Statistical matching analysis for complex survey data with applications

Pier Luigi Conti*, Daniela Marella** and Mauro Scanu***[1]

*Dipartimento di Scienze Statistiche, Sapienza Università di Roma

**Dipartimento di Scienze della Formazione, Università "Roma Tre"

***ISTAT, Italian National Statistical Institute, Roma, Italy

Online Supplementary Material

Proof of Proposition 1. To prove statement $C1$, observe first that $E_{P_h}\left[\widehat{N}_h(x)\right] = Np(x)$, and

$$V_{P_h}\left(\widehat{N}_h(x)\right) = \sum_{i=1}^{N}\left(\frac{1}{\pi_{i,h}} - 1\right) I_{(x_i=x)} + \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\pi_{ij,h}}{\pi_{i,h}\pi_{j,h}} - 1\right) I_{(x_i=x)} I_{(x_j=x)}, \quad h = 1, 2.$$

Assumption A5 implies that the sampling design P_h is of asymptotically maximal entropy, so that the inequalities

$$\left|\frac{\pi_{ij,h}}{\pi_{i,h}\pi_{j,h}} - 1\right| \leqslant \frac{C}{N} \quad i \neq j \tag{1}$$

hold, C being an absolute constant (cfr. Hájek (1981), p. 74). From (1) and Chebyshev inequality, statement $C1$ follows.

The proof of statement $C2$ is based on the same arguments as in Conti (2014) (Lemmas 1-4 and Proposition 1). First of all, from a first order Taylor expansion and statement $C1$, it is seen that the asymptotic law of $W_{h,N}^{x}(y)$ coincides with the asymptotic law of

$$\frac{\sqrt{N}}{p(x)}\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{D_{i,h}}{\pi_{i,h}} I_{(x_i=x)}\left(I_{(y_{hi}\leqslant y)} - F_{hN}(y\,|x)\right) - \frac{1}{N}\sum_{i=1}^{N} I_{(x_i=x)}\left(I_{(y_{hi}\leqslant y)} - F_{hN}(y\,|x)\right)\right\}$$

[1]email: pierluigi.conti@uniroma1.it, daniela.marella@uniroma3.it, scanu@istat.it
Corresponding author: Daniela Marella - Dipartimento di Scienze della Formazione, Università "Roma Tre" - daniela.marella@uniroma3.it

$$= \frac{\sqrt{N}}{p(x)} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(\frac{D_{i,h}}{\pi_{i,h}} - 1 \right) I_{(x_i=x)} \left(I_{(y_{hi} \leqslant y)} - F_{hN}(y\,|x) \right) \right\}.$$

Define now

$$\begin{aligned} Z_{i,N} &= I_{(x_i=x)} \left(I_{(y_{hi} \leqslant y)} - F_{hN}(y\,|x) \right) - \pi_{i,h} \frac{\sum_{i=1}^{N} (1-\pi_{i,h}) I_{(x_i=x)} (I_{(y_{hi} \leqslant y)} - F_{hN}(y\,|x))}{\sum_{i=1}^{N} \pi_{i,h}(1-\pi_{i,h})} \\ S_N^2 &= \sum_{i=1}^{N} \left(\frac{1}{\pi_{i,h}} - 1 \right) Z_{i,N}^2. \end{aligned}$$

Using the same arguments as in Conti (2014) (Lemmas 1-2), and taking into account that

$$\mathbb{E}\left[I_{(x_i=x)} \left(I_{(y_{hi} \leqslant y)} - F_{hN}(y\,|x) \right) \right] = 0$$

it is immediate to see that, as N goes to infinity,

$$\frac{1}{N} \sum_{i=1}^{N} \frac{1}{\pi_{i,h}} I_{(x_i=x)} \left(I_{(y_{hi} \leqslant y)} - F_{hN}(y\,|x) \right) \overset{a.s.}{\to} 0 \tag{2}$$

$$\frac{1}{N} \sum_{i=1}^{N} (1-\pi_{i,h}) I_{(x_i=x)} \left(I_{(y_{hi} \leqslant y)} - F_{hN}(y\,|x) \right) \overset{a.s.}{\to} 0 \tag{3}$$

$$Z_{i,N} - I_{(x_i=x)} \left(I_{(y_{hi} \leqslant y)} - F_{hN}(y\,|x) \right) \overset{a.s.}{\to} 0 \tag{4}$$

$$\frac{S_N^2}{N} \overset{a.s.}{\to} (\varsigma_h - 1) p(x) F_h(y\,|x)(1 - F_h(y\,|x)). \tag{5}$$

At this point the claimed result can be obtained by using the same technique as in Conti (2014) (Lemmas 3-4 and Proposition 1). Independence of $W_1^x(\cdot)$ and $W_2^x(\cdot)$ follows from independence of $W_{1,N}^x(\cdot)$ and $W_{2,N}^x(\cdot)$ for each N.

Finally, statement $C3$ is a consequence of $C2$ and the Glivenko-Cantelli theorem.

Proof of Proposition 2. From Theorem 4.2 in Gietl and Reffel (2013), it follows that if IPF converges than the solution matrix of IPF continuously depends on the starting matrix and on the marginals. Hence, using Proposition 1 and $C3$, $\widehat{p}^*(y_1^{l_1}, y_2^{l_2}|x)$ converges in probability to the solution of IPF procedure that uses $p_{ST}(y_1^{l_1}, y_2^{l_2}|x)$s as entries of the starting matrix, and with

marginals $p_h(\cdot|x)$, *i.e.* to $p^*(y_1^{l_1}, y_2^{l_2}|x)$, say. This proves the first claim of Proposition 2. The same arguments also proves that $p_N^*(y_1^{l_1}, y_2^{l_2}|x)$ converge a.s. to $p^*(y_1^{l_1}, y_2^{l_2}|x)$, from which the second claim of Proposition 2 follows.

Proof of Proposition 3. The technique of proof is identical to that of Proposition 3 in Conti *et al.* (2015), by applying the Skorokhod representation theorem to the processes $W_{h,N}^x(\cdot)$, $W_h^x(\cdot)$ in Proposition 1. As far as the asymptotic variance is concerned, define first the sets

$$T_1^x = \left\{(y_1,y_2) : K_+^x(y_1,y_2) = F_2(y_2|x)\right\},\ T_2^x = \left\{(y_1,y_2) : K_+^x(y_1,y_2) = F_2(\gamma_{y_1}(a_x)|x)\right\},$$

$$T_3^x = \left\{(y_1,y_2) : K_+^x(y_1,y_2) = F_1(y_1|x)\right\},\ T_4^x = \left\{(y_1,y_2) : K_+^x(y_1,y_2) = F_1(\delta_{y_2}(b_x)|x)\right\},$$

$$S_0^x = \left\{(y_1,y_2) : K_-^x(y_1,y_2) = 0\right\},\ S_1^x = \left\{(y_1,y_2) : K_-^x(y_1,y_2) = F_1(y_1|x) + F_2(y_2|x) - 1\right\},$$

$$S_2^x = \left\{(y_1,y_2) : K_-^x(y_1,y_2) = F_1(\delta_{y_2}(b_x)|x) + F_2(y_2|x) - 1\right\},$$

$$S_3^x = \left\{(y_1,y_2) : K_-^x(y_1,y_2) = F_1(y_1|x) + F_2(\gamma_{y_1}(a_x)|x) - 1\right\},$$

$$S_4^x = \left\{(y_1,y_2) : K_-^x(y_1,y_2) = F_1(\delta_{y_2}(b_x)|x) + F_2(\gamma_{y_1}(a_x)|x) - 1\right\}$$

and the functions

$$\begin{aligned}
\tau_1^x(y_1,y_2) &= \left\{I_{((y_1,y_2)\in T_3^x)} - I_{((y_1,y_2)\in S_1^x)} - I_{((y_1,y_2)\in S_3^x)}\right\} \\
\tau_2^x(y_1,y_2) &= \left\{I_{((y_1,y_2)\in T_4^x)} - I_{((y_1,y_2)\in S_2^x)} - I_{((y_1,y_2)\in S_4^x)}\right\} \\
\tau_3^x(y_1,y_2) &= \left\{I_{((y_1,y_2)\in T_1^x)} - I_{((y_1,y_2)\in S_1^x)} - I_{((y_1,y_2)\in S_2^x)}\right\} \\
\tau_4^x(y_1,y_2) &= \left\{I_{((y_1,y_2)\in T_2^x)} - I_{((y_1,y_2)\in S_3^x)} - I_{((y_1,y_2)\in S_4^x)}\right\} \\
\beta^x(J;\, a,\, b) &= \min(J(a\,|x),\, J(b\,|x)) - J(a\,|x)\, J(b\,|x);\quad J = F_1,\, F_2 \\
R^x(y_1,y_2) &= K_-^x(y_1,y_2) - K_+^x(y_1,y_2)
\end{aligned}$$

Then, $V(F_1, F_2;\, x)$ possesses the form stated in Proposition 3, with

$$V_1(F_1,F_2;\, x) = \int_{\mathbb{R}^4} F_2(y_2|x)\, F_2(z_2|x)\, \beta^x(F_1;\, y_1, z_1)\, dR^x(y_1,y_2)\, dR^x(z_1,z_2)$$

$$
\begin{aligned}
&+ \int_{\mathbb{R}4} \tau_1^x(y_1,y_2)\,\tau_1^x(z_1,\,z_2)\beta^x(F_1;\,y_1,z_1)\,d[F_1(y_1|x)F_2(y_2|x)]\,d[F_1(z_1|x)F_2(z_2\,|x)] \\
&+ \int_{\mathbb{R}^4} \tau_2^x(y_1,y_2)\,\tau_2^x(z_1,z_2)\beta^x(F_1;\,\delta_{y_2}(b_x),\,\delta_{z_2}(b_x))\,d[F_1(y_1|x)F_2(y_2|x)]\,d[F_1(z_1|x)F_2(z_2|x)] \\
&+2\int_{\mathbb{R}^4} F_2(y_2|x)\,\tau_1^x(z_1,z_2)\beta^x(F_1;\,y_1,z_1)\,d[R^x(y_1,y_2)\,d[F_1(z_1|x)F_2(z_2|x)] \\
&+2\int_{\mathbb{R}^4} F_2(y_2|x)\,\tau_2^x(z_1,z_2)\beta^x(F_1;\,y_1,\delta_{z_2}(b_x))\,dR^x(y_1,y_2)\,d[F_1(z_1|x)F_2(z_2|x)] \\
+\;\; &2\int_{\mathbb{R}^4} \tau_1^x(y_1,y_2)\,\tau_2^x(z_1,z_2)\beta^x(F_1;\,y_1,\delta_{z_2}(b_x))\,d[F_1(y_1|x)F_2(y_2|x)]\,d[F_1(z_1|x)F_2(z_2\,|x)] \qquad (6)
\end{aligned}
$$

and

$$
\begin{aligned}
&V_2(F_1,F_2;\,x) = \int_{\mathbb{R}^4} F_1(y_1|x)\,F_1(z_1|x)\,\beta^x(F_2;\,y_2,z_2)\,dR^x(y_1,y_2)\,dR^x(z_1,z_2) \\
&+ \int_{\mathbb{R}^4} \tau_3^x(y_1,y_2)\,\tau_3^x(z_1,z_2)\beta^x(F_2;\,y_2,\,z_2)\,d[F_1(y_1|x)F_2(y_2|x)]\,d[F_1(z_1|x)F_2(z_2|x)] \\
&+ \int_{\mathbb{R}^4} \tau_4^x(y_1,y_2)\,\tau_4^x(z_1,z_2)\beta^x(F_2;\,\gamma_{y_1}(a_x),\,\gamma_{z_1}(a_x))\,d[F_1(y_1|x)F_2(y_2|x)]\,d[F_1(z_1|x)F_2(z_2|x)] \\
&+2\int_{\mathbb{R}^4} F_1(y_1|x)\,\tau_3^x(z_1,z_2)\beta^x(F_2;\,y_2,z_2)\,d[R^x(y_1,y_2)\,d[F_1(z_1|x)F_2(z_2|x)] \\
&+2\int_{\mathbb{R}^4} F_1(y_1|x)\,\tau_4^x(z_1,z_2)\beta^x(F_2;\,y_2,\gamma_{z_1}(a_x))\,dR^x(y_1,y_2)\,d[F_1(z_1|x)F_2(z_2\,|x)] \\
+\;\; &2\int_{\mathbb{R}^4} \tau_3(y_1,y_2)\,\tau_4^x(z_1,z_2)\beta^x(F_2;\,y_2,\,\gamma_{z_1}(a_x))\,d[F_1(y_1|x)F_2(y_2|x)]\,d[F_1(z_1|x)F_2(z_2|x)]. \qquad (7)
\end{aligned}
$$

Proof of Proposition 4. A first order Taylor expansion shows that the asymptotic distribution of $\sqrt{N}(\widehat{p}_h(x) - p_N(x))$ coincides with the asymptotic distribution of

$$
\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{D_{i,h}}{\pi_{i,h}} - 1\right)(I_{(x_i=1)} - p_N(x)), \quad h = 1,\,2. \qquad (8)
$$

Define next:

$$
Z_{i,N} \;\; = \;\; (I_{(x_i=1)} - p_N(x)) - \pi_{i,h}\frac{\sum_{i=1}^{N}(1-\pi_{i,h})(I_{(x_i=1)} - p_N(x))}{\sum_{i=1}^{N}\pi_{i,h}(1-\pi_{i,h})},
$$

$$S_N^2 \;=\; \sum_{i=1}^{N}\left(\frac{1}{\pi_{i,h}}-1\right)Z_{i,N}^2.$$

It is immediate to see that, as N goes to infinity,

$$Z_{i,N}-(I_{(x_i=1)}-p_N(x)) \overset{a.s.}{\to} 0, \quad \frac{1}{N}S_N^2 \overset{a.s.}{\to} p(x)(1-p(x))(\varsigma_h-1).$$

Hence, the same arguments as in the proof of Proposition 1 lead to:

$$\sqrt{N}(\widehat{p}_h(x)-p_N(x)) \overset{d}{\to} N(0,\tau^2(\varsigma_h-1)p(x)(1-p(x))), \quad h=1,\,2. \tag{9}$$

From the independence of $\widehat{p}_1(x)$ and $\widehat{p}_2(x)$, Proposition 4 follows.

Finally, the asymptotically optimal value of τ is the value of τ minimizing $\tau^2(\varsigma_1-1)+(1-\tau)^2(\varsigma_2-1)$, that turns out to be equal to $(\varsigma_1-1)/(\varsigma_1+\varsigma_2-2)$.

Proof of Proposition 6. First of all, the equality

$$\left(\widehat{n}_1^{-1}+\widehat{n}_2^{-1}\right)^{-1/2}\left(\widehat{\Delta}_H-\Delta(F_{1N},\,F_{2N})\right)=\widetilde{L}_{1N}+\widetilde{L}_{2N} \tag{10}$$

holds, where

$$\widetilde{L}_{1N} \;=\; \left(\widehat{n}_1^{-1}+\widehat{n}_2^{-1}\right)^{-1/2}\sum_{k=1}^{K}\widehat{p}_{12}(x^k)\left(\widehat{\Delta}^{x^k}-\Delta^{x^k}(F_{1N},\,F_{2N})\right) \tag{11}$$

$$\widetilde{L}_{2N} \;=\; \left(\widehat{n}_1^{-1}+\widehat{n}_2^{-1}\right)^{-1/2}\sum_{k=1}^{K}\Delta^{x^k}(F_{1N},\,F_{2N})\left(\widehat{p}_{12}(x^k)-p_N(x^k)\right). \tag{12}$$

From Propositions 3 - 5 it is argued that (11), (12) are asymptotically independent, as N increases. Hence, we just have to study separately the asymptotic behavior of (11) and (12). As far as (11) is concerned, using the symbol $\sim$ for asymptotic equivalence, in view of Propositions

1 - 5 we have:

$$\begin{aligned}
\widetilde{L}_{1N} &\sim \left(\widehat{n}_1^{-1} + \widehat{n}_2^{-1}\right)^{-1/2} \sum_{k=1}^{K} p(x^k) \left(\widehat{\Delta}^{x^k} - \Delta^{x^k}(F_{1N},\, F_{2N})\right) \\
&\sim \sum_{k=1}^{K} \sqrt{\frac{\widehat{n}_1(x^k)\widehat{n}_2(x^k)}{\widehat{n}_1(x^k) + \widehat{n}_2(x^k)}} \sqrt{\frac{\widehat{n}_1(x^k) + \widehat{n}_2(x^k)}{\widehat{n}_1 + \widehat{n}_2}} \left(\widehat{\Delta}^{x^k} - \Delta^{x^k}(F_{1N},\, F_{2N})\right) \\
&\sim \sum_{k=1}^{K} \sqrt{p(x^k)} \left(\widehat{n}_1(x^k)^{-1} + \widehat{n}_2(x^k)^{-1}\right)^{-1/2} \left(\widehat{\Delta}^{x^k} - \Delta^{x^k}(F_{1N},\, F_{2N})\right) \\
&\stackrel{d}{\to} N\left(0,\, \sum_{k=1}^{K} p(x^k) V(F_1, F_2;\, x^k)\right). \qquad (13)
\end{aligned}$$

The quantity $\widetilde{L}_{2N}$ can be dealt with similarly, from which the proposition follows.

References

Conti, P. L. (2014). On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. *Sankhya*, **76**, 234–259.

Conti, P. L., Marella, D., and Scanu, M. (2015). How far from identifiability? A systematic overview of the statistical matching problem in a non-parametric framework. *Communications in Statistics - Theory and Methods.* DOI : 10.1080/03610926.2015.1010005.

Gietl, C. and Reffel, F. P. (2013). Continuity of f-Projections and Applications to the Iterative Proportional Fitting Procedure. *Preprint Nr. 13/2013 Institut für Mathematik, Universitäu Augsburg.*

Hájek, J. (1981). *Sampling from a finite population.* Marcel Dekker, New York.