Cautious statistical modelling for categorical data under epistemic and ontic data imprecision

Julia Plass, Supervision: Prof. Thomas Augustin

Department of Statistics, Ludwig-Maximilians University

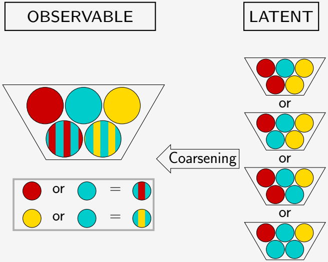
14^{th} of September 2015

21. DStatG Nachwuchsworkshop, Hamburg

Epistemic vs. ontic interpretation (Couso, Dubois, Sánchez, 2014)

Epistemic imprecision:

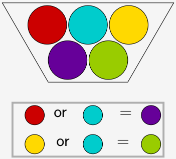
"Imprecise observation of something precise"



⇒ Truth is hidden due to the underlying coarsening mechanism

Ontic imprecision:

"Precise observation of something imprecise"

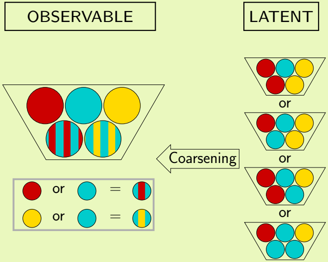


⇒ Truth is represented by coarse observation

Examples of data under epistemic imprecision

Epistemic imprecision:

"Imprecise observation of something precise"



$\Rightarrow$ Truth is hidden due to the underlying coarsening mechanism

Examples:

- Matched data sets with partially overlapping variables
- Coarsening as anonymization technique
- Missing data as special case

Here: PASS-data
$\mathcal{Y}$: income, X: UBII

$$\Omega_{\mathcal{Y}} = \{<, \geq, \text{na}\}$$
$$\Omega_X = \{0 \text{ (no)}, 1 \text{ (yes)}\}$$

Basic problem (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

OBSERVABLE

LATENT

coarse data $\mathcal{Y}$

Observation model $\mathcal{Q}$

error-freeness

latent variable Y

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

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

for j=1,...,K-1

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

for reference category K

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

(multinomial logit model)

Main goal:

Maximum-Likelihood estimation of $\boldsymbol{\gamma} = (\boldsymbol{q}_{\mathscr{Y}|\boldsymbol{xy}}^T, \ \boldsymbol{\pi}_{\boldsymbol{y}}^T)^T$

Basic idea (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

OBSERVABLE

LATENT

Use random-set perspective and determine ML estimator

$\hat{p}_{x\mathscr{y}} = \hat{P}(\mathcal{Y} = \mathscr{y} | X = x)$

$\longrightarrow$ $\boxed{\hat{p}_{x\mathscr{y}} = \frac{n_{x\mathscr{y}}}{n_x}}$

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

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

and the **invariance of the likelihood** under parameter transformations:

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

$$\hat{\pi}_{xy} \in \left[\frac{n_{x\{y\}}}{n_x}, \frac{\sum_{\mathscr{y} \ni y} n_{x\mathscr{y}}}{n_x}\right]$$

$$\hat{q}_{\mathscr{y}|xy} \in \left[0, \frac{n_{x\mathscr{y}}}{n_{x\{y\}} + n_{x\mathscr{y}}}\right]$$

OBSERVABLE

LATENT

Use random-set perspective and determine ML estimator

$\hat{p}_{x\mathscr{y}} = \hat{P}(\mathcal{Y} = \mathscr{y} | X = x)$

$\longrightarrow \boxed{\hat{p}_{x\mathscr{y}} = \frac{n_{x\mathscr{y}}}{n_x}}$

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

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

and the **invariance of the likelihood** under parameter transformations:

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

$$\hat{\pi}_{xy} \in \left[\frac{n_{x\{y\}}}{n_x}, \frac{\sum_{\mathfrak{y} \ni y} n_{x\mathfrak{y}}}{n_x}\right]$$

$$\hat{q}_{\mathfrak{y}|xy} \in \left[0, \frac{n_{x\mathfrak{y}}}{n_{x\{y\}} + n_{x\mathfrak{y}}}\right]$$

Basic idea (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

OBSERVABLE

LATENT

Use random-set perspective and determine ML estimator

$\hat{p}_{x\mathscr{y}} = \hat{P}(\mathcal{Y} = \mathscr{y} | X = x)$

$\longrightarrow \boxed{\hat{p}_{x\mathscr{y}} = \frac{n_{x\mathscr{y}}}{n_x}}$

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

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

and the **invariance of the likelihood** under parameter transformations:

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

$$\hat{\pi}_{xy} \in \left[\frac{n_{x\{y\}}}{n_x}, \frac{\sum_{\mathfrak{y}\ni y} n_{x\mathfrak{y}}}{n_x}\right]$$

$$\hat{q}_{\mathfrak{y}|xy} \in \left[0, \frac{n_{x\mathfrak{y}}}{n_{x\{y\}}+n_{x\mathfrak{y}}}\right]$$

Basic idea (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

OBSERVABLE

LATENT

Use random-set perspective and determine ML estimator

$\hat{p}_{x\mathscr{y}} = \hat{P}(\mathcal{Y} = \mathscr{y} | X = x)$

$\longrightarrow \boxed{\hat{p}_{x\mathscr{y}} = \frac{n_{x\mathscr{y}}}{n_x}}$

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

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

and the **invariance of the likelihood** under parameter transformations:

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

$$\hat{\pi}_{xy} \in \left[\frac{n_{x\{y\}}}{n_x}, \frac{\sum_{\mathbf{y} \ni y} n_{x\mathbf{y}}}{n_x}\right]$$

$$\hat{q}_{\mathbf{y}|xy} \in \left[0, \frac{n_{x\mathbf{y}}}{n_{x\{y\}} + n_{x\mathbf{y}}}\right]$$

$\downarrow$

Illustration (PASS data)

$\hat{\pi}_{0<} \in [0.41,\ 0.64]$ $\quad$ $\hat{\pi}_{1<} \in [0.10,\ 0.34]$

$\hat{\beta}_{<0} \in [-0.37,\ 0.59]$ $\quad$ $\hat{\beta}_{<} \in [-1.83,\ -1.25]$

Reliable incorporation of auxiliary information

$\hat{\pi}_{x<} \in \left[\frac{n_{x<}}{n_x}, \ \frac{n_{x<}+n_{x\mathrm{na}}}{n_x}\right]$

Partial identification

(Manski, 2003)

point identification

Sensitivity analysis

(Vansteelandt, Goetghebeur, Kenward, Molenberghs, 2006)

Reliable incorporation of auxiliary information

$\hat{\pi}_{x<} \in \left[\frac{n_{x<}}{n_x}, \; \frac{n_{x<}+n_{x\text{na}}}{n_x}\right]$

Partial identification
(Manski, 2003)

point identification

Sensitivity analysis
(Vansteelandt, Goetghebeur, Kenward, Molenberghs, 2006)

$\downarrow$ e.g.

$\text{R} = \frac{q_{na|x\geq}}{q_{na|x<}}$

CAR: R=1
(Heitjan, Rubin, 1991)
$q_{\text{na}|x<} = q_{\text{na}|x\geq}$

$\downarrow$

$\hat{\pi}_{x<} = \frac{n_{x<}}{n_{x<}+n_{x\geq}}$

Auxiliary information:

Refined estimators:

Reliable incorporation of auxiliary information

$\hat{\pi}_{x<} \in \left[\frac{n_{x<}}{n_x}, \frac{n_{x<}+n_{x\text{na}}}{n_x}\right]$

Partial identification

(Manski, 2003)

point identification

Sensitivity analysis

(Vansteelandt, Goetghebeur, Kenward, Molenberghs, 2006)

e.g.

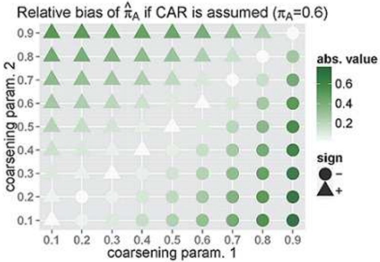
$\text{R} = \frac{q_{na|x\geq}}{q_{na|x<}}$

CAR: R=1

(Heitjan, Rubin, 1991)

$q_{\text{na}|x<} = q_{\text{na}|x\geq}$

Auxiliary information:



Refined estimators:

$\hat{\pi}_{x<} = \frac{n_{x<}}{n_{x<}+n_{x\geq}}$

Reliable incorporation of auxiliary information

$\hat{\pi}_{x<} \in \left[\frac{n_{x<}}{n_x}, \; \frac{n_{x<}+n_{x\text{na}}}{n_x}\right]$

Partial identification
(Manski, 2003)

point identification

Sensitivity analysis
(Vansteelandt, Goetghebeur, Kenward, Molenberghs, 2006)

e.g.

$\text{R} = \frac{q_{na|x\geq}}{q_{na|x<}}$

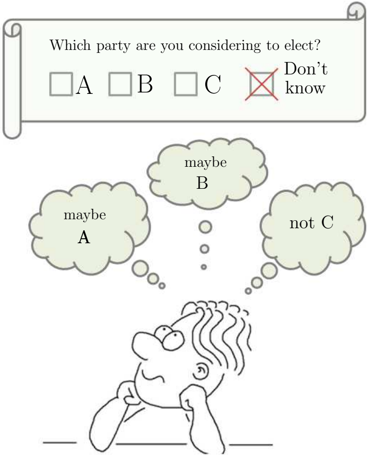
CAR: R=1
(Heitjan, Rubin, 1991)
$q_{\text{na}|x<} = q_{\text{na}|x\geq}$

$\hat{\pi}_{x<} = \frac{n_{x<}}{n_{x<}+n_{x\geq}}$

e.g.

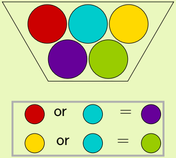
Auxiliary information: $\text{R} = \frac{q_{na|x\geq}}{q_{na|x<}} < 1$

Refined estimators: $\hat{\pi}_{x<} \in \left[\frac{n_{x<}}{n_x}, \; \frac{n_{x<}}{n_{x<}+n_{x\geq}}\right]$



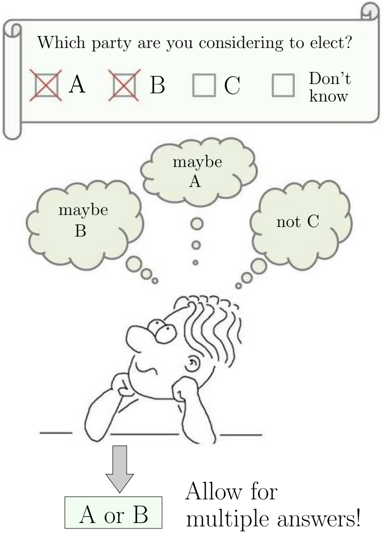
Ontic imprecision:

"Precise observation of something imprecise"



$\Rightarrow$ Truth is represented by coarse observation

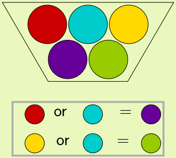
Example of data under ontic imprecision



Allow for multiple answers!

Ontic imprecision:

"Precise observation of something imprecise"



$\Rightarrow$ Truth is represented by coarse observation

Analysis for ontic case (Plass, Fink, Schöning, Augustin, 2015)

General analysis:

- Interpretation of coarse answers as ontic sets (random sets) (Couso, Dubois, Sánchez, 2014)
- Regard coarse answers like "A or B" as own categories
- Extension of state space S of Y to $S^* = \mathcal{P}(S) \setminus \emptyset$ of Y^*

Example: Multinomial logistic regression

For each category $s \in S^* = \{1, \ldots, m-1\}$, $m = |S^*|$, probabilities of response Y^* given covariates $\boldsymbol{x}_i$ are modelled by

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

with $\tilde{\boldsymbol{x}}_i^T = (1, \boldsymbol{x}_i^T)$ and for reference category m by

$$P^*(Y_i^* = m \,|\, \mathbf{x}_i) = \left(1 + \Sigma_{r=1}^{m-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r^*)\right)^{-1} .$$

Illustration by the GLES'13 data

- Y: first vote (reference category S)
- X: religious denomination, most important information source

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

$\Rightarrow$ Own categories for coarse categories
$\Rightarrow$ remarkable differences partly associated with a change in sign

Illustration by the GLES'13 data

- Y: first vote (reference category S)
- X: religious denomination, most important information source

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

⇒ Own categories for coarse categories
⇒ remarkable differences partly associated with a change in sign

Illustration by the GLES'13 data

- Y: first vote (reference category S)
- X: religious denomination, most important information source

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

$\Rightarrow$ Own categories for coarse categories

$\Rightarrow$ remarkable differences partly associated with a sign change

Summary and outlook

EPISTEMIC

- Obtain MLE referring to the latent variable via the observation model $\mathcal{Q}$
- Inclusion of auxiliary information via further restrictions on $\mathcal{Q}$

ONTIC

- Coarse categories as own categories
  $\Rightarrow$ Change in state space
- Statistical methods do not change, only interpretation

Next steps:

- Bayesian approach
- Likelihood-based hypothesis tests, uncertainty regions
- Other "deficiency" processes

- Ontic imprecision in covaraites
- Adaptation to ordinal scale

References

Couso, Dubois, Sánchez.
Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables, Springer, 2014.

Heitjan, Rubin.
Ignorability and Coarse Data, Annals of Statistics, 1991.

Manski.
Partial Identification of Probability Distributions, Springer, 2003.

Plass, Fink, Schöning, Augustin.
Statistical modelling in surveys without neglecting the undecided: multinomial logistic regression models and imprecise classification trees under ontic data imprecision, ISIPTA, 2015.

Plass, Augustin, Cattaneo, Schollmeyer.
Statistical modelling under epistemic data imprecision: some results on estimating multinomial distributions and logistic regression for coarse categorical data, ISIPTA, 2015.

Vansteelandt, Goetghebeur, Kenward, Molenberghs.
Ignorance and uncertainty regions as inferential tools in a sensitivity analysis, Stat. Sin., 2006.