Categorical regression analysis for coarse data

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

*Department of Statistics, Ludwig-Maximilians University and
**School of Mathematics and Physical Sciences, University of Hull

&

02nd of December 2016

Coarse data:

Data are not observed in the resolution originally intended in the subject matter context

Categorical regression analysis:

- Modelling the (not necessarily causal) relation between some covariates X (input variables) and a dependent categorical variable Y (output variable)
- Here considering . . .
  . . . Y is partly only observed in a coarse(ned) way ($\mathcal{Y}$)
  . . . precisely observed covariates

Outline: Common features

1.) Stressing the distinction between ontic and epistemic data imprecision

2.) "Disambiguation" strategy

3.) Incorporation of coarsening assumptions
   - error freeness
   - superset assumption
   - coarsening at random
   - subgroup independence

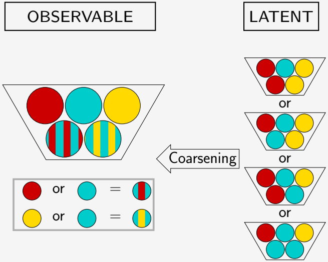
COMPARISON 1:

Distinction between epistemic and ontic data imprecision

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

Epistemic imprecision:

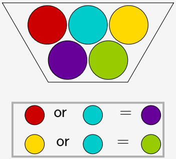
"Imprecise observation of something precise"



⇒ Truth is hidden due to the underlying coarsening mechanism

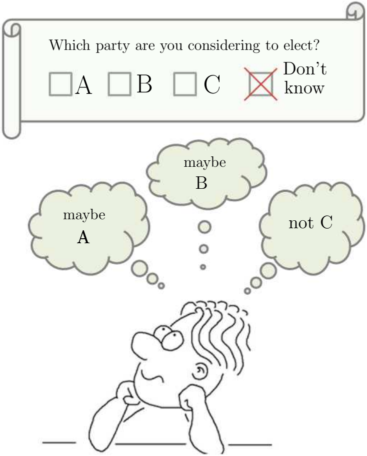
Ontic imprecision:

"Precise observation of something imprecise"



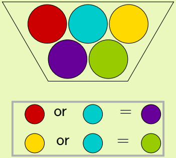
⇒ Truth is represented by coarse observation

Example of data under ontic imprecision

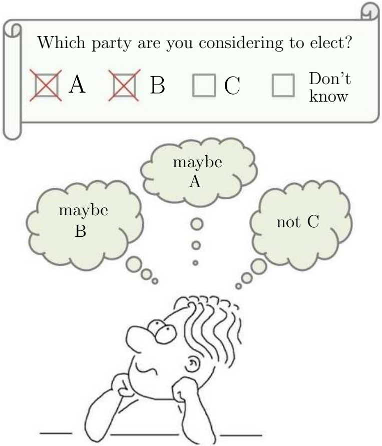


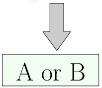
Ontic imprecision:

"Precise observation of something imprecise"



$\Rightarrow$ Truth is represented by coarse observation

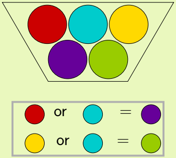




Allow for multiple answers!

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^*
- Multi-label classification

Example: Multinomial logistic regression

For each category $s \in S^* = \{1, \dots, m-1\}$, $m = |S^*|$, 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

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

- 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

- 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

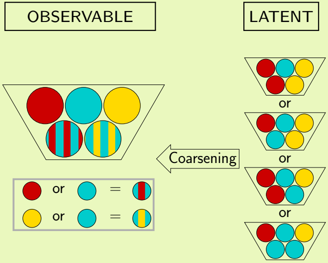
Now: Epistemic data imprecision

COMPARISON 2: Disambiguation strategy

A first comparison of the disambiguation strategy

Epistemic imprecision:

"Imprecise observation of something precise"



⇒ Truth is hidden due to the underlying coarsening mechanism

You

- Machine Learning
- Simultaneous model identification and data disambiguation
- Generalized loss function

We

- Survey statistics
- First: information, then: inference
- Likelihood approach

Hüllermeier (2014) in short

Extension principle:

- Consider all models that are compatible with the observations
- All models are assessed as equally plausible

Basic idea:

- Accounting for model assumptions
- "Model induction and data disambiguation go hand in hand"
- Instead of "ambiguation" of the learning algorithm (extension principle), "ambiguation" of the loss functions

  $\Rightarrow$ Disambiguation strategy: The most plausible precise value is the one that minimizes the generalized loss function

LATENT

$\pi_{xy} :=$

$P(Y_i = y | X_i = x)$

(error-freeness)

Observation model

$q_{\mathfrak{y}|xy} :=$

$P(\mathcal{Y}_i = y | X_i = x, Y_i = y)$

OBSERVABLE

$p_{x\mathfrak{y}} :=$

$P(\mathcal{Y}_i = \mathfrak{y} | X_i{=}x)$

LATENT

$\vartheta = (\pi_{\mathsf{xy}}^{\mathsf{T}},\ \mathsf{q}_{\mathfrak{y}|\mathsf{xy}}^{\mathsf{T}})^{\mathsf{T}}$

OBSERVABLE

$\mathsf{p}_{\mathsf{x}\mathfrak{y}} := \mathsf{P}(\mathcal{Y}_{\mathsf{i}} = \mathfrak{y} | \mathsf{X}_{\mathsf{i}} = \mathsf{x})$

Cautious ML estimation (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

LATENT

$\vartheta = (\pi_{xy}^T,\ q_{\mathfrak{y}|xy}^T)^T$

OBSERVABLE

$p_{x\mathfrak{y}} := P(\mathcal{Y}_i = \mathfrak{y} | X_i = x)$

unique $\arg_{p_{x\mathfrak{y}}} L(p_{x\mathfrak{y}}|\text{data}) = \hat{p}_{x\mathfrak{y}} = \frac{n_{x\mathfrak{y}}}{n_x}$

1.) Determine MLE of observed variable distribution

Cautious ML estimation (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

LATENT		OBSERVABLE	
$\vartheta^{(1)}$, $\vartheta^{(2)}$, $\vartheta^{(3)}$, $\dots$	$\longrightarrow$	$p_{x\mathfrak{y}} := P(\mathcal{Y}_i = \mathfrak{y} \mid X_i = x)$	unique $\arg_{p_{x\mathfrak{y}}} L(p_{x\mathfrak{y}} \mid \text{data}) = \hat{p}_{x\mathfrak{y}} = \frac{n_{x\mathfrak{y}}}{n_x}$

$$\Phi : [0,1]^{\dim(\Theta_{lat})} \longrightarrow [0,1]^{\dim(\Theta_{obs})}$$

1.) Determine MLE of observed variable distribution

2.) Use connection between both worlds

$$p_{x\mathfrak{y}} = \sum_{y \in \mathfrak{y}} \left(\pi_{xy} \cdot q_{\mathfrak{y}|xy} \right).$$

LATENT	OBSERVABLE	
$\vartheta^{(1)}$, $\vartheta^{(2)}$, $\vartheta^{(3)}$, $\ldots$	$p_{x\mathfrak{y}} := P(\mathcal{Y}_i = \mathfrak{y} \mid X_i = x)$	$\arg_{\vartheta^{(1)}} L(\vartheta^{(1)} \mid \text{data})$ $\arg_{\vartheta^{(2)}} L(\vartheta^{(2)} \mid \text{data})$ $\arg_{\vartheta^{(3)}} L(\vartheta^{(3)} \mid \text{data})$ $\ldots$

$$\Phi : [0,1]^{\dim(\Theta_{lat})} \longrightarrow [0,1]^{\dim(\Theta_{obs})}$$

1.) Determine MLE of observed variable distribution

2.) Use connection between both worlds

$$p_{x\mathfrak{y}} = \sum_{y \in \mathfrak{y}} \left(\pi_{xy} \cdot q_{\mathfrak{y}|xy} \right).$$

3.) Use invariance of the likelihood

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

Cautious ML estimation (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

LATENT	OBSERVABLE	
$\vartheta^{(CAR)} \longrightarrow$	$p_{x\mathcal{y}} := P(\mathcal{Y}_i = \mathcal{y} \mid X_i = x)$	unique $\arg_{\vartheta^{(CAR)}} L(\vartheta^{(CAR)} \mid \text{data})$

$$\Phi : [0,1]^{\dim(\Theta_{lat})} \longrightarrow [0,1]^{\dim(\Theta_{obs})}$$

1.) Determine MLE of observed variable distribution

2.) Use connection between both worlds

$$p_{x\mathcal{y}} = \sum_{y \in \mathcal{y}} \left(\pi_{xy} \cdot q_{\mathcal{y}|xy} \right).$$

3.) Use invariance of the likelihood

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

Estimation of regression coefficients

Final comparison of disambiguation strategy

You

- Model assumptions via the specification of the loss function
- Learning the data and the model simultaneously

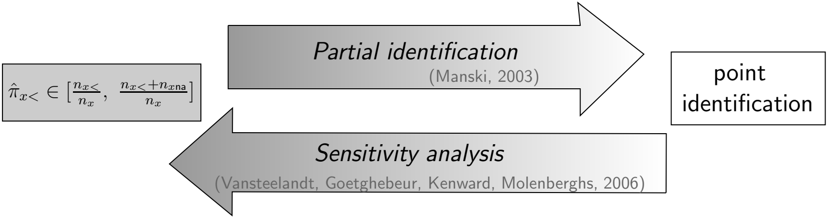
We

- Including model assumptions via the response function
- No learning of the disambigation process
- Only external assumptions about the coarsening behaviour

COMPARISON 3:
Incorporation of coarsening assumptions

Error freeness, superset assumption, CAR, SI

Reliable incorporation of auxiliary information



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.

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

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

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

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.

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

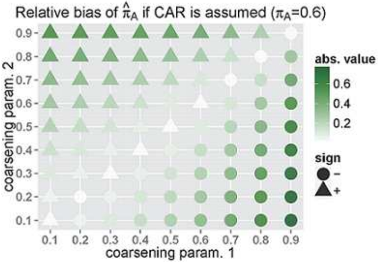
CAR: R=1
(Heitjan, Rubin, 1991)

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

↓

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

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.

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

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

Comparison: Incorporation of assumptions

You

- Superset assumption
- Model assumptions

We