Testing of coarsening mechanisms: Coarsening at random versus subgroup independence

Julia Plass, Marco E. G. V. Cattaneo, Georg Schollmeyer, and Thomas Augustin

Abstract Since coarse(ned) data naturally induce set-valued estimators, analysts often assume coarsening at random (CAR) to force them to be single-valued. Using the PASS data as an example, we re-illustrate the impossibility to test CAR and contrast it to another type of uninformative coarsening called subgroup independence (SI). It turns out that SI is testable.

Key words: coarse data, missing data, coarsening at random (CAR), hypothesis testing, likelihood-ratio test

1 The problem of testing coarsening mechanisms

Traditional statistical methods dealing with missing data (e.g. EM algorithm, imputation techniques) require identifiability of parameters, which frequently tempts analysts to make the *missing at random* (MAR) assumption ([7]) simply for pragmatic reasons without justifications in substance (e.g. [5]). Since MAR is not testable (e.g. [8]), this way to proceed is especially alarming.
Looking at the problem in a more general way, incomplete observations may be included not only in the sense of missing, but also coarse(ned) data. In this way, additionally to fully observed and unobserved, also partially observed values are considered.[1] In the context of coarse data, the *coarsening*

Julia Plass · Georg Schollmeyer · Thomas Augustin
Department of Statistics, LMU Munich
e-mail: {julia.plass,georg.schollmeyer,augustin}@stat.uni-muenchen.de

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

[1] When dealing with coarse data, it is important to distinguish between epistemic data imprecision, considered here, and ontic data imprecision (cf. [2]).

This is the unformatted version of: doi:10.1007/978-3-319-42972-4_51

at random (CAR) assumption (e.g. [4]) is the analogue of MAR. Although the impossibility of testing CAR is already known from literature, providing an intuitive insight into this point will be a first goal of this paper. Apart from CAR, we focus on another, in a sense dual, assumption that we called *subgroup independence* (SI) in [11]. In our categorical setting (cf. Section 2), SI not only makes parameters identifiable, but is also testable as demonstrated here. Thus, we elaborate the substantial difference in the testability of CAR and SI and start with illustrating both assumptions by a running example based on the PASS data in Section 2 ([14]). In Section 3 we sketch the crucial argument of the estimation and show how the generally set-valued estimators are refined by implying CAR or SI. Testability of both assumptions is discussed in Section 4, where a likelihood-ratio test is suggested for SI.

2 Coarsening models: CAR and SI

Throughout this paper, we refer to the case of a coarse categorical response variable Y and one precisely observed binary covariate X. The results may be easily transferred to cases with more than one arbitrary categorical covariates by using dummy variables and conditioning on the then emerged subgroups. For sake of conciseness, the example refers to the case of a binary Y, where coarsening corresponds to missingness, but all results are applicable in a general categorical setting.

We approach the problem of coarse data in our setting by distinguishing between a latent and an observed world: A random sample of a categorical response variable $Y_1, \dots, Y_n$ with realizations $y_1, \dots, y_n$ in sample space Ω_Y is part of the latent world. The basic goal consists of estimating the individual probabilities $\pi_{xy} = P(Y_i = y | X_i = x)$ given the precise values of a categorical covariate X with sample space Ω_X. Unfavorably, the values of Y can only be observed partially and thus the realizations $\mathfrak{y}_1, \dots, \mathfrak{y}_n$ of a sample $\mathcal{Y}_1, \dots, \mathcal{Y}_n$ of a random object $\mathcal{Y}$ within sample space $\Omega_\mathcal{Y} = \mathcal{P}(\Omega_Y) \setminus \emptyset$ constitute the observed world, with $\mathfrak{y}_i \ni y_i$.[2] A connection between both worlds, and thus between π_{xy} and $p_{x\mathfrak{y}} = P(\mathcal{Y}_i = \mathfrak{y} | X_i = x)$, is established via an observation model governed by the coarsening parameters $q_{\mathfrak{y}|xy} = P(\mathcal{Y}_i = \mathfrak{y} | X_i = x, Y_i = y)$ with $\mathfrak{y} \in \Omega_\mathcal{Y}$, $y \in \Omega_Y$, and $x \in \Omega_X$. As the dimension of these coarsening parameters increases considerably with $|\Omega_X|$ and $|\Omega_Y|$, for reasons of conciseness, we mainly confine ourselves to the discussion of the example 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. Assuming only error-freeness, we generally refrain from making strict assumptions on $q_{\mathfrak{y}|xy}$. In contrast to this, under CAR and SI the coarsening parameters are strongly restricted.

[2] This error-freeness implies that Y is an almost sure selector of $\mathcal{Y}$ (in the sense of e.g. [9]).

Heitjan and Rubin ([6]) 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. [7]). 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 $\mathfrak{y}$. The strong limitation of this assumption is illustrated by the running example generally introduced in the following box.

Running example (Table 1 shows the summary of the data)

- German Panel Study "Labour Market and Social Security" ([14]) (PASS, wave 5, 2011)
- Y: income < 1000€ (a) or ≥ 1000€ (b) $\Rightarrow y \in \{a,\ b\}$
- $\mathcal{Y}$: some respondents give no suitable answer ($\{a,b\}$: $y = a$ or $y = b$) $\Rightarrow \mathfrak{y} \in \{\{a\},\{b\},\{a,b\}\} \Rightarrow$ coarse answer $\{a,b\}$ is missing observation
- X: receipt of Unemployment Benefit II (UBII), $x \in \{0$ (no), 1 (yes)$\}$

Table 1 Data of the PASS example

UBII	Income	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	$\{a\}$	$n_{1\{a\}} = 36$	$n_1 = 87$
	$\{b\}$	$n_{1\{b\}} = 42$	
	$\{a,b\}$	$n_{1\{a,b\}} = 9$	

Referring to the example, under CAR, which coincides here with MAR,[3] the probability of giving no suitable answer is taken to be independent of the true income category in both subgroups split by UBII, i.e.

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

Generally, CAR could be quite problematic in this context, as practical experiences show that reporting missing or coarsened answers is notably common in specific income groups (e.g. [10]).

If the data are missing not at random (MNAR) ([7]), commonly the missingness process is modelled by including parametric assumptions (e.g. [7], [6])

[3] The PASS data provide income in different levels of coarseness induced by follow-up questions for non-respondents. For sake of simplicity, we consider only the income question explained in the box, but the study provides also coarse ordinal data in the general sense.

or a cautious procedure is chosen ending up in set-valued estimators (cf. e.g. [3], [17], [11]). For the categorical case, there is a special case of MNAR, in which single-valued estimators are obtained without parametric assumptions. For motivating this case, one can further differentiate MNAR, distinguishing between the situation where missingness depends on both the values of the response Y and the covariate X and the situation where it depends on the values of Y only. Referring to the related coarsening case, the latter case corresponds to SI investigated in [11]. This independence from the covariate value shows, beside CAR, an alternative kind of uninformative coarsening. Again, one should use this assumption cautiously: Under SI, giving a coarse answer is taken to be independent of the UBII given the value of Y, i.e.

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

Mostly, this turns out to be doubtful, as the receipt of the UBII influences the income, which typically has an impact on the non-response to the income question.

3 Estimation: General approach, CAR and SI

This section recalls some important aspects of an approach developed in [11] by sketching the basic idea of the therein considered cautious, likelihood-based estimation technique. The resulting estimators are not only given for the general case, but also when the assumptions in focus are included.

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. Since the counts $(n_{x\mathfrak{y}})_{x\in\Omega_X,\mathfrak{y}\in\Omega_\mathcal{Y}}$ are multinomially distributed, the unique MLE is obtained by the relative frequencies of the respective categories, coarse categories treated as own categories. Secondly, we connect the parameters of both worlds by a mapping Φ. For the binary case with $x \in \{0,\ 1\}$ one obtains $\Phi:\ [0,\ 1]^6 \to [0,\ 1]^4$ with

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

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 set-valued estimators $\hat{\pi}_{xy}$ and $\hat{q}_{\mathfrak{y}|xy}$, namely

$$\hat{\pi}_{xa} \in \left[\frac{n_{x\{a\}}}{n_x},\ \frac{n_{x\{a\}}+n_{x\{a,b\}}}{n_x}\right], \quad \hat{q}_{\{a,b\}|xy} \in \left[0,\ \frac{n_{x\{a,b\}}}{n_{x\{y\}}+n_{x\{a,b\}}}\right], \tag{2}$$

with $x \in \{0,\ 1\}$ and $y \in \{a,\ b\}$. Points in these sets are constrained by the relationships in Φ. In the spirit of the methodology of *partial identification* ([8]), these sets may be refined by including assumptions about the coarsening justified from the application standpoint. Very strict assumptions may induce point identified parameters, as estimation under CAR or SI in the categorical case shows.[4]

Including CAR, i.e. restricting the set of possible coarsening mechanisms to $q_{\{a,b\}|xa} = q_{\{a,b\}|xb}$ with $x \in \{0,\ 1\}$, induces an injective mapping Φ leading to the point-valued estimators

$$\hat{\pi}^{CAR}_{xa} = \frac{n_{x\{a\}}}{n_{x\{a\}} + n_{x\{b\}}}, \quad \hat{q}^{CAR}_{\{a,b\}|xa} = \hat{q}^{CAR}_{\{a,b\}|xb} = \frac{n_{x\{a,b\}}}{n_x}. \tag{3}$$

Thus, under this type of uninformative coarsening, $\hat{\pi}_{xa}$ corresponds here 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. Under rather weak regularity conditions, namely $\pi_{0a} \neq \pi_{1a}$, $\pi_{0a} \notin \{0, 1\}$, and $\pi_{1a} \notin \{0, 1\}$ for $x \in \{0, 1\}$, also under SI the mapping Φ becomes injective (cf. [12]) in our categorical setting. Hence, point-valued estimators

$$\begin{aligned}
\hat{\pi}^{SI}_{xa} &= \frac{n_{x\{a\}}}{n_x} \frac{n_0\, n_{1\{b\}} - n_{0\{b\}}\, n_1}{n_{0\{a\}}\, n_{1\{b\}} - n_{0\{b\}}\, n_{1\{a\}}}, \\
\hat{q}^{SI}_{\{a,b\}|xa} &= \frac{n_{0\{a,b\}}\, n_{1\{b\}} - n_{0\{b\}}\, n_{1\{a,b\}}}{n_0\, n_{1\{b\}} - n_{0\{b\}}\, n_1}, \\
\hat{q}^{SI}_{\{a,b\}|xb} &= \frac{n_{0\{a,b\}}\, n_{1\{a\}} - n_{0\{a\}}\, n_{1\{a,b\}}}{n_0\, n_{1\{a\}} - n_{0\{a\}}\, n_1}
\end{aligned} \tag{4}$$

are obtained, provided they are well-defined and inside $[0,\ 1]$.

4 Testing

Due to the substantial bias of $\hat{\pi}_{xy}$ if CAR or SI are wrongly assumed (cf. e.g. [12]), testing these assumptions is of particular interest. Although it is already established that it is not possible to test whether the CAR condition holds (e.g. [8]), it may be insightful, in particular in the light of Section 4.2, to address this impossibility in the context of the example.

[4] 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 [13].

4.1 Testing of CAR

A closer consideration of (3) already indicates that CAR can never be rejected without including additional assumptions about the coarsening. This point is illustrated in Fig. 1 by showing the interaction between points in the intervals in (2). Thus, this uninformative coarsening – in the sense that all coarse observations are ignored – is always a possible scenario included in the general set-valued estimators in (2).
Exemplary for subgroup 0, under CAR we obtain $\hat{\pi}_{0a}^{CAR} = 0.09$, $\hat{q}_{\{a,b\}|0a}^{CAR} = \hat{q}_{\{a,b\}|0b}^{CAR} = 0.18$, which may not be excluded from the general estimators $\hat{\pi}_{0a} \in [0.07,\ 0.26]$, $\hat{q}_{\{a,b\}|0a} \in [0, 0.71]$ and $\hat{q}_{\{a,b\}|0b} \in [0, 0.20]$ unless further assumptions as e.g. "respondents from the high income group tend to give coarse answers more likely" are justified.
Nevertheless, there are several approaches that show how testability of CAR is achieved by distributional assumptions (e.g. [4]), e.g. the naive Bayes assumption ([5]), or by the inclusion of instrumental variables (cf. [1]).

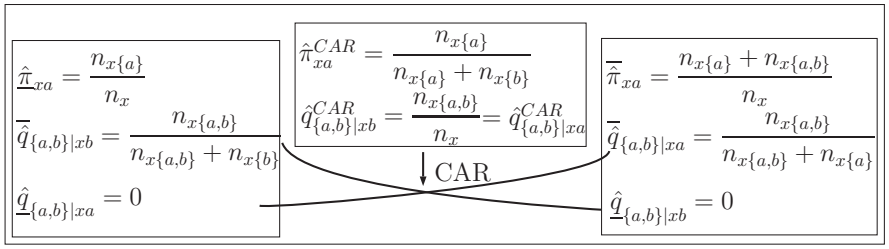

Fig. 1 Since the relationships expressed via Φ in (1) have to be met, only specific points from the set-valued estimators in (2) are combinable, ranging from $(\underline{\hat{\pi}}_{xa},\ \underline{\hat{q}}_{\{a,b\}|xa},\ \overline{\hat{q}}_{\{a,b\}|xb})$ to $(\overline{\hat{\pi}}_{xa},\ \overline{\hat{q}}_{\{a,b\}|xa},\ \underline{\hat{q}}_{\{a,b\}|xb})$ with the CAR case always included.

4.2 Testing of SI

Applying the estimators in (4) to the example, one obtains $\hat{\pi}_{0a}^{SI} = 0.42$, $\hat{\pi}_{1a}^{SI} = 0.40$, $\hat{q}_{\{a,b\}|0a}^{SI} = \hat{q}_{\{a,b\}|1a}^{SI} = -0.04$, and $\hat{q}_{\{a,b\}|0b}^{SI} = \hat{q}_{\{a,b\}|1b}^{SI} = 0.20$ partly outside $[0,\ 1]$. This shows that there are data situations that might hint to (partial) incompatibility with SI. In general for the categorical case, a statistical test for the following hypotheses can be constructed:

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

As test statistic we can use the likelihood ratio (e.g. [16])

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

here with $\vartheta=(\pi_{0a},\pi_{1a},q_{\{a,b\}|0a},q_{\{a,b\}|1a},q_{\{a,b\}|0b},q_{\{a,b\}|1b})^T$.[5] In fact, recent simulation studies corroborate the decrease of Λ with deviation from SI (cf. [12]). The sensitivity of Λ with regard to the test considered here is also illustrated informally in Fig. 2 by depicting Φ in (1) for two data situations, where only the second one gives evidence against SI. The gray line symbolizes all arguments satisfying SI, while the bold line represents all arguments maximizing the likelihood (i.e. all values in (2) compatible with each other). The intersection of both lines represents the values in (4), and if it is included in the domain of Φ (cf. first case of Fig. 2), the same maximal value of the likelihood is obtained regardless of including SI or not, resulting in $\Lambda=1$. An intersection outside the domain (cf. second case of Fig. 2) induces a lower value of the likelihood under SI, also reflected in $\Lambda<1$. For the example one obtains $\Lambda\approx 0.93$ and thus there is a slight evidence against SI based on a direct interpretation of the likelihood ratio, while setting a general decision rule depending on a significance level α remains as an open problem.

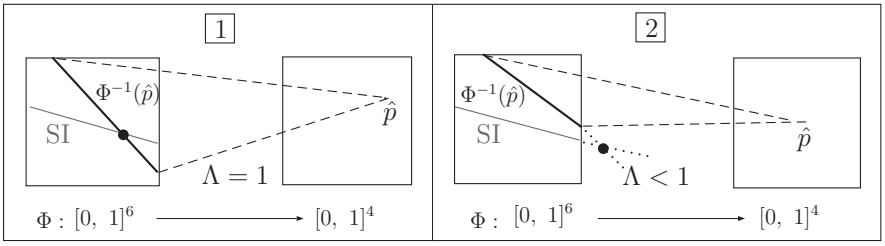

Fig. 2 The impact on Λ of two substantially differing data situations is illustrated.

5 Conclusion

We focused on the testability of CAR and SI by investigating the compatibility of the estimators (3) and (4) with the observed data. While CAR is generally not testable, SI may be tested and a "pure likelihood" approach was proposed. To obtain a statistical test for SI at a fixed level of significance α, we want to determine the (asymptotic) distribution of $-2\log\Lambda$ under H_0

[5] While the denominator of Λ can be obtained using any values in (2) compatible with each other, the numerator must in general be calculated by numerical optimization. Alternatives to this statistic include a test decision based on uncertainty regions ([15]).

next, which is expected to deviate from the χ^2-distribution of the standard case. Furthermore, a generalized version of SI – in the sense of assuming particular coarsening parameters to be known multiples of each other – will allow for a more flexible application of this hypothesis test.

Acknowledgements We thank 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 are grateful to two anonymous reviewers for their very helpful remarks and valuable suggestions.

References

1. Breunig C (2015) Testing missing at random using instrumental variables. Humboldt University, Collaborative Research Center 649, `https://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2015-016.pdf`
2. Couso I, Dubois D (2014) Statistical reasoning with set-valued information: Ontic vs. epistemic views. Int J Approx Reason 55:1502–1518
3. Denoeux T (2014) Likelihood-based belief function: justification and some extensions to low-quality data. Int J Approx Reason 55:1535–1547
4. Jaeger M (2005) Ignorability for categorical data. Ann Stat 33:1964–1981
5. Jaeger M (2006) On testing the missing at random assumption. In: Fürnkranz J, Scheffer T, Spiliopoulou M (eds) Machine Learning: ECML 2006. Springer
6. Heitjan D, Rubin D (1991) Ignorability and coarse data. Ann Stat 19:2244–2253
7. Little R, Rubin D (2002) Statistical Analysis with Missing Data. 2nd edn, Wiley
8. Manski C (2003) Partial Identification of Probability Distributions. Springer
9. Nguyen H (2006) An Introduction to Random Sets. CRC Press
10. Korinek A, Mistiaen J, Ravallion M (2006) Survey nonresponse and the distribution of income. J Econ Inequal 4:33–55
11. Plass J, Augustin T, Cattaneo M, 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, Quaeghebeur E (eds) ISIPTA '15. SIPTA
12. Plass J, Augustin T, Cattaneo M, Schollmeyer G (2016) Statistical modelling under epistemic data imprecision, LMU Munich, `http://www.statistik.lmu.de/~jplass/forschung.html`
13. Stanghellini E, Vantaggi B (2013) Identification of discrete concentration graph models with one hidden binary variable. Bernoulli 19:1920–1937
14. Trappmann M, Gundert S, Wenzig C, Gebhardt D (2010) PASS: a household panel survey for research on unemployment and poverty. Schmollers Jahrb 130:609–623
15. Vansteelandt S, Goetghebeur E, Kenward M, Molenberghs G (2006) Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. Stat Sin 16:953–979
16. Wilks S (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Stat 9:60–62
17. Zaffalon M, Miranda E (2009) Conservative inference rule for uncertain reasoning under incompleteness. J Artif Intell Res 34:757–821