Reliable incorporation of auxiliary information in regression analysis with a coarse categorical response variable

Sommerklausur 2016, Holzhausen

Julia Plass, Thomas Augustin,
Marco Cattaneo, Georg Schollmeyer

10^{th} of June 2016

Motivation of the problem

- Different kinds of uncertainty
  $\Rightarrow$ due to sampling (confidence intervals,...)
  $\Rightarrow$ due to incompleteness (is often forgotten!)

Motivation of the problem

- Different kinds of uncertainty
  $\Rightarrow$ due to sampling (confidence intervals,...)
  $\Rightarrow$ due to incompleteness (is often forgotten!)

- **Missingness mechanism can not be tested**
  - **But:** common procedures (e.g. imputation) require strong (point-identifying) assumptions on the missingness $\Rightarrow$ bias
  - **Why not?:** Proper reflection of the available information on the missingness

    (e.g. Manksi (2016), Kenward, Goetghebeur, Molenberghs (2001), Nguyen (2006))

Motivation of the problem

- Different kinds of uncertainty
  $\Rightarrow$ due to sampling (confidence intervals,...)
  $\Rightarrow$ due to incompleteness (is often forgotten!)

- **Missingness mechanism can not be tested**
  - **But:** common procedures (e.g. imputation) require strong (point-identifying) assumptions on the missingness $\Rightarrow$ bias
  - **Why not?:** Proper reflection of the available information on the missingness
    (e.g. Manksi (2016), Kenward, Goetghebeur, Molenberghs (2001), Nguyen (2006))

$\Rightarrow$ Here: Focus on the tradeoff between information and reliability in the context of **coarse categorical data**

Outline

1. Where do coarse categorical data typically arise?
2. How to deal with coarse categorical data?
3. How to incorporate auxiliary information?
4. Are there possibilities to test on point identifying assumptions (coarsening at random, subgroup independence)?

Coarse categorical data

Coarse data:

- Data are not observed in the resolution originally intended
- Here: imprecise observation of something precise

  (epistemic interpretation, cf. Couso, Dubois, Sánchez, 2014)

Examples:

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

Here:

- response variable is coarse only
- categorical response variable and caegorical covariate(s)

The PASS data example

- German Panel Study "Labour market and social security" (IAB)
- X: self-employed, $\Omega_X = \{0 \text{ (no)}, 1 \text{ (yes)}\}$
- $\mathcal{Y}$: observed income, Y: latent true income, where $\Omega_{\mathcal{Y}} \subseteq \mathcal{P}(\Omega_Y) \setminus \emptyset$

The PASS data example

- German Panel Study "Labour market and social security" (IAB)
- X: self-employed, $\Omega_X = \{0 \text{ (no)}, 1 \text{ (yes)}\}$
- $\mathcal{Y}$: observed income, Y: latent true income, where $\Omega_{\mathcal{Y}} \subseteq \mathcal{P}(\Omega_Y) \setminus \emptyset$

Example 1: ordinal scaled Y

$$
\begin{aligned}
\Omega_Y &= \{< 500, \geq 500, < 2000, \geq 2000\} \\
\Omega_{\mathcal{Y}} &= \{\{< 500\}, \{\geq 500\}, \{< 2000\}, \{\geq 2000\}, \\
&\quad \{< 500, \geq 500\}^{<1000}, \{< 2000, \geq 2000\}^{\geq 1000}, \mathsf{na}\}
\end{aligned}
$$

The PASS data example

- German Panel Study "Labour market and social security" (IAB)
- X: self-employed, $\Omega_X = \{0 \text{ (no)}, 1 \text{ (yes)}\}$
- $\mathcal{Y}$: observed income, Y: latent true income, where $\Omega_{\mathcal{Y}} \subseteq \mathcal{P}(\Omega_Y) \setminus \emptyset$

Example 1: ordinal scaled Y

$$\begin{aligned} \Omega_Y &= \{< 500, \geq 500, < 2000, \geq 2000\} \\ \Omega_{\mathcal{Y}} &= \{\{< 500\}, \{\geq 500\}, \{< 2000\}, \{\geq 2000\}, \\ &\quad \{< 500, \geq 500\}^{<1000}, \{< 2000, \geq 2000\}^{\geq 1000}, \text{na}\} \end{aligned}$$

Example 2: nominal scaled Y (reduces to missing case)

$$\begin{aligned} \Omega_Y &= \{<, \geq\}, \quad (\text{abbr. for: } < 1000\ , \geq 1000) \\ \Omega_{\mathcal{Y}} &= \{\{<\}, \{\geq\}, \overbrace{\{<, \geq\}}^{\text{na}}\} \end{aligned}$$

Basic problem

OBSERVABLE

coarse data

$\mathcal{Y}$

$p_{\mathscr{y}} = P(\mathcal{Y}_i = \mathscr{y})$

LATENT

latent variable

Y

$\pi_{ij} = P(Y_i = j)$

Basic problem

OBSERVABLE

LATENT

coarse data

$\mathcal{Y}$

$p_{\mathscr{y}} = P(\mathcal{Y}_i = \mathscr{Y})$

Observation model $\mathcal{Q}$

error-freeness

$q_{\mathscr{y}|y} =$

$P(\mathcal{Y}_i = \mathscr{Y} | Y_i = y)$

latent variable

Y

$\pi_{ij} = P(Y_i = j)$

Basic problem

OBSERVABLE

LATENT

coarse data

$\mathcal{Y}$

$p_{\mathscr{y}} = P(\mathcal{Y}_i = \mathscr{Y})$

Observation model $\mathcal{Q}$

error-freeness

$q_{\mathscr{y}|y} =$

$P(\mathcal{Y}_i = \mathscr{Y} | Y_i = y)$

latent variable

Y

$\pi_{ij} = P(Y_i = j)$

Main goal:

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

Basic problem

OBSERVABLE

LATENT

coarse data

$\mathcal{Y}$

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

Observation model $\mathcal{Q}$

error-freeness

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

latent variable

Y (nominal scaled)

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

$\pi_{iK} = 1 - \pi_{i1} - \ldots - \pi_{iK-1}$

(multinomial logit model)

Main goal:

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

Basic problem

OBSERVABLE

LATENT

coarse data

$\mathcal{Y}$

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

Observation model $\mathcal{Q}$

error-freeness

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

latent variable

Y (ordinal scaled)

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

π_{ij} with $P(Y_i \leq j | \mathbf{x}_i)$

$= \frac{\exp(\theta_j + \mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\theta_j + \mathbf{x}_i^T \boldsymbol{\beta})}$

for reference category K

$\pi_{iK} = 1 - \pi_{i1} - \ldots - \pi_{iK-1}$

(cummulative logit model)

Main goal:

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

Basic idea (Plass, Augustin, Cattaneo, Schollmeyer, ISIPTA '15)

OBSERVABLE

LATENT

Determine ML estimator

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

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

Basic idea (Plass, Augustin, Cattaneo, Schollmeyer, ISIPTA '15)

OBSERVABLE

LATENT

Determine ML estimator

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

Use the **connection between** p and γ

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

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

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

OBSERVABLE

LATENT

Determine ML estimator

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

Use the **connection between** p and γ

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

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

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

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

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

OBSERVABLE

LATENT

Determine ML estimator

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

Use the **connection between** p and γ

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

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

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

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

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

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

Illustration (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

$$\underline{\hat{\pi}}_{x<} = \frac{n_{x\{<\}}}{n_x}$$

$$\underline{\hat{q}}_{na|x<} = 0$$

$$\overline{\hat{q}}_{na|x\geq} = \frac{n_{xna}}{n_{xna} + n_{x\{\geq\}}}$$

• • •

$$\underline{\hat{\pi}}_{x<} = \frac{n_{x\{<\}} + n_{xna}}{n_x}$$

$$\overline{\hat{q}}_{na|x<} = \frac{n_{xna}}{n_{xna} + n_{x\{<\}}}$$

$$\underline{\hat{q}}_{na|x\geq} = 0$$

Illustration (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

$$\hat{\underline{\pi}}_{x<} = \frac{n_{x\{<\}}}{n_x}$$

$$\hat{\underline{q}}_{na|x<} = 0$$

$$\hat{\overline{q}}_{na|x\geq} = \frac{n_{xna}}{n_{xna} + n_{x\{\geq\}}}$$

$$\hat{\underline{\pi}}_{x<} = \frac{n_{x\{<\}} + n_{xna}}{n_x}$$

$$\hat{\overline{q}}_{na|x<} = \frac{n_{xna}}{n_{xna} + n_{x\{<\}}}$$

$$\hat{\underline{q}}_{na|x\geq} = 0$$

• • •

- Results for subgroup $x = 0$ (not self-employed)

$$\hat{\pi}_{0<} \in [0.16, 0.36], \quad \hat{q}_{na|0\geq} \in [0, 0.23], \quad \hat{q}_{na|0<} \in [0, 0.55]$$

Illustration (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

$$\underline{\hat{\pi}}_{x<} = \frac{n_{x\{<\}}}{n_x}$$

$$\underline{\hat{q}}_{na|x<} = 0$$

$$\overline{\hat{q}}_{na|x\geq} = \frac{n_{xna}}{n_{xna} + n_{x\{\geq\}}}$$

$$\underline{\hat{\pi}}_{x<} = \frac{n_{x\{<\}} + n_{xna}}{n_x}$$

$$\overline{\hat{q}}_{na|x<} = \frac{n_{xna}}{n_{xna} + n_{x\{<\}}}$$

$$\underline{\hat{q}}_{na|x\geq} = 0$$

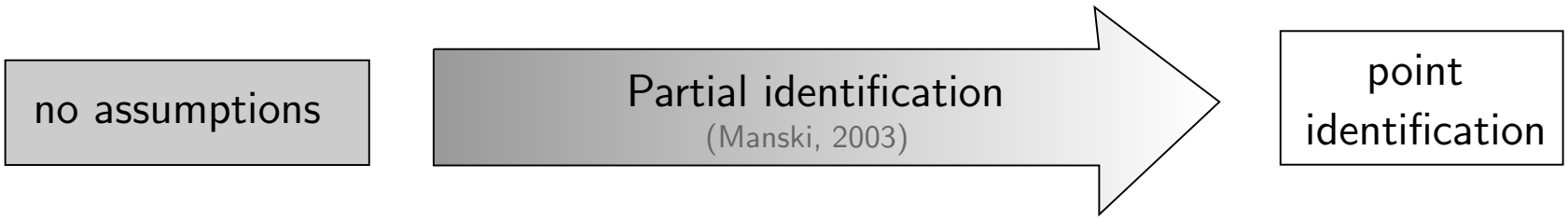
• • •

- Results for subgroup $x = 0$ (not self-employed)

$$\hat{\pi}_{0<} \in [0.16, 0.36], \quad \hat{q}_{na|0\geq} \in [0, 0.23], \quad \hat{q}_{na|0<} \in [0, 0.55]$$

- Results in terms of the regression coefficients:
  $\hat{\beta}_{<0} \in [-1.66, -0.59]$ (Intercept), $\hat{\beta}_{<} \in [0.61, 1.15]$

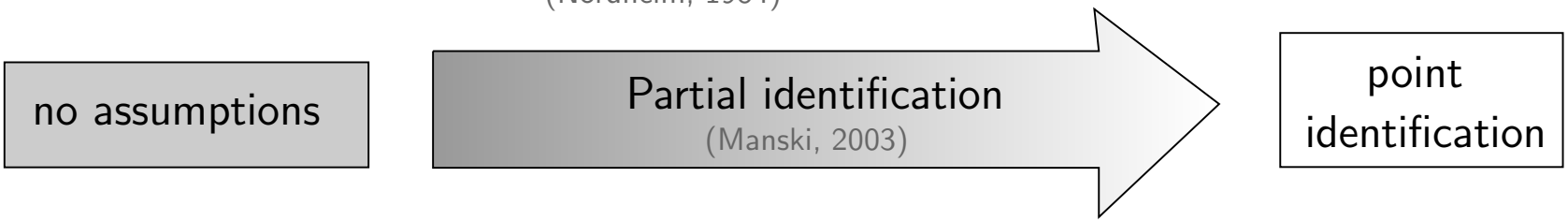
Reliable incorporation of auxiliary information



Reliable incorporation of auxiliary information

include assumption, e.g. via $R = \frac{q_{na|\geq}}{q_{na|<}}$

(Nordheim, 1984)



Reliable incorporation of auxiliary information

include assumption, e.g. via $R = \frac{q_{na|\geq}}{q_{na|<}}$

(Nordheim, 1984)

no assumptions

Partial identification
(Manski, 2003)

point identification

e.g.

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

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

Auxiliary information:

Refined estimators:

Reliable incorporation of auxiliary information

include assumption, e.g. via $R = \frac{q_{na|\geq}}{q_{na|<}}$

(Nordheim, 1984)

no assumptions → Partial identification (Manski, 2003) → point identification

Auxiliary information:

e.g. → $R < 1$

e.g. → CAR: R=1 (Heitjan, Rubin, 1991)

Refined estimators:

$$\hat{\pi}_{x<} \in \left[\frac{n_{\{<\}}}{n}, \frac{n_{\{<\}}}{n_{\{<\}}+n_{\{\geq\}}}\right]$$

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

Coarsening at random & subgroup independence

	coarsening at random (CAR) (Heitjan, Rubin, 1991)	subgroup independence (SI)
Generally	For each fixed $\mathscr{Y}$, $q_{\mathscr{Y}\mid y}$ takes the same values for all y that are consistent with $\mathscr{Y}$	Coarsening does not depend on the value of the covariate
Example	$q_{\text{na}\mid 0<} = q_{\text{na}\mid 0\geq}$ $q_{\text{na}\mid 1<} = q_{\text{na}\mid 1\geq}$	$q_{\text{na}\mid 0<} = q_{\text{na}\mid 1<}$ and $q_{\text{na}\mid 0\geq} = q_{\text{na}\mid 1\geq}$

Testing SI (Plass, Cattaneo, Schollmeyer, Augustin, 2016)

- Data situations that might hint to incompatibility with SI
- Hypotheses:

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

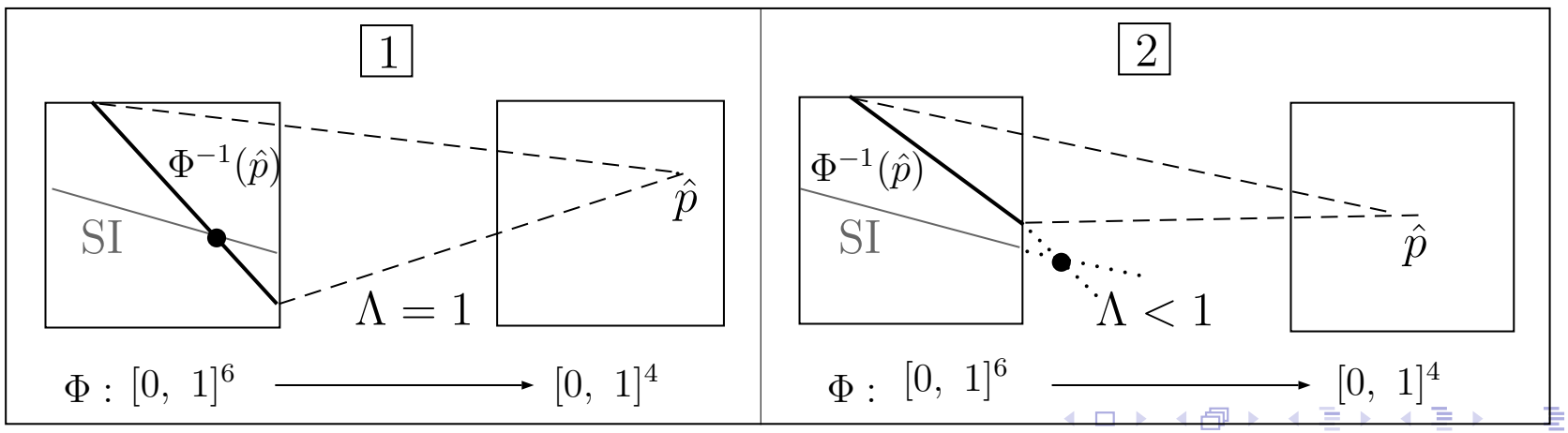
Testing SI (Plass, Cattaneo, Schollmeyer, Augustin, 2016)

- Data situations that might hint to incompatibility with SI
- Hypotheses:

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

- Test statistic based on the Likelihood ratio Λ may be constructed:

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



Summary and outlook

- We studied a cautious regression approach with categorical precisely observed covariates in case of a
  - ... coarse response of nominal scale (multinomial logit model)
  - ... coarse response of ordinal scale (cummulative model)
- Include auxiliary information via restrictions on the coarsening
- Support information via hypothesis tests on the coarsening

Next steps:

- Decision rule for test on SI
- Comparison to traditional approaches
- Application of the approach to other problems (e.g. propensity score matching)

References


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


Heitjan, Rubin.
Ignorability and Coarse Data, Ann. Stat., 1991.


Kenward, Goetghebeur, Molenberghs.
Sensitivity analysis for incomplete categorical data, Stat. Modelling, 2001


Manski.
Credible interval estimates for official statistics with survey nonresponse, J Econometrics , 2016.


Nordheim.
Inference from nonrandomly missing categorical data:An example from a genetic study on Turner's syndrome, J. Am. Stat. Assoc., 1984.


Nguyen.
An introduction to random sets, CRC press, 2006


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.


Plass, Cattaneo, Schollmeyer, Augustin.
Tesing of coarsening mechanisms: Coarsening at random versus subgroup independence, SMPS, 2016


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