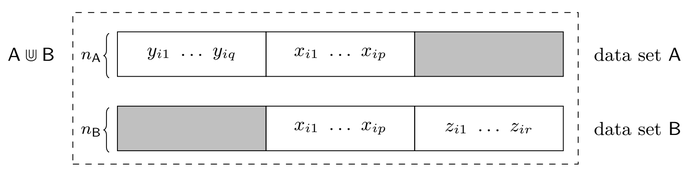
Imprecise Imputation for Statistical Matching

Eva Endres, Paul Fink and Thomas Augustin
Department of Statistics, Ludwig-Maximilians-Universität München (LMU Munich), Germany

the general framework of statistical matching (e.g. D'Orazio et al., 2006)

starting point



assumptions

$A \cup B$ contains $n_A + n_B$ i.i.d. observation, following a common distribution $P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z})$; $\mathcal{I}_A \cap \mathcal{I}_B = \emptyset$

aim

joint information about either $(\mathbf{Y}, \mathbf{Z})$ or $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$

micro approach: construction of complete synthetic data file

macro approach: estimation of the joint distribution

challenge

unidentified parameters are not estimable on $A \cup B$

common approaches

1. assumption of conditional independence
2. utilization of auxiliary information
3. take "uncertainty" (ambiguity) into account
   → estimate lower and upper bounds

missing data interpretation

nonparametric random hot deck imputation (micro approach)

imprecise imputation I

original data

y_1	y_2	x_1	x_2	z_1	z_2
1	2	1	0		
0	2	0	0		
		1	0	0	0
		1	0	1	1
		0	0	1	2

domain imputation (cf. Ramoni & Sebastiani, 2001)

y_1	y_2	x_1	x_2	z_1	z_2
1	2	1	0	$\{0,1\}$	$\{0,1,2\}$
0	2	0	0	$\{0,1\}$	$\{0,1,2\}$
$\{0,1\}$	$\{0,1,2\}$	1	0	0	0
$\{0,1\}$	$\{0,1,2\}$	1	0	1	1
$\{0,1\}$	$\{0,1,2\}$	0	0	1	2

with donor classes

variable-wise imputation

y_1	y_2	x_1	x_2	z_1	z_2
1	2	1	0	$\{0,1\}$	$\{0,1\}$
0	2	0	0	$\{1\}$	$\{2\}$
$\{1\}$	$\{2\}$	1	0	0	0
$\{1\}$	$\{2\}$	1	0	1	1
$\{0\}$	$\{2\}$	0	0	1	2

case-wise imputation

(y_1, y_2)	(x_1, x_2)	(z_1, z_2)
$(1,2)$	$(1,0)$	$\{(0,0),(1,1)\}$
$(0,2)$	$(0,0)$	$\{(1,2)\}$
$\{(1,2)\}$	$(1,0)$	$(0,0)$
$\{(1,2)\}$	$(1,0)$	$(1,1)$
$\{(0,2)\}$	$(0,0)$	$(1,2)$

imprecise imputation II

domains

$x_1, x_2, y_1, z_1 \in \{0,1\}$
$y_2, z_2 \in \{0,1,2\}$

Y_1: sex
Z_1: pregnancy

logical constraint

$$y_1 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$$

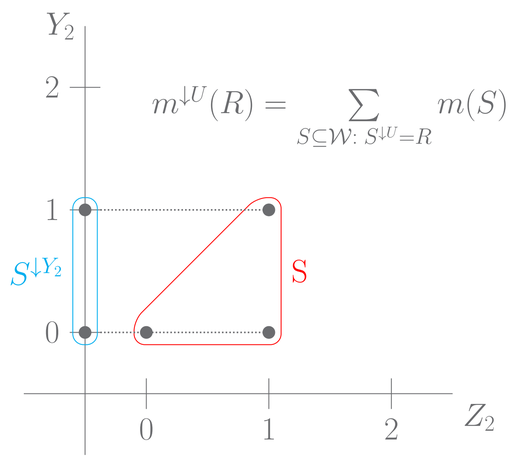
$$z_1 = \begin{cases} 1 & \text{not pregnant} \\ 0 & \text{pregnant} \end{cases}$$

constraint: no pregnant men

$(y_1, y_2, x_1, x_2, z_1, z_2)$
$\{(1,2,1,0,0,0); (1,2,1,0,0,1); (1,2,1,0,0,2); (1,2,1,0,1,0); (1,2,1,0,1,1); (1,2,1,0,1,2)\}$
$\{(0,2,0,0,0,0); (0,2,0,0,0,1); (0,2,0,0,0,2); (0,2,0,0,1,0); (0,2,0,0,1,1); (0,2,0,0,1,2)\}$
$\{(0,0,1,0,0,0); (0,1,1,0,0,0); (0,2,1,0,0,0); (1,0,1,0,0,0); (1,1,1,0,0,0); (1,2,1,0,0,0)\}$
$\{(0,0,1,0,1,1); (0,1,1,0,1,1); (0,2,1,0,1,1); (1,0,1,0,1,1); (1,1,1,0,1,1); (1,2,1,0,1,1)\}$
$\{(0,0,0,0,1,2); (0,1,0,0,1,2); (0,2,0,0,1,2); (1,0,0,0,1,2); (1,1,0,0,1,2); (1,2,0,0,1,2)\}$

estimation on (partially) set-valued data

marginalization



$$m^{\downarrow U}(R) = \sum_{S \subseteq \mathcal{W}:\, S^{\downarrow U} = R} m(S)$$

focusing (Dubois & Prade, 1992)

Let $R \in \mathcal{P}(\mathcal{W})$ be the event of interest, and $Q \subseteq \mathcal{W}$ the interested subclass. For $\mathrm{Pl}(Q) > 0$, the focused belief function and the focused plausibility function are given by

$$\mathrm{Bel}(R|Q) = \frac{\mathrm{Bel}(R \cap Q)}{\mathrm{Bel}(R \cap Q) + \mathrm{Pl}(\bar{R} \cap Q)}$$

$$\mathrm{Pl}(R|Q) = \frac{\mathrm{Pl}(R \cap Q)}{\mathrm{Pl}(R \cap Q) + \mathrm{Bel}(\bar{R} \cap Q)}$$

example on case-wise imputed data

marginals

$\mathrm{Bel}(Z_1 = \{1\}) = m_{Z_1}(\{1\}) = 0.6$
$\mathrm{Pl}(Z_1 = \{1\}) = m_{Z_1}(\{1\}) + m_{Z_1}(\{0,1\}) = 0.8$

$$m_{Z_1}(R) = \begin{cases} 0.2 & \text{if } R = \{0\} \\ 0.6 & \text{if } R = \{1\} \\ 0.2 & \text{if } R = \{0,1\} \\ 0 & \text{else} \end{cases}$$

$\mathrm{Bel}(Y_1 = \{1\}, Z_1 = \{1\}) = m_{Y_1,Z_1}(\{(1,1)\}) = 0.2$
$\mathrm{Pl}(Y_1 = \{1\}, Z_1 = \{1\}) = m_{Y_1,Z_1}(\{(1,1)\}) + m_{Y_1,Z_1}(\{(1,0);(1,1)\}) = 0.4$

$$m_{Y_1,Z_1}(R) = \begin{cases} 0.4 & \text{if } R = \{(0,1)\} \\ 0.2 & \text{if } R = \{(1,0)\} \\ 0.2 & \text{if } R = \{(1,1)\} \\ 0.2 & \text{if } R = \{(1,0);(1,1)\} \\ 0 & \text{else} \end{cases}$$

conditionals via focusing

$\mathrm{Bel}(Z_1 = \{1\}|Z_2 \subseteq \{0,1\}) = \dfrac{\mathrm{Bel}(\{(1,0);(1,1)\})}{\mathrm{Bel}(\{(1,0);(1,1)\}) + \mathrm{Pl}(\{(0,0);(0,1)\})} = 0.2/(0.2+0.4) = 1/3$

$\mathrm{Bel}(\{(1,0);(1,1)\}) = m_{Z_1,Z_2}(\{(0,0);(0,1);(1,0);(1,1)\}) = 0.2$
$\mathrm{Pl}(\{(0,0);(0,1)\}) = m_{Z_1,Z_2}(\{(0,0);(0,1);(1,0);(1,1)\}) + m_{Z_1,Z_2}(\{(0,0)\}) = 0.2 + 0.2 = 0.4$

$\mathrm{Pl}(Z_1 = \{1\}|Z_2 \subseteq \{0,1\}) = \dfrac{\mathrm{Pl}(\{(1,0);(1,1)\})}{\mathrm{Pl}(\{(1,0);(1,1)\}) + \mathrm{Bel}(\{(0,0);(0,1)\})} = 0.4/(0.4+0.2) = 2/3$

$\mathrm{Bel}(\{(0,0);(0,1)\}) = m_{Z_1,Z_2}(\{(0,0)\}) = 0.2$
$\mathrm{Pl}(\{(1,0);(1,1)\}) = m_{Z_1,Z_2}(\{(0,0);(0,1);(1,0);(1,1)\}) + m_{Z_1,Z_2}(\{(1,1)\}) = 0.2 + 0.2 = 0.4$

$R = \{(1,0);(1,1);(1,2)\}$
$\bar{R} = \{(0,0);(0,1);(0,2)\}$
$Q = \{(0,0);(0,1);(1,0);(1,1)\}$

references

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*, Wiley, Chichester, United Kingdom.
Ramoni, M. and Sebastiani, P. (2001). Robust learning with missing data, *Machine Learning* **45**(2): 147–170.
Dubois, D. and Prade, H. (1992). Evidence, knowledge, and belief functions, *International Journal of Approximate Reasoning* **6**(3): 295–319.