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

Julia Plass*, Thomas Augustin*, Marco Cattaneo**, Georg Schollmeyer*

*Department of Statistics, Ludwig-Maximilians Universität Munich and **Department of Mathematics, University of Hull

Distinction between Epistemic and Ontic Interpretation (Couso, Dubois, Sánchez, 2014, Springer)

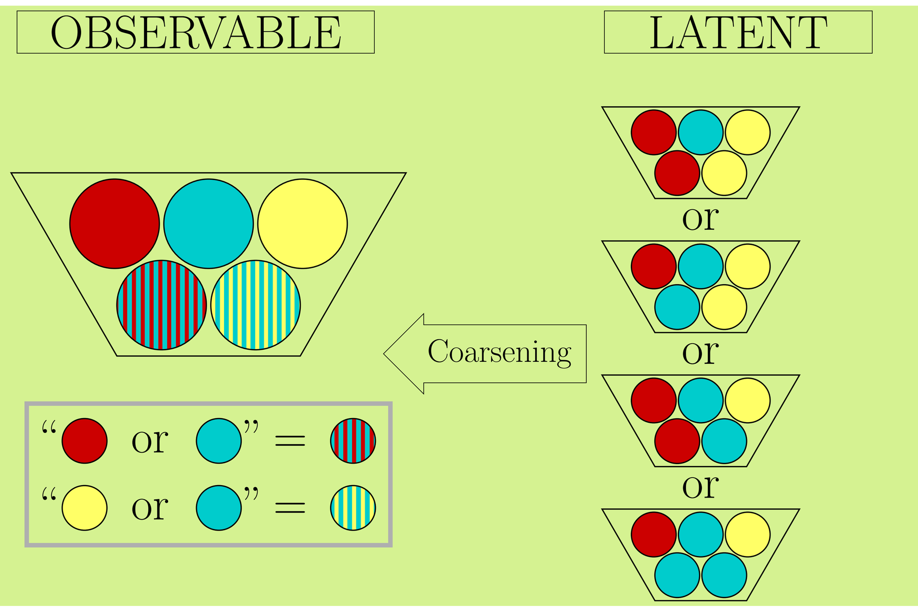
Ontic poster in session on Wednesday

Epistemic data imprecision:

- Imprecise observation of something precise
- Actually precise values may only be observed in a coarse form, due to an underlying coarsening mechanism

Examples:

- Missing data as a special case
- Coarsening deliberately applied as an anonymization technique
- Matched data sets with not completely identical categories

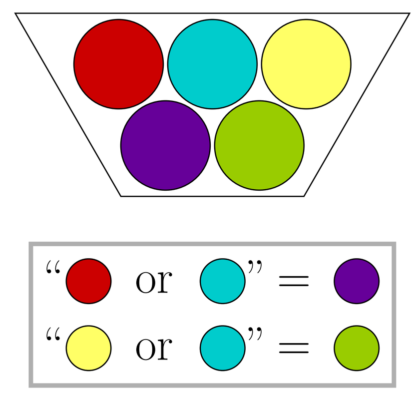


Ontic data imprecision:

- Precise observation of something imprecise
- Truth is represented by coarse observations

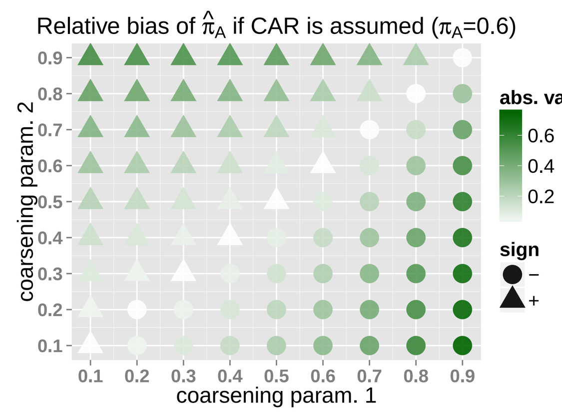
Example:

Answers of indecisive respondents



Already existing approaches

- Still common to impose strict assumptions and thus **to enforce precise results** ⇒ Problem: results may be substantially biased (cf. Figure; π_A is parameter of interest, CAR is only satisfied if coarsening parameter 1 = coarsening parameter 2)
- There is a variety of different **set-valued approaches** aiming at a proper reflection of the available information
  - using a Bayesian point of view (e.g. de Cooman, Zaffalon, Artif. Intell)
  - via random sets (Nguyen, 2006, An Introduction to Random Sets)
  - via likelihood-based belief function (Denoeux, 2014, IJAR)
  - via the profile likelihood (e.g. Cattaneo, Wiencierz, 2012, IJAR)
  - **Here:** Likelihood-based approach, strongly influenced by the methodology of partial identification, coarse categorical data only



Basic idea

Considered cases:
1.) Homogeneous case (i.i.d. case)
2.) Case with categorical covariates X (here the representative case of one binary covariate is considered)

OBSERVABLE

$\mathcal{Y}$ coarse data

Only realizations $\mathscr{y}_1, \dots, \mathscr{y}_n$ of a sample $\mathcal{Y}_1, \dots, \mathcal{Y}_n$ of a random variable $\mathcal{Y}$ in sample space $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \emptyset$ are observed.

↓

$p_{\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i), \ i = 1, \dots, n$

Estimators $\hat p_{\mathscr{y}}$ and $\hat p_{\mathscr{y}}$ can be directly obtained via maximum-likelihood estimation:

Likelihood for parameters $\boldsymbol{p} = (p_1, \dots, p_{|\Omega_{\mathcal{Y}}|-1})^T$
$L(\boldsymbol{p}) \propto \prod_{\mathscr{y} \in \Omega_{\mathcal{Y}}} p_{\mathscr{y}}^{n_{\mathscr{y}}}$ is *uniquely* maximized by

$\boxed{\hat p_{\mathscr{y}} = \frac{n_{\mathscr{y}}}{n},} \quad \mathscr{y} \in \{1, \dots, |\Omega_{\mathcal{Y}}| - 1\}$

and thus $\hat p_{|\Omega_{\mathcal{Y}}|} = 1 - \sum_{m=1}^{|\Omega_{\mathcal{Y}}|-1} \hat p_m$.

For fixed x, $(n_{x1}, \dots n_{x|\Omega_{\mathcal{Y}}|}) \sim M(n_x, (p_{x1}, \dots, p_{x|\Omega_{\mathcal{Y}}|}))$ with conditional probabilities $p_{x\mathscr{y}} = P(\mathcal{Y} = \mathscr{y} | X = x)$.

Likelihood for parameters $\boldsymbol{p} = (p_{x1}, \dots, p_{x|\Omega_{\mathcal{Y}}|-1})^T$
$L(\boldsymbol{p}) \propto \prod_{\mathscr{y} \in \Omega_{\mathcal{Y}}} p_{0\mathscr{y}}^{n_{0\mathscr{y}}} \prod_{\mathscr{y} \in \Omega_{\mathcal{Y}}} p_{1\mathscr{y}}^{n_{1\mathscr{y}}}$ is *uniquely* maximized by

$\boxed{\hat p_{x\mathscr{y}} = \frac{n_{x\mathscr{y}}}{n_x},} \quad$ for $x \in \{0, 1\}$.

⟵ Observation model $\mathcal{Q}$

Here: error-freeness is assumed

↓

coarsening mechanism
$q_{\mathscr{y}|y} = P(\mathcal{Y} = \mathscr{y} | Y = y)$
$q_{\mathscr{y}|xy} = P(\mathcal{Y} = \mathscr{y} | X = x, Y = y)$

Use the **connection between $\boldsymbol{p}$ and $\boldsymbol{\gamma}$**

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

and the **invariance of the likelihood** under parameter transformations, i.e.:

$$\boxed{\hat\Gamma = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}}$$

$\Phi(\cdot)$ in general is not injective
$\Rightarrow \hat\Gamma$ is set-valued

LATENT

Y latent variable

$Y_1, \dots, Y_n$ is a random sample of a categorical response variable in sample space $\Omega_Y = \{1, \dots, j, \dots, K\}$.

↓

Main goal: Estimation of $\pi_{ij} = P(Y_i = j)$

$\pi_{i1} = \pi_1, \dots, \pi_{iK} = \pi_K$

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

for categories $j = 1, \dots, K-1$ and

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

(multinomial logit model)

↓

$\boldsymbol{\gamma} = (\boldsymbol{q}_{\mathscr{y}|\boldsymbol{y}}^T, \ \boldsymbol{\pi}_{\boldsymbol{y}}^T)^T$

$\boldsymbol{\gamma} = (\boldsymbol{q}_{\mathscr{y}|\boldsymbol{xy}}^T, \ \boldsymbol{\pi}_{\boldsymbol{xy}}^T)^T$

$\hat\pi_y \in \left[\frac{n_{\{y\}}}{n}, \ \frac{\sum_{\mathscr{y} \ni y} n_{\mathscr{y}}}{n}\right]$
$\hat q_{\mathscr{y}|y} \in \left[0, \ \frac{n_{\mathscr{y}}}{n_{\{y\}} + n_{\mathscr{y}}}\right]$

$\hat\pi_{xy} \in \left[\frac{n_{x\{y\}}}{n_x}, \ \frac{\sum_{\mathscr{y} \ni y} n_{x\mathscr{y}}}{n_x}\right]$
$\hat q_{\mathscr{y}|xy} \in \left[0, \ \frac{n_{x\mathscr{y}}}{n_{x\{y\}} + n_{x\mathscr{y}}}\right]$

Illustration 1:

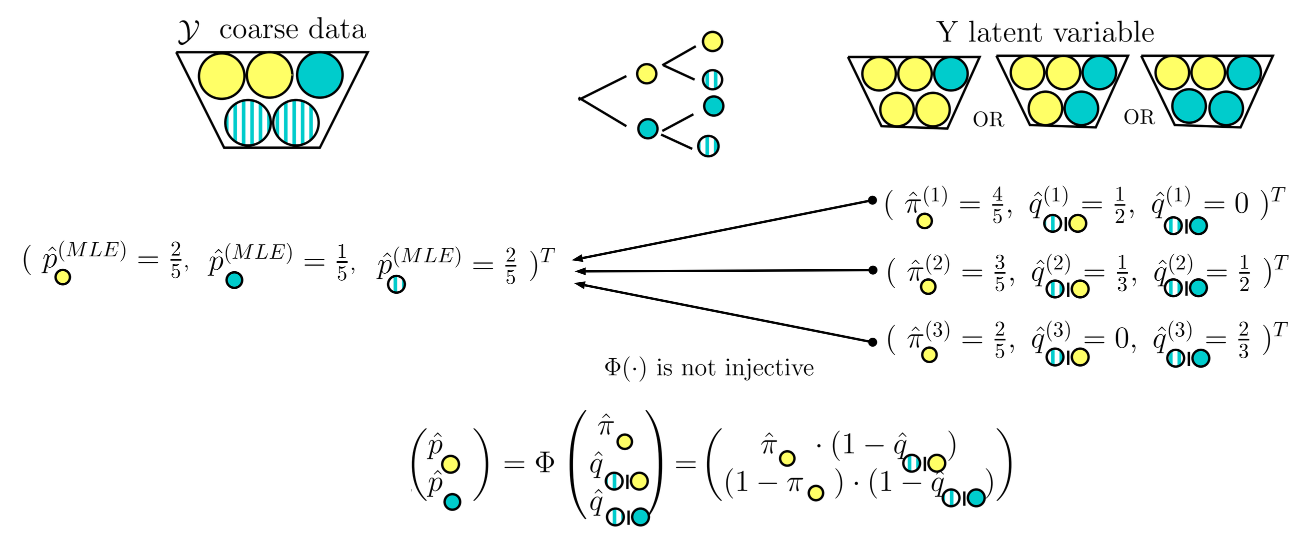


Illustration 2: PASS-Data

(German panel study "Labour market and Social Security"; Trappmann, Gundert, Wenzig, Gebhardt, 2010, Schmollers Jahrbuch)

- Here: $\Omega_Y = \{< 1000€ (<), \geq 1000€ (\geq)\}$
- Some respondents give no suitible answer ("na"; i.e. coarse answer "$<$ 1000€ or $\geq$ 1000€"): $\Omega_{\mathcal{Y}} = \{<, \geq, \text{na}\}$

i.i.d. case

$n_< = 238, \ n_\geq = 835, \ n_{\text{na}} = 338 \Rightarrow \hat\pi_< \in \left[\frac{238}{1411}, \ \frac{238+338}{1411}\right]$

Categorical covariate Unemployment Benefit II (UBII)

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

$\hat\pi_{0<} \in \left[\frac{130}{319}, \ \frac{130+75}{319}\right]$

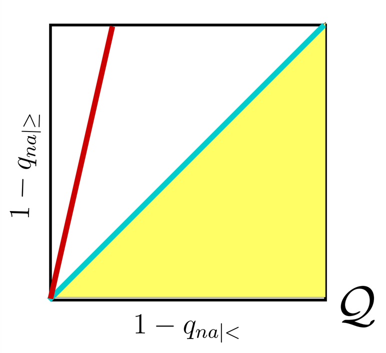
$\hat\pi_{1<} \in \left[\frac{108}{1092}, \ \frac{108+263}{1092}\right]$ or

$\hat\beta_{<0} \in [-0.37, \ 0.59]$

$\hat\beta_{<1} \in [-1.83, \ -1.25]$

Reliable Incorporation of Auxiliary Information

Starting from point-identifying assumptions, we use sensitivity parameters to allow inclusion of partial knowledge.

Assumption	Coarsening at random (CAR) and its generalization $R = \frac{1-q_{na\|\geq}}{1-q_{na\|<}}$	Subgroup independent coarsening (SIC) and its generalizations $R_1 = \frac{1-q_{na\|0<}}{1-q_{na\|1<}}$ & $R_2 = \frac{1-q_{na\|0\geq}}{1-q_{na\|1\geq}}$
Illustration	 **CAR** $q_{na\|<} = q_{na\|\geq}$ i.e. probability of "na" does not depend on true income category **Assumptions about exact value of R** e.g. $R=1$ $R=4$, where R=1 corresponds to CAR. **Rough evaluation of R** e.g. $R \leq 1$: low income group has a higher tendency to report in a precise way	**SIC** $q_{na\|0<} = q_{na\|1<}$ and $q_{na\|0\geq} = q_{na\|1\geq}$ Reporting "na" does not depend on the receipt of the UBII **Assumption about the exact value of R_1 and R_2** Knowledge about the relative magnitude of precise observations in both subgroups allows more flexible inclusion of information $R_1 = R_2 = 1$ represents SIC **Rough evaluation of R_1 and R_2** Partial knowledge about the relative magnitude
Estimators	$\hat\pi_< = \frac{n_<}{n_< + n_\geq}$ $\hat\pi_< = \frac{n_< \cdot 4}{n_\geq + n_< \cdot 4}$ $\hat\pi_< = [\frac{n_<}{n}, \ \frac{n_<}{n_< + n_\geq}]$	Exemplary for SIC $\hat\pi_{0<} = \frac{n_{0<}}{n_0} \frac{n_{1\geq} n_0 - n_1 n_{0\geq}}{n_{0<} n_{1\geq} - n_{0\geq} n_{1<}}$ $\hat\pi_{1<} = \frac{n_{1<}}{n_1} \frac{n_{1\geq} n_0 - n_1 n_{0\geq}}{n_{0<} n_{1\geq} - n_{0\geq} n_{1<}}$

Summary and Outlook

- Via the observation model $\mathcal{Q}$ maximum-likelihood estimators referring to the latent variable may be obtained for both cases
  - ... the homogeneous case
  - ... the case with categorical covariates
- Proper inclusion of auxiliary information via restrictions on $\mathcal{Q}$

Next steps:

- Likelihood-based hypothesis tests and uncertainty regions for coarse categorical data
- Inclusion of auxiliary information by sets of priors
- Consideration of other "deficiency" processes
- Extension to metric covariates?