Incoherence correction strategies in statistical matching

Andrea Capotorti and Barbara Vantaggi

- We deal with the managing of inconsistencies inside the Statistical Matching (integration of sources) framework;
  - when logical relations among the variables are present incoherence can arise in the probability evaluations
- different methods can be used to remove such incoherences:
  - maximize the "partial likelihood function" on the base of observed data;
  - least committal imprecise probability extensions;
  - specific precise "distances" minimization.

Statistical Matching: integration of sources in a coherent setting

$(X_1, Y_1), ..., (X_{n_A}, Y_{n_A})$ and $(X_{n_A+1}, Z_{n_A+1}), ..., (X_{n_A+n_B}, Z_{n_A+n_B})$ two random samples (with a finite range) related to two sources A and B concerning the same population of interest, and drawn according to the same sampling scheme.
We can elicit from the two files the relevant probability values

$$\mathcal{Y}_{j|i} = P_{Y|(X=x_i)}(Y = y_j)\ \mathcal{Z}_{k|i} = P_{Z|(X=x_i)}(Z = z_k)\ \mathcal{X}_i = P_X(X = x_i)$$

that even separately coherent, whenever there are some logical constraints among the variables Y and Z, could induce an incoherent **whole assessment**

$$(\mathcal{E}, \mathbf{p}) \text{ with } \begin{array}{l} \mathcal{E} = \left\{ \begin{array}{r} (X = x_i),\ (Y = y_j)|(X = x_i),\ (Z = z_k)|(X = x_i) \\ \text{for any } x_i, y_j, z_k \end{array} \right\}, \\ \mathbf{p} = \{\mathcal{X}_i, \mathcal{Y}_{j|i}, \mathcal{Z}_{k|i}\}_{i,j,k} \quad . \end{array}$$

Anyhow, incoherence can *localize* only in association to elements of $\mathcal{E}$ with the same conditioning event $(X = x_i)$.

Coherent Extension

To adjust the initially incoherent assessment $(\mathcal{E}, \mathbf{p})$ it is possible to determine a coherent sub-assessment $(\mathcal{G}, \mathbf{p}_{|\mathcal{G}})$ with maximal cardinality and coherently extend it to the rest $\mathcal{F} = \mathcal{E} \setminus \mathcal{G}$ by the generalized Bayesian updating scheme obtaining an imprecise sub-assessment

$$(\mathcal{F}, [\underline{\mathbf{p}_{\mathcal{F}}}, \overline{\mathbf{p}_{\mathcal{F}}}]).$$

Note that inference on decision targets can be performed again through the generalized Bayesian updating scheme but applied to imprecise evaluations.

Whenever too vague, inference bounds can be eventually reduced to coherent cores, i.e. *total* coherent subintervals with highest degree of support.

Minimization of (pseudo)distances among probability distributions

An other way to correct incoherence is to minimize one of most widely adopted divergencies among conditional assessments $\mathbf{p} = [p_1, \ldots, p_n]$ and $\mathbf{q} = [q_1, \ldots, q_n]$ on the same set of conditional events $\mathcal{E}$:

$$L1(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |q_i - p_i|; \; L2(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} (q_i - p_i)^2; \; KL(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} (q_i \ln(q_i/p_i) - q_i + p_i)$$

... but for partial conditional probability assessments $\mathbf{p} \in (0, 1)^n$ on $\mathcal{E}$ recently we tailored the following "discrepancy"

$$\Delta(\mathbf{p}, \boldsymbol{\alpha}) = \sum_{i | \alpha(H_i) > 0} \alpha(H_i) \left(q_i \ln \frac{q_i}{p_i} + (1 - q_i) \ln \frac{(1 - q_i)}{(1 - p_i)} \right),$$

with $\mathbf{q}_{\boldsymbol{\alpha}}$ induced by a probability mass function $\boldsymbol{\alpha}$, but for Statistical Matching ...

A mixture of discrepancies

... it is better to minimize the following discrepancy reformulation:

$$\Delta_{mix}(\mathbf{p}, \{\boldsymbol{\alpha}_i\}_i) = \sum_i \varkappa_i \left[\sum_j \left(q_{j|i}^{\alpha_i} \ln \frac{q_{j|i}^{\alpha_i}}{\mathcal{Y}_{j|i}} + (1 - q_{j|i}^{\alpha_i}) \ln \frac{(1 - q_{j|i}^{\alpha_i})}{(1 - \mathcal{Y}_{j|i})} \right) + \right.$$
$$\left. + \sum_k \left(q_{k|i}^{\alpha_i} \ln \frac{q_{k|i}^{\alpha_i}}{\mathcal{Z}_{k|i}} + (1 - q_{k|i}^{\alpha_i}) \ln \frac{(1 - q_{k|i}^{\alpha_i})}{(1 - \mathcal{Z}_{k|i})} \right) \right]$$

where each distribution α_i works just on the sample space spanned by the conditional events $\{(Y = y_j)|(X = x_i), (Z = z_k)|(X = x_i)\}$, it is constrained to fulfill the normalizing condition $\alpha_i(X = x_i) = \varkappa_i$, and generates the conditional probabilities

$$q_{j|i}^{\alpha_i} = \frac{\alpha_i(Y = y_j)}{\alpha_i(X = x_i)} \quad q_{k|i}^{\alpha_i} = \frac{\alpha_i(Z = z_k)}{\alpha_i(X = x_i)}.$$

A practical example

We applied the three methodologies (likelihood maximization, coherent extension and distances minimizations) to real data representing a subset of employees with three categorical variables (Age, Educational Level and Professional Status) discretized into:
A_1=15-17 y.o., A_2=18-22 y.o., A_3=23-64 y.o., $A_4 = (\geq 65)$ y.o;
E_1=None or comp. sch., E_2=Voc. sch., E_3=Second. sch., E_4=Degree;
S_1=Manager, S_2=Clerk, S_3=Worker.
We observed the following conditional assessment ($--$ denotes impossible configuration):

	A_1	A_2	A_3	A_4
$P(\cdot)$	0.0065	0.0238	0.9594	0.0104
$P(S_1\|\cdot)$	$--$	$--$	0.1616	0.6667
$P(S_2\|\cdot)$	$--$	0.2273	0.3913	0.1111
$P(S_3\|\cdot)$	1	0.7727	0.4293	0.2222
$P(E_1\|\cdot)$	1	0.4242	0.3419	0.6667
$P(E_2\|\cdot)$	0	0.1818	0.0918	0
$P(E_3\|\cdot)$	$--$	0.3940	0.4176	0.2
$P(E_4\|\cdot)$	$--$	$--$	0.1422	0.1333

.

$P(\cdot|A_4)$ is not coherent since from logical constraints it follows $E_1 \wedge S_1 = \emptyset$ and $E_1 \subseteq S_3$ while we have $P(E_1|A_4) + P(S_1|A_4) + P(S_3|A_4) > 1$ and $P(E_1|A_4) > P(S_1|A_4)$.

Several incoherence correction with associated inference results for the target $S3|E4$

	$S_1\|A_4$	$S_2\|A_4$	$S_3\|A_4$	$E_1\|A_4$	$E_2\|A_4$	$E_3\|A_4$	$E_4\|A_4$	$S_3\|E_4$
p	0.6667	0.1111	0.2222	0.6667	0	0.2000	0.1333	$\emptyset$
$L1_{\|\mathcal{F}}$	0.2222	-	0.6667	0.6667	-	-	-	[0,0.6285]
$L1_{\|A4}$	0.5266	0	0.4734	0.4734	0	0.2836	0.2431	[0,0.6234]
$L2_{\|A4}$	0.5333	0.0389	0.4278	0.4278	0.0389	0.3	0.2333	[0,0.6238]
$KL_{\|A4}$	0.4856	0.1179	0.3965	0.3965	0.1179	0.2914	0.1942	[0,0.6257]
Δ_{mix}	0.4985	0.0939	0.4077	0.4077	0.0939	0.2943	0.2042	[0,0.6252]
ML	0.4286	0.0714	0.5000	0.5000	0	0.3000	0.2000	[0,0.6254]
$IP_{\mathcal{E}\setminus\mathcal{F}}$	[0 , 0.2222]	-	[0.6667 0.8889]	-	-	-	-	[0,0.6386]
core								[0.0017,0.6286]
$IP_{\mathcal{E}\setminus\{\cdot\|A_4\}}$	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0,0.6607]
core								[0,0.6349]

where $\mathcal{F} = \{E_1|A_4, S_1|A_4, S_3|A_4\}$ is the minimal cardinality subset of $\mathcal{E}$ associated to incoherent values, and...

Conclusion

- $L1_{|\mathcal{F}}$ ($L1$ minimization for sub-ass. on $\mathcal{F}$) and $IP_{\mathcal{E}\setminus\mathcal{F}}$ (coherent imprecise extension induced by sub-ass. on $\mathcal{E}\setminus\mathcal{F}$) perform quite well: even though a drastic change on the probability values, they induce quite reasonable inference bounds;
- $L1_{|A4}$ ($L1$ minimization for sub-ass. $P(\cdot|A_4)$) and ML (maximum likelihood estimation) give similar results and in particular they leave to 0 the probability of $E_2|A_4$ since the absence of observations in the original data;
- others adjustments induced by (pseudo)distances minimizations for sub-ass. $P(\cdot|A_4)$ have all quite similar behaviors;
- Δ_{mix} has the advantage of automatically localize of the scenarios where the adjustment can be performed;
- the wider imprecise correction $IP_{\mathcal{E}\setminus\{\cdot|A_4\}}$ (coherent imprecise extension induced by sub-ass. on $\mathcal{E}\setminus\{\cdot|A_4\}$), surely performs worst.