Measuring uncertainty in statistical matching for discrete distributions

Pier Luigi Conti

Dipartimento di Scienze Statistiche - Sapienza Università di Roma

Daniela Marella

Dipartimento di Scienze dell'Educazione - Università Roma Tre

Statistical Matching: introductory aspects

The goal of *statistical matching* (sometimes called "data fusion") is to combine information available in different sample surveys from the same target population, in order to produce a synthetic dataset containing all variates separately observed in such different sample surveys.

Formally, let (X, Z, Y) be a three-dimensional random variable (r.v.), and let A and B be two samples of size n_A, n_B, respectively, composed by i.i.d. copies of (X, Z, Y). Because of the adopted observation process, only (X, Y) are observed in A, and only (X, Z) are observed in B. Hence, the r.v. Z (Y) is *missing* in A (B).

In symbols, actual sample observations are denoted by

- Sample A: $(X_1^A, Y_1^A), \ldots, (X_{n_A}^A, Y_{n_A}^A)$;
- Sample B: $(X_1^B, Z_1^B), \ldots, (X_{n_B}^B, Z_{n_B}^B)$.

Statistical matching consists in producing a complete data set, where the variables X, Y, Z of interest are simultaneously recorded.

No parametric assumptions are made on the joint distribution function (d.f.) of X, Y, Z (and on its marginal d.f.s, as well).

Since no joint observations of X, Y, Z are available, *sample data are unable to identify the joint d.f. of* (X, Y, Z).

Two main approaches to statistical matching have been considered.

1. Techniques based on the conditional independent assumption between Y and Z given X (CIA);
2. Techniques using external auxiliary information on the statistical relationship between Y and Z (e.g. an additional sample C where (X, Z, Y) are jointly observed is available).

Identification problem: if either the CIA is a misspecified assumption or external auxiliary information is not available or "too weak", the statistical model for (X, Z, Y) is not identifiable (*model uncertainty*).

Goals of this talk

1. Investigation of the model uncertainty when X, Y, Z are ordered categorical variates.
2. Definition of an overall measure of uncertainty for non-identifiable models.
3. Evaluation of the effect on model uncertainty due to the introduction of logical constraints. Such constraints include restrictions on the support of the joint distribution of (Z, Y) given X. An estimator of the model uncertainty measure is proposed, and its asymptotic behavior is studied.

Basic assumption. X, Y, Z are discrete r.v.s. Furthermore, X possesses I categories, Y possesses J categories, and Z possesses K categories.

With no loss of generality, we can assume that $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$, are the (ordered) categories taken by X, Y and Z, respectively.

Symbols used

$\gamma_{jk|i}$ = conditional probability $Pr(Y = j,\, Z = k|X = i)$;

$\phi_{j|i}$ = conditional probability $Pr(Y = j|X = i)$;

$\psi_{k|i}$ = marginal probability $Pr(Z = k|X = i)$

Conditionally on $X = i$, the distribution functions (d.f.'s) of (Y, Z), Y, and Z, are equal to

$$H_{j,k|i} = \sum_{j=1}^{j} \sum_{z=1}^{k} \gamma_{yz|i},$$

$$F_{j|i} = \sum_{y=1}^{j} \phi_{y|i},$$

$$G_{k|i} = \sum_{z=1}^{k} \psi_{z|i},$$

respectively, as $j = 1, \ldots, J$, $k = 1, \ldots, K$, $i = 1, \ldots, I$.

Model uncertainty

The sampling mechanism is unable to identify the joint distribution of (X, Y, Z), but only a *class of possible distributions* of (X, Y, Z). Roughly speaking, this produces *uncertainty* about the actual distribution of (X, Y, Z) in the above mentioned class, even when the marginal distributions of (X, Y) and (X, Z) are known.

Of course, this happens because the sampling mechanism is actually unable to identify the conditional distribution of (Y, Z) given X. This is the actual reason for the lack of identifiability of the distribution of (X, Y, Z). Hence, considering uncertainty about the conditional distribution of (Y, Z) given X is equivalent to consider uncertainty on the distribution of the triple (X, Y, Z).

When no further information on the distribution of $(X,\, Y,\, Z)$ is available, the bivariate (conditional) d.f. $H_{jk|i}$ lies in between the lower bound

$$H^{-}_{jk|i} = \max(0,\ F_{j|i} + G_{k|i} - 1)$$

and the upper bound

$$H^{+}_{jk|i} = \min(F_{j|i},\ G_{k|i}).$$

In symbols:

$$H^{-}_{jk|i} = \max(0, F_{j|i}+G_{k|i}-1) \leq H_{jk|i} \leq \min(F_{j|i}, G_{k|i}) = H^{+}_{jk|i}$$

for each triple j, k, i.

Measures of uncertainty

The lower bound $H^-_{jk|i}$ and the upper bound $H^+_{jk|i}$ allow one to construct a a measure of *pointwise uncertainty* for $H_{jk|i}$:

$$H^+_{jk|i} - H^-_{jk|i}$$

namely the length of the interval $[H^-_{jk|i},\ H^+_{jk|i}]$ $H_{jk|i}$ lies in.

Next step is to summarize all pointwise measures of uncertainty above defined (one for each triple $(j,\ k,\ i)$ into a unique, *overall measure*.

To this purpose, we may take the average length

$$\Delta = \int_{\mathbb{R}^3} (H^{+}_{jk|i} - H^{-}_{jk|i})\, dT(i,\, j,\, k)$$

where $T(i,\, j,\, k)$ is a weight function on $\mathbb{R}^3$, *i.e.* a measure having total mass 1.

A "natural" choice consists in taking

$$d\,T(i,j,k) = d\,F(j|i)\, d\,G(k|i)\, d\,Q(i) = \phi_{j|i}\, \psi_{k|i}\, \xi_i.$$

This distribution is "natural" because: i) it is the simplest choice given the available d.f.s $F(j|i)$, $G(k|i)$, $Q(i)$ and makes the integral in Δ easily computable in many cases; ii) among all the possible associations between Y and Z, we consider a neutral position, *i.e.* we do not give preference to any specific positive or negative association.

On the basis of the above remarks, a conditional measure of uncertainty is

$$\Delta^{x=i} = \sum_{j=1}^{J} \sum_{k=1}^{K} (H^{+}_{jk|i} - H^{-}_{jk|i})\, \phi_{j|i}\, \psi_{k|i} \tag{1}$$

As a matter of fact, by averaging (1) with respect to X, we obtain the overall measure of uncertainty Δ:

$$\Delta = \sum_{i=1}^{I} \Delta^{x=i}\, \xi_i. \tag{2}$$

The unconditional uncertainty measure is a weighted mean of conditional uncertainty measures. Then, the larger $\Delta^{x=i}$s, the more uncertain the data generating statistical model.

Evaluating uncertainty under constraints

Uncertainty about the statistical model can be reduced by extra-sample auxiliary information. Such an external knowledge reduces the uncertainty characterizing the statistical matching problem, since some models for (X, Z, Y) become "impossible" and must be excluded from the set of plausible distribution functions.

Several kinds of auxiliary information can be (realistically) introduces. A few examples are listed in the sequel.

$-$ Constraints on the association between Y and Z. For instance, Y and Z are positively quadrant dependent (conditionally on X): $H_{jk|i} \geq F_{j|i}\, G_{k|i}$ for every $j,\ k,\ i$.

$-$ The correlation coefficient between Y and Z (conditionally on X) ranges on an interval $(\rho^-,\ \rho^+)$, with $\rho^- > -1$ and/or $\rho^+ > 1$.

$-$ *Logical constraints* between Y and Z: some pairs $(k,\ k)$ do not lie on the support of $(Y,\ Z)$ (conditionally on X).

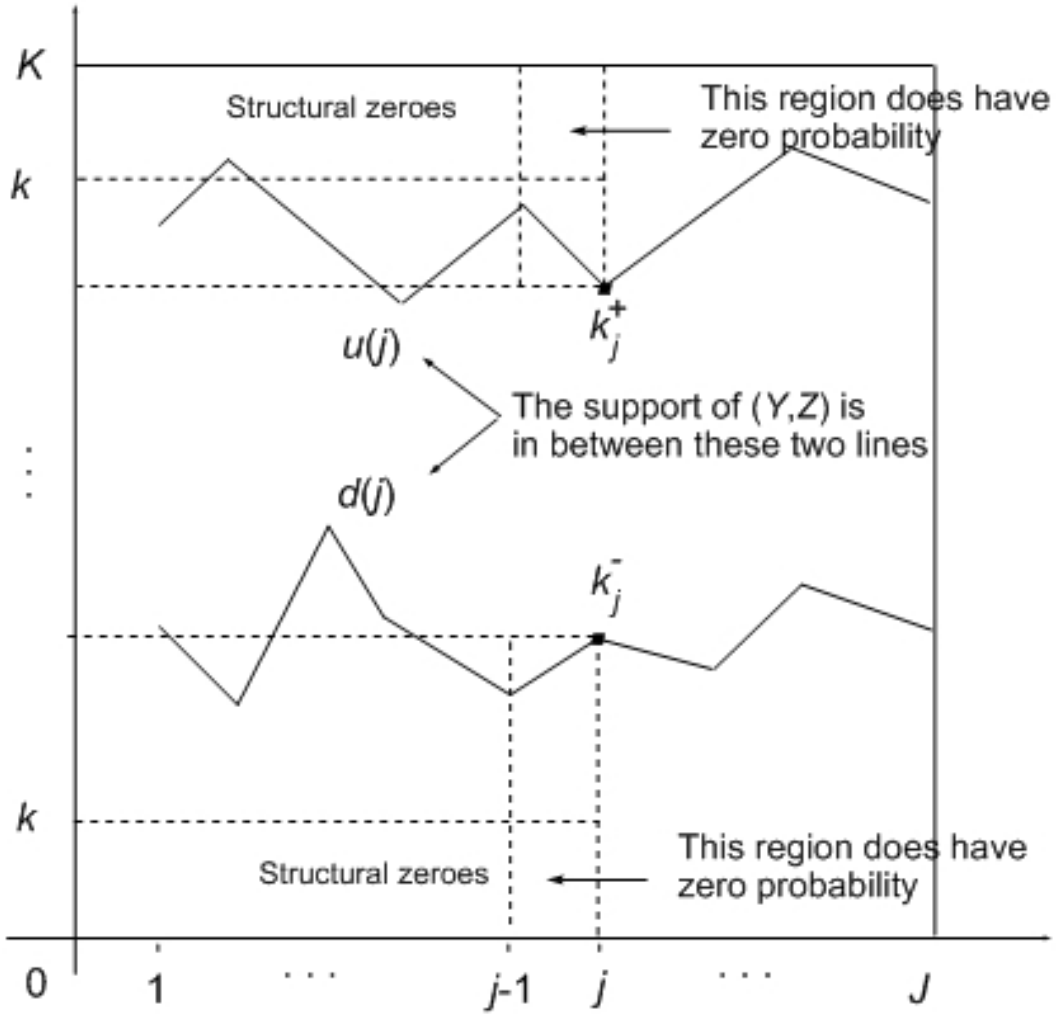
In this talk, we focus on *logical constraints* between Y and Z (given X). Due to the discrete nature of $(X,\ Y,\ Z)$, logical constraints are equivalent to *structural zeroes*.

In the sequel, we will concentrate on structural zeros that reduce the support of Y and Z in a "regular way", useful to manage uncertainty.

From an intuitive point of view, the presence of structural zeros is a form of information that should reduce the uncertainty about $H_{jk|i}$. More formally, the main idea is to use structural zeroes in order improve the lower bound $H^{-}_{jk|i}$ and/or the upper bound $H^{+}_{jk|i}$, and hence to reduce the measure of uncertainty previously defined.

The kind of "regular regions of structural zeroes" considered in the present talk in depicted in Figure 1.

Figure 1: *Structural zeros in a Y-regular domain.*



For each $j \in \{1, \ldots, J\}$, define the two integers:

$$k_j^+ = \text{largest integer } k \text{ such that } \gamma_{jk|i} > 0;$$
$$k_j^- = \text{smallest integer } k \text{ such that } \gamma_{jk|i} > 0.$$

Of course, there exist integers j_1, j_2 such that $k_{j_1}^+ = K$ and $k_{j_2}^- = 1$.

Analogously, for each $k \in \{1, \ldots, K\}$, define the two integers:

$$j_k^+ = \text{largest integer } k \text{ such that } \gamma_{jk|i} > 0;$$
$$j_k^- = \text{smallest integer } k \text{ such that } \gamma_{jk|i} > 0.$$

Again, there exist integers k_1, k_2 such that $j_{k_1}^+ = J$ and $j_{k_2}^- = 1$.

The support of $(Y,\, Z)$ (given X) is *Y-regular* if, for all $j = 1,\, \ldots,\, J$,

$$\gamma_{jk|i} = 0 \;\forall\, k > k_j^+, \quad \gamma_{jk|i} = 0 \;\forall\, k < k_j^-. \tag{3}$$

Similarly, the support of $(Y,\, Z)$ (given X) is *Z-regular* if, for all $k = 1,\, \ldots,\, K$,

$$\gamma_{jk|i} = 0 \;\forall\, j > j_k^+, \quad \gamma_{jk|i} = 0 \;\forall\, j < j_k^-. \tag{4}$$

The algorithm to compute the lower bound $H^{-}_{jk|i}$ and the upper bound $H^{+}_{jk|i}$ is described below.

Step 0 *Take* $j = 1$, *and compute* k_j^-, k_j^+. *Define next* $F_{0|i} = 0$, $G_{0|i} = 0$, *and*

$$H^{y+}_{jk|i} = \begin{cases} 0 \quad if \quad k < k_j^- \\ \min(F_{j|i},\, G_{k|i}) \quad if \quad k_j^- \le k \le k_j^+ \\ F_{j|i} \quad if \quad k > k_j^+ \end{cases}$$

and

$$H^{y-}_{jk|i} = \begin{cases} 0 \quad if \quad k < k_j^- \\ \max(0,\, F_{j|i} + G_{k|i} - 1) \quad if \quad k_j^- \le k \le k_j^+ \\ F_{j|i} \quad if \quad k > k_j^+ \end{cases}.$$

Replace j *by* $j + 1$. *Go to Step 1.*

Step 1 *If $j > J$, then go to Step 3. Otherwise, go to Step 2.*

Step 2 *Compute k_j^-, k_j^+. Define*

$$H_{jk|i}^{y+} = \begin{cases} H_{j-1k|i}^{y+} & if \;\; k < k_j^- \\ \min(F_{j|i},\, G_{k|i}) & if \;\; k_j^- \le k < k_j^+ \\ \min(F_{j|i},\, G_{k|i},\, F_{j|i} - F_{j-1|i} + H_{j-1k|i}^{y+}) & if \;\; k > k_j^+ \end{cases}$$

and

$$H_{jk|i}^{y-} = \begin{cases} \max(0,\, F_{j|i} + G_{k|i} - 1,\, H_{j-1k|i}^{-}) & if \;\; k \le k_j^- \\ \max(0,\, F_{j|i} + G_{k|i} - 1) & if \;\; k_j^- \le k \le k_j^+ \\ \max(0,\, F_{j|i} + G_{k|i} - 1,\, F_{j|i} - F_{j-1|i} + H_{j-1k|i}^{y-}) & if \;\; k > \end{cases}$$

Replace j by $j+1$. Go to Step 1.

Step 3 *Stop. $H_{jk|i}^{y-}$ and $H_{jk|i}^{y+}$ are the lower and upper bounds for $H_{jk!i}$, respectively.*

The same reasoning holds also when the support of $(Y,\, Z)$ given X is Z-regular: it is enough to exchange the role of Y and Z. Let us denote by $H^{z-}_{jk|i}$ and $H^{z+}_{jk|i}$ be the lower and upper bounds for $H_{jk!i}$, respectively, in case of Z-regularity. Finally, define:

$$H^{+}_{jk|i} = \begin{cases} \min(H^{y+}_{jk|i},\, H^{z+}_{jk|i}) \;\; if \;\; the\ support\ of\ (Y,\, Z\,|X) \\ \qquad\qquad\qquad\qquad\qquad\qquad is\ both\ Y\ and\ Z-regular \\ H^{y+}_{jk|i} \qquad\quad if \;\; the\ support\ of\ (Y,\, Z\,|X) \\ \qquad\qquad\qquad\qquad\qquad\qquad is\ only\ Y-regular \\ H^{z+}_{jk|i} \qquad\quad if \;\; the\ support\ of\ (Y,\, Z\,|X) \\ \qquad\qquad\qquad\qquad\qquad\qquad is\ only\ Z-regular \end{cases}$$

and

$$H^{-}_{jk|i} = \begin{cases} \max(H^{y-}_{jk|i},\, H^{z-}_{jk|i}) & if \quad the\ support\ of\ (Y,\, Z\,|X) \\ & \quad is\ both\ Y\ and\ Z-regular \\ H^{y-}_{jk|i} & if \quad the\ support\ of\ (Y,\, Z\,|X) \\ & \quad is\ only\ Y-regular \\ H^{z-}_{jk|i} & if \quad the\ support\ of\ (Y,\, Z\,|X) \\ & \quad is\ only\ Z-regular \end{cases}$$

The main result obtained is that the lower and upper bounds computed in the presence of "regular structural zeroes" improve the ones previously computed:

$$H^{-}_{jk|i} \geq \max(0,\ F_{j|i} + G_{k|i} - 1)$$
$$H^{+}_{jk|i} \leq \min(F_{j|i},\ G_{k|i})$$

Hence, *using the new bonds* $H^{-}_{jk|i}$ *and* $H^{+}_{jk|i}$ *produces measures of uncertainty* $\Delta^{x=i}$, Δ *smaller than those obtained in the absence of structural zeroes.*

Statistical inference on uncertainty measures under constraints

The uncertainty measures introduced can be estimated on the basis of the sample data. The notation used is the following.

Sample A: $(X_1^A, Y_1^A), \ldots, (X_{n_A}^A, Y_{n_A}^A)$;
Sample B: $(X_1^B, Z_1^B), \ldots, (X_{n_B}^B, Z_{n_B}^B)$.

$n_{A,i}^x$ ($n_{B,i}^x$): number of sample observations is sample A (B) such that $X = i$

$n_{A,ij}^{xy}$ ($n_{B,ik}^{xz}$): number of observations in sample A (B) such that $X = i$ and $Y = j$ ($X = i$ and $Z = k$)

The probabilities ξ_i, $\phi_{j|i}$, $\psi_{k|i}$ can be first estimated by the corresponding sample proportions

$$\widehat{\xi}_i = \frac{n^x_{A,i} + n^x_{B,i}}{n_A + n_B}, \quad i = 1, \ldots, I;$$

$$\widehat{\phi}_{j|i} = \frac{n^{xy}_{A,ij}}{n^x_{A,i}}, \quad i = 1, \ldots, I, \ j = 1, \ldots, J;$$

$$\widehat{\psi}_{k|i} = \frac{n^{xz}_{B,ik}}{n^x_{B,i}}, \quad i = 1, \ldots, I, \ k = 1, \ldots, K.$$

The c.d.f.s $F_{j|i}$, $G_{k|i}$ can be next estimated by the corresponding empirical distribution functions (e.d.f.s):

$$\widehat{F}_{j|i} = \frac{n^{xy}_{A,i1} + \cdots + n^{xy}_{A,ij}}{n^{x}_{A,i}}, \quad i = 1, \ldots, I, \ j = 1, \ldots, J;$$

$$\widehat{G}_{k|i} = \frac{n^{xz}_{B,i1} + \cdots + n^{xz}_{B,ik}}{n^{x}_{AB,i}}, \quad i = 1, \ldots, I, \ k = 1, \ldots, K.$$

As a consequence, the upper end lower bound for $H_{jk|i}$, $H^{+}_{jk|i}$, $H^{-}_{jk|i}$, can be simply estimated by replacing the actual d.f.s $F_{j|i}$s, $G_{k|i}$ with the corresponding e.d.f.s $\widehat{F}_{j|i}$s, $\widehat{G}_{k|i}$s. Hence, the conditional and unconditional measures of uncertainty can be estimated by

$$\widehat{\Delta}^{x=i} = \sum_{j=1}^{J}\sum_{k=1}^{K}\left(\widehat{H}^{+}_{jk|i} - \widehat{H}^{-}_{jk|i}\right)\,\widehat{\phi}_{j|i}\,\widehat{\psi}_{k|i}, \tag{5}$$

$$\widehat{\Delta} = \sum_{i=1}^{I}\widehat{\Delta}^{x=i}\,\widehat{\xi}_{i} \tag{6}$$

respectively.

The properties of estimators (5), (6) are listed below.

Consistency

$$\widehat{\Delta}^{x=i} \overset{a.s.}{\to} \Delta^{x=i} \ \ as\ n_A \to \infty,\ n_B \to \infty, \ \ i = 1, \ldots, I;$$
$$\widehat{\Delta} \overset{a.s.}{\to} \Delta \ \ as\ n_A \to \infty,\ n_B \to \infty.$$

Asymptotic normality The rescaled measures of uncertainty

$$\sqrt{\frac{n^x_{A,i} n^x_{B,i}}{n^x_{A,i} + n^x_{B,i}}}(\widehat{\Delta}^{x=i} - \Delta^{x=i}), \quad \sqrt{\frac{n_A n_B}{n_A + n_B}}(\widehat{\Delta} - \Delta)$$

possess normal asymptotic distribution σ^2_i, σ^2, respectively, as n_A, n_B tend to infinity.

The asymptotic variances σ_i^2s, σ^2 do have a complicate form, depending on the "true" $F_{j|i}$s, $G_{k|i}$s . However, they can be consistently estimated by bootstrap method, that works as follows.

1. Generate from the e.d.f. of sample A a bootstrap sample of size n_A.

2. Generate from the e.d.f. of sample B a bootstrap sample sample of size n_B.

3. Use samples generated in steps 1, 2 to compute the "bootstrap version" $\widetilde{\Delta}^{x=i}$ of $\widehat{\Delta}^{x=i}$.

Steps 1-3 are repeated M times, so that the M bootstrap values

$\widetilde{\Delta}_m^{x=i}$, $m = 1, \ldots, M$ are obtained. Let $\overline{\Delta}^{x=i}$ be their average, and let S_M^{2x} be their variance:

$$\overline{\Delta}^{x=i} = \frac{1}{M}\sum_{m=1}^{M} \widetilde{\Delta}_m^{x=i}, \quad S_M^{2x} = \frac{1}{M-1}\sum_{m=1}^{M} (\widetilde{\Delta}_m^{x=i} - \overline{\Delta}^{x=i})^2.$$

As an estimate of σ_i^2, we may take

$$\widehat{\sigma}_{i,M}^2 = \frac{n_{A,x}\, n_{B,x}}{n_{A,x} + n_{B,x}} S_M^{2x}. \tag{7}$$

From (7) it is also easy to construct an estimate of the unconditional variance σ^2.

The above results are useful to construct point and interval estimates of the uncertainty measures $\triangle^{x=i}$, $\triangle$. They are also useful to test the hypothesis that the class of bivariate d.f.s with upper bounds $H^{+}_{jk|i}$s and lower bounds $H^{-}_{jk|i}$ is "narrow", when structural zeroes are considered.