.e. $\vartheta = \Phi(\gamma)$, by virtue of the invariance of the likelihood we can thirdly determine the MLE of γ as the inverse image of $\hat{\vartheta}$ under the function Φ. Since the mapping Φ is generally not injective, there are several $\hat{\gamma}$, all leading to the same maximum value of the log-likelihood. Thus, we obtain the set-valued estimator

$$\hat{\Gamma} = \{\hat{\gamma} \mid \Phi(\hat{\gamma}) = \hat{\vartheta}\} \tag{2}$$

Table 3: Estimation of the parameters of the latent world (**Example 1**).

$\hat{\pi}_{\mathbf{x}<}$	$\hat{q}_{na\vert\mathbf{x}<}$	$\hat{q}_{na\vert\mathbf{x}\geq}$
$\hat{\pi}_{00<} \in [0.37,\ 0.76]$	$\hat{q}_{\text{na}\vert 00<} \in [0,\ 0.51]$	$\hat{q}_{\text{na}\vert 00\geq} \in [0,\ 0.62]$
$\hat{\pi}_{01<} \in [0.22,\ 0.63]$	$\hat{q}_{\text{na}\vert 01<} \in [0,\ 0.66]$	$\hat{q}_{\text{na}\vert 01\geq} \in [0,\ 0.53]$
$\hat{\pi}_{10<} \in [0.27,\ 0.60]$	$\hat{q}_{\text{na}\vert 10<} \in [0,\ 0.55]$	$\hat{q}_{\text{na}\vert 10\geq} \in [0,\ 0.49]$
$\hat{\pi}_{11<} \in [0.22,\ 0.55]$	$\hat{q}_{\text{na}\vert 11<} \in [0,\ 0.61]$	$\hat{q}_{\text{na}\vert 11\geq} \in [0,\ 0.43]$

by replacing the left hand side of (1) by the MLEs $\hat{p}_{\mathbf{x}\mathfrak{y}}$ of the observed world, already calculated in the first step, and the right hand side by the empirical analogues of the respective parameters.

Throughout the paper, instead of giving the set-valued estimator $\hat{\Gamma}$ in (2) itself, we illustrate it by building its one-dimensional projections. Thus, estimators for the single components of γ are obtained, here represented as

$$\hat{\pi}_{\mathbf{x}y} \in \left[\frac{n_{\mathbf{x}\{y\}}}{n_{\mathbf{x}}}, \frac{\sum_{\mathfrak{y}\ni y} n_{\mathbf{x}\mathfrak{y}}}{n_{\mathbf{x}}}\right], \quad \hat{q}_{\mathfrak{y}\vert\mathbf{x}y} \in \left[0, \frac{n_{\mathbf{x}\mathfrak{y}}}{n_{\mathbf{x}\{y\}} + n_{\mathbf{x}\mathfrak{y}}}\right], \tag{3}$$

for all $\mathbf{x} \in \Omega_X,\ y \in \Omega_Y$ and all $\mathfrak{y} \in \Omega_{\mathcal{Y}}$ such that $\{y\} \subsetneq \mathfrak{y}$, with $n_{\mathbf{x}} > 0$ and $\frac{0}{0} := 1$. It is important to keep in mind that points in these intervals are constrained by the restrictions in (1). The result in (3) can be shown to correspond to the one obtained from cautious data completion, plugging in all potential precise sample outcomes compatible with the observations [cf. 2, §7.8].

For sake of illustration, we apply this approach to **Example 1**, where four subgroups result from splitting by the different values of the two covariates, hence we consider $\mathbf{x} \in \Omega_X = \{\text{“00”}, \text{“01”}, \text{“10”}, \text{“11”}\}$ interpreted as “age=0, Abitur=0”, “Abitur=0, age=1”, “Abitur=1, age=0” and “Abitur=1, age=1”, respectively. Using Table 1 and referring to the data of the first subgroup, one uniquely obtains

$$\hat{p}_{00<} = \frac{n_{00<}}{n_{00}} = \frac{97}{262}, \quad \hat{p}_{00\geq} = \frac{n_{00\geq}}{n_{00}} = \frac{63}{262} \quad \text{and} \quad \hat{p}_{00na} = \frac{n_{00na}}{n_{00}} = \frac{102}{262},$$

with $n_{00} = n_{00<} + n_{00\geq} + n_{00na}$. There are indeed multiple $\hat{\gamma}$, i.e. estimated combinations of coarsening parameters and latent variable distributions, that are compatible with the restriction in (1) and thus lead to this estimated observed variable distribution. Different scenarios for the estimation of $\pi_{00<}$ are conceivable ranging from attributing all coarse categories “na” to “$\geq$” to including them all in category “$<$”,[5] thus obtaining (cf. (3))

$$\hat{\pi}_{00<} \in [\underline{\hat{\pi}}_{00<}, \overline{\hat{\pi}}_{00<}] \quad \text{with} \quad \underline{\hat{\pi}}_{00<} = \frac{97}{262} \approx 0.37 \quad \text{and} \quad \overline{\hat{\pi}}_{00<} = \frac{97+102}{262} \approx 0.76\,.$$

The resulting estimators (i.e. the one-dimensional projections of $\hat{\Gamma}$) in **Example 1** are shown in Table 3.

[47] presented an approach based on the profile likelihood to describe statistical evidence with missing data without imposing untestable assumptions, hence allowing for an alternative way to achieve the results in (3). Compared to the global log-likelihood $l(\cdot)$ in dependence of all parameters, the profile log-likelihood is a function of the parameter of interest only and arises from the global log-likelihood by considering all other parameters

[5] This is technically related to the Dempster-Shafer Theory [cf. 36].

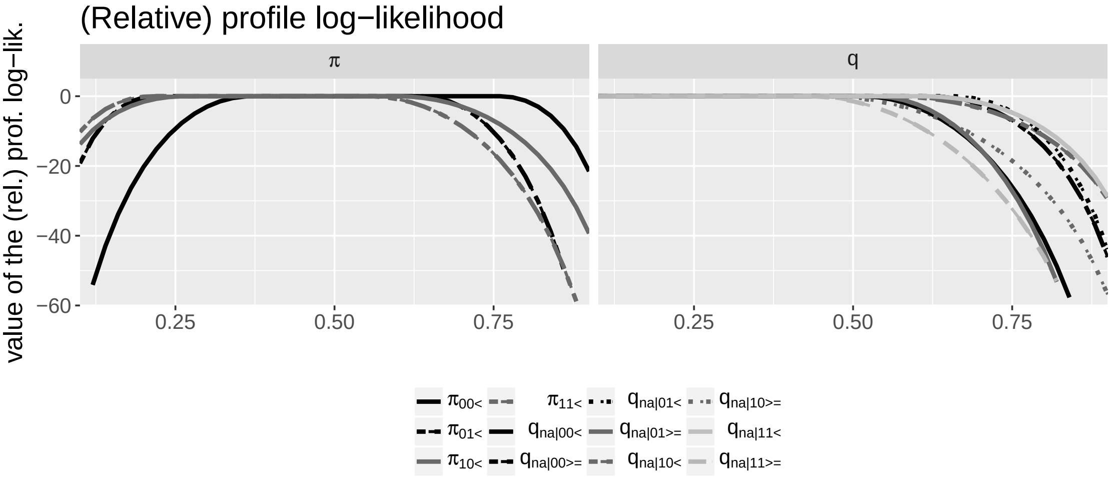


Figure 2: Referring to the the data of **Example 1**, the (relative) profile log-likelihood function for every parameter in γ is depicted.

as nuisance parameters [cf., e.g. 31, p. 80]. In our case, a specific parameter $\pi_{\mathbf{x}y}$ or $q_{\mathfrak{y}|\mathbf{x}y}$, $\mathbf{x} \in \Omega_X,\ y \in \Omega_Y,\ \mathfrak{y} \in \Omega_{\mathcal{Y}}$ might be of interest and the profile log-likelihood follows as

$$l(\pi_{\mathbf{x}y}) = \max_{\xi} l(\pi_{\mathbf{x}y}, \xi) \quad \text{or} \quad l(q_{\mathfrak{y}|\mathbf{x}y}) = \max_{\xi} l(q_{\mathfrak{y}|\mathbf{x}y}, \xi) \tag{4}$$

with nuisance parameters ξ corresponding to γ without $\pi_{\mathbf{x}y}$ and $q_{\mathfrak{y}|\mathbf{x}y}$, respectively. Thus, we can graphically represent the profile log-likelihood by varying the values of the parameter of interest on a grid and evaluating the log-likelihood at each fixed value for the parameter of interest and the nuisance parameters maximizing the log-likelihood in this case. Figure 2 shows the (relative) profile log-likelihood for **Example 1**, obtained by shifting the profile log-likelihood by the maximum value of the log-likelihood function along the y-axis. The range of the plateau characterizes the maximum likelihood estimator for the parameter of interest and hence is in accordance with the results in Table 3. The explicit formula for the profile log-likelihood for $\pi_{\mathbf{x}y}$ is given in [6].

3.1.2 Maximum likelihood estimators for the regression coefficients

Whenever a saturated model is used, the reparametrization in terms of the regression coefficients means no reduction of the dimension and the link function $g(\pi_{\mathbf{x}y})$ is bijective. Since it is also continuous, [cf., e.g. 11, p. 304], the bounds of the estimated regression coefficients can be calculated as a direct transformation of the bounds of the latent variable distribution (cf. Section 3.1.1).

To illustrate the procedure, we refer to the data situation of **Example 1**, where the logit model with the response function

$$\pi_{\mathbf{x}<} = P(Y = \text{“} < \text{”}\ |\mathbf{x}) = \frac{\exp(\beta_0 + d(\mathbf{x})^T \boldsymbol{\beta})}{1 + \exp(\beta_0 + d(\mathbf{x})^T \boldsymbol{\beta})} \tag{5}$$

for the category of interest, here “<”, and

$$\pi_{\mathbf{x}\geq} = \frac{1}{1 + \exp(\beta_0 + d(\mathbf{x})^T \boldsymbol{\beta})}, \tag{6}$$

Table 4: Regression estimates obtained without parametric assumptions (Example 1). interactions: $\hat{\beta}_{12} \in [-2.76,\ 3.64]$ (cautious estimation), $\hat{\beta}_{12} = 0.63$ (traditional)

cautious estimation	$\hat{\beta}_0 \in [-0.53,\ 1.15]$	$\hat{\beta}_1 \in [-2.16,\ 0.92]$	$\hat{\beta}_2 \in [-2.42,\ 1.08]$
traditional procedure	$\hat{\beta}_0 = 0.43$	$\hat{\beta}_1 = -0.85$	$\hat{\beta}_2 = -0.94$

for the reference category, here "$\geq$", is appropriate. Equivalently, the logit model can be described by the link function

$$g(\pi_{\mathbf{x}<}) = \ln\Big(\frac{\pi_{\mathbf{x}<}}{1-\pi_{\mathbf{x}<}}\Big) = \beta_0 + d(\mathbf{x})^T\boldsymbol{\beta}\ . \qquad (7)$$

Considering a saturated model, we specify the linear predictor as $\beta_0 + \beta_1 \cdot \text{Abitur} + \beta_2 \cdot \text{age} + \beta_{12} \cdot \text{age*Abitur}$. The bounds of the four regression coefficients are then determined by transforming the bounds of the four estimators $\hat{\pi}_{00<}$, $\hat{\pi}_{01<}$, $\hat{\pi}_{10<}$ and $\hat{\pi}_{11<}$, hence obtaining

$$\begin{aligned}
\hat{\beta}_0 &\in \Big[\ln\Big(\frac{\hat{\underline{\pi}}_{00<}}{1-\hat{\underline{\pi}}_{00<}}\Big),\ \ln\Big(\frac{\overline{\hat{\pi}}_{00<}}{1-\overline{\hat{\pi}}_{00<}}\Big)\Big]\ , \qquad (8)\\
\hat{\beta}_1 &\in \Big[\ln\Big(\frac{\hat{\underline{\pi}}_{10<}}{1-\hat{\underline{\pi}}_{10<}}\Big) - \overline{\hat{\beta}}_0,\ \ln\Big(\frac{\overline{\hat{\pi}}_{10<}}{1-\overline{\hat{\pi}}_{10<}}\Big) - \underline{\hat{\beta}}_0\Big]\\
\hat{\beta}_2 &\in \Big[\ln\Big(\frac{\hat{\underline{\pi}}_{01<}}{1-\hat{\underline{\pi}}_{01<}}\Big) - \overline{\hat{\beta}}_0,\ \ln\Big(\frac{\overline{\hat{\pi}}_{01<}}{1-\overline{\hat{\pi}}_{01<}}\Big) - \underline{\hat{\beta}}_0\Big]\ ,\\
\hat{\beta}_{12} &\in \Big[\ln\Big(\frac{\hat{\underline{\pi}}_{11<}}{1-\hat{\underline{\pi}}_{11<}}\Big) - \overline{\hat{\beta}}_1 - \overline{\hat{\beta}}_2 - \underline{\hat{\beta}}_0,\ \ln\Big(\frac{\overline{\hat{\pi}}_{11<}}{1-\overline{\hat{\pi}}_{11<}}\Big) - \underline{\hat{\beta}}_1 - \underline{\hat{\beta}}_2 - \overline{\hat{\beta}}_0\Big]\ .
\end{aligned}$$

For **Example 1** the cautious regression estimates are given in Table 4, where they can also be compared to the results from a traditional procedure[6] assuming uninformative coarsening (in the sense of coarsening at random; more details follow in Section 5). Although the estimates from the traditional procedure are generally included in the result from the cautious estimation, they do not express the lack of knowledge about the coarsening mechanism, also pretending specific signs.

3.2 The non-saturated model

We now study non-saturated regression models, where parametric assumptions are included in the regression model in the sense that certain interactions are set equal to zero. In this way, the number of parameters that have to be estimated is reduced and the regression coefficients are generally no longer able to reproduce the latent variable distribution. We focus on the setting with the binary response variable of **Example 1**, thus choosing the response function in (5) and (6) and link function in (7) again, but now the vector of regression coefficients does not contain any interactions, i.e. $\beta_{12} = 0$. In Section 4, we discuss to which extent the obtained results can be transferred to coarse data, not reducing to the missing data problem.

In this here considered setting, we now present both methods to determine cautious regression estimators, which were already briefly announced in Section 2.3: At first, we turn to the two-step method, which allows for a direct comparison to the procedure and results of the saturated model, and hence we can investigate the impact of the parametric

[6] First of all, we calculated the estimated latent variable distribution under coarsening at random [cf., e.g. 33, Equation (10)] and then transformed it via (8).

assumption on the regression model. Furthermore, this way gives a first insight into the type of possible situations that have to be distinguished, also including cases where the two-step method is unrewarding. For that reason, we subsequently also present the direct method. The general roles and advantages of the two methods are only then discussed in Section 4.

Due to the inclusion of parametric assumptions on the regression model, we can no longer rely on a bijective link function, justifying the direct transformation of the bounds of the latent variable distribution (cf. (8)). Nevertheless, a two-step procedure can still be useful, firstly estimating the latent variable distribution (cf. Section 3.1.1), thus applying (3) to e.g. obtain $\hat{\underline{\pi}}_{00<}$ and $\hat{\overline{\pi}}_{00<}$, and secondly trying to minimize/maximize the regression parameters under the condition that this estimated latent variable distribution (cf. Section 3.1.1) can be produced. This leads us to the following optimization problem, here referring to $\pi_{\mathbf{x}<} = h(\beta_0 + \beta_1 \cdot \text{Abitur} + \beta_2 \cdot \text{age})$ of **Example 1** with the response function in (5) and (6) and presented for the determination of the bounds of the effect of Abitur, i.e. $\underline{\beta}_1$ and $\overline{\beta}_1$:

$$\begin{aligned}
&\beta_1 \to \text{min/max} \quad \text{given} && \qquad\qquad (9)\\
&\hat{\underline{\pi}}_{00<} \le \frac{\exp(\beta_0)}{1+\exp(\beta_0)} \le \hat{\overline{\pi}}_{00<}, && \hat{\underline{\pi}}_{10<} \le \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} \le \hat{\overline{\pi}}_{10<},\\
&\hat{\underline{\pi}}_{01<} \le \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)} \le \hat{\overline{\pi}}_{01<}, && \hat{\underline{\pi}}_{11<} \le \frac{\exp(\beta_0+\beta_1+\beta_2)}{1+\exp(\beta_0+\beta_1+\beta_2)} \le \hat{\overline{\pi}}_{11<}\ .
\end{aligned}$$

In fact, in more general cases it is not sufficient to include inequalities for the bounds of the estimated latent variable distribution only. This and related consequences will be discussed in Section 4. By using the link function in (7), this optimization problem can be transformed into one with linear constraints:

$$\begin{aligned}
&\beta_1 \to \text{min/max} \quad \text{given} && \qquad\qquad (10)\\
&\ln\Big(\frac{\hat{\underline{\pi}}_{00<}}{1-\hat{\underline{\pi}}_{00<}}\Big) \le \beta_0 \le \ln\Big(\frac{\hat{\overline{\pi}}_{00<}}{1-\hat{\overline{\pi}}_{00<}}\Big), && \ln\Big(\frac{\hat{\underline{\pi}}_{10<}}{1-\hat{\underline{\pi}}_{10<}}\Big) \le \beta_0+\beta_1 \le \ln\Big(\frac{\hat{\overline{\pi}}_{10<}}{1-\hat{\overline{\pi}}_{10<}}\Big),\\
&\ln\Big(\frac{\hat{\underline{\pi}}_{01<}}{1-\hat{\underline{\pi}}_{01<}}\Big) \le \beta_0+\beta_2 \le \ln\Big(\frac{\hat{\overline{\pi}}_{01<}}{1-\hat{\overline{\pi}}_{01<}}\Big), && \ln\Big(\frac{\hat{\underline{\pi}}_{11<}}{1-\hat{\underline{\pi}}_{11<}}\Big) \le \beta_0+\beta_1+\beta_2 \le \ln\Big(\frac{\hat{\overline{\pi}}_{11<}}{1-\hat{\overline{\pi}}_{11<}}\Big)\ .
\end{aligned}$$

Considering optimization problems as in (9) or (10) with the objective function chosen as the respective regression coefficient of interest, the following types of results have to be distinguished, where $\hat{\underline{\pi}}_{\mathbf{x}y}$ and $\hat{\overline{\pi}}_{\mathbf{x}y}$ represent the estimated bounds obtained without parametric assumptions on the regression model (cf. Section 3.1.1), while $\hat{\underline{\pi}}^*_{\mathbf{x}y}$ and $\hat{\overline{\pi}}^*_{\mathbf{x}y}$ denote the bounds achievable under the parametric assumptions[7]:

1. There is a solution.

   (a) Regression estimators are obtainable that are able to produce the estimated bounds of the latent variable distribution calculated without parametric assumptions (i.e. $\hat{\pi}^*_{\mathbf{x}y} \in [\hat{\underline{\pi}}_{\mathbf{x}y},\ \hat{\overline{\pi}}_{\mathbf{x}y}]$).

[7] For instance, the bounds $\hat{\underline{\pi}}^*_{10<}$ and $\hat{\overline{\pi}}^*_{10<}$ are determined by choosing $\beta_0 + \beta_1$ as objective function in the optimization problem (10). Generally, we use the superscript "*" only when we explicitly want to distinguish the respective parameter/estimator from the one without parametric assumptions on the regression model.

Table 5: Regression estimates with parametric assumptions (Example 1).

cautious estimation	$\hat{\beta}_0 \in [-0.53,\ 1.15]$	$\hat{\beta}_1 \in [-1.84,\ 0.92]$	$\hat{\beta}_2 \in [-1.68,\ 1.08]$
traditional procedure	$\hat{\beta}_0 = 0.35$	$\hat{\beta}_1 = 0.05$	$\hat{\beta}_2 = 0.00$

(b) The resulting regression estimators can only represent tighter bounds of the estimated latent variable distribution (i.e. $\hat{\pi}^*_{\mathbf{x}y} \in [\underline{\hat{\pi}}^*_{\mathbf{x}y},\ \overline{\hat{\pi}}^*_{\mathbf{x}y}]$ with $\underline{\hat{\pi}}^*_{\mathbf{x}y} > \underline{\hat{\pi}}_{\mathbf{x}y}$ and/or $\overline{\hat{\pi}}^*_{\mathbf{x}y} < \overline{\hat{\pi}}_{\mathbf{x}y}$), hence the inequalities are not satisfied with equality.

2. There is no solution.[8]

By rearranging the system of inequalities in (10), we can derive the following necessary and sufficient condition for the existence of a solution of the linear optimization problem (situation 1):

$$\begin{aligned}
\ln\left(\frac{\underline{\hat{\pi}}_{11<}}{1-\underline{\hat{\pi}}_{11<}}\right) + \ln\left(\frac{\underline{\hat{\pi}}_{00<}}{1-\underline{\hat{\pi}}_{00<}}\right) &\leq \ln\left(\frac{\overline{\hat{\pi}}_{10<}}{1-\overline{\hat{\pi}}_{10<}}\right) + \ln\left(\frac{\overline{\hat{\pi}}_{01<}}{1-\overline{\hat{\pi}}_{01<}}\right) \qquad (11)\\
\ln\left(\frac{\underline{\hat{\pi}}_{10<}}{1-\underline{\hat{\pi}}_{10<}}\right) + \ln\left(\frac{\underline{\hat{\pi}}_{01<}}{1-\underline{\hat{\pi}}_{01<}}\right) &\leq \ln\left(\frac{\overline{\hat{\pi}}_{11<}}{1-\overline{\hat{\pi}}_{11<}}\right) + \ln\left(\frac{\overline{\hat{\pi}}_{00<}}{1-\overline{\hat{\pi}}_{00<}}\right)
\end{aligned}$$

It turns out that **Example 1** is classified to situation 1, and more specifically to 1(a); the corresponding results for cautious regression estimates are given in Table 5. Again, we can conclude that results from a traditional procedure (assuming coarsening at random) have to be treated with caution: While this approach would suggest no effect of age, avoiding specific coarsening assumptions could also indicate a negative or positive effect. Comparing the results with the ones from Table 4 gives some indication about the impact of the parametric assumption on the regression model.

In the saturated model, the regression estimators are obtainable by a simple transformation (cf. Section 3.1), hence they reveal the same information as the estimated parameters determining the latent variable distribution. This is not the case in the non-saturated model, where further restrictions are included induced by the loss of flexibility from the lack of several interactions. Consequently, under parametric assumptions on the regression model tighter bounds for the regression estimators may result, but they are never wider. In this way, there is a synergy of the uncertainty associated to the coarse data problem and the one due to the parametric assumption on the regression model, which we study next for situation 1 by comparing the estimation of the coarsening parameter from the saturated model to the one from the non-saturated model.

For this purpose, we can exploit the relation between the parameters of the observed and the latent world expressed by (1). When the optimization problem in (10) is solvable (i.e. in situation 1), then the estimators of the latent variable distribution fit to the data in the sense that the estimators for the parameters of the observed world, i.e. $\hat{p}_{\mathbf{x}\boldsymbol{\mathfrak{y}}}$, $\mathbf{x} \in \Omega_X$, $\boldsymbol{\mathfrak{y}} \in \Omega_{\mathcal{Y}}$, are unaffected by the parametric assumptions and are still calculated by the (conditional) relative frequency (cf. Section 3.1.1). Since in situation 1(a) also the estimated bounds of the latent variable distribution coincide with the ones obtained without parametric assumptions, the estimated bounds of the coarsening parameters remain unchanged by the parametric assumption as well. Thus, for **Example 1**, $\underline{\hat{q}}_{\boldsymbol{\mathfrak{y}}|\mathbf{x}y}$ and $\overline{\hat{q}}_{\boldsymbol{\mathfrak{y}}|\mathbf{x}y}$

[8] In general, corresponding optimization problems are not solvable in the precise case either: Here, the parametric assumption on the regression model is too strong and hence prevents that the estimated response probability can be reproduced by means of the regression estimators.

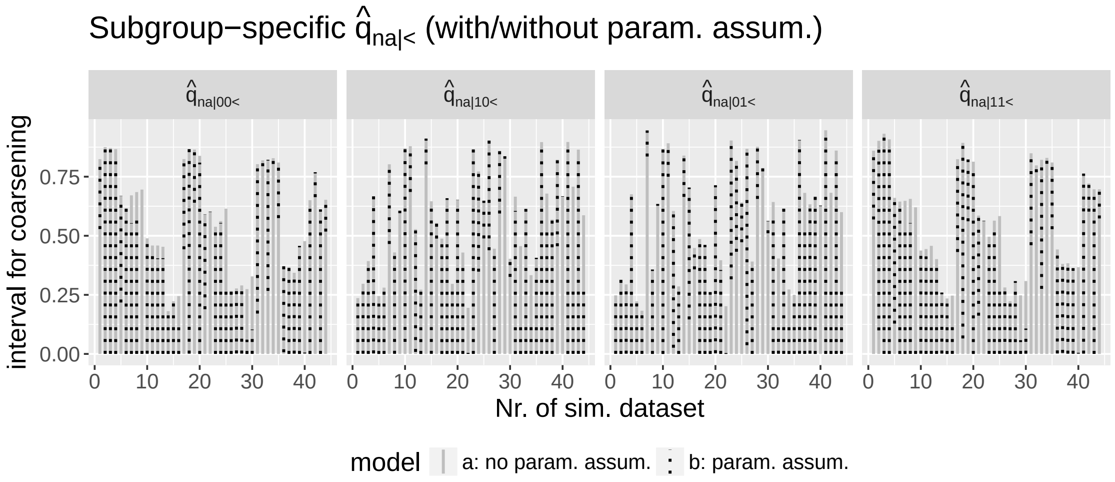


Figure 3: We restrict ourselves to data situations 1(b): In some cases the parametric assumption on the regression model induces a noticeable reduction of the coarsening intervals, while in others that are close to situation 1(a) the refinement is hardly recognizable.

can still be inferred from Table 3, even if parametric assumptions are included. In situation 1(b), by applying the relation in (1) for the binary case and solving for the coarsening parameters the following estimated bounds are achievable:

$$\hat{q}_{na|\mathbf{x}<} \in \left[1-\frac{\hat{p}_{\mathbf{x}<}}{\hat{\underline{\pi}}^*_{\mathbf{x}<}},\ \frac{\hat{\overline{\pi}}^*_{\mathbf{x}<}-\hat{p}_{\mathbf{x}<}}{\hat{\overline{\pi}}^*_{\mathbf{x}<}}\right] \quad \text{and} \quad \hat{q}_{na|\mathbf{x}\geq} \in \left[1-\frac{\hat{p}_{\mathbf{x}\geq}}{\hat{\underline{\pi}}^*_{\mathbf{x}\geq}},\ \frac{\hat{\overline{\pi}}^*_{\mathbf{x}\geq}-\hat{p}_{\mathbf{x}\geq}}{\hat{\overline{\pi}}^*_{\mathbf{x}\geq}}\right], \tag{12}$$

with $\frac{0}{0} := 1$. Whenever $\hat{\underline{\pi}}^*_{\mathbf{x}<} = \hat{\underline{\pi}}_{\mathbf{x}<}$, then $\hat{\underline{\pi}}^*_{\mathbf{x}<} = \hat{p}_{\mathbf{x}<}$ is valid, such that the lower bound of $\hat{q}_{na|\mathbf{x}<}$ stays zero and is thus not refined (while analogous conclusions can be made for the lower bound of $\hat{q}_{na|\mathbf{x}\geq}$). Due to $\hat{\underline{\pi}}^*_{\mathbf{x}y} \geq \hat{\underline{\pi}}_{\mathbf{x}y}$ and/or $\hat{\overline{\pi}}^*_{\mathbf{x}y} \leq \hat{\overline{\pi}}_{\mathbf{x}y}$, the bounds in (12) are generally not wider than those received without parametric assumptions. This is in line with the tenor in [17], who holds the view that model selection and the "disambiguation" of the incomplete data should go "hand in hand" in the sense that precise values that are consistent with the observation, but appear to be implausible under the model assumption, should no longer be under consideration. However, on the other hand from taking the model assumptions seriously several difficulties may occur, as the problem of possible ill-conditioning of the obtained set-valued estimators under such strong parametric assumptions, shortly discussed for the case of linear regression in Schollmeyer and Augustin [35, Section 6.1 and Appendix A therein].

We further investigate how the parametric assumption on the regression model may affect the estimated coarsening parameters in situation 1(b) by simulating different data situations, arising from assuming the same marginal distribution for the covariates as in **Example 1** and then varying the parameters of the observed variable distribution on a grid of values. Figure 3 shows the development of the intervals for the estimated coarsening $\hat{q}_{na|\mathbf{x}<}$ under the parametric assumption for those datasets that are classified into situation 1(b). As a by-product of this simulation study, we gain a first insight about the frequency of the different situations: From the 100 data sets we considered, 35 were classified into situation 1(a), 44 into situation 1(b) and 21 into situation 2. This already indicates that the number of cases where the optimization problem is not solvable is not

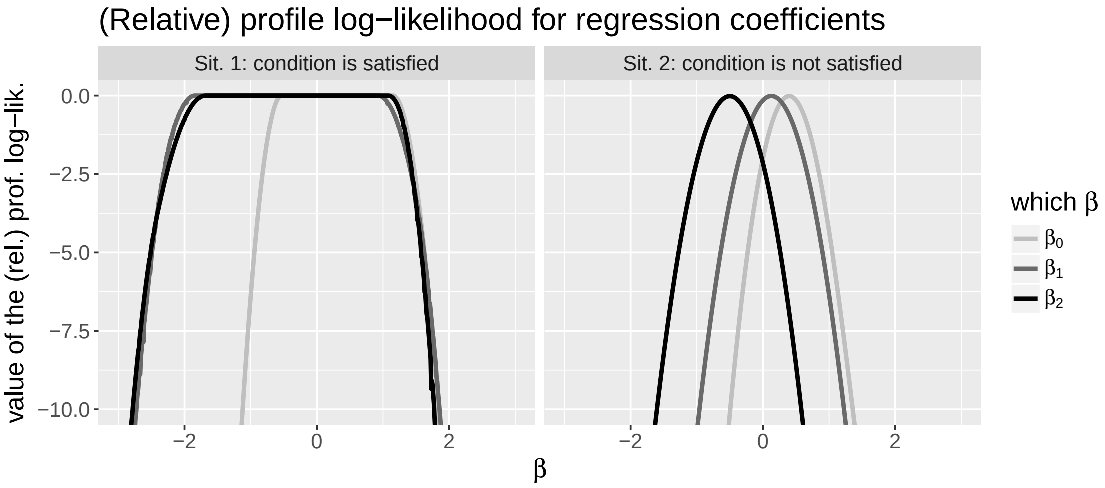


Figure 4: The left-hand part refers to the data situation of **Example 1** classified into Situation 1(a). An arbitrary data situation where the condition is not satisfied is underlying the right-hand part.

negligible, which leads us to continue with investigating the direct approach next.

We consider the (relative) profile log-likelihood again, now not in dependence of a specific $\pi_{\mathbf{x}y}$ (cf. (4)), but of the regression coefficient of interest. The global log-likelihood $l(\beta_0, \beta_1, \beta_2, q_{na|00<}, q_{na|00\geq}, \dots, q_{na|11\geq})$ is obtained from the one depending on the parameters determining the latent variable distribution and the coarsening parameters by replacing all parameters $\pi_{\mathbf{x}y}$ by the chosen response function. The profile log-likelihood function of e.g. β_1 is then given by

$$l(\beta_1) = \max_{\xi} l(\beta_1, \xi) \, , \tag{13}$$

taking β_0, β_2, and all coarsening parameters as nuisance parameters ξ.
Figure 4 gives the (relative) profile log-likelihood for two data situations, one corresponds to the one in **Example 1** and is thus in accordance with the condition in (11), while the other is not ($n_{00<} = 60$, $n_{00\geq} = 10$, $n_{00na} = 10$, $n_{10<} = 30$, $n_{10\geq} = 40$, $n_{10na} = 5$, $n_{01<} = 20$, $n_{01\geq} = 50$, $n_{01na} = 2, n_{11<} = 40$, $n_{11\geq} = 10$, $n_{11na} = 5$). The ranges of the plateaus within the left plot corroborate the respective intervals for the regression estimators presented in Table 5.[9] It appears that precise maximum likelihood estimators are obtained, when the condition is not satisfied, while otherwise imprecision is still inherent (cf. Figure 3).
This systematic difference with regard to the nature of the result (imprecise versus precise results in situation 1 and 2, respectively) represents a particularity ascribable to the interaction of the parametric assumption on the regression model and the coarse data problem: While the parametric assumption on the regression model generally brings us into situation 2, whenever all data are precisely observed, the availability of coarse data and the associated flexibility due to the variety of possible underlying precise data scenarios can allow to “repair” the incompatibility with the observed data. This gives us the opportunity not only to assess whether the observed data fit to the model assumptions, but also to

[9] This is invisible to the naked eye, but the results from numerical optimization are quite exact.

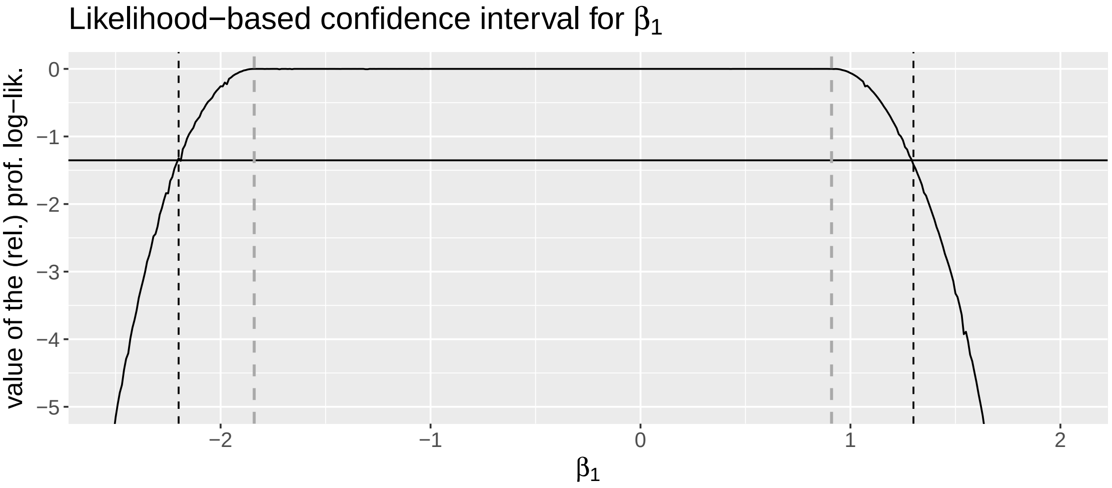


Figure 5: While the δ-cut is symbolized by the solid line, the black dashed lines mark the bounds of the confidence interval, here with $\alpha = 0.1$. The extent of the sampling uncertainty is visible by comparing these bounds with the bounds of the maximum-likelihood estimator characterized by the gray lines.

Table 6: Likelihood-based confidence intervals for the regression coefficients (**Example 1**).

for β_0 : $[-0.75,\ 1.40]$	for β_1 : $[-2.20,\ 1.29]$	for β_2 : $[-2.11,\ 1.35]$

actively decide about the inclusion of additional coarsening or model assumptions, when the solvability of the optimization problem represents our claim.

3.3 Likelihood-based confidence intervals

Taking the cautious analysis seriously, the recognition of the sampling error induced by the absence of an infinite sample is crucial. There have already been several proposals to attach value to both sources of uncertainty and confidence intervals for the latent variable distribution have been constructed [also cf. 18, 16, 37, 43]. To give confidence intervals for the regression parameters, we can tie on one of the here presented methods and either rely on a two-step method by reparametrizing the confidence intervals for the latent variable distribution via the relation formalized by the link function or base our considerations on the profile-log likelihood, where we here decide for the second option. These likelihood-based confidence intervals are appealing due to their (compared to Wald intervals) better performance in case of a small sample size [cf. e.g. 27].

Generally, likelihood-based confidence intervals are constructed by cutting the (relative) profile (log-)likelihood function at level δ with $\delta = (-0.5\chi^2_{1,1-\alpha})$ [cf., e.g. 44]. The confidence interval is then specified by regarding all parameters of interest whose value of the profile likelihood is larger than the value of δ. Likelihood-based confidence intervals in the presence of coarse data are already studied for $\pi_{\mathbf{x}y}$, $\mathbf{x} \in \Omega_X,\ y \in \Omega_Y$, relying on the profile likelihood presented in Section 3.1 [cf. 6, 47]. By referring to the (relative) profile (log-)likelihood for the regression coefficients, we can analogously proceed and define asymptotic $(1-\alpha)$ confidence intervals by using these δ-cuts.

In Figure 5, we exemplify the construction of likelihood-based confidence intervals for the Abitur effect β_1 by using the data in **Example 1**. The result with regard to the other coefficients can be inferred from Table 6. By comparing these intervals with the ones in Table 5 an impression about the magnitude of the sampling uncertainty can be gained.

4 Studying the data application with coarse data in the strict sense (Example 2)

Since we up to now focused on a setting reducing to the missing data situation, a discussion from a more general viewpoint and an illustrative study of a situation with coarse data as present in **Example 2** is of interest. In the saturated model, the cautious regression estimators can generally be determined by a two-step procedure that gives us the cautious regression estimators by transforming the bounds of the latent variable distributions in a direct and easy way. In the non-saturated model, the preferable method (cf. Section 2.3) is not that clear. Thus, we now address the advantages and limitations of both ways, throughout turning to a non-saturated model.
To account for the ordinal structure of the response variable in **Example 2**, we base our analysis on the cumulative logit model [cf., e.g. 11, p. 334–337]. This model is based on the notion that the ordinal response categories are received due to the impossibility to collect the values of a latent continuous variable $\tilde{Y}$, thus introducing a second layer of latency. For this variable a regression model $\tilde{Y} = -d(\mathbf{x})^T\boldsymbol{\beta} + \epsilon$ with $\epsilon \sim F$ is assumed, where F is the logistic distribution function. The connection to our categorical variable of interest Y is given by $Y = y^{(l)} \iff \beta_{0y^{(l-1)}} < \tilde{Y} \leq \beta_{0y^{(l)}}$, $l = 1, \ldots, m$, where $y^{(l)}$ is the lth category within the ordered categories $y^{(1)}, \ldots, y^{(l)}, \ldots, y^{(m)}$, and $-\infty = \beta_{0y^{(0)}} < \beta_{0y^{(1)}} < \cdots < \beta_{0y^{(m)}} = \infty$. In this way, the intercepts are increasing with the order of the respective category. While the intercepts are category-specific, the regression coefficients $\boldsymbol{\beta}$ are not in this model, also referred to as proportional-odds assumption. The ordinal structure is included by basing the analysis on the cumulative probabilities describing the distribution function $F(\cdot)$, hence considering the response function

$$P(Y \leq y^{(l)} \ |\mathbf{x}) = F(\beta_{0y^{(l)}} + d(\mathbf{x})^T\boldsymbol{\beta}), \quad \text{with} \tag{14}$$

$$F(\beta_{0y^{(l)}} + d(\mathbf{x})^T\boldsymbol{\beta}) = \frac{\exp(\beta_{0y^{(l)}} + d(\mathbf{x})^T\boldsymbol{\beta})}{1 + \exp(\beta_{0y^{(l)}} + d(\mathbf{x})^T\boldsymbol{\beta})}, \quad \text{and with}$$

$$\pi_{\mathbf{x}y^{(l)}} = \ P(Y = y^{(l)} \ |\mathbf{x}) = F(\beta_{0y^{(l)}} + d(\mathbf{x})^T\boldsymbol{\beta}) - F(\beta_{0y^{(l-1)}} + d(\mathbf{x})^T\boldsymbol{\beta}), \quad l = 1, \ldots, m,$$

[cf., e.g. 11, p. 335].

In the context of **Example 1** we already noticed that the proposed two-step method is unrewarding, whenever we are in situation 2. Now, we will additionally find that even when we are in situation 1, this procedure not necessarily simplifies the calculation as it did in **Example 1**. For given values of the covariates $\mathbf{x} \in \Omega_X$, the optimization problem considered in connection with **Example 1** only included estimated bounds for one parameter, i.e. only for $\pi_{\mathbf{x}<}$. Since a given $\hat{\pi}_{\mathbf{x}<}$, $\mathbf{x} \in \Omega_X$, refers to a specific precise scenario uniquely determining the compatible coarsening estimators $\hat{q}_{na|\mathbf{x}<}$ and $\hat{q}_{na|\mathbf{x}<}$, in situation 1 we can be sure that the cautious regression estimators obtained by the two-step method (cf. (9) and (10)) indeed maximize the respective profile log-likelihood. To fully determine the distribution in generalized probability theory, it is not sufficient to have the probability assessments on each elementary event only, but knowledge for all subsets is needed [cf., e.g., 36]. Thus, relying on the cumulative logit model, we have to include inequalities for each subset Q of Ω_Y, where the lower and upper bounds (of the confidence in Q in a given group $\mathbf{x} \in \Omega_X$) can (again) be calculated by the estimated belief and plausibility of Q, respectively. [10] While the theory behind is out of the scope of this paper, a quick look at **Example 2**, where this leads to $2^7 \cdot 4 \cdot 2 + 5 = 1029$ inequalities[11], already clarifies that a way through the optimization problem may no longer simplify the calculation.[12] Additionally, it is not possible anymore to transform the obtained constraints, such as

$$\begin{aligned}
\hat{\underline{\pi}}_{00c} \leq \quad & \frac{\exp(\beta_{0c})}{1+\exp(\beta_{0c})} - \frac{\exp(\beta_{0b})}{1+\exp(\beta_{0b})} - \frac{\exp(\beta_{0a})}{1+\exp(\beta_{0a})} \quad \leq \hat{\overline{\pi}}_{00c} \\
\hat{\underline{\pi}}_{10c} \leq \quad & \frac{\exp(\beta_{0c}+\beta_1)}{1+\exp(\beta_{0c}+\beta_1)} - \frac{\exp(\beta_{0b}+\beta_1)}{1+\exp(\beta_{0b}+\beta_1)} - \frac{\exp(\beta_{0a}+\beta_1)}{1+\exp(\beta_{0a}+\beta_1)} \quad \leq \hat{\overline{\pi}}_{10c} \\
\hat{\underline{\pi}}_{01c} \leq \quad & \frac{\exp(\beta_{0c}+\beta_2)}{1+\exp(\beta_{0c}+\beta_2)} - \frac{\exp(\beta_{0b}+\beta_2)}{1+\exp(\beta_{0b}+\beta_2)} - \frac{\exp(\beta_{0a}+\beta_2)}{1+\exp(\beta_{0a}+\beta_2)} \quad \leq \hat{\overline{\pi}}_{01c} \\
\hat{\underline{\pi}}_{11c} \leq \quad & \frac{\exp(\beta_{0c}+\beta_1+\beta_2)}{1+\exp(\beta_{0c}+\beta_1+\beta_2)} - \frac{\exp(\beta_{0b}+\beta_1+\beta_2)}{1+\exp(\beta_{0b}+\beta_1+\beta_2)} - \\
& \frac{\exp(\beta_{0a}+\beta_1+\beta_2)}{1+\exp(\beta_{0a}+\beta_1+\beta_2)} \quad \leq \hat{\overline{\pi}}_{11c}
\end{aligned}$$

when choosing Q to be "c" as example[13] (cf. (14)), into linear ones, further preventing a facilitation of computation.

Next, we turn to the direct method. The log-likelihood for the regression coefficients can again be written down by relying on the log-likelihood $l(\pi_{00a}, \ldots, q_{\{abcdefg\}|g})$ and replacing the latent variable distribution by the respective connection to the regression coefficients, for the cumulative model given by the response function in (14). In Figure 6 the (smoothed) (relative) profile log-likelihood functions for all regression parameters are depicted, where we here refer to one possible data scenario that is compatible with the data in Table 2.[14] From a substance matter view this is sufficient, since the results from all data scenarios closely resemble each other. The maximum likelihood estimators for the regression coefficients are again received by considering the maxima/maximum of the respective function. Due to numerical problems that occur in the optimization we can again not be sure about the kind of results, i.e. whether the optimum is indeed unique – as Figure 6 suggests – or not. Solving these computational challenges should be part of further research.

[10] In this way, we calculate the estimated belief of a specific Q, by including all respondents that report categories in $\mathcal{P}(\Omega_Y)$ that support Q for sure and hence are fully contained within Q, while the estimated plausibility accounts for all respondents giving answers that possibly support and thus intersect Q [cf. 36, 7]. This only extends the special case of **Example 1**, where only singletons Q were considered, but the calculation of the lower and upper bound corresponded to the estimated belief and plausibility (of query set "<" in the respective group $\mathbf{x} \in \Omega$), cf. Footnote 5.

[11] Cross-classifying the two covariates gives us four groups ("00", "10", "01" and "11") and hence we consider four inequalities (as in **Example 1**) for every of the 2^7 subsets of Y. Additionally we obtain inequalities for the lower and upper bounds, respectively and five further inequalities are given by $\beta_{0a} < \beta_{0b} < \cdots < \beta_{0f}$ induced by the cumulative logit model.

[12] Even if some constraints may be eliminated – theory [cf., e.g. 36] tells us that we e.g. do not need inequalities for the empty set and the ones for a set and its complement are equivalent – a high number of inequalities remains.

[13] This can be similarly written down for the other subsets Q.

[14] We attribute a higher selection probability to scenarios that are similar to the true one.

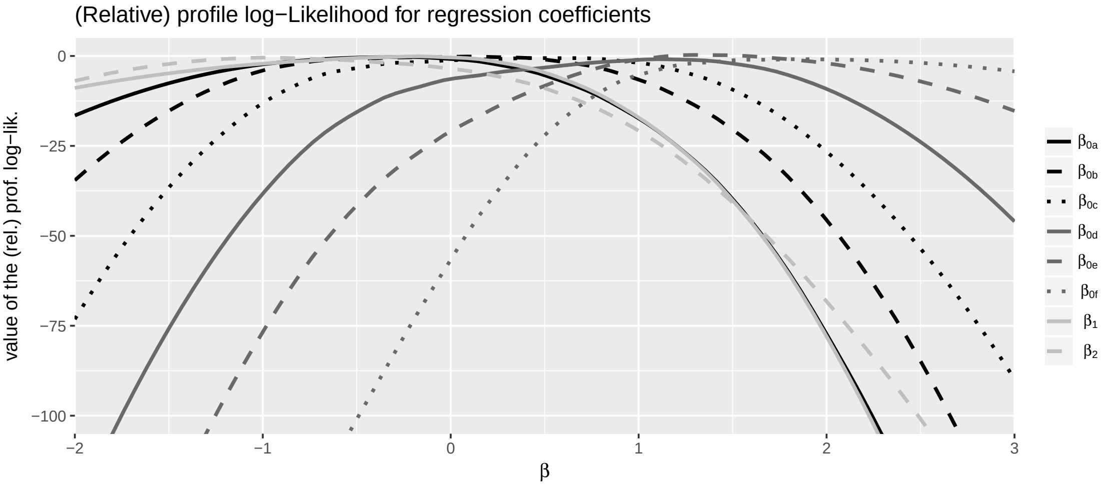


Figure 6: Relying on the data in Table 2, for all regression coefficients the respective profile-likelihood is shown.

To sum up, whenever a saturated model is of interest, basing considerations on a two-step method gives us direct formulas to calculate the cautious regression estimates. Referring to non-saturated models, the (nonparametrich) latent variable distribution and the regression coefficients do not bear the same information anymore; however, we could indeed find a way to rely on a two-step method. Although the two-step method of that kind showed to be helpful to investigate the role of the parametric assumption on the regression model in the "disambiguation" of the coarse data, it should and can be applied only in particular situations: In a setting with a binary response variable (as in **Example 1**), the two-step method turned out to be very simple – in the sense that we obtain a manageable number of linear constraints. However, when we are in situation 2 (for setting of **Example 1**, we could derive a proper criterion), we have to draw on the direct method also in these simple cases. Depending on the setting and the chosen response function, the direct method may lead to technical difficulties (as already met in context of **Example 2**), here left as an open problem.

5 Incorporation of auxiliary information

Although results obtained from a cautious analysis as described in Section 3 and Section 4 at a first glance may be regarded as practically unappealing due to an unsatisfactory information content, one should generally avoid conjuring information just to force an ability to act. However, there are frequently situations where some tenable auxiliary information about the incompleteness is obtainable, refining the results in the spirit of partial identification and sensitivity analysis [e.g. 21, 23]. For the missing-data problem, literature already reveals some possibilities to incorporate (partial) knowledge, mostly by restricting either the distribution of the incompleteness or the response propensities [e.g. 24]. By formulating constraints on $q_{\mathfrak{Y}|\mathbf{x}y}$, we concentrate on the first option in the context of coarse data. For this purpose, we start by considering two specific, quite strict, assumptions: Coarsening at random (CAR) and subgroup independence (SI). Afterwards, we look at generalizations to have a medium to include also other kind of knowledge, including weak knowledge about the coarsening process.

Heitjan and Rubin [13] introduced the concept of CAR, which requires constant coarsening probabilities $q_{\mathfrak{y}|y}$ regardless of the true underlying value y as long as it matches with the fixed observed value $\mathfrak{y}$. Adapting this assumption for our contingency table framework, the requirement has to be valid for all subgroups split by the considered covariates. An alternative type of coarsening is characterized by the independence from the corresponding covariate values. In [33] we called this assumption subgroup independence (SI) and studied it in more detail in the setting considered there.
Nordheim [29] suggests a possibility to generalize the MAR assumption by including the ratio between missing mechanisms into the analysis of non-randomly missing and misclassified data. In [32] we applied this idea by making assumptions about the coarsening ratios

$$R_{\mathbf{x},y,y',\mathfrak{y}} = \frac{q_{\mathfrak{y}|\mathbf{x}y}}{q_{\mathfrak{y}|\mathbf{x}y'}}, \quad \mathfrak{y} \in \Omega_{\mathcal{Y}},\ y,\ y' \in \mathfrak{y},\ \mathbf{x} \in \Omega_X \ , \tag{15}$$

defined for all pairs of directly successive categories y and y', where the special case of CAR is expressed by setting all these ratios equal to 1. Analogously, assumptions about the ratios

$$R_{\mathbf{x},\mathbf{x}',y,\mathfrak{y}} = \frac{q_{\mathfrak{y}|\mathbf{x}y}}{q_{\mathfrak{y}|\mathbf{x}'y}}, \quad \mathfrak{y} \in \Omega_{\mathcal{Y}},\ y \in \mathfrak{y},\ \mathbf{x}, \mathbf{x}' \in \Omega_X \ , \tag{16}$$

defined for all $\mathbf{x}$ and $\mathbf{x}'$ with two directly successive covariate values and equal other covariate values may be imposed, with $R_{\mathbf{x},\mathbf{x}',y\mathfrak{y}} = 1$, $\forall \mathbf{x},\ \mathbf{x}' \in \Omega_X,\ \mathfrak{y} \in \Omega_{\mathcal{Y}},\ y \in \Omega_Y$ representing the case of SI [cf. 33]. If all coarsening ratios in (15) were known, the parameter of interest, i.e. all parameters determining the latent variable distribution, would be point-identified, hence a particular coarsening scenario would be considered. In this way, these coarsening ratios can be regarded as sensitivity parameters in the sense of [43]. In specific cases this is also valid for the coarsening ratios in (16), studied in more detail in [33].
In most practical cases it is unrealistic to claim knowledge about the exact value of the ratios. Nevertheless, it seems quite realistic that former studies or substance-matter considerations allow rough statements about the magnitude of the ratios. In order to investigate how to include such weak knowledge about the coarsening process into the cautious estimation of the regression coefficients presented in the previous sections, we start by taking a closer look at some results under a specific partial assumption in the setting of **Example 1**, before we discuss some more general partial assumptions in context of **Example 2**.

Example 1: Frequently, there are situations, where assumptions as "respondents with a high income rather tend to give no answer compared to the ones with a low income" might be justified from an application standpoint. This weak knowledge about the missingness can be formalized as $q_{na|\mathbf{x}<} < q_{na|\mathbf{x}\geq}$ or $R_{\mathbf{x},<,\geq,na} \in [0,1[$. Consequently, we can still rely on the consideration of the (relative) profile log-likelihood by simply adding this linear constraint on the coarsening parameters into the original optimization problem. Figure 7 shows the obtained (relative) profile likelihood functions, also indicating the δ-cut for the construction of asymptotic 90% confidence intervals. By comparing the results in Table 7, giving the estimated regression coefficients and respective confidence intervals under the auxiliary information about the missingness, to the ones without auxiliary information in Table 5 and Table 6, one notes a remarkable refinement of the results.

Example 2: Assumptions of that kind can also be included in the presence of coarse data in the strict sense, hence incorporating for instance $q_{\{a,b,c\}|\mathbf{x}a} < q_{\{a,b,c\}|\mathbf{x}b} < q_{\{a,b,c\}|\mathbf{x}c}$.

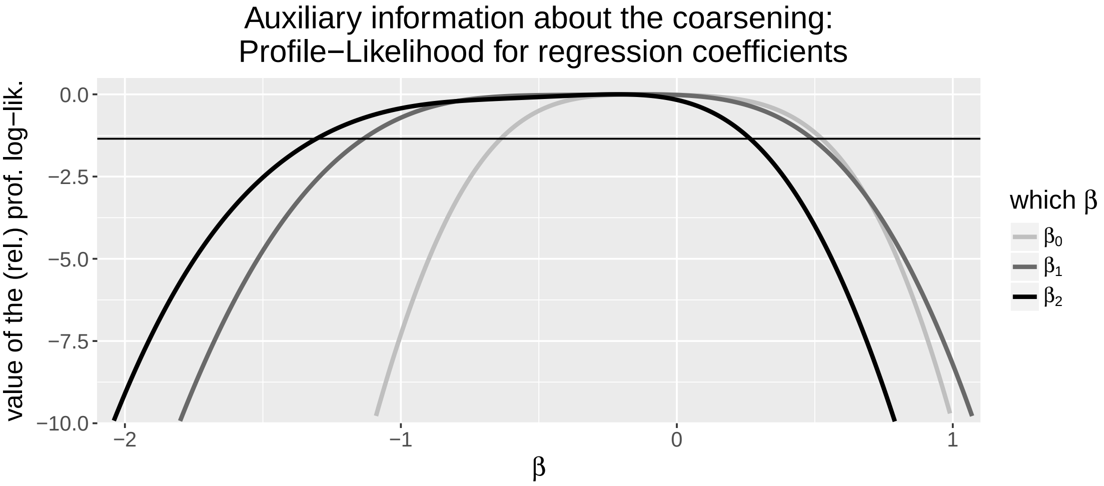


Figure 7: Based on the auxiliary information $q_{na|\mathbf{x}<} < q_{na|\mathbf{x}\geq}$ and the data of **Example 1**, the (relative) profile log-likelihood is determined. The δ-cut is marked by the horizontal line.

Table 7: Reliable regression estimates and confidence intervals under $q_{na|\mathbf{x}<} < q_{na|\mathbf{x}\geq}$ (**Example 1**).

point estimation	$\hat{\beta}_0 \in [-0.53,\ 0.35]$	$\hat{\beta}_1 \in [-0.73,\ 0.05]$	$\hat{\beta}_2 \in [-0.85,\ 0.00]$
confidence interval for ...	$\beta_0 :\ [-0.74,\ 0.64]$	$\beta_1 :\ [-1.17,\ 0.52]$	$\beta_2 :\ [-1.35,\ 0.34]$

More generally, $R_{\mathbf{x},y,y',\mathcal{Y}}$ (or analogously $R_{\mathbf{x},\mathbf{x}',y,\mathcal{Y}}$) can be assumed to be in the interval $[\underline{R},\ \overline{R}]$ with $\underline{R}, \overline{R} \in \mathbb{R}_0^+$, where one can practically incorporate this information by adding the linear constraints $q_{\mathcal{Y}|\mathbf{x}y} \geq q_{\mathcal{Y}|\mathbf{x}y'} \cdot \underline{R}$ and $q_{\mathcal{Y}|\mathbf{x}y} \leq q_{\mathcal{Y}|\mathbf{x}y'} \cdot \overline{R}$ into the optimization problem. As a special case, there are several practical situations where CAR or SI is principally conceivable, but their exact satisfaction is rather questionable. Then the inclusion of specific neighborhood assumptions [as e.g. addressed in 24, for MAR] is desirable, requiring that the coarsening probabilities lie in the environment of the CAR or SI case. This corresponds to choosing $R_{\mathbf{x},y,y',\mathcal{Y}}$ or $R_{\mathbf{x},\mathbf{x}',y,\mathcal{Y}}$ to lie within the interval $[\frac{1}{\tau_1},\ \tau_2]$, where $\tau_1,\ \tau_2 \geq 1$ specify the neighborhood. Further research should be devoted to the incorporation of auxiliary information in terms of comparable statements about the ratios (as e.g. $R_{\mathbf{x},a,b,\{a,b,c\}} \leq R_{\mathbf{x},b,c,\{a,b,c\}}$) leading to bilinear constraints and the investigation of the impact of auxiliary information under the three situations (situation 1(a), (b), 2).

6 Concluding remarks

Most reports containing survey results, also including publications in official statistics, at best point to the fact that non-sampling errors occured, but totally neglect to quantify them [cf. 24]. This practice is especially undesirable since it not only bluffs certainty leading to misinterpretation of results, but may also conduce to a substantial bias. Consequently, communication of the underlying uncertainty should be part of every trustworthy data analysis. Frequently, a considerable contribution to the non-sampling error is ascribable to the item nonresponse problem, which we tackled here by addressing the more general situation of coarse data.

We explicitly departed from the goal of forcing a particular coarsening scenario to achieve point-identified parameters. Allowing for partially identified parameters enables the user to make an analysis driven by the available information about the coarsening process, instead of – maybe unfoundedly chosen – optimization criteria or point-identifying coarsening assumptions. By generalizing the coarsening at random and the subgroup independence assumptions, we could reveal a practical possibility how the user can include frequently available rough statements about the coarsening to refine the results obtained from an analysis based on no assumptions about the coarsening at all.

Aiming at a reliable categorical regression analysis in the presence of coarse data, two different methods to determine cautious regression coefficients have been discussed in the light of data examples: The first one is based on a two-step procedure, which turned out to simplify things only in specific situations, such as cases with a binary response variable, and is even then not always rewarding. Studying this procedure gave rise to various types of results (situation 1(a), 1(b), 2). In this way, we figured out that the parametric assumption on the regression model can induce a principally differing impact on the estimated coarsening parameters, from no effect, via tighter bounds, through to point-identified parameters. The second method, here called direct method, relies on the (relative) profile log-likelihood, where the estimated bounds of the regression coefficients are given by considering the set of all maxima. This procedure is natural, always applicable – although the computation of the (relative) profile log-likelihood may be challenging – and offers a simple way to construct confidence intervals. Having a closer look at response functions of further categorical regression models and discussing the appropriateness of both methods in this context should be part of further research.

We applied all findings to the PASS data. A comparison of the results of our cautious approach to the ones of a traditional method relying on coarsening at random showed that sometimes even certainty about the sign of the regression estimates would be pretended by the latter procedure. Depending on the research question, our results might be assessed as too little informative, especially if the confidence intervals are the focus of interest. But this does not justify to return to traditional methods, which here would pretend certainty about even the sign of the regression coefficients in some cases. Thus, a possibly small content of information should not be regarded as a weakness of an approach based on the methodology of partial identification, but associated to sparse additional knowledge. Although the gain of information achieved by the explicit collection of coarse data is comparably small in our case, which is ascribable to the low proportion of coarse compared to missing answers, the used questionnaire design for requesting the income of the PASS study is recommendable, especially for sensitive topics.

The cautious likelihood approach for the latent variable distribution turns out to be a fruitful field of study for further research: The connection between the latent and the observed world gives the opportunity to transfer existing likelihood-based methods for precise categorical data [as e.g. statistical tests, as e.g. in 33] to the setting of coarse data. Another promising topic is the application of our cautious approach to other problems relying on strong assumptions. A direct reference is conjectured for misclassification, propensity score matching and statistical matching, where starting points are already provided in [26], [38] (who studied an approach based on partial identification to estimate treatment effects without considering propensity scores), [9], respectively. Propensity score matching and statistical matching traditionally rely on strict assumptions, namely the strongly ignorable treatment assignment as well as the conditional independence assumption, respectively, where a cautious strategy would allow for a relaxation of these prerequisites.

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. Furthermore, we appreciate the help of Paul Fink, who discussed data disclosure issues with us. The first author thanks the LMUMentoring program, providing financial support for young, female researchers.

References

[1] Arpino, B., De Cao, E., and Peracchi, F. (2014). Using panel data for partial identification of human immunodeficiency virus prevalence when infection status is missing not at random. *J. R. Statist. Soc. A*, 177:587–606.

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

[3] Beyer, O., Hüser, A., Rudloff, K., and Rüst, M. (2014). Statistische Geheimhaltung: Rechtliche Grundlagen und fachliche Regelungen der Statistik der Bundesagentur für Arbeit. Accessed: 2017-10-13.

[4] Brack, G. (2017). Wie viele Flüchtlinge sind ohne Schulabschluss? Accessed: 2017-10-13.

[5] Brücker, H. and Schupp, J. (2017). Annähernd zwei Drittel der Geflüchteten haben einen Schulabschluss. Accessed: 2017-10-13.

[6] Cattaneo, M. and Wiencierz, A. (2012). Likelihood-based Imprecise Regression. *International Journal of Approximate Reasoning*, 53:1137–1154.

[7] Couso, I., Dubois, D., and Sánchez, L. (2014). *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables.* Springer.

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

[9] D'Orazio, M., Di Zio, M., and Scanu, M. (2006). Statistical matching for categorical data: Displaying uncertainty and using logical constraints. *Journal of Official Statistics*, 22:137–157.

[10] Drechsler, J., Kiesl, H., and Speidel, M. (2015). MI double feature: Multiple imputation to address nonresponse and rounding errors in income questions. *Austrian Journal of Statistics*, 44:59–71.

[11] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, Methods and Applications.* Springer.

[12] Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5:475–492.

[13] Heitjan, D. and Rubin, D. (1991). Ignorability and coarse data. *Ann. Statist.*, 19:2244–2253.

[14] Heumann, C. (2004). *Monte Carlo Methods for Missing Data in Generalized Linear and Generalized Linear Mixed Models.* Habilitation (post-doctoral thesis). Ludwig-Maximilians Universität, München.

[15] Hoeren, D. (2017). 59 Prozent der Flüechtlinge haben keinen Schulabschluss. Accessed: 2017-10-13.

[16] Horowitz, J. and Manski, C. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Am. Statist. Ass.*, 95:77–84.

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

[18] Imbens, G. and Manski, C. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

[19] Jackson, D., White, I., and Leese, M. (2010). How much can we learn about missing data?: an exploration of a clinical trial in psychiatry. *J. R. Statist. Soc. A*, 173:593–612.

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

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

[22] Little, R. and Rubin, D. (2014). *Statistical Analysis with Missing Data.* 2nd edition, Wiley.

[23] Manski, C. (2003). *Partial Identification of Probability Distributions.* Springer.

[24] Manski, C. (2015). Credible interval estimates for official statistics with survey nonresponse. *Journal of Econometrics*, 191:293–301.

[25] McLachlan, G. and Peel, D. (2004). *Finite mixture models.* Wiley.

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

[27] Neale, M. and Miller, M. (1997). The use of likelihood-based confidence intervals in genetic models. *Behavior Genetics*, 27:113–120.

[28] Nguyen, H. (2006). *An Introduction to Random Sets.* CRC.

[29] Nordheim, E. (1984). Inference from nonrandomly missing categorical data: An example from a genetic study on Turner's syndrome. *J. Am. Statist. Ass.*, 79:772–780.

[30] Olson, K. (2013). Do non-response follow-ups improve or reduce data quality?: a review of the existing literature. *J. R. Statist. Soc. A*, 176:129–145.

[31] Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press.

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

[33] Plass, J., Cattaneo, M., Schollmeyer, G., and Augustin, T. (2017). On the testability of coarsening assumptions: A hypothesis test for subgroup independence. *International Journal of Approximate Reasoning*, 90:292–306.

[34] Rubin, D. (1976). Inference and missing data. *Biometrika*, 63:581–592.

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

[36] Shafer, G. (1976). *A Mathematical Theory of Evidence.* Princeton University Press.

[37] Stoye, J. (2009a). More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315.

[38] Stoye, J. (2009b). Partial identification and robust treatment choice: an application to young offenders. *Journal of Statistical Theory and Practice*, 3:239–254.

[39] Tamer, E. (2010). Partial identification in econometrics. *Annual Review Economics*, 2:167–195.

[40] Tanna, G. (2017). Missing data analysis in practice t. Raghunathan, 2016 Boca Raton, Chapman and Hall–CRC 230 pp.,£ 52.99 ISBN 978-1-482-21192-4. *J. R. Statist. Soc. A*, 180:684–685.

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

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

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

[44] Venzon, D. and Moolgavkar, S. (1988). A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 37:87–94.

[45] Verbeke, G. and Molenberghs, G. (2009). *Linear Mixed Models for Longitudinal Data.* Springer.

[46] Wood, A., White, I., and Hotopf, M. (2006). Using number of failed contact attempts to adjust for non-ignorable non-response. *J. R. Statist. Soc. A*, 169(3):525–542.

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

International Journal of Approximate Reasoning 90 (2017) 292–306

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

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*

ARTICLE INFO

Article history:
Received 10 January 2017
Received in revised form 4 July 2017
Accepted 25 July 2017
Available online 19 September 2017

Keywords:
Coarse data
Missing data
Coarsening at random (CAR)
Likelihood-ratio test
Partial identification
Sensitivity analysis

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 first re-illustrate the impossibility to test CAR and then contrast it to another type of coarsening called subgroup independence (SI). 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. The results are illustrated by the data of the German Panel Study "Labour Market and Social Security" (PASS).

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. e.g. [17]) simply for pragmatic reasons without justifications in substance (cf. e.g. [15]). Since MAR is not testable without strong additional assumptions (e.g. [18]) and wrongly including MAR may induce a substantial bias, this way to proceed is especially alarming.

Beside missing data, there are further kinds of deficient data, such as data affected by measurement errors/misclassification (cf. e.g. [11]) or coarse(ned) data (cf. e.g. [12]) where only subsets of the complete data sample space are observed, known to include the unobserved, precise value.[1] Throughout the paper, we consider coarse data, including missing data as special case, thus addressing partially observed values, explicitly excluding the erroneous observation of a variable, disregarding measurement errors/misclassification. For instance, coarse data may arise in data sets where coarsening is

[☆] This paper is part of the Virtual special issue on Soft methods in probability and statistics, edited by Barbara Vantaggi, Maria Brigida Ferraro, Paolo Giordani. 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 [25].

[*] Corresponding author.
E-mail address: julia.plass@stat.uni-muenchen.de (J. Plass).

[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. [5]).

http://dx.doi.org/10.1016/j.ijar.2017.07.014

deliberately applied as anonymization technique or matched data sets with not completely identical categories. In the context of coarse data, the *coarsening at random* (CAR) (cf. [12]) assumption is the analogue of MAR. Although the impossibility of testing CAR is already known from literature (cf. e.g. [14]), providing an intuitive insight into this point will be a first goal of our paper. Apart from CAR, we focus on another, in a sense dual, assumption that we called *subgroup independence* (SI) in [22] 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. [10], [13]), others end up with set-valued results (cf. e.g. [3], [16], [22]). A general view is given by Couso and Dubois [6], distinguishing between different types of likelihoods, the visible, the latent and the total likelihood. Here, we use the cautious approach developed in [22], which refers to the latent likelihood and is – just as e.g. [19,8] (in the context of misclassification) and [28] – strongly influenced by the methodology of *partial identification* (cf. [18]). Thus, according to the spirit of partial identification, instead of being forced to make often 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 traditional approaches 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 that 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 the context of testing.

Throughout the paper, we refer to the case of a coarse categorical response variable Y and a 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 the framework is also applicable in the general categorical setting.

For this categorical setting, we characterize cases where SI makes parameters not only 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 the hypotheses, illustrate the sensitivity of the test statistic 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" (PASS), 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 may be refined by assuming 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. All results of this paper are given for a general categorical setting, but the running example refers to the illustrative case of binary data. To emphasize the general applicability of our approach, we briefly discuss further examples in Section 8, also addressing potential limitations. Finally, Section 9 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 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 the probabilities π_{xy} and $p_{x\mathfrak{y}} = P(\mathcal{Y}=\mathfrak{y}|X=x)$, 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 the individual i $(i=1,\dots,n)$ via the values x and y exclusively, but also require distinct parameters in the sense of Rubin (cf. e.g. [17]) as well as error-freeness,[2] i.e. $\mathfrak{y}\ni y$, explicitly excluding the case of misclassification.

An essential part of our argumentation is based on comparing the dimensions of the parameter space of the latent world Θ_{lat} and the parameter space of the observed world Θ_{obs}. While $\theta_{lat}\in\Theta_{lat}$ describes the latent variable distribution π_{xy}

[2] This implies that Y is a selector of $\mathcal{Y}$ (in the sense of e.g. [20, p. 43]).

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	$\{a\}$	$n_{1\{a\}} = 36$	$n_1 = 87$
	$\{b\}$	$n_{1\{b\}} = 42$	
	$\{a,b\}$	$n_{1\{a,b\}} = 9$	

and the coarsening parameters $q_{\mathfrak{y}|xy}$, $\mathfrak{y} \in \Omega_{\mathcal{Y}}$, $x \in \Omega_X$, $y \in \Omega_Y$, the parameter $\theta_{obs} \in \Theta_{obs}$ represents the observed variable distribution $p_{x\mathfrak{y}}$. We choose one of the minimal possible parametrizations, in order to be clear about the dimension of the parameter spaces, generally obtained as

$$\begin{aligned} dim(\Theta_{lat}) &= \overbrace{k \cdot (m-1)}^{\text{latent variable distr.}} + \overbrace{k \cdot m \cdot (2^{m-1}-1)}^{\text{coarsening param.}}, \\ dim(\Theta_{obs}) &= \overbrace{k \cdot (2^m - 2)}^{\text{observed variable distr.}}, \end{aligned} \tag{1}$$

with $k = |\Omega_X|$ and $m = |\Omega_Y|$. Due to the restriction that probabilities sum up to one, we refrain from the incorporation of $q_{\mathfrak{y}|xy}$ with $\mathfrak{y} = \{y\}$, $x \in \Omega_X$, $y \in \Omega_Y$, thus starting from index $z = 2$ in the calculation of the number of coarsening parameters in one subgroup[3]: $\sum_{z=2}^{m} z \cdot \binom{m}{z} = m \cdot (2^{m-1} - 1)$. For the same reason, for each subgroup x, only $(m-1)$ and $(2^m - 2)$ parameters π_{xy} and $p_{x\mathfrak{y}}$ determine the latent variable distribution and the observed variable distribution, respectively, where $|\Omega_{\mathcal{Y}}| = 2^m - 1$.

As the number of the coarsening parameters increases considerably with k and m, for reasons of conciseness, we start by mainly confining ourselves to the discussion of a running example[4] considering binary variables. While we denote the different categories of X by numbers, letters are used to refer to the categories of Y. In this way, the example addresses 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. Consequently, defining

$$\begin{aligned} \theta_{lat} &= (\pi_{0a},\ q_{\{a,b\}|0a},\ q_{\{a,b\}|0b},\ \pi_{1a},\ q_{\{a,b\}|1a},\ q_{\{a,b\}|1b})^T \quad \text{and} \\ \theta_{obs} &= (p_{0\{a\}},\ p_{0\{b\}},\ p_{1\{a\}},\ p_{1\{b\}})^T , \end{aligned} \tag{2}$$

we obtain $dim(\Theta_{lat}) = 6$ and $dim(\Theta_{obs}) = 4$ as dimensions of the respective parameter spaces. The example is introduced in the following box:

Running example:
The German Panel Study "Labour Market and Social Security" (PASS, [31], 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 € (category a) or ≥ 1000 € (category 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.

Although we repeatedly make use of this binary example, all results are applicable for the general categorical case with k subgroups and m categories of variable Y. Thus, the example is only used to simplify the understanding of the basic points, while the main contributions of this paper, i.e. considerations regarding identifiability and testability as well as the proposed hypothesis test, refer to the general categorical setting. To stress the generality, we later briefly illustrate a case not automatically reducing to the missing data situation in the end, also discussing the complexities inherent to the applications with arbitrary finite sample spaces (cf. Section 8).

[3] The binomial coefficient $\binom{m}{z}$ gives the number of z-element subsets of Ω_Y, where for each z-element subset exactly z coarsening parameters are needed.

[4] Another application of the cautious likelihood approach used here is studied in [26] in the context of small area estimation, relying on the data of the German General Social Survey.

3. Coarsening models

Considering our categorical setting, we look at two ways of assuming the coarsening process to be uninformative in the sense that certain variables do not play any role: The coarsening can be independent of the value of the response variable or of the covariate(s), thus ending up in CAR (cf. Section 3.1) or SI (cf. Section 3.2), respectively.

3.1. Coarsening at random and its generalized version

Heitjan and Rubin ([12]) 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. [17]). 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 naturally adapted 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 [9].

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.

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

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 (cf. e.g. [30]).

A generalization (extending Nordheim's [21] 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): it consists in assuming the values of the ratios of coarsening parameters for given subgroups and coarse observations, i.e.

$$R_{x,y,y'\mathfrak{y}} = \frac{q_{\mathfrak{y}|xy}}{q_{\mathfrak{y}|xy'}}, \tag{3}$$

defined for all subgroups $x \in \Omega_X$ and all compatible $y, y' \in \Omega_Y$ and $\mathfrak{y} \in \Omega_{\mathcal{Y}}$, where y and y' are directly successive[5] (cf. [24]). In the missing data situation of our running example, we assume the values of the ratios

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

where $R_{0,a,b,\{a,b\}} = R_{1,a,b,\{a,b\}} = 1$ represents the special case of CAR/MAR. In most cases, it might be difficult to justify knowledge about the exact value of the ratios, but former studies or material considerations may naturally 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. [22]) also underlying our paper.

3.2. Subgroup independence and its generalized version

If the data are missing not at random (MNAR) [17], commonly the missingness process is modeled by including parametric assumptions (e.g. [12]), or a cautious procedure is chosen ending up in set-valued estimators (cf. e.g. [7], [22], [34]). For the categorical setting, it turns out that there is a special case of MNAR, in which single-valued estimators can be obtained without additional 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 setting, the latter case corresponds to SI sketched in [22], and studied in detail here. This independence from the covariate value shows, beside CAR, an alternative kind of coarsening assumption.

Again, one should generally 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.

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

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

[5] Considering categories without inherent order, an arbitrary order has to be chosen.

A generalization, in the following called *generalized subgroup independence* (gSI), consists in assuming the values of the ratios

$$R_{x,x',y,\mathfrak{y}} = \frac{q_{\mathfrak{y}|xy}}{q_{\mathfrak{y}|x'y}}\,, \tag{4}$$

defined for all compatible $y \in \Omega_Y$ and $\mathfrak{y} \in \Omega_{\mathcal{Y}}$ (apart from $\mathfrak{y} = \{y\}$) and directly successive (cf. Footnote 5) covariate values $x,\ x' \in \Omega_X$ (cf. [24]). In the example, the values of the ratios

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

are assumed, where assuming $R_{0,1,a,\{a,b\}} = R_{0,1,b,\{a,b\}} = 1$ corresponds to SI. 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, (g)CAR and (g)SI

This section recalls some important aspects of our approach developed in [22] 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 for the point-identifiability of 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, the MLE is uniquely obtained by the relative frequencies of the respective categories (cf. [27]), coarse categories treated as own categories. Secondly, we connect the parameters of both worlds by a mapping

$$\begin{aligned} \Phi:\ &\Theta_{lat} \to \Theta_{obs}, \\ &\theta_{lat} \mapsto \theta_{obs} \end{aligned} \tag{5}$$

expressing the observation process, where Θ_{lat} and Θ_{obs} are the parameter space of the latent and the observed world, respectively. The mapping Φ can be shown to be separable into independent components Φ_x corresponding to subgroup x, $x \in \Omega_X$.

For our example, we obtain

$$\Phi_x \begin{pmatrix} \pi_{xa} \\ q_{\{a,b\}|xa} \\ q_{\{a,b\}|xb} \end{pmatrix} = \begin{pmatrix} \pi_{xa}\cdot(1 - q_{\{a,b\}|xa}) \\ (1-\pi_{xa})\cdot(1 - q_{\{a,b\}|xb}) \end{pmatrix} = \begin{pmatrix} p_{x\{a\}} \\ p_{x\{b\}} \end{pmatrix}, \tag{6}$$

$x \in \{0, 1\}$, 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. In this way, we obtain the set-valued estimator

$$\hat{\Gamma} = \{\hat{\theta}_{lat} \mid \Phi(\hat{\theta}_{lat}) = \hat{\theta}_{obs}\}, \tag{7}$$

with $\hat{\theta}_{lat}$ and $\hat{\theta}_{obs}$ as the MLE's of θ_{lat} and θ_{obs}, respectively.[6] This set-valued estimator can also be illustrated by building the one dimensional projections, which are intervals: in the situation of the example

$$\hat{\pi}_{xa} \in \left[\frac{n_{x\{a\}}}{n_x},\ \frac{n_{x\{a\}} + n_{x\{a,b\}}}{n_x}\right], \quad \hat{q}_{\{a,b\}|xy} \in \left[0,\ \frac{n_{x\{a,b\}}}{n_{x\{y\}} + n_{x\{a,b\}}}\right], \tag{8}$$

with $x \in \{0,\ 1\}$ and $y \in \{a,\ b\}$. Points in these intervals are constrained by the relationships in Φ. The obtained set-valued estimator in (7), and thus the corresponding projections, may be refined by including assumptions about the coarsening

[6] This result is strictly related to the one obtained from cautious data completion (cf. e.g. [1], §7.8.), by plugging in all potential precise values compatible with the observations.

justified from the application standpoint (in the spirit of [18]).[7] Very strict assumptions may induce point-identified parameters, as estimation under CAR or SI in the categorical case shows.[8]

4.2. Basic argument of studying the identifiability

Discussing identifiability, we consider the general case with $k = |\Omega_X|$ and $m = |\Omega_Y|$, using the setting of the example only for reasons of illustration. In Sections 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 degrees of freedom under the assumption in focus (here generally noted as *aspt*), i.e.

$$df^{aspt} = dim(\Theta_{obs}) - dim(\Theta_{lat}^{aspt}), \tag{9}$$

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 equal to $k \cdot (2^m - 2)$ (cf. Equation (1)) independently of whether the assumption of CAR/gCAR or SI/gSI is included.[9]

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. (9)). 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 (also cf. Equation (1) and its explanation). 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 in the example, i.e. by restricting the set of possible coarsening mechanisms to $q_{\{a,b\}|xa} = q_{\{a,b\}|xb}$ with $x \in \{0,\ 1\}$, we obtain the point-valued estimators

$$\hat{\pi}_{xa}^{CAR} = \frac{n_{x\{a\}}}{n_{x\{a\}} + n_{x\{b\}}}, \quad \hat{q}_{\{a,b\}|xa}^{CAR} = \hat{q}_{\{a,b\}|xb}^{CAR} = \frac{n_{x\{a,b\}}}{n_x}. \tag{10}$$

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 $\hat{q}_{\{a,b\}|xa} = \hat{q}_{\{a,b\}|xb}$ 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 receive point-valued estimators for the general version as well. For fixed values of the ratios in (3), the parameters of main interest π_{xy} are point-identified, wherefore the ratios may be regarded as sensitivity parameters in the sense of [16]. 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 these assumptions (cf. [22]).

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 $dim(\Theta_{lat}^{SI}) = k \cdot (m-1) + m \cdot (2^{m-1} - 1)$ (as compared to Equation (1)). Solving (cf. (9)) in this setting

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

for k, we obtain the condition

$$k \geq \frac{m \cdot (2^{m-1} - 1)}{2^m - m - 1}, \tag{11}$$

that has to be satisfied to concede point-valued estimators.

In this paper we focus on the setting where $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$ with all categories observable. But frequently, especially in cases with a high number of categories for the variable Y, there are naturally data situations where only specific coarse

[7] An approach that aims at refining the results under total ignorance is e.g. given in [34], where the conservative inference rule is presented as a compromise between a too optimistic (i.e. assuming CAR) and a too pessimistic (i.e. assuming total ignorance) knowledge about the coarsening process. Note that a different setting is studied there, considering coarse covariates instead of a coarse response variable.

[8] 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 [29].

[9] For every of the k subgroups, $|\Omega_{\mathcal{Y}}| - 1 = |\mathcal{P}(\Omega_Y) \setminus \{\emptyset\}| - 1$ parameters of the observed world have to be estimated (cf. Section 2).

categories, i.e. a strict subset of $\mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$, can be observed and 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}, so that the minimum number of subgroups needed for point-identifiability generally can no longer be expressed in terms of m exclusively. In particular, in the prominent missing data case, which is of high practical relevance, 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 in order to have point-identifiability, since in this case

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

In the setting of our example, there are two subgroups available, which corresponds to the lower bound in (11), such that the respective condition is satisfied. This is in line with the result that under rather weak regularity conditions, namely $\pi_{0a} \neq \pi_{1a}$,[10] $\pi_{0a} \notin \{0, 1\}$, and $\pi_{1a} \notin \{0, 1\}$ for $x \in \{0, 1\}$, under SI the mapping Φ becomes injective (a proof is given in [23, p. 17, 20]). Hence, we obtain point-valued estimators

$$\begin{aligned} \hat{\pi}_{xa}^{SI} &= \frac{n_{x\{a\}}}{n_x} \, \frac{n_0\, n_{1\{b\}} - n_{0\{b\}}\, n_1}{n_{0\{a\}}\, n_{1\{b\}} - n_{0\{b\}}\, n_{1\{a\}}}, \\ \hat{q}_{\{a,b\}|xa}^{SI} &= \frac{n_{0\{a,b\}}\, n_{1\{b\}} - n_{0\{b\}}\, n_{1\{a,b\}}}{n_0\, n_{1\{b\}} - n_{0\{b\}}\, n_1}, \\ \hat{q}_{\{a,b\}|xb}^{SI} &= \frac{n_{0\{a,b\}}\, n_{1\{a\}} - n_{0\{a\}}\, n_{1\{a,b\}}}{n_0\, n_{1\{a\}} - n_{0\{a\}}\, n_1}, \end{aligned} \tag{12}$$

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 (4), the estimator in (7) 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. [23, p. 15, 18]), testing these assumptions is 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. [18, 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 (10) 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 arising from (7). Spoken for the situation of the example: The coarsening scenario where respondents from the low income category and respondents from 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 estimator in (7).

For the example, under CAR we obtain

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

which may not be excluded from the set-valued estimator, and also the corresponding 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, too.

Nevertheless, there are several approaches that show how testability of MAR is achieved by the inclusion of additional assumptions (e.g. [14]), where the results probably could be extended to CAR. For instance, testability of MAR can be achieved

[10] The case of $\pi_{0a} = \pi_{1a}$ represents the homogeneous case, where multiple solutions result (cf. [22], p. 254).

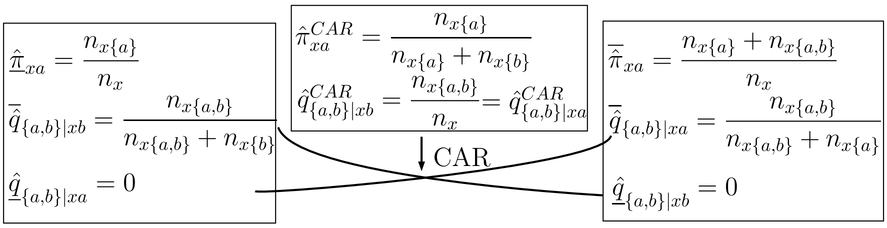

$\hat{\underline{\pi}}_{xa} = \frac{n_{x\{a\}}}{n_x}$
$\hat{\overline{q}}_{\{a,b\}|xb} = \frac{n_{x\{a,b\}}}{n_{x\{a,b\}} + n_{x\{b\}}}$
$\hat{\underline{q}}_{\{a,b\}|xa} = 0$
$\hat{\pi}^{CAR}_{xa} = \frac{n_{x\{a\}}}{n_{x\{a\}} + n_{x\{b\}}}$
$\hat{q}^{CAR}_{\{a,b\}|xb} = \frac{n_{x\{a,b\}}}{n_x} = \hat{q}^{CAR}_{\{a,b\}|xa}$
CAR
$\hat{\overline{\pi}}_{xa} = \frac{n_{x\{a\}} + n_{x\{a,b\}}}{n_x}$
$\hat{\overline{q}}_{\{a,b\}|xa} = \frac{n_{x\{a,b\}}}{n_{x\{a,b\}} + n_{x\{a\}}}$
$\hat{\underline{q}}_{\{a,b\}|xb} = 0$

Fig. 1. Since the relationships expressed via Φ in (6) have to be met, only specific points from the estimators in (8) are combinable, ranging from $(\hat{\underline{\pi}}_{xa}, \hat{\underline{q}}_{\{a,b\}|xa}, \hat{\overline{q}}_{\{a,b\}|xb})$ to $(\hat{\overline{\pi}}_{xa}, \hat{\overline{q}}_{\{a,b\}|xa}, \hat{\underline{q}}_{\{a,b\}|xb})$ with the CAR case always included.

under the availability of instrumental variables that are required to be conditionally independent from the missingness given the response variable and covariates and additionally assuming bounded completeness (cf. [2]). Another approach of that kind is for instance given in [15], where distributional constraints on the structure of a network are incorporated. Generally, the challenge remains to distinguish between cases, where MAR is justifiably rejected/not rejected, and cases where the included additional assumptions were wrongly made, so that the test decision is meaningless.

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 (11) as a necessary condition to be able to obtain point-valued estimators at all. In this sense, we cannot generally obtain point-valued estimators as in the case of CAR. Similarly, also when studying the testability of SI, two cases have to be distinguished: The case of $df^{SI} < 0$, where SI cannot 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)[11] estimators in (12) 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 (12) to the example. We obtain the unconstrained estimates

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

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 SI, representing a too strict coarsening rule in certain observed data situations, wherefore SI might be testable.

Although we will present the test statistic only then in Section 6.1 (cf. (14)), we can – at least if we restrict to the standard case with sufficiently many subgroups – already prepare its main underlying idea: Comparing the maximal likelihood under SI and the maximal likelihood achieved under refraining from strict coarsening assumptions and using those mentioned in Section 2 only, allows us to distinguish the two cases pointing to the two possible test decisions. Case 1: The likelihood optimized under SI achieves the computational maximum obtained by $\Phi^{-1}(\hat{\theta}_{obs})$, where Φ^{-1} is the inverse of Φ. In this situation the value of our likelihood-based test statistic will result in the test decision that SI cannot be rejected. Case 2: 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.[12] Then, our test statistic indeed will react sensitively to the reduction of the likelihood value and will lead to a rejection of SI if this reduction is large enough in the light of the significance level α. This differentiation between the two cases gives us the opportunity to test on SI, while we always end up in case 1 if CAR is included into the likelihood optimization making testability impossible (cf. Section 5.1). In the next section, especially in Fig. 2, we will seize on the two cases characterizing the two possible test decisions, where the sensitivity of the deviation between the maximum value of the likelihood with and without SI will be exploited in the likelihood ratio test.

If the criterion given in (11) 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. General aspects: hypotheses, test statistic and test decision

If sufficient subgroups are available in the sense that the condition in (11) is met, a statistical test for the following hypotheses can be constructed in the categorical case:

[11] Probability restrictions are not included.

[12] In our example, the unconstrained estimators in (12), which are the unique inverse image of the MLE's $\hat{p}_{x\{a\}}$ and $\hat{p}_{x\{b\}}$ under (an extension of) the injective function Φ, are partly outside the interval $[0,\ 1]$.

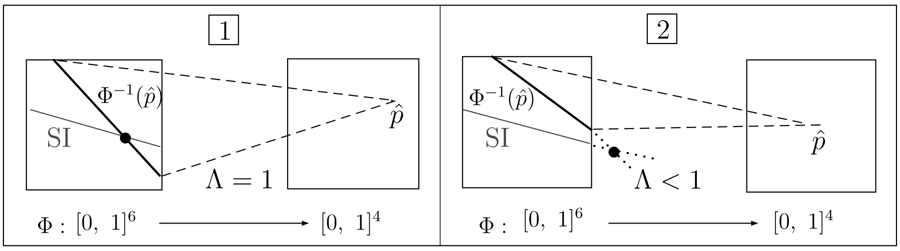


Fig. 2. The impact on Λ of two substantially differing data situations is illustrated.

Table 2
Distribution of T under H_0 in dependence of k and m.

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

$$\begin{aligned} H_0 &: q_{\mathfrak{y}|xy} = q_{\mathfrak{y}|x'y} \text{ for all } \mathfrak{y} \in \Omega_{\mathcal{Y}},\ x, x' \in \Omega_X,\ y \in \Omega_Y, \\ H_1 &: q_{\mathfrak{y}|xy} \neq q_{\mathfrak{y}|x'y} \text{ for some } \mathfrak{y} \in \Omega_{\mathcal{Y}},\ x, x' \in \Omega_X,\ y \in \Omega_Y. \end{aligned} \tag{13}$$

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 (13) can be based on the classical test statistic (e.g. [33])

$$T = -2 \cdot \ln(\Lambda(\mathfrak{y}_1, \ldots, \mathfrak{y}_n, x_1, \ldots, x_n)) \tag{14}$$

with likelihood ratio

$$\Lambda(\mathfrak{y}_1, \ldots, \mathfrak{y}_n, x_1, \ldots, x_n) = \frac{\sup_{H_0} L(\theta_{lat} || \mathfrak{y}_1, \ldots, \mathfrak{y}_n, x_1, \ldots, x_n)}{\sup_{H_0 \cup H_1} L(\theta_{lat} || \mathfrak{y}_1, \ldots, \mathfrak{y}_n, x_1, \ldots, x_n)}, \tag{15}$$

(cf., e.g. (2)).[13] While the denominator of Λ can be obtained by using any point in (7) (e.g. θ^{CAR}, which generally cannot be excluded from (7), cf. Section 5.1), the numerator must in general be calculated by numerical optimization. In fact, simulation studies corroborate the decrease of Λ with deviation from SI (cf. [23, p. 19]). The sensitivity of Λ with regard to the test considered here is also illustrated informally in Fig. 2 by depicting Φ in (5) for two data situations with binary variables, 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 points in (7)). The intersection of both lines represents the values in (12), and if it is included in the domain of Φ (cf. left case of Fig. 2), 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. right case of Fig. 2) 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.

Next, we aim at determining a general decision rule depending on the significance level α. In the 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. e.g. [33]). 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). It can be easily checked that condition (11) corresponds to

$$k > \frac{m}{2}, \tag{16}$$

when $m \geq 4$. While the quantile $\chi^2_{df,1-\alpha}$, with $df = df^{SI}$, gives the critical value in case of $df^{SI} > 0$, the critical value is calculated based on a specific asymptotic distribution in case of $df^{SI} = 0$, investigated in the next section. Table 2 shows the distribution of the test statistic under the null hypothesis for a given number of subgroups and categories of the variable of interest.

[13] Alternatives to this statistic would include the construction of uncertainty regions, in the spirit of [32], and then apply the duality between tests and confidence regions.

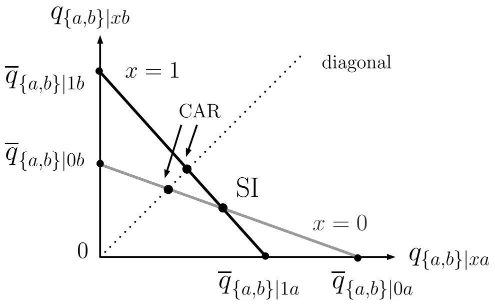

$q_{\{a,b\}|xb}$
$\overline{q}_{\{a,b\}|1b}$
$x=1$
diagonal
CAR
$\overline{q}_{\{a,b\}|0b}$
SI
$x=0$
0
$q_{\{a,b\}|xa}$
$\overline{q}_{\{a,b\}|1a}$
$\overline{q}_{\{a,b\}|0a}$

Fig. 3. The gray and black solid lines symbolize all coarsening parameters within Γ (cf. (7)) 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.

6.2. *The test decision in the special case of $df^{SI} = 0$*

In order to derive the distribution of the test statistic in the special case of $df^{SI} = 0$, it shows to be beneficial to restate the hypotheses in (13) in terms of the parameters of the observed world first. In this way, we will be able to clearly distinguish between the boundary and the non-boundary cases, which will be of great importance in this context. The special case of $df = 0$ is achieved in the setting with binary variables addressed in the example, which we will investigate now in more detail. It can be easily checked that the binary setting (i.e. $k = m = 2$) represents the only case with $df = 0$. Thus, one should mainly be concerned with non-testability (whenever $df^{SI} < 0$) and basing the decision on $\chi^2_{df^{SI},1-\alpha}$ (whenever $df^{SI} > 0$).

Considering the setting of the example, one can write the hypotheses as

$$H_0^*: (p_{0\{a\}} \cdot p_{1\{a,b\}} - p_{1\{a\}} \cdot p_{0\{a,b\}}) \cdot (p_{0\{b\}} \cdot p_{1\{a,b\}} - p_{1\{b\}} \cdot p_{0\{a,b\}}) \leq 0$$

$$H_1^*: (p_{0\{a\}} \cdot p_{1\{a,b\}} - p_{1\{a\}} \cdot p_{0\{a,b\}}) \cdot (p_{0\{b\}} \cdot p_{1\{a,b\}} - p_{1\{b\}} \cdot p_{0\{a,b\}}) > 0.$$

To explain the conditions therein, Fig. 3 shows informally the subgroup specific coarsening parameters $q_{\{a,b\}|xa}$ and $q_{\{a,b\}|xb}$ ranging from 0 to

$$\overline{q}_{\{a,b\}|xa} = \frac{p_{x\{a,b\}}}{p_{x\{a,b\}} + p_{x\{a\}}} \,, \qquad \overline{q}_{\{a,b\}|xb} = \frac{p_{x\{a,b\}}}{p_{x\{a,b\}} + p_{x\{b\}}} \tag{17}$$

respectively, $x \in \{0,\ 1\}$, where the interactions between $q_{\{a,b\}|xa}$ and $q_{\{a,b\}|xb}$ can be inferred from Fig. 1. The assumption of SI is only achievable, if both lines intersect, i.e.

$$\overline{q}_{\{a,b\}|1b} - \overline{q}_{\{a,b\}|0b} \geq 0 \text{ and } \overline{q}_{\{a,b\}|1a} - \overline{q}_{\{a,b\}|0a} \leq 0 \,, \tag{18}$$

or the other way round. After replacing the upper bounds for the coarsening parameters in (18) by (17) and making some little rearrangements, it turns out that an intersection requires

$$p_{0\{b\}} \cdot p_{1\{a,b\}} - p_{1\{b\}} \cdot p_{0\{a,b\}} \geq 0 \text{ and } p_{0\{a\}} \cdot p_{1\{a,b\}} - p_{1\{a\}} \cdot p_{0\{a,b\}} \leq 0 \,,$$

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^*, Fig. 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.

By referring to the hypothesis H_0^*, 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 Fig. 3 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 (wrongly) rejected. Against this, according to Chernoff ([4]), in the boundary case

$$T \underset{H_0}{\overset{a}{\sim}} 0.5 \cdot \delta_0 + 0.5 \cdot \chi_1^2 \,, \tag{19}$$

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 (19). 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 chosen

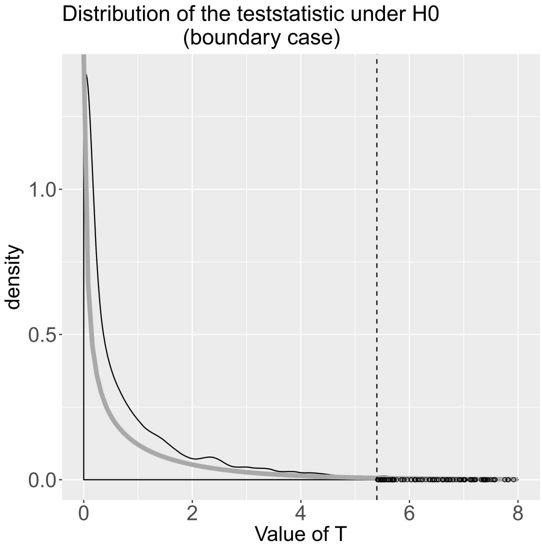


Fig. 4. For an exemplary boundary case, the (smoothed) empirical distribution of the test statistic T under H_0 (black line) is compared to the theoretical asymptotic distribution (gray line).

as $2 \cdot \alpha$, thus obtaining the critical value $\chi^2_{1,1-2\cdot\alpha}$.[14] 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 $p_{1\{a,b\}} = 0.4$. Fig. 4 shows the theoretical asymptotic distribution in (19) as well as the (smoothed) empirical distribution of the obtained values for the test statistic, where both lines are quite close indeed. The vertical line 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 Fig. 4), we obtain the estimated type I error of ≈ 0.0110, basically complying with the level α.

7. Generalized version of the test

By using the ratios $R_{x,x',y,\mathfrak{y}}$ in (4), the hypothesis test for SI may be generalized straightforwardly for gSI. For this purpose, we introduce the hypotheses

$$\begin{aligned} H_0 &: q_{\mathfrak{y}|xy} = R_{x,x',y,\mathfrak{y}} \cdot q_{\mathfrak{y}|x'y}, \text{ for all } \mathfrak{y} \in \Omega_{\mathcal{Y}},\ x, x' \in \Omega_X,\ y \in \Omega_Y, \\ H_1 &: q_{\mathfrak{y}|xy} \neq R_{x,x',y,\mathfrak{y}} \cdot q_{\mathfrak{y}|x'y}, \text{ for some } \mathfrak{y} \in \Omega_{\mathcal{Y}},\ x, x' \in \Omega_X,\ y \in \Omega_Y. \end{aligned} \tag{20}$$

As a test statistic we again utilize T in (14), where the numerator of the likelihood ratio Λ in (15) 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 consider the PASS data example and the ratios in (4). Thus, we focus on the hypotheses

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

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 get $T = 9.2$, exceeding the obtained critical value of ≈ 5.4 (given by the $(1 - 2 \cdot \alpha)$-quantile of the χ^2_1-distribution, with $\alpha = 0.01$), so that H_0 can be rejected.

Fig. 5 gives an overview of the test decision for testing various hypothesis 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 line, which indicates the critical value, are rejected by the likelihood-ratio test based on $\alpha = 0.01$. Thus, subgroup independence (with $R_{0,1,a,\{a,b\}} = R_{0,1,b,\{a,b\}} = 1$,

[14] 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-tests have the (worst-case) asymptotic distribution given in (19).

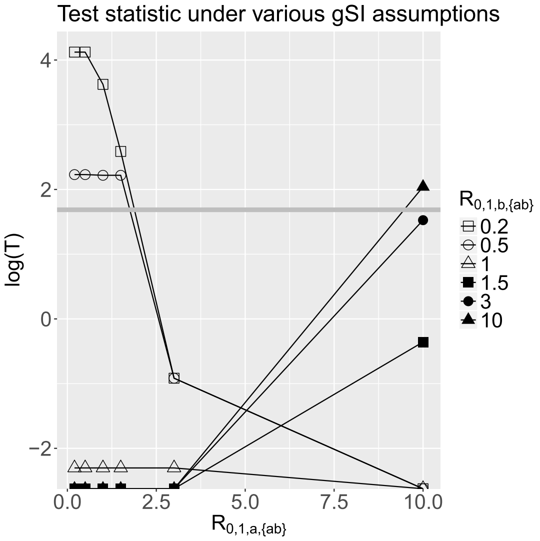


Fig. 5. The figure gives some indication of the test decision for a selection of coarsening scenarios, where the horizontal line marks the critical value. All other lines represents the value of the test statistic in dependence of $R_{0,1,a,\{a,b\}}$ for a given value of $R_{0,1,b,\{a,b\}}$, where only 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.

Table 3
Dimensions in case of $k = 3$ values of Y and $m = 3$ subgroups.

θ_{lat}^{SI}	$dim(\Theta_{lat}^{SI})$
$\pi_{0a}, \pi_{0b}, \pi_{1a}, \pi_{1b}, \pi_{2a}, \pi_{2b}$ $q_{\{a,b\}\|0a}, q_{\{a,b\}\|0b}, q_{\{a,c\}\|0a}, q_{\{a,c\}\|0c}, q_{\{b,c\}\|0b}, q_{\{b,c\}\|0c},$ $q_{\{a,b,c\}\|0a}, q_{\{a,b,c\}\|0b}, q_{\{a,b,c\}\|0c}$	$6\ (= k \cdot (m-1))+$ $9\ (= m \cdot (2^{m-1} - 1))$
θ_{obs}	$dim(\Theta_{obs})$
$p_{0\{a\}}, p_{0\{b\}}, p_{0\{c\}}, p_{0\{a,b\}}, p_{0\{a,c\}}, p_{0\{b,c\}},$ $p_{1\{a\}}, p_{1\{b\}}, p_{1\{c\}}, p_{1\{a,b\}}, p_{1\{a,c\}}, p_{1\{b,c\}},$ $p_{2\{a\}}, p_{2\{b\}}, p_{2\{c\}}, p_{2\{a,b\}}, p_{2\{a,c\}}, p_{2\{b,c\}}$	$18\ (= k \cdot (2^m - 2))$

cf. (4)) is represented by a point falling below the line, so 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, is above the line, resulting in a rejection of H_0. Interpreting the dependencies depicted in Fig. 5 as a whole, the null hypothesis is rejected if both ratios are jointly either relatively small or large. 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 $R_{x,x',y,\mathfrak{y}}$ 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. $R_{0,1,a,\{a,b\}} \in [0.2,\ 1.5]$ and $R_{0,1,b,\{a,b\}} \in [0.2,\ 0.5]$ (cf. Fig. 5).[15]

8. Non-binary data: illustrations and discussion of limitations

Despite the general representation of all results of this paper, in the context of the illustration we focused on a binary setting, reducing to the missing data problem. To make the coarse data structure clearly visible, we briefly exemplify more general categorical settings now. Thereby, we start by considering a response variable with three possible values, i.e. $\Omega_Y = \{a, b, c\}$, e.g. denoting three income categories that are either precisely observed, partly observed or completely unobserved.[16] To make the parameters identifiable under SI and to guarantee testability, according to Equation (11), at least three subgroups are required in this case, such that the here considered covariate “receipt of UBII” would not be sufficient,

[15] This idea of testing on partial assumptions reminds of the hypothesis test by Nordheim [21], who formalized hypotheses about the latent variable distribution (not about the coarsening parameters) and included $R_{x,y,y'\mathfrak{y}}$ (not $R_{x,x',y,\mathfrak{y}}$) into the respective test statistic.

[16] The questioning technique in the PASS data leads to data of that kind obtaining for instance coarse categorical data with a value like “either (< 500 €) or (≥ 500 € and ≤ 1000 €)” induced by a nonresponse to a later question (also cf. [24]).

Table 4
Minimum number of subgroups k for a given m.

m	2	3, 4, 5	6, 7	8, 9	...	20	...	50	...
minimum k	2	3	4	5	...	11	...	26	...

but a covariate as e.g. “age” with categories “≤ 30 years”, “> 30 and ≤ 40 years”, “> 40 years” could be employed. Using this covariate, i.e. $x \in \{0, 1, 2\}$, we obtain θ_{lat}^{SI} and θ_{obs} and the dimensions of Θ_{lat} and Θ_{obs} as given in Table 3. In this way, we compare the test statistic determined by (14) to the $(1-\alpha)$-quantile of $\chi^2_{df^{SI}}$ with $df^{SI} = 18 - 15$.

The minimum number of subgroups already turned out to be a restriction that should not be neglected. A first impression about the minimum number of necessary subgroups can be gained by considering Table 4 (also cf. the condition in (16)). If a rather high number of categories of the response variable were possible, as e.g. $m = 20$, already eleven subgroups would be necessary, and an explosion of the number of parameters would follow (in this case $dim(\Theta_{lat}^{SI}) = 10,485,949$ and $dim(\Theta_{obs}) = 11,534,314$).

Nevertheless, from a practical viewpoint both points do not have to be regarded as a dramatic limitation: Firstly, in the context of survey questionnaires, most categorical variables reveal a small number of categories, thus mostly regarding cases with a very small k. Secondly, in cases of a comparably high number of categories of Y the observed values $\mathfrak{y}$ might rather be in $\tilde{\Omega}_{\mathcal{Y}} \subsetneq \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$, where in most practical situations $|\tilde{\Omega}_{\mathcal{Y}}|$ might even be remarkably reduced compared to the cardinality of $\Omega_{\mathcal{Y}}$ considered here (also cf. Section 4.4). Thus, the number of parameters that have to be estimated substantially reduces. Thirdly, most surveys mainly provide categorical data, such that the inclusion of several categorical covariates should be reasonable in most cases, inducing a remarkable increase[17] of the available subgroups. Nevertheless, especially in rather small datasets, a high number of subgroups may induce the drawback of observing only few units per subgroups. Thus, giving confidence intervals is of great importance to communicate the uncertainty arising from this point.

9. Conclusion

We studied the (non-)testability of the dual assumptions CAR and SI, as well as the extended assumptions gCAR and gSI, in a categorical setting. 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. Moreover, we elaborated the criterion of the minimum number of subgroups required to obtain also point-valued estimators in the case of SI and gSI at all. The estimates of the 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 specific case where the calculation of the critical value has to be based on a mixture distribution, referring to the common χ^2-distribution with the number of degrees of freedom achieved in the estimation problem under SI is appropriate in all other cases (cf. Section 8). 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, 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 cannot 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 can be shown to be incompatible with some data situations since SI may require too strong coarsening rules for each given subgroup, which means that it is testable.

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 set-valued estimator does not respond sensitively. 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 estimator in (7).

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, and also of the two anonymous reviewers of the SMPS submission [25] underlying this paper. The first author thanks the LMU Mentoring program, providing financial support for young, female researchers.

[17] For instance considering three binary covariates (coded by 0 and 1 respectively), would already lead to $2^3 = 8$ subgroups, obtained by splitting by “0,0,0”, “0,0,1”, ..., “1,1,1”.

Appendix A. Visual depiction of H_0^* over a grid of parameters (cf. Section 6.2)

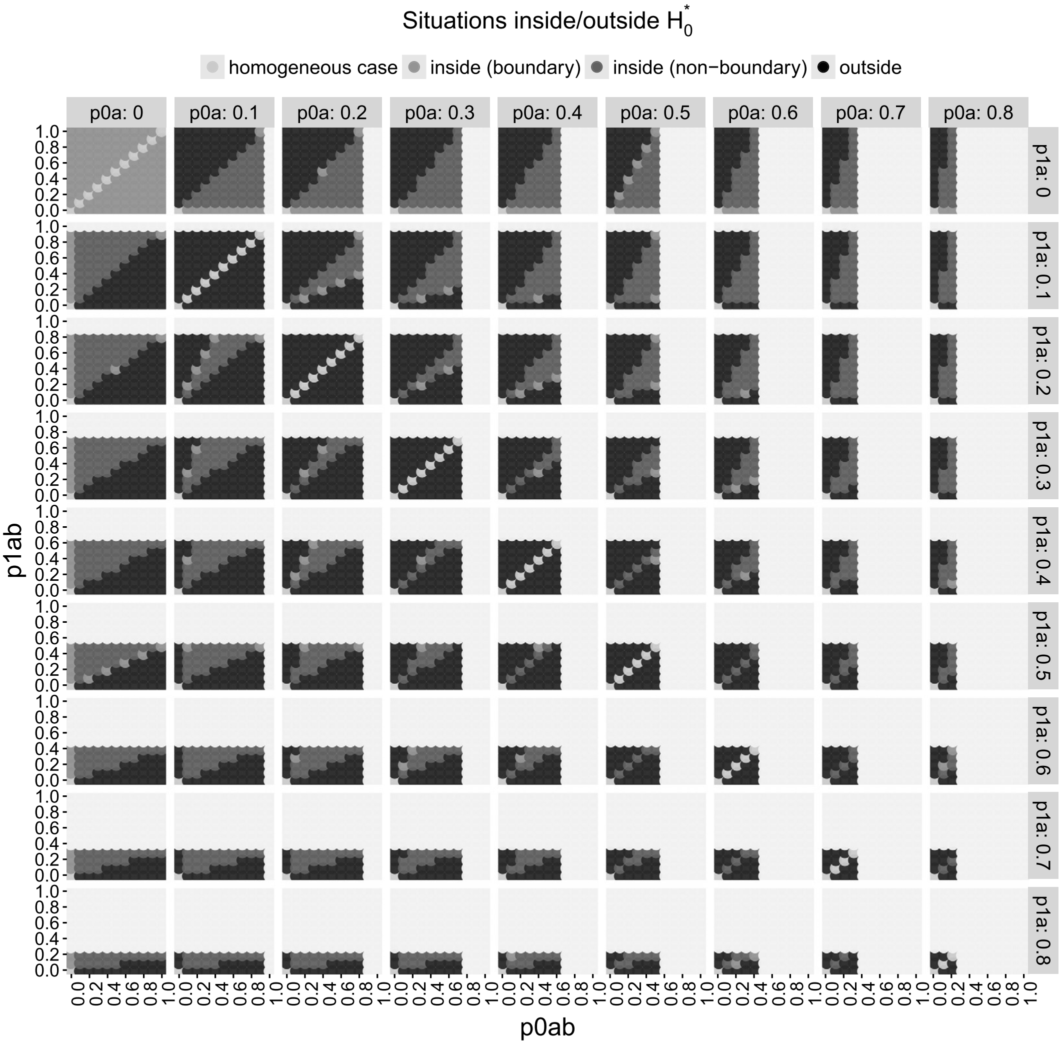


Fig. 6. On a grid of values for the observed variable distribution different cases are distinguished: While the boundary case contains all combinations with either $p_{0\{a\}} \cdot p_{1\{a,b\}} = p_{1\{a\}} \cdot p_{0\{a,b\}}$ or $p_{0\{b\}} \cdot p_{1\{a,b\}} = p_{1\{b\}} \cdot p_{0\{a,b\}}$, joint equality is attained in the i.i.d. case. Moreover, it is differentiated between combinations that are (non-boundary) inside and outside H_0^*. Impossible cases, where the sum of probabilities exceeds one, are not marked by points.

References

[1] 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.

[2] C. Breunig, Testing missing at random using instrumental variables, J. Bus. Econ. Stat. (2017), http://dx.doi.org/10.1080/07350015.2017.1302879, accepted for publication.

[3] M. Cattaneo, A. Wiencierz, Likelihood-based imprecise regression, Int. J. Approx. Reason. 53 (2012) 1137–1154.

[4] H. Chernoff, On the distribution of the likelihood ratio, Ann. Math. Stat. 25 (1954) 573–578.

[5] I. Couso, D. Dubois, Statistical reasoning with set-valued information: ontic vs. epistemic views, Int. J. Approx. Reason. 55 (2014) 1502–1518.

[6] 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, in: Intelligent Systems and Computing Series, Springer, 2016, pp. 141–148.

[7] T. Denœux, Likelihood-based belief function: justification and some extensions to low-quality data, Int. J. Approx. Reason. 55 (2014) 1535–1547.

[8] M. Di Zio, B. Vantaggi, Partial identification in statistical matching with misclassification, Int. J. Approx. Reason. 82 (2016) 227–241.

[9] R. Gill, P. Grünwald, An algorithmic and a geometric characterization of coarsening at random, Ann. Stat. 36 (2008) 2409–2422.

[10] 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.

[11] P. Gustafson, Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments, CRC Press, 2003.

[12] D. Heitjan, D. Rubin, Ignorability and coarse data, Ann. Stat. 19 (1991) 2244–2253.

[13] E. Hüllermeier, Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization, Int. J. Approx. Reason. 55 (2014) 1519–1534.

[14] M. Jaeger, Ignorability for categorical data, Ann. Stat. 33 (2005) 1964–1981.

[15] 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, in: Lecture Notes in Artificial Intelligence, Springer, 2006, pp. 671–678.

[16] M. Kenward, E. Goetghebeur, G. Molenberghs, Sensitivity analysis for incomplete categorical data, Stat. Model. 1 (2001) 31–48.

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

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

[19] F. Molinari, Partial identification of probability distributions with misclassified data, J. Econom. 144 (2008) 81–117.

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

[21] E. Nordheim, Inference from nonrandomly missing categorical data: an example from a genetic study on Turner's syndrome, J. Am. Stat. Assoc. 79 (1984) 772–780.

[22] 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.

[23] J. Plass, T. Augustin, M. Cattaneo, G. Schollmeyer, Statistical Modelling Under Epistemic Data Imprecision, Tech. rep. LMU Munich, 2017, http://jplass.userweb.mwn.de/forschung.html.

[24] J. Plass, T. Augustin, M. Cattaneo, G. Schollmeyer, C. Heumann, Reliable categorical regression analysis for non-randomly coarsened data, preliminary version of a technical report available at http://jplass.userweb.mwn.de/forschung.html, 2017.

[25] 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, in: Intelligent Systems and Computing Series, Springer, 2016, pp. 415–422.

[26] J. Plass, A. Omar, T. Augustin, Towards a cautious modelling of missing data in small area estimation, in: ISIPTA '17, Proceedings of Machine Learning Research 62, 2017, accepted for publication.

[27] C. Rao, Maximum likelihood estimation for the multinomial distribution, Sankhya 18 (1957) 139–148.

[28] G. Schollmeyer, T. Augustin, Statistical modeling under partial identification: distinguishing three types of identification regions in regression analysis with interval data, Int. J. Approx. Reason. 56 (2015) 224–248.

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

[30] R. Tourangeau, T. Yan, Sensitive questions in surveys, Psychol. Bull. 133 (2007) 859–883.

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

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

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

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

PMLR: Proceedings of Machine Learning Research, vol. 62, 253-264, 2017 ISIPTA '17

Towards a Cautious Modelling of Missing Data in Small Area Estimation

Julia Plass JULIA.PLASS@STAT.UNI-MUENCHEN.DE
Aziz Omar AZIZ.OMAR@STAT.UNI-MUENCHEN.DE
Thomas Augustin AUGUSTIN@STAT.UNI-MUENCHEN.DE
Department of Statistics, LMU Munich,
Germany (Plass, Omar, Augustin)
Department of Mathematics, Insurance and Applied Statistics, Helwan University
Egypt (Omar)

Abstract

In official statistics, the problem of sampling error is rushed to extremes when not only results on sub-population level are required, which is the focus of Small Area Estimation (SAE), but also missing data arise. When the nonresponse is wrongly assumed to occur at random, the situation becomes even more dramatic, since this potentially leads to a substantial bias. Even though there are some treatments jointly considering both problems, they are all reliant upon the guarantee of strong assumptions on the missingness. For that reason, we aim at developing cautious versions of well known estimators from SAE by exploiting the results from a recently suggested likelihood approach, capable of including tenable partial knowledge about the nonresponse behaviour in an adequate way. We generalize the synthetic estimator and propose a cautious version of the so-called LGREG-synthetic estimator in the context of design-based estimators. Then, we elaborate why the approach above does not directly extend to model-based estimators and proceed with some first studies investigating different missingness scenarios. All results are illustrated through the German General Social Survey 2014, also including area-specific auxiliary information from the German Federal Statistical Office's data report.

Keywords: small area estimation; LGREG-synthetic estimator; missing data; partial identification; sensitivity analysis; likelihood; logistic regression; logistic mixed model; German General Social Survey.

1. Introduction

Survey methodology distinguishes between sampling and non-sampling error (cf., e.g., Biemer, 2010). Sampling error occurs when only a subset, but not the whole population can be included in a survey, yet the aim is to generalize the results beyond the units that have been sampled. Sampling error is especially severe if the population is composed of several sub-populations and the samples drawn from these sub-populations are not large enough to permit a satisfying precision on sub-population level. A set of methods has been introduced to tackle such situations and is referred to as *Small Area Estimation* (SAE). The main approach of SAE is to use additional data sources, such as administrative records and census data, as auxiliary data in an attempt to increase the effective sample size (cf., e.g., Münnich et al., 2013; Rao and Molina, 2015).

A common non-sampling error encountered in inference is item-nonresponse. Applying the EM-algorithm and Multiple Imputations are the recent practices (cf., e.g., Little and Rubin, 2014). Both techniques force point-identifiability, i.e. uniqueness of parameters, by requiring the assumption that the missingness is occurring randomly (MAR), i.e. independently of the true underlying value of the variable of interest given covariates. Since the MAR assumption is generally not testable and wrongly imposing it may cause a substantial bias, results have to be treated with caution.

According to the methodology of partial identification in the spirit of Manski (2003), one does not have to insist on strong assumptions to obtain a result at all. Allowing for partially identified parameters enables to incorporate tenable knowledge only. In this way, one receives imprecise – but credible – results, which are refined if additional knowledge about the missingness is available. In this context, there are already several approaches refraining from strong assumptions on the missingness process (cf., e.g., Couso and Dubois, 2014; Denœux, 2014). These cautious procedures also represent a popular field of research of the ISIPTA symposia (cf.,e.g., Cattaneo and Wiencierz, 2012; Schollmeyer and Augustin, 2015; Utkin and Coolen, 2011). Since we may not conjure information about the missingness process or make other strong modelling assumptions (cf., e.g., Couso and Sánchez, 2016; Hüllermeier, 2014), uncertainty due to nonresponse has to be interpreted as lack of knowledge. Thus, approaches, explicitly communicating the associated uncertainty, are indispensable. In the context of official statistics this point was recently stressed by Manski (2015).

Since nonresponse may seriously reduce the already small sample size in SAE jointly considering both issues is especially challenging. As far as we know, already existing approaches dealing with nonresponse in SAE are based on strong assumptions on the missingness process, as MAR or the missing not at random (NMAR) assumption plus strict distributional assumptions. Thus, considering a cautious approach for dealing with nonresponse in SAE represents the core of this paper. To pursue this goal, in Section 2 we start by introducing the notation for the setting considered here followed by an introduction to our application using the German General Social Survey. Afterwards, we give a basic overview about prominent design-based estimators applicable in our situation in Section 3. Two design-based estimators, the classical synthetic estimator and the LGREG-synthetic estimator, are generalized in Section 4. While cautious versions are given for the case of including no missingness assumptions at all, the case of including weak assumptions is considered for both estimators by relying on the cautious likelihood approach developed in Plass et al. (2015). In Section 5 the results are illustrated by means of the application example. In Section 6 we discuss why our approach cannot be directly extended to prominent model-based estimators and then perform a first sensitivity analysis. Section 7 concludes by summarizing the major points and giving some remarks on further research.

2. Setting

Technically, our setting is as follows: Let the population U under study have a total size of N units, and be divided into M non-overlapping domains (areas) U_i, each containing units j, $j = 1, \dots, N_i$ with N_i as the size of U_i, $i = 1, \dots, M$. Let Y be a binary variable of interest that is assumed to have a relation with a set of k precisely observed categorical covariates $X_1, \dots, X_k$ through a certain model. Cross classifying the categorical covariates forms a k-dimension table with a total number of cells v, where the g-th cell – representing the g-th subgroup of the population – contains known joint absolute frequency $X_i^{[g]}$, $g = 1, \dots, v$, $i = 1, \dots, M$. To infer about π_i, the probability of a certain category of Y in area i, a sample s of size n is selected, such that a sample s_i of size n_i is selected from area i with $\sum_{i=1}^{M} n_i = n$. Within s_i, sample units j, $j = 1, \dots, n_i$ ($j \in s_i$) are selected with inclusion probability $1/w_{ij}$, where w_{ij} are the usual sample weights. Sample values of the covariates, denoted by $x_{1ij}, \dots, x_{kij}$, are assumed to be completely observed, while of sample values of Y, denoted by y_{ij}, some are missing. Accordingly, s_i is partitioned into $s_{i,obs}$ and $s_{i,mis}$ that refer to sample units with observed and unobserved values of Y, respectively. If we additionally split by g, the samples are denoted by $s_i^{[g]}$, $s_{i,obs}^{[g]}$ and $s_{i,mis}^{[g]}$.

Application example: To illustrate the setting (and later on the results), we rely on the German General Social Survey (GGSS) (GESIS Leibniz Institute for the Social Sciences, 2016). We are interested in the area-specific ratio of people at risk of poverty, where German federal states are the areas completely partitioning the overall domain "Germany" (i.e. $M = 17$)[1]. We construct a binary response variable with values "poor" and "rich" by comparing the collected equivalent income measured on the OECD modified scale with the poverty risk threshold given by 60% of the median net equivalent income, i.e. 986.65€ for year 2014 (DESTATIS, Statistisches Bundesamt, 2016b). The poverty variable shows 454 missing values. As covariates, we use the highest school leaving certificate, which – for ease of presentation – is dichotomized, distinguishing between categories "no Abitur"[2] and "Abitur" only, as well as sex.[3] We base the analysis on the sample with $|s| = 3466$, $|s_{obs}| = 3012$, $|s_{mis}| = 454$. The German Federal Statistical Office' data report (DESTATIS, Statistisches Bundesamt, 2016a) provides area-specific totals $X_i^{[g]}$, $i = 1, \ldots, M$, $g = 1, \ldots, v$, split by the values of the covariates, i.e. the absolute frequencies of the four subgroups "male-no Abitur", "male-Abitur", "female - no Abitur " and "female - Abitur" in area i.

3. Theoretical Background of Design-Based Estimators

SAE techniques result in producing estimators $\hat{\pi}_i$ for area of interest i, $i = 1, \ldots, M$, that are either design-based or model-based.[4] In this paper, we mainly refer to design-based estimators, while we consider model-based ones in Section 6 only. Design-based estimators are either direct estimators that only use data from the targeted area, or indirect estimators that rely on data from other areas as well. This is justified under the assumption of similarity between the areas made to *borrow strength* from other areas.

The Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) $\hat{\pi}_{i,HT} = \frac{1}{N_i}\sum_{j=1}^{n_i} w_{ij}y_{ij}$ for an area i, well known in sampling theory, provides a method to estimate the mean of subpopulation (area) i, thereby accounting for the different sampling probabilities of respondents by sampling weights. The so-called *synthetic estimator* from SAE is a design-based indirect estimator, which is built upon the HT estimator, incorporating not only information from the area of interest, but averaging over all M areas. Thus, the area specific probability π_i is estimated as

$$\hat{\pi}_{i,\text{SYN}} \equiv \hat{\pi}_{\text{SYN}} = \frac{1}{N}\sum_{i=1}^{M}\sum_{j\in s_i} w_{ij}y_{ij} = \frac{1}{N}\sum_{i=1}^{M} N_i \cdot \hat{\pi}_{i,HT} \ , \ \forall i = 1, \ldots, M \ . \tag{1}$$

Since there is no distinction between areas and sample information is included about the response variable only, it merely serves as a basis for further estimators.

[1] Although Germany is divided into 16 federal states, the GGSS differentiates between 17 ones, additionally distinguishing between "former East-Berlin" and "former West-Berlin".

[2] The "Abitur" is the general qualification for university entrance in Germany.

[3] Since there should not be any regional differences with regard to covariate sex, the reason for the inclusion of this covariate rather lies in the interest of illustrating the subgroup specific analysis in a proper way than in an increase of explanatory power in the subject matter context.

[4] While properties of design-based estimators (e.g. bias and variance) are evaluated under sampling distribution over all samples with population parameters held fixed, model-based estimators usually condition on the selected sample, and inference regarding them is carried out with respect to the underlying model (cf., e.g., Rao and Molina, 2015).

An estimator that employs sample data as well as area specific auxiliary information on the joint totals $X_{1i}, \ldots, X_{ki}$ is the GREG-synthetic estimator (cf. Särdnal et al., 1992), where we here use its logistic version, the *LGREG-synthetic estimator* (cf. Lehtonen and Veijanen, 1998). Applying the LGREG-synthetic estimator is split into two steps:

First, the regression coefficients β_0, β_1,. . ., β_k are estimated by means of a standard logistic regression model linking π_{ij}, i.e. the probability for individual j, $j = 1, \ldots, n_i$ in s_i, $i = 1, \ldots, M$, to have the value $y_{ij} = 1$, to the linear predictor containing the individual auxiliary information, here always assuming that all interactions are incorporated.[5] Referring to the application example, we consider two covariates, hence the model includes β_0, β_1, β_2 and an interaction $\beta_{1:2}$, expressing the joint effect of both covariates. According to the aim of borrowing strength, one obtains global regression coefficients. From the estimated global regression coefficients, by applying the response function of a standard logistic regression model, we receive global predictions that only depend on the values of the covariate, but are independent of the area. To stress this, we write $\hat{\pi}^{[g]}$, $g = 1, \ldots, v$, instead of $\hat{\pi}_{ij}$ in our case of categorical covariates. The calculation of these predictions becomes simpler here: Due to the strict monotonicity of the response function, the categorical nature of the covariates and the inclusion of all interactions, a unique relation between the regression coefficients and the predictions can be shown (as, e.g., addressed in Plass et al., 2017). Consequently, we can directly calculate the subgroup specific predictions by

$$\hat{\pi}^{[g]} = \sum_{i=1}^{M} \sum_{j \in s_i^{[g]}} \frac{y_{ij}}{n^{[g]}} \,, \tag{2}$$

with $n^{[g]}$ denoting the cell-count in subgroup g, $g = 1, \ldots, v$.

Second, area-specific information is used: In our setting, the original LGREG-estimator (cf., e.g., Lehtonen and Veijanen, 1998, p.52) for a certain area of interest i can be expressed as

$$\hat{\pi}_{i,LGREG} = \sum_{g=1}^{v} \Big(\overbrace{\sum_{j \in s_{i,g}} w_{ij} y_{ij}}^{\text{HT-part}} + \overbrace{\hat{\pi}^{[g]} \, (X_i^{[g]} - \sum_{j \in s_{i,g}} w_{ij})}^{\text{correction term}} \Big) / N_i \,. \tag{3}$$

It can be understood as the HT estimator corrected by a term accounting for under- and overrepresentation of certain constellations of covariates in the sample, present in case of $X_i^{[g]} > \sum_{j \in s_{i,g}} w_{ij}$ and $X_i^{[g]} < \sum_{j \in s_{i,g}} w_{ij}$, resprectively. The subgroup specific representation in (3) will turn out to be beneficial in context of developing a cautious version (cf. Section 4.2 and 4.3).

4. Cautious Versions of Design-based Estimators under Nonresponse

Since the already established ways of dealing with nonresponse in SAE require strong assumptions, we aim at improving the presented prominent estimators by striving for a proper reflection of the available information on the missingness process. For this purpose, we use the framework of the cautious approach developed for the more general case of coarse[6] categorical data in Plass et al.

[5] This is quite natural in this context, since only then the full information about the subgroup specific information, also provided by the auxiliary information in terms of totals, is used.

[6] The data problem only distinguishes between fully observed and completely unobserved values, while coarse data additionally include partial observations, e.g. in the sense of grouped data (cf. Heitjan and Rubin, 1991).

(2015) and further extended in Plass et al. (2017) to practically frame the inclusion of auxiliary information. We start by recalling the basic elements of this approach in the following section.

4.1 A Cautious Approach for Dealing with Nonresponse

An observation model $\mathcal{Q}$ is used as a medium to frame the procedure of incorporating auxiliary information on the incompleteness. Restricting to the missing data problem and a binary response variable and considering the problem for subgroup g, $g = 1, \ldots, v$, the model $\mathcal{Q}^{[g]}$ is determined by the set of missingness parameters $q_{na|y}^{[g]}$, i.e. the probability associated with refusing the answer ("na"), given a certain subgroup g and the true value $y \in \{0, 1\}$ of the response variable.[7] In the spirit of partial identification, one can start by incorporating "no" assumptions[8] on $q_{na|y}^{[g]}$, then restricting these missingness parameters successively by certain conceivable conditions. The cautious approach includes this observation model into a classical categorical likelihood problem. For this purpose, a connection between the parameters $\pi^{[g]}$ and $p_{\mathfrak{y}}^{[g]}$ is established via the observation model, where $p_{\mathfrak{y}}^{[g]}$ refers to the observed value $\mathfrak{y} \in \{0, 1, na\}$, thus treating the missing values as a category of its own. The invariance of the likelihood allows to rewrite the log-likelihood in terms of $p_{\mathfrak{y}}^{[g]}$, which can be uniquely maximized in terms of the parameters of interest by relying on the theorem of total probability, receiving

$$\begin{aligned} \ell(\pi^{[g]},\, q_{na|0}^{[g]},\, q_{na|1}^{[g]}) = & n_1^{[g]}\Big(\ln(\pi^{[g]}) + \ln(1 - q_{na|1}^{[g]})\Big) + n_0^{[g]}\Big(\ln(1 - \pi^{[g]}) + \ln(1 - q_{na|0}^{[g]})\Big) \\ & + n_{na}^{[g]}\Big(\ln(\pi^{[g]} q_{na|1}^{[g]} + (1 - \pi^{[g]}) q_{na|0}^{[g]})\Big)\ , \end{aligned} \tag{4}$$

where $n_1^{[g]}$, $n_0^{[g]}$ and $n_{na}^{[g]}$ refer to the respective observed cell counts within subgroup g, which later on have to be replaced by appropriate sample weights. By maximizing the log-likelihood in (4), we determine the generally set-valued[9] estimators, whose one-dimensional projections can be represented by the lower and upper bounds of intervals, namely $\underline{\hat{\pi}}^{[g]}$, $\overline{\hat{\pi}}^{[g]}$, $\underline{\hat{q}}_{na|0}^{[g]}$, $\overline{\hat{q}}_{na|0}^{[g]}$, $\underline{\hat{q}}_{na|1}^{[g]}$ and $\overline{\hat{q}}_{na|1}^{[g]}$. Thereby, $\underline{\hat{\pi}}^{[g]}$ is attained under $\overline{\hat{q}}_{na|0}^{[g]}$ and $\underline{\hat{q}}_{na|1}^{[g]}$, while $\overline{\hat{\pi}}^{[g]}$ is associated with $\underline{\hat{q}}_{na|0}^{[g]}$ and $\overline{\hat{q}}_{na|1}^{[g]}$.

By considering $q_{na|1}^{[g]} = R \cdot q_{na|0}^{[g]}$, with missing ratio $R \in \mathcal{R} \subseteq \mathbb{R}_0^+$ (also cf. Nordheim (1984)),[10] and $\mathcal{R}$ as the set of missing ratios, assumptions about the missingness can be incorporated. Specific values of R are associated with a particular missingness scenario, thus point-identifying $\pi^{[g]}$. For instance, $R = 1$ represents the missingness scenario under gMAR[11], requiring $q_{na|1}^{[g]} = q_{na|0}^{[g]}$. Partial (weak) assumptions, like incorporating $R \in \mathcal{R}$ into (4), thus refine the result obtained from the log-likelihood optimization without the inclusion of any missingness assumptions. Since it can be shown that $\underline{\hat{\pi}}^{[g],\mathcal{R}}$, $\overline{\hat{q}}_{na|0}^{[g],\mathcal{R}}$ and $\underline{\hat{q}}_{na|1}^{[g],\mathcal{R}}$ as well as $\overline{\hat{\pi}}^{[g],\mathcal{R}}$, $\underline{\hat{q}}_{na|0}^{[g],\mathcal{R}}$ and $\overline{\hat{q}}_{na|1}^{[g],\mathcal{R}}$, i.e. the bounds under the partial assumptions expressed by $\mathcal{R} = [\underline{R}, \overline{R}]$, are achieved under missingness ratios $\underline{R}$ and $\overline{R}$, respectively, one does not have to optimize the log-likelihood for all values in $[\underline{R},\ \overline{R}]$, but optimizing under $\underline{R}$ and $\overline{R}$ is sufficient. While $\mathcal{R} = [0,\ 1]$ corresponds to $q_{na|1}^{[g]} \leq q_{na|0}^{[g]}$, a cautious version of

[7] Referring to the framework of analyzing contingency tables, it is natural to drop the reference to individual j.

[8] In fact, we confine ourselves to very general assumptions detailed in Plass et al. (2017).

[9] The mapping relating $\hat{\pi}^{[g]}$ to $\hat{p}_{\mathfrak{y}}^{[g]}$ is generally not injective.

[10] Here we consider a different R than in Plass et al. (2015).

[11] Conditioning on subgroup g generalizes the typical MAR assumption.

gMAR is given by $\mathcal{R} = [\max(0,\ 1-\tau),\ 1+\tau], \tau \geq 0$, where the degree of cautiousness is given by the definition of the neighborhood τ (cf. Plass et al., 2017).

4.2 Cautious SAE: Including no Missingness Assumptions

In case of considering $\mathcal{R} = \mathbb{R}_0^+$, i.e. incorporating no assumption on the missingness, the result of the cautious likelihood approach (Plass et al., 2015, p. 251) can be shown to correspond to the one obtained from cautious data completion, plugging in all potential precise sample outcomes compatible with the observations (cf. Augustin et al., 2014, §7.8). Thus, here the lower and upper bound of the synthetic estimator in (1) can be calculated in this case by considering the extreme cases of regarding all missing values as $y_{ij} = 0$, $\forall j \in s_{i,mis}$, $i = 1, \ldots, M$, or all as $y_{ij} = 1$, $\forall j \in s_{i,mis}$, $i = 1, \ldots, M$:

$$\underline{\hat{\pi}}_{i,SYN} = \frac{1}{N}\sum_{i=1}^{M}\sum_{j\in s_{i,obs}} w_{ij}y_{ij}\ , \qquad \overline{\hat{\pi}}_{i,SYN} = \frac{1}{N}\sum_{i=1}^{M}\Big(\sum_{j\in s_{i,obs}} w_{ij}y_{ij} + \sum_{j\in s_{i,mis}} w_{ij}\Big)\ . \tag{5}$$

In order to study the bounds $\underline{\hat{\pi}}_{i,LGREG}$ and $\overline{\hat{\pi}}_{i,LGREG}$, it turns out to be beneficial to break the summation over all areas into a term for area i^* [12] of interest and a summation over all other areas $i \neq i^*$. With the regularity condition that sampling weights within area i are equal such that $w_{ij} = w_i, \forall j = 1, \ldots, n_i$, and defining $n^{[g]}$ and $n_i^{[g]}$ to be respectively the number of units in s and s_i existing in subgroup g, $g = 1, \ldots, v$, $i = 1, \ldots, M$, we can rewrite $\hat{\pi}_{i^*,LGREG}$ in (3) as

$$\sum_{g=1}^{v}\Bigg(\Big(\sum_{\substack{i=1\\ i\neq i^*}}^{M}\sum_{j\in s_i^{[g]}}\frac{y_{ij}}{n^{[g]}}\Big)\Big(X_{i^*}^{[g]} - n_{i^*}^{[g]}w_{i^*}\Big) + \sum_{j\in s_{i^*}^{[g]}}\frac{y_{i^*j}}{n^{[g]}}\Big(X_{i^*}^{[g]} - w_{i^*}(n_{i^*}^{[g]} + n^{[g]})\Big)\Bigg)/N_{i^*}\ , \tag{6}$$

$$\text{with} \sum_{j\in s_i^{[g]}}\frac{y_{ij}}{n^{[g]}} = \sum_{j\in s_{i,obs}^{[g]}}\frac{y_{ij}}{n^{[g]}} + \sum_{j\in s_{i,mis}^{[g]}}\frac{y_{ij}}{n^{[g]}} \quad \text{and} \quad \sum_{j\in s_{i^*}^{[g]}}\frac{y_{i^*j}}{n^{[g]}} = \sum_{j\in s_{i^*,obs}^{[g]}}\frac{y_{i^*j}}{n^{[g]}} + \sum_{j\in s_{i^*,mis}^{[g]}}\frac{y_{i^*j}}{n^{[g]}}\ ,$$

when missing data are included. The problem consists of finding the values of y_{ij} for the nonrespondents that minimize (maximize) Equation (6). Since Equation (6) is a sum of subgroup specific quantities, optimization for each subgroup g, $g = 1, \ldots, v$, separately is sufficient. Provided that $X_{i^*}^{[g]} \geq n_{i^*}^{[g]}w_{i^*}$, we can directly infer that the term referring to the areas $i \neq i^*$ is minimized (maximized) if all the y_{ij}'s, $j \in s_{i,mis}$ are equal to zero (one). Otherwise, the other extreme allocation of zeros and ones is chosen to obtain the minimum (maximum). Analogous considerations can be accomplished in the term associated with area i^*, now based on the condition $X_{i^*}^{[g]} \geq w_{i^*}(n_{i^*}^{[g]} + n^{[g]})$.

4.3 Cautious SAE: First Attempts to Include (Partial) Missingness Assumptions

When partial assumptions in the sense of $R \in [\underline{R}, \overline{R}]$ are tenable, it is useful to express the cautious synthetic estimator and the LGREG-synthetic estimator in terms of $\hat{\pi}^{\mathcal{R}}$, $\hat{q}^{\mathcal{R}}_{na|0}$ and $\hat{q}^{\mathcal{R}}_{na|1}$ obtained by optimizing a log-likelihood as given in (4) under the constraints expressed by R. By again splitting

[12] Whenever a differentiation between quantities summing up over all regions and quantities referring to a specific region is needed, we explicitly write i^* for the region under consideration.

$j \in s_i$ into $j \in s_{i,obs}$ and $j \in s_{i,mis}$, the lower bound for the synthetic estimator is received as[13]

$$\underline{\hat{\pi}}^{\mathcal{R}}_{SYN} = \frac{1}{N} \sum_{i=1}^{M} \Big(\sum_{j \in s_{i,obs}} w_{ij} y_{ij} + \underline{\hat{q}}^{\mathcal{R}}_{na|i1} \cdot \underline{\hat{\pi}}^{\mathcal{R}}_{i} \cdot \sum_{j \in s_i} w_{ij} \Big) \,, \tag{7}$$

where $\underline{\hat{q}}^{\mathcal{R}}_{na|i1} \cdot \underline{\hat{\pi}}^{\mathcal{R}}_{i} \cdot \sum_{j \in s_i} w_{ij}$ is the – here smallest – estimated weighted number of nonrespondents with $y_{ij} = 1$, $j \in s_{i,mis}$, under the missingness assumption in focus. Thereby, the included estimators are received by refraining from a subgroup specific consideration, thus regarding $\ell(\pi^{\mathcal{R}},\ q^{\mathcal{R}}_{na|0},\ q^{\mathcal{R}}_{na|1})$ instead of $\ell(\pi^{[g],\mathcal{R}},\ q^{[g],\mathcal{R}}_{na|0},\ q^{[g],\mathcal{R}}_{na|1})$ (cf. (4)). Analogously, $\overline{\hat{\pi}}^{\mathcal{R}}_{SYN}$ is achieved by using $\overline{\hat{q}}^{\mathcal{R}}_{na|i1}$ and $\overline{\hat{\pi}}^{\mathcal{R}}_{i}$ within (7).

To derive the cautious LGREG-synthetic estimator described by $\underline{\hat{\pi}}^{\mathcal{R}}_{i^*,LGREG}$ and $\overline{\hat{\pi}}^{\mathcal{R}}_{i^*,LGREG}$, we base our presentation on the lower bound, while the upper bound is obtained analogously vice versa. Basically, there are two ways to generalize the LGREG-synthetic estimator to a cautious version: One could either consider one overall likelihood or make consistent use of the fact that the LGREG-synthetic estimator is a combination of two estimators, a global one motivated by the idea of "borrowing strength" and another one referring to area i^*. Here, we address the second possibility, while the first one should be studied in further research. For this purpose, we start by maximizing two log-likelihoods, namely $\ell(\pi^{[g],\mathcal{R}},\ q^{[g],\mathcal{R}}_{na|0},\ q^{[g],\mathcal{R}}_{na|1})$ and $\ell(\pi^{[g],\mathcal{R}}_{i^*},\ q^{[g],\mathcal{R}}_{na|i^*0},\ q^{[g],\mathcal{R}}_{na|i^*1})$, under $\underline{R}$ and $\overline{R}$ to derive the respective projections of the generally set-valued estimators. In a next step, we then approach the calculation of $\underline{\hat{\pi}}^{\mathcal{R}}_{i^*,LGREG}$ by including those estimators that minimize

$$\sum_{g=1}^{v} \Bigg(\overbrace{\sum_{j \in s^{[g]}_{i^*,obs}} w_{i^*} y_{i^*j} + \hat{q}^{[g],\mathcal{R}}_{na|i^*1} \hat{\pi}^{[g],\mathcal{R}}_{i^*} \cdot \sum_{j \in s^{[g]}_{i^*}} w_{i^*j}}^{\text{HT-part}} + \overbrace{\hat{\pi}^{[g],\mathcal{R}} (X^{[g]}_{i^*} - n^{[g]}_{i^*} w_{i^*})}^{\text{correction term}} \Bigg) / N_{i^*} \,, \tag{8}$$

which is a version of the classical LGREG-synthetic estimator in Equation (3), where the HT-part is represented in terms of $\hat{\pi}^{[g],\mathcal{R}}_{i^*}$ and $\hat{q}^{[g],\mathcal{R}}_{na|i^*1}$, guaranteeing for the partial assumptions under consideration. Due to the distinct estimation of $\pi^{[g]}$ and $\pi^{[g]}_{i^*}$, we now try to take the associated dependence into account: The interrelation between both estimators may be clearly inferred by considering the representations

$$\hat{\pi}^{[g]}_{i^*} = \Big(\sum_{j \in s^{[g]}_{i^*}} y_{ij} \Big) / n^{[g]}_{i^*} \quad \text{and} \quad \hat{\pi}^{[g]} = \Big(\sum_{\substack{i=1 \\ i \neq i^*}}^{M} \sum_{j \in s^{[g]}_{i}} y_{ij} + \sum_{j \in s^{[g]}_{i^*}} y_{ij} \Big) / n^{[g]} \tag{9}$$

(here for ease of representation given without splitting into $s_{i,obs}$ and $s_{i,mis}$), both including respondents from area i^*.[14] Whenever $X^{[g]}_{i^*} > n^{[g]}_{i*}$, we achieve $\underline{\hat{\pi}}^{\mathcal{R}}_{i^*,LGREG}$ if $\underline{\hat{\pi}}^{[g],\mathcal{R}}_{i^*}$, $\underline{\hat{q}}^{[g],\mathcal{R}}_{na|i^*1}$, $\underline{\hat{\pi}}^{[g],\mathcal{R}}$ are taken in (8). This choice is possible in this case, since individuals $j \in s^{[g]}_{i^*}$ are assumed to have the

[13] For more details see the preliminary version of a technical report available at http://jplass.userweb.mwn.de/forschung.html.

[14] While in (6) a splitting into terms for area i^* and areas $i \neq i^*$ was achieved, this cannot be accomplished here. Note that $\sum_{\substack{i=1 \\ i \neq i^*}}^{M} \sum_{j \in s^{[g]}_{i}} \frac{y_{ij}}{n^{[g]}}$ and $\sum_{j \in s^{[g]}_{i^*}} \frac{y_{i^*j}}{n^{[g]}}$, appearing in Equation (6), are different from (9) and cannot be regarded as estimated probabilities due to the different reference in numerator and denominator.

same values within both estimated probabilities in (9). Considering the situation of $X_{i^*}^{[g]} < n_{i*}^{[g]}$, this is not the case. While $\hat{\pi}^{[g],\mathcal{R}}$ is supposed to be maximal, $\hat{\pi}_{i^*}^{[g],\mathcal{R}}$ and $\hat{q}_{na|i^*1}^{[g],\mathcal{R}}$ should be minimal to minimize (8). To proceed, we give a reasonable way out of this situation. Thereby, we distinguish between the case (i), where the correction term in (8) is of greater importance compared to the HT-part and case (ii), considering the opposite situation.

Case (i): The lower bound of the LGREG-synthetic estimator should be obtained by selecting $\overline{\hat{\pi}}^{[g],\mathcal{R}}$. In this way, for all individuals $j \in s_{i^*}^{[g]}$ the lowest possible scenario compatible with the partial knowledge is assumed, such that the inclusion of $\overline{\hat{\pi}}^{[g],\mathcal{R}}$ and $\overline{\hat{q}}_{na|i^*1}^{[g],\mathcal{R}}$ directly follows. This is supported by Equation (6), indicating that bounds of $\hat{\pi}_{i^*}^{[g],\mathcal{R}}$ are included instead of estimators referring to a scenario between.[15]

Case (ii): $\underline{\hat{\pi}}^{[g],\mathcal{R}}$, $\underline{\hat{\pi}}_{i^*}^{[g],\mathcal{R}}$ and $\underline{\hat{q}}_{na|i^*1}^{[g],\mathcal{R}}$ are incorporated for $\underline{\hat{\pi}}_{i^*,LGREG}^{\mathcal{R}}$, while $\underline{\hat{\pi}}^{[g],\mathcal{R}}$ is improvable by assuming the upper missingness scenario for individuals from $i \neq i^*$. A practical compromise is the inclusion of a pooled estimator

$$\hat{\pi}_{\text{pooled}}^{[g]} = \left(\overline{\hat{\pi}}_{i\neq i^*}^{[g]} \cdot n_{i\neq i^*}^{[g]} + \underline{\hat{\pi}}_{i^*}^{[g]} \cdot n_{i^*}^{[g]}\right) / n^{[g]} \ , \tag{10}$$

to receive $\underline{\hat{\pi}}_{i^*,LGREG}^{\mathcal{R}}$, where $\hat{\pi}_{i\neq i^*}^{[g],\mathcal{R}}$ can also be obtained from the cautious log-likelihood calculated based on all data except from area i^*. Analogously, a pooled version can be determined for the calculation of $\overline{\hat{\pi}}_{i^*,LGREG}^{\mathcal{R}}$.

Because of the under-/overweighting of certain subgroups in the sample, automatically some $(X_{i^*}^{[g]} - n_{i^*}^{[g]} w_{i^*})$ will be positive and others negative, such that the distinction of different cases can not be avoided. The development of a criterion evaluating the "importance" of the HT-term and the correction term used in our argument should be part of further research. Thereby, also the results and conditions from Section 4.2 should be taken into consideration. Up to then, we choose the minimum of the results from case (i) and (ii) to obtain a suggestion for $\underline{\hat{\pi}}_{i^*,LGREG}^{\mathcal{R}}$.

5. Results from the Application Example

The area-specific poverty rate is the focus of our illustration explained in Section 2. Yet, we explicitly avoid making conclusions on the poverty in a substance matter sense, considering this application as a first illustration of technical aspects of the elaborated cautious estimators only. Here, additionally to the case without assuming anything about the missingness process, we studied the weak assumption that rich respondents tend to refuse the income question more often compared to poor ones, i.e. $R \in [0,\ 1]$ (assum. 1), as well as a cautious version of MAR, here incorporating $R \in [0.3,\ 1.7]$ (assum. 2). Although subgroup specific assumptions were feasible in the context of the LGREG-synthetic estimator, we here impose the same missingness assumption on all subgroups.

By applying Equations (7) and (8) to the (weighted) marginal sample data,[16] we can calculate the cautious synthetic estimator and the LGREG-synthetic estimator for the different situations of

[15] From Equation (6) we could conclude that either all or no virtual values $y_{ij}, j \in s_{i^*,mis}$, have to be equal to 1 to obtain $\underline{\hat{\pi}}_{i^*,LGREG}$ and $\overline{\hat{\pi}}_{i^*,LGREG}$ in the case of no assumptions. If partial assumptions are included, this applies in the sense that this does not have to be satisfied for all, but for the minimum/maximum number of virtual values that is consistent with the partial missing assumption ending up with $\underline{\hat{\pi}}_{i^*}^{[g],\mathcal{R}}$or $\overline{\hat{\pi}}_{i^*}^{[g],\mathcal{R}}$.

[16] In the GGSS, respondents from East-Germany are oversampled, such that weights are required in the analysis (0.564 (East Germany), 1.205 (West Germany), cf. Koch et al. (1994)).

	no assum.	assum. 1	assum. 2
$[\underline{\hat{\pi}}_{SYN},\ \overline{\hat{\pi}}_{SYN}]$	$[0.167,\ 0.300]$	$[0.167,\ 0.193]$	$[0.175,\ 0.208]$

Table 1: Bounds for the synthetic estimator under various missingness assumptions

	no assum.		assum. 1		assum. 2	
Federal state	$\underline{\hat{\pi}}_{i,LGREG}$	$\overline{\hat{\pi}}_{i,LGREG}$	$\underline{\hat{\pi}}_{i,LGREG}$	$\overline{\hat{\pi}}_{i,LGREG}$	$\underline{\hat{\pi}}_{i,LGREG}$	$\overline{\hat{\pi}}_{i,LGREG}$
BW	0.129	0.366	0.129	0.210	0.141	0.224
BY	0.088	0.233	0.088	0.133	0.091	0.141
HB	0.077	0.405	0.115	0.193	0.125	0.206
HH	0.009	0.196	0.014	0.075	0.019	0.083

Table 2: Bounds for the LGREG-synthetic estimator under various missingness assumptions

partial knowledge (cf. Table 1 and Table 2, respectively). The practically weak assumptions already induce a remarkable refinement of the intervals obtained under no assumptions.[17]. Due to the separate likelihood optimization that in some cases led us to the pooled version, including different bounds for i^* and $i \neq i^*$, the lower bound from "no assum." and "assum. 1" do not necessarily have to coincide here. This gives rise to an overall likelihood approach that admittedly refrains from "borrowing strength" within the missingness process, but implicitely accounts for interrelations.

6. First Studies Towards a Cautious Model-based Estimator under Nonresponse

Until now, we focused on models dealing with the small sample size by incorporating observations from other areas on the one hand and area-specific auxiliary information on the other hand. To account for between-area variation beyond that explained by auxiliary variables, model-based estimators relying on mixed models establish a basis. Model-based estimators incorporate data from different areas through a model that depends on the level of aggregation of the auxiliary variables. The well known Fay-Herriot (FH) area-level model, introduced by Fay III and Herriot (1979) for linear regression, has been further developed for categorical regression by MacGibbon and Tomberlin (1989). By relying on the logistic mixed model, they include area specific random effects $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ into the linear predictor of a standard logistic regression model. Based on this model, we can make predictions contributing to the final model-based estimators.

Since we aim at applying the cautious likelihood approach, we consider the likelihood in the mixed model context first. Generally, the marginal likelihood of the i-th area is received by averaging over the probability distribution of the random effects u_i (cf., e.g., Booth and Hobert, 1999). Since thereby almost always analytically intractable integrals are involved, numerical methods are required for the maximization. Consequently, the cautious likelihood approach is stretched to the limits of its direct applicability if model-based estimators are of interest.

Nevertheless, we proceed with some studies to get a first impression about the predictions obtained from a mixed model if refrained from strong assumptions on the missingness process. Since

[17] We use the official abbreviations of the federal states, here BW and BY for Baden-Wuerttemberg and Bavaria, and HB and HH for the federal city states (hanse town (H)) Bremen and Hamburg.

the random effects u_i and the regression coefficients are estimated simultaneously with the aid of approximation methods, we can no longer establish a direct connection between the subgroup specific probabilities and the regression coefficients, as we did in Section 4. Hence, we here start with a first sensitivity analysis, estimating $\beta_0, \ldots, \beta_k$ and u_i under different types of missingness mechanisms. Since for a part of our research question, i.e. getting a first impression about the bounds of the estimated random effects, an area-specific missingness behaviour is of high interest, we simplify the databases classifying the federal states into four regions ("northeast",..., "southwest"), thus substantially reducing the scenarios that have to be considered within a corresponding missing type. Moreover restricting to the covariate "Abitur" (yes/no), we investigate the impact of two different missing types over a grid of values: The first missing type requires independence of the covariates, whereas the second type depends on the covariate and the area.

While the estimated random effects tend to show no systematic reaction to different missingness scenarios, the regression estimates[18] attain the bounds in the extreme missingness situations. Consequently, by focusing on the scenarios that either regard all or no missing values as $y_{ij} = 1$, we apparently can at least give an estimator based on the best-worst-case estimation of the regression coefficients, here denoted by $\hat{\pi}^{\beta} \in [\underline{\hat{\pi}}^{\beta}, \overline{\hat{\pi}}^{\beta}]$. For this purpose, we use $\hat{\beta}_0, \ldots, \hat{\beta}_k, \hat{u}_i$ obtained for the extreme cases to determine the individual prediction bounds. Again, in our categorical case it turns out to be sufficient to calculate the bounds of $\hat{\pi}^{[g],\beta}$, now not only split by the values of the covariate, but also the region. Using $\hat{\pi}^{[g],\beta}$ and the area-specific totals $X_i^{[g]}$, the bounds of a model-based estimator, relying on the best-worst estimation of β, can be calculated.

7. Conclusion

By exploiting the cautious likelihood approach (cf. Plass et al., 2015), we considered an opportunity to adapt the LGREG-synthetic estimator for nonresponse, without the need of strict and often practically untenable assumptions about the missingness process. The included observation model is a powerful medium to make use of frequently available, partial assumptions about the missingness, where results from the application example corroborated that very weak assumptions may already suffice to substantially refine the results obtained without the inclusion of any missingness assumptions. Further research should be devoted to a more extensive consideration of the here proposed method characterized by separate likelihood optimizations. Although some first investigations of cautious model-based estimators were accomplished, due to the technically different situation, a more detailed study should be part of future research. In addition, comparing the magnitude of both principally differing sources of uncertainty induced by the problems in focus (i.e. sampling uncertainty as well as lack of knowledge associated to SAE and nonresponse, respectively) is notably worthwhile. For this purpose, uncertainty regions (cf. Vansteelandt et al., 2006), covering both types of uncertainties, should be investigated.

Acknowledgments

The first author thanks the LMUMentoring program, providing financial support for young, female researchers. The second author thanks the government of Egypt and the German Academic

[18] cf. figure in the prelim. version of a technical report mentioned in footnote 9.

Exchange Service (DAAD) for their joint financial support. We are very grateful for the helpful remarks of all three anonymous reviewers and especially appreciate the constructive suggestions of one rather critical reviewer, improving the presentation.

References

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

P. Biemer. Total survey error: Design, implementation, and evaluation. *Public Opin. Q.*, 74:817–848, 2010.

J. Booth and J. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.*, 61:265–285, 1999.

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

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

I. Couso and L. Sánchez. Machine learning models, epistemic set-valued data and generalized loss functions: An encompassing approach. *Inf. Sci. (Ny)*, 358:129–150, 2016.

T. Denœux. Likelihood-based belief function: Justification and some extensions to low-quality data. *Int. J. Approx. Reason.*, 55:1535–1547, 2014.

DESTATIS, Statistisches Bundesamt. Micro-census 2014 – DESTATIS: Results: Federal states, year, sex, general school education, 2016a. https://www.genesis.destatis.de [accessed: 04.02.2017].

DESTATIS, Statistisches Bundesamt. EU-SILC 2014 – DESTATIS: Living conditions, risk of poverty, 2016b. https://www.destatis.de [accessed: 04.02.2017].

R. Fay III and R. Herriot. Estimates of income for small places: an application of James-Stein procedures to census data. *J. Am. Stat. Assoc.*, 74:269–277, 1979.

GESIS Leibniz Institute for the Social Sciences. German General Social Survey – ALLBUS 2014. GESIS Data Archive, Cologne, 2016. ZA5242 Data file Version 1.0.0.

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

D. Horvitz and D. Thompson. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.*, 47(260):663–685, 1952.

E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *Int. J. Approx. Reason.*, 55:1519–1534, 2014.

A. Koch, S. Gabler, and M. Braun. Konzeption und Durchführung der “Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften” (ALLBUS) 1994. *ZUMA-Arbeitsbericht*, 94, 1994.

R. Lehtonen and A. Veijanen. Logistic generalized regression estimators. *Surv. Methodol.*, 24: 51–56, 1998.

R. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2nd edition, 2014.

B. MacGibbon and T. Tomberlin. Small area estimation of proportions via empirical Bayes techniques. *Surv. Methodol.*, 15:237–252, 1989.

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

C. Manski. Credible interval estimates for official statistics with survey nonresponse. *J. Econometrics*, 191:293–301, 2015.

R. Münnich, J. Burgard, and M. Vogt. Small Area-Statistik: Methoden und Anwendungen. *AStA Wirtschafts-und Sozialstatistisches Archiv*, 6:149–191, 2013.

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, 1984.

J. Plass, T. Augustin, M. Cattaneo, and 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, and E. Quaeghebeur, editors, *Proc ISIPTA '15*, pages 247–256. SIPTA, 2015.

J. Plass, T. Augustin, M. Cattaneo, G. Schollmeyer, and C. Heumann. Reliable categorical regression analysis for non-randomly coarsened data. Preliminary version of a technical report available at http://jplass.userweb.mwn.de/forschung.html, 2017.

J. Rao and I. Molina. *Small Area Estimation*. Wiley, 2nd edition, 2015.

C. Särdnal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer, 1992.

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

L. Utkin and F. Coolen. Interval-valued regression and classification models in the framework of machine learning. In F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger, editors, *Proc ISIPTA '11*, pages 371–380. SIPTA, 2011.

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

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, §8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, 29.05.2018
Ort, Datum

Julia Plaß