Supplement to "Identification, data combination, and the risk of disclosure"

(*Quantitative Economics*, Vol. 9, No. 1, March 2018, 395–440)

Tatiana Komarova
Department of Economics, London School of Economics and Political Science

Denis Nekipelov
Department of Economics, University of Virginia

Evgeny Yakovlev
New Economic School

Appendix A: Construction of individual identifiers

This section overviews existing techniques for the construction of individual identifiers, especially in the case of numerical variables.

The key element of our identification argument is based on the construction of the identifying variables Z^y and Z^x such that we can merge some or all observations in the disjoint data bases to enable estimation of the econometric model of interest. While we took the existence of these variables as given, their construction in itself is an important issue and there is a vast literature in applied statistics and computer science that is devoted to the analysis of the broken record linkage. For completeness of the analysis in our paper we present some highlights from that literature.

In general the task of merging disjoint data bases is a routine necessity in may practical applications. In many cases there do exist perfect cross-data-base identifiers of individual entries. There could be multiple reasons why that is the case. For instance, there could be errors in data entry and processing, wrong variable formatting, and duplicate data entry. The idea that has arisen in Newcombe, Kennedy, Axford, and James (1959) and was later formalized in Fellegi and Sunter (1969) was to treat the record linkage problem as a problem of classification of record subsets into matches, nonmatches, and uncertain cases. This classification is based on defining the similarity metric between each two records. Then given the similarity metric one can compute the probability of a particular pair of records being a match or non-match. The classification of pairs is then performed by fixing the probability of erroneous identification of a nonmatched

Tatiana Komarova: t.komarova@lse.ac.uk
Denis Nekipelov: denis@virginia.edu
Evgeny Yakovlev: eyakovlev@nes.ru

https://doi.org/10.3982/QE568

pair of records as a match and a matched pair of records as a nonmatch by minimizing the total proportion of pairs that are uncertain. This matching technique is based on the underlying assumption of randomness of records being broken. As a result, using the sample of perfectly matched records one can recover the distribution of the similarity metric for the matched and unmatched pairs of records. Moreover, as in hypothesis testing, one needs to fix the probability of record misidentification. Finally, the origin of the similarity metric remains arbitrary.

A large fraction of the further literature was devoted to, on one hand, development of classes of similarity metrics that accommodate nonnumeric data and, on the other hand, development of fast and scalable record classification algorithms. For obvious reasons, measuring the similarity of string data turns out to be the most challenging. Edit distance (see Gusfield (1997) for instance) is a metric that can be used to measure the string similarity. The distance between the two strings is determined as the minimum number of insert, delete, and replace operations required to transform one string into another. Another measure developed in Jaro (1989) and elaborated in Winkler (1999) is based on the length of matched strings, and the number of common characters and their position within the string. In its modification it also allows for the prefixes in the names and is mainly intended to link relatively short strings such as individual names. Alternative metrics are based on splitting strings into individual "tokens" that are substrings of a particular length and then analyzing the power of sets of overlapping and nonoverlapping tokens. For instance, the Jaccard coefficient is based on the relative number of overlapping and overall tokens in two strings. More advanced metrics include the TF/IDF metric that is based on the term frequency (TF) or the number of times the term (or token) appears in the document (or string) and the inverse document frequency (IDF) or the number of documents containing the given term. The structure of the TF/IDF-based metric construction is outlined in Salton and Harman (2003). The distance measures may include a combination of the edit distance and the TF/IDF distance such as a fuzzy match similarity metric as described in Chaudhuri, Ganjam, Ganti, and Motwani (2003).

Given a specific definition of the distance, the practical aspects of matching observations will entail calibration and application of a particular technique for matching observations. The structure of those techniques is based on, first, the assumption regarding the data structure and the nature of the record errors. Second, it depends on the availability of known matches, and, thus, allows empirical validation of a particular matching technique. When such a validation sample is available, one can estimate the distribution of the similarity measures for matched and nonmatched pairs for the validation sample. Then using the estimated distribution, one can assign the matches for the pairs outside the validation sample. When one can use numeric information in addition to the string information, one can use hybrid metrics that combine the known properties of numeric data entries and the properties of string entries.

Ridder and Moffitt (2007) overview some techniques for purely numeric data combination in the absence of validation subsamples that may incorporate distributional assumptions on the "similar" numeric variables. For instance, joint normality assumption with a known sign of correlation can allow one to invoke likelihood-based techniques for record linkage.

Appendix B: More details on the risk of disclosure and choice of threshold sequences

Propositions 2 and 3 demonstrate that the compliance of the decision rule generated by a particular threshold sequence with a given bound guarantee for the disclosure risk depends on the rate at which the threshold sequence converges toward zero as the sizes of $\mathcal{D}^y$ and $\mathcal{D}^x$ increase. Informally, consider two threshold sequences α_N and α_N^*, where the former converges to zero much faster than the latter so that $\frac{\alpha_N^*}{\alpha_N} \to \infty$. Clearly, for large enough sizes of the data sets $\mathcal{D}^y$ and $\mathcal{D}^x$, the sequence α_N^* not only allows more observations to be included in the combined data set, but also gives a greater number of possible combined data sets. In fact, all observations with the values of the constructed identifiers z_i^x between $\frac{1}{\alpha_N^*}$ and $\frac{1}{\alpha_N}$ are rejected by the decision rule implied by the sequence α_N but could be approved by the decision rule implied by the sequence α_N^*. In addition, the sequence α_N^* is much more liberal in its definition of the proximity between the identifiers z_j^y and z_i^x. As a result, the decision rule implied by the sequence α_N^* generates larger combined data sets. Because the matching information in $(-\frac{1}{\alpha_N}, -\frac{1}{\alpha_N^*}) \cup (\frac{1}{\alpha_N^*}, \frac{1}{\alpha_N})$ is less reliable than that in $(-\infty, -\frac{1}{\alpha_N}) \cup (\frac{1}{\alpha_N}, \infty)$ and because linkages for observations with larger distances between the identifiers are decreasingly reliable, the sequence α_N^* results in a larger proportion of incorrect matches. The effect can be so significant that even for arbitrarily large data sets the probability of making a data combination error does not approach 0. In Proposition 2, where nondisclosure is not guaranteed and the probability of making a data combination error of the first kind approaches 0 as N^y and N^x increase, thresholds used for the decision rule shrink to zero faster than those in Proposition 3, where nondisclosure is guaranteed.

In the remarks below we consider cases when the tails of the distributions of identifiers are geometric or exponential.

Remark (Absence of disclosure guarantees). Here we consider cases when the tails of the distributions of identifiers are geometric or exponential.

(a) Suppose that for small enough $\alpha > 0$, we have $\phi(\alpha) = b_1\alpha^{c_1}$, $b_1, c_1 > 0$ and $\psi(\alpha) = b_2\alpha^{c_2}$, $b_2, c_2 > 0$. If $\alpha_N > 0$ is chosen in such a way that

$$\alpha_N = o\left(\frac{1}{(N^x)^{\frac{1}{c_2+2}}}\right) \tag{S1}$$

as $N^y \to \infty$, then

$$\inf_{x\in\mathcal{X}, y\in\mathcal{Y}} \inf_{\mathcal{D}^x, \mathcal{D}^y} \inf_{i,j} p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) \to 1 \quad \text{as } N^y \to \infty$$

and, thus, nondisclosure is not guaranteed.

(b) Alternatively, suppose that for small enough $\alpha > 0$, we have $\phi(\alpha) = b_1 e^{-c_1/\alpha}$, $b_1, c_1 > 0$ and $\psi(\alpha) = b_2 e^{-c_2/\alpha}$, $b_2, c_2 > 0$. If $\alpha_N \to 0$ is chosen in such a way that

$$\lim_{N^y\to\infty} N^x e^{-\frac{c_2}{\alpha_N}} \alpha_N = 0, \tag{S2}$$

then

$$\inf_{x\in\mathcal{X},y\in\mathcal{Y}}\inf_{\mathcal{D}^x,\mathcal{D}^y}\inf_{i,j}p_{ij}^N\left(x,y,\mathcal{D}^x,\mathcal{D}^y\right)\to 1 \quad \text{as } N^y\to\infty$$

and, thus, nondisclosure is not guaranteed.

For instance, sequences $\alpha_N=\frac{a}{(N^x)^d}$ when $a,d>0$ satisfy this condition.

Proof. (a) Let us check that if a sequence α_N is chosen as in (S1), then it satisfies (10). In other words, let us check that

$$\frac{N^x}{\alpha_N^{c_1}}\int_{\frac{1}{\alpha_N}}^{\infty}\left(\left(\frac{1}{z-\alpha_N}\right)^{c_2}-\left(\frac{1}{z+\alpha_N}\right)^{c_2}\right)\frac{1}{z^{c_1+1}}\,dz\to 0 \quad \text{as } N^y\to\infty.$$

Indeed,

$$\begin{aligned}&\frac{N^x}{\alpha_N^{c_1}}\int_{\frac{1}{\alpha_N}}^{\infty}\left(\left(\frac{1}{z-\alpha_N}\right)^{c_2}-\left(\frac{1}{z+\alpha_N}\right)^{c_2}\right)\frac{1}{z^{c_1+1}}\,dz\\&=\frac{N^x}{\alpha_N^{c_1}}\int_{\frac{1}{\alpha_N}}^{\infty}\left(1-\left(1-\frac{2\alpha_N}{z+\alpha_N}\right)^{c_2}\right)\frac{(z-\alpha_N)^{-c_2}}{z^{c_1+1}}\,dz.\end{aligned}$$

If α_N is small enough, then for all $z\geq\frac{1}{\alpha_N}$, it holds that $1-(1-\frac{2\alpha_N}{z+\alpha_N})^{c_2}\leq q_1\frac{\alpha_N}{z+\alpha_N}$ for some constant $q_1>0$. Therefore, if α_N is small enough, then for all $z\geq\frac{1}{\alpha_N}$ we have

$$\left(1-\left(1-\frac{2\alpha_N}{z+\alpha_N}\right)^{c_2}\right)\frac{(z-\alpha_N)^{-c_2}}{z^{c_1+1}}\leq q_2\frac{\alpha_N}{z^{c_1+c_2+2}}$$

for some constant $q_2>0$. Finally, note that

$$\frac{q_2N^x}{\alpha_N^{c_1-1}}\int_{\frac{1}{\alpha_N}}^{\infty}\frac{1}{z^{c_1+c_2+2}}\,dz=\frac{q_2N^x}{1+c_1+c_2}\alpha_N^{c_2+2}\to 0 \quad \text{as } N^y\to\infty$$

if α_N is chosen as in (S1).

(b) Let us check that if a sequence α_N is chosen as in (S2), then it satisfies (10). In other words, let us check that

$$N^xe^{\frac{c_1}{\alpha_N}}\int_{\frac{1}{\alpha_N}}^{\infty}\left(e^{-c_2(z-\alpha_N)}-e^{-c_2(z+\alpha_N)}\right)e^{-c_1z}\,dz\to 0 \quad \text{as } N^y\to\infty.$$

Indeed,

$$N^xe^{\frac{c_1}{\alpha_N}}\int_{\frac{1}{\alpha_N}}^{\infty}\left(e^{-c_2(z-\alpha_N)}-e^{-c_2(z+\alpha_N)}\right)e^{-c_1z}\,dz=N^xe^{-\frac{c_2}{\alpha_N}}\frac{e^{c_2\alpha_N}-e^{-c_2\alpha_N}}{c_1+c_2}.$$

Note that for some constant $r>0$,

$$e^{c_2\alpha_N}-e^{-c_2\alpha_N}\leq r\alpha_N.$$

Now it is clear that if α_N is chosen as in (S2), then (10) holds. □

Remark (Disclosure guarantees).

(a) Suppose for $\alpha \in (0, \bar{\alpha})$, $\phi(\alpha) = b_1 \alpha^{c_1}$, $b_1, c_1 > 0$, and $\psi(\alpha) = b_2 \alpha^{c_2}$, $b_2, c_2 > 0$. Let the sequence of $\alpha_N \to 0$ (as $N^y \to \infty$) be chosen in such a way that

$$\liminf_{N^y \to \infty} \alpha_N \left(N^x\right)^{\frac{1}{c_2+2}} > 0. \tag{S3}$$

Then nondisclosure is guaranteed.

(b) Suppose for $\alpha \in (0, \bar{\alpha})$, $\phi(\alpha) = b_1 e^{-c_1/\alpha}$, $b_1, c_1 > 0$ and $\psi(\alpha) = b_2 e^{-c_2/\alpha}$, $b_2, c_2 > 0$. Let the sequence of $\alpha_N \to 0$ (as $N^y \to \infty$) be chosen in such a way that

$$\liminf_{N^y \to \infty} N^x e^{-\frac{c_2}{\alpha_N}} \alpha_N > 0. \tag{S4}$$

Then nondisclosure is guaranteed.

For instance, sequences $\alpha_N = \frac{a}{\log N^x}$, when $a > c_2$, satisfy this condition (in this case, $\lim_{N^y \to \infty} N^x e^{-\frac{c_2}{\alpha_N}} \alpha_N = \infty$).

Proof. (a) Let us check that if a sequence α_N is chosen as in (S3), then it satisfies (11). In other words, let us check that

$$\liminf_{N^y \to \infty} b_2 c_1 \frac{N^x}{\alpha_N^{c_1}} \int_{\frac{1}{\alpha_N}}^{\infty} \left(\left(\frac{1}{z - \alpha_N} \right)^{c_2} - \left(\frac{1}{z + \alpha_N} \right)^{c_2} \right) \frac{1}{z^{c_1+1}} \, dz > 0.$$

Use $((\frac{1}{z-\alpha_N})^{c_2} - (\frac{1}{z+\alpha_N})^{c_2}) \frac{1}{z^{c_1+1}} = (1 - (1 - \frac{2\alpha_N}{z+\alpha_N})^{c_2}) \frac{(z-\alpha_N)^{-c_2}}{z^{c_1+1}}$ and note that if α_N is small enough, then for all $z \geq \frac{1}{\alpha_N}$,

$$1 - \left(1 - \frac{2\alpha_N}{z + \alpha_N}\right)^{c_2} \geq \tilde{q}_1 \frac{\alpha_N}{z + \alpha_N}$$

for some constant $\tilde{q}_1 > 0$. Therefore, if α_N is small enough, then for all $z \geq \frac{1}{\alpha_N}$ we have

$$\left(1 - \left(1 - \frac{2\alpha_N}{z + \alpha_N}\right)^{c_2}\right) \left(\frac{1}{z - \alpha_N}\right)^{c_2} \frac{1}{z^{c_1+1}} \geq \tilde{q}_2 \frac{\alpha_N}{z^{c_1+c_2+2}}$$

for some constant $\tilde{q}_2 > 0$. Finally, note that if α_N is chosen as in (S3), then

$$\liminf_{N^y \to \infty} \tilde{q}_2 b_2 c_1 \frac{N^x}{\alpha_N^{c_1-1}} \int_{\frac{1}{\alpha_N}}^{\infty} \frac{1}{z^{c_1+c_2+2}} \, dz = \liminf_{N^y \to \infty} \tilde{q}_2 b_2 c_1 \frac{N^x}{1 + c_1 + c_2} \alpha_N^{c_2+2} > 0.$$

(b) Let us check that if a sequence α_N is chosen as in (S4), then it satisfies (11). In other words, we want to check that

$$\liminf_{N^y \to \infty} c_1 N^x e^{\frac{c_1}{\alpha_N}} \int_{\frac{1}{\alpha_N}}^{\infty} \left(e^{-c_2(z-\alpha_N)} - e^{-c_2(z+\alpha_N)}\right) e^{-c_1 z} \, dz > 0.$$

Note that

$$N^x e^{\frac{c_1}{\alpha_N}} \int_{\frac{1}{\alpha_N}}^{\infty} \left(e^{-c_2(z-\alpha_N)} - e^{-c_2(z+\alpha_N)}\right) e^{-c_1 z}\, dz = N^x e^{-\frac{c_2}{\alpha_N}} \frac{e^{c_2\alpha_N} - e^{-c_2\alpha_N}}{c_1 + c_2}$$

and for some constant $\tilde{r} > 0$,

$$e^{c_2\alpha_N} - e^{-c_2\alpha_N} \geq \tilde{r}\alpha_N.$$

Thus, if α_N is chosen as in (S4), then (11) holds. □

It can be seen in the preceding two remarks that the rates of the threshold sequences used for the decision rule can be described in terms of the size of the data set $\mathcal{D}^x$ alone rather than both $\mathcal{D}^y$ and $\mathcal{D}^x$. This is quite intuitive because in the data base in Assumption 2 we assumed that $\mathcal{D}^y$ contains the subset of individuals from the data base $\mathcal{D}^x$ and, hence, $\mathcal{D}^x$ is larger. The size of the larger data set is the only factor determining how many potential matches from this data set we are able to find for any observation in the smaller data set without using any additional information from the identifiers.

Appendix C: Proof of Proposition 5

Fix $\widetilde{\theta} \in \Theta_\infty$. Let $\pi \in \Pi^\infty$ be such that $\widetilde{\theta}$ minimizes

$$Q(\theta, \pi) \equiv g_\pi(\theta)' W_0 g_\pi(\theta).$$

We can find a sequence $\{\pi^N(\cdot,\cdot)\}$ that converges to π uniformly over all y and all x. Let θ_N be any value that minimizes

$$Q_N\left(\theta, \pi^N\right) \equiv g^N(\theta)' W_0 g^N(\theta)$$

for the chosen $\pi^N(\cdot,\cdot)$. Clearly, $\theta_N \in \Theta_N$. Let us show that $\theta_N \to \widetilde{\theta}$.

First, we establish that $\sup_{\theta\in\Theta} |Q_N(\theta, \pi^N) - Q(\theta, \pi)| \to 0$. Note that

$$\begin{aligned} Q_N\left(\theta, \pi^N\right) - Q(\theta, \pi) = {} & \left(g^N(\theta) - g_\pi(\theta)\right)' W_0 \left(g^N(\theta) - g_\pi(\theta)\right) \\ & + 2 g_\pi(\theta)' W_0 \left(g^N(\theta) - g_\pi(\theta)\right). \end{aligned}$$

Therefore,

$$\begin{aligned} \sup_{\theta\in\Theta} \left| Q_N\left(\theta, \pi^N\right) - Q(\theta, \pi) \right| \leq {} & \sup_{\theta\in\Theta} \left\| g^N(\theta) - g_\pi(\theta) \right\|^2 \|W_0\| \\ & + 2 \sup_{\theta\in\Theta} \left\| g_\pi(\theta) \right\| \sup_{\theta\in\Theta} \left\| g^N(\theta) - g_\pi(\theta) \right\| \|W_0\|. \end{aligned}$$

Conditions (15) imply that $\sup_{\theta\in\Theta} \|g_\pi(\theta)\| < \infty$. Thus, we only need to establish that $\sup_{\theta\in\Theta} \|g^N(\theta) - g_\pi(\theta)\| \to 0$. Using condition (18), we can show that $g^N(\theta)$ can be represented as the sum of four terms,

$$g^N(\theta) = A_{N1} + A_{N2} + B_{N1} + B_{N2},$$

where

$$\begin{aligned}
\frac{A_{N1}}{1-\pi} = {} & \int\int\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_j^y - z_i^x|<\alpha_N} h(x_i)\rho(y_j, x_i; \theta) \\
& \times f_{Y,X|Z^y,Z^x}\big(y_j, x_i|z_j^y, z_i^x\big) f_{Z^y,Z^x}\big(z_j^y, z_i^x\big)\, dz_j^y\, dz_i^x\, dy_j\, dx_i \\
& \times \bigg(\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_j^y - z_i^x|<\alpha_N} \\
& \times f_{Y,X|Z^y,Z^x}\big(y_j, x_i|z_j^y, z_i^x\big) f_{Z^y,Z^x}\big(z_j^y, z_i^x\big)\, dz_j^y\, dz_i^x\, dy_j\, dx_i\bigg)^{-1}, \\
A_{N2} = {} & \int\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_j^y - z_i^x|<\alpha_N} o_{yx}(1) h(x_i)\rho(y_j, x_i; \theta) \\
& \times f_{Y,X|Z^y,Z^x}\big(y_j, x_i|z_j^y, z_i^x\big) f_{Z^y,Z^x}\big(z_j^y, z_i^x\big)\, dz_j^y\, dz_i^x\, dy_j\, dx_i \\
& \times \bigg(\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_j^y - z_i^x|<\alpha_N} \\
& \times f_{Y,X|Z^y,Z^x}\big(y_j, x_i|z_j^y, z_i^x\big) f_{Z^y,Z^x}\big(z_j^y, z_i^x\big)\, dz_j^y\, dz_i^x\, dy_j\, dx_i\bigg)^{-1}, \\
\frac{B_{N1}}{\pi} = {} & \int\int\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_i^x - z_j^y|<\alpha_N} h(x_i)\rho(y_j, x_i; \theta) f_{Y,Z^y}\big(y_j, z_j^y\big) \\
& \times f_{X,Z^x}\big(x_i, z_i^x\big)\, dz_j^y\, dz_i^x\, dy_j\, dx_i \\
& \times \bigg(\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_i^x - z_j^y|<\alpha_N} f_{Y,Z^y}\big(y_j, z_j^y\big) f_{X,Z^x}\big(x_i, z_i^x\big)\, dz_j^y\, dz_i^x\, dy_j\, dx_i\bigg)^{-1}, \\
B_{N2} = {} & \int\int\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_i^x - z_j^y|<\alpha_N} o_{yx}(1) h(x_i)\rho(y_j, x_i; \theta) f_{Y,Z^y}\big(y_j, z_j^y\big) \\
& \times f_{X,Z^x}\big(x_i, z_i^x\big)\, dz_j^y\, dz_i^x\, dy_j\, dx_i \\
& \times \bigg(\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_i^x - z_j^y|<\alpha_N} f_{Y,Z^y}\big(y_j, z_j^y\big) f_{X,Z^x}\big(x_i, z_i^x\big)\, dz_j^y\, dz_i^x\, dy_j\, dx_i\bigg)^{-1},
\end{aligned}$$

and terms $o_{yx}(1)$ do not depend on θ and are such that $\sup_{y_j \in \mathcal{Y}, x_i \in \mathcal{X}} |o_{yx}(1)| \to 0$ as $\alpha_N \to 0$.

Proposition 4 implies that $E[h(X)\rho(Y, X; \theta)||Z^x| > \frac{1}{\alpha}, |Z^x - Z^y| < \alpha] = E[h(X) \times \rho(Y, X; \theta)]$. Therefore,

$$A_{N1} = (1-\pi)E\big[h(X)\rho(Y, X; \theta)\big]$$

and, thus,

$$g^N(\theta) - g_\pi(\theta) = A_{N2} + B_{N1} + B_{N2} - \pi E^*\big[h(X)\rho(\widetilde{Y}, X; \theta)\big].$$

Note that

$$\begin{aligned}\sup_{\theta\in\Theta}\|A_{N2}\| &\le \sup_{y_j,x_i}\big|o_{yx}(1)\big|\cdot E\Big[\sup_{\theta\in\Theta}\big\|h(X)\rho(Y,X;\theta)\big\|\,\big|\,\big|Z^x\big|>\frac{1}{\alpha},\big|Z^x-Z^y\big|<\alpha\Big]\\ &= \sup_{y_j,x_i}\big|o_{yx}(1)\big|\cdot E\Big[\sup_{\theta\in\Theta}\big\|h(X)\rho(Y,X;\theta)\big\|\Big]\to 0\end{aligned}$$

as $\alpha_N\to 0$. From Assumption 3(iv), for small α_N the denominator in B_{N1}/π is the sum

$$\begin{aligned}&\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}\big(o_{xz^x}(1)+o_{z^y y}(1)+o_{z^y y}(1)o_{xz^x}(1)\big)\\ &\quad\times g_2\big(z_j^y\big)g_1\big(z_i^x\big)f_Y(y_j)f_X(x_i)\,dz_j^y\,dz_i^x\,dy_j\,dx_i\\ &\quad+\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}g_2\big(z_j^y\big)g_1\big(z_i^x\big)\,dz_j^y\,dz_i^x\end{aligned}$$

and, similarly, the numerator is the sum

$$\begin{aligned}&\int\int h(x_i)\rho(y_j,x_i;\theta)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i\cdot\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}g_2\big(z_j^y\big)g_1\big(z_i^x\big)\,dz_j^y\,dz_i^x\\ &\quad+\int\int h(x_i)\rho(y_j,x_i;\theta)\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}\big(o_{xz^x}(1)+o_{z^y y}(1)+o_{z^y y}(1)o_{xz^x}(1)\big)\\ &\quad\times g_2\big(z_j^y\big)g_1\big(z_i^x\big)f_Y(y_j)f_X(x_i)\,dz_j^y\,dz_i^x\,dy_j\,dx_i,\end{aligned}$$

where $o_{yz^y}(1)$ and $o_{xz^x}(1)$ do not depend on θ and are such that

$$\sup_{|z_j^y|>\frac{1}{\alpha_N}-\alpha_N}\ \sup_{y_j}|o_{yz^y}(1)|\to 0\quad\text{and}\quad\sup_{|z_i^x|>\frac{1}{\alpha_N}}\ \sup_{x_i}|o_{xz^x}(1)|\to 0$$

as $\alpha_N\to 0$. Then $B_{N1}-\pi E^*[h(X)\rho(\widetilde{Y},X;\theta)]$ is the sum of the two terms

$$\pi E^*\big[h(X)\rho(\widetilde{Y},X;\theta)\big]\cdot\left(\frac{C_{N1}}{C_{N1}+\int\int D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}-1\right)\tag{S5}$$

and

$$\pi\cdot\frac{\int\int h(x_i)\rho(y_j,x_i;\theta)D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}{C_{N1}+\int\int D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i},\tag{S6}$$

where

$$C_{N1}=\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}g_2\big(z_j^y\big)g_1\big(z_i^x\big)\,dz_j^y\,dz_i^x,$$

$$D_{N1}(y_j,x_i)=\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}\big(o_{xz^x}(1)+o_{z^y y}(1)+o_{z^y y}(1)o_{xz^x}(1)\big)g_2\big(z_j^y\big)g_1\big(z_i^x\big)\,dz_j^y\,dz_i^x.$$

The supremum over $\theta \in \Theta$ of the norm of the term in (S5) is bounded from above by

$$\pi E^*\Big[\sup_{\theta\in\Theta}\big\|h(X)\rho(\widetilde{Y}, X; \theta)\big\|\Big] \cdot \left|\frac{C_{N1}}{C_{N1} + \int\int D_{N1}(y_j, x_i) f_Y(y_j) f_X(x_i)\, dy_j\, dx_i} - 1\right|.$$

Because

$$\big|D_{N1}(y_j, x_i)\big| \leq \sup_{|z_i^x| > \frac{1}{\alpha_N}} \sup_{|z_j^y| > \frac{1}{\alpha_N} - \alpha_N} \sup_{y_j, x_i} \big|o_{yz^y xz^x}(1)\big| \cdot C_{N1}$$

with $\sup_{|z_i^x| > \frac{1}{\alpha_N}} \sup_{|z_j^y| > \frac{1}{\alpha_N} - \alpha_N} \sup_{y_j, x_i} |o_{yz^y xz^x}(1)| \to 0$, then

$$\frac{C_{N1}}{C_{N1} + \int\int D_{N1}(y_j, x_i) f_Y(y_j) f_X(x_i)\, dy_j\, dx_i} \to 1 \quad \text{as } \alpha_N \to 0.$$

Hence, (S5) converges to 0 uniformly over $\theta \in \Theta$.

The supremum over $\theta \in \Theta$ of the norm of the term in (S6) is bounded from above by

$$\pi \cdot \frac{\int\int \sup_{\theta\in\Theta}\big\|h(x_i)\rho(y_j, x_i; \theta)\big\|\big|D_{N1}(y_j, x_i)\big| f_Y(y_j) f_X(x_i)\, dy_j\, dx_i}{C_{N1} + \int\int D_{N1}(y_j, x_i) f_Y(y_j) f_X(x_i)\, dy_j\, dx_i}$$

$$\leq \pi \cdot \sup_{|z_i^x| > \frac{1}{\alpha_N}} \sup_{|z_j^y| > \frac{1}{\alpha_N} - \alpha_N} \sup_{y_j, x_i} \big|o_{yz^y xz^x}(1)\big| \cdot \frac{C_{N1} \cdot E^*\Big[\sup_{\theta\in\Theta}\big\|h(X)\rho(\widetilde{Y}, X; \theta)\big\|\Big]}{C_{N1} + \int\int D_{N1}(y_j, x_i) f_Y(y_j) f_X(x_i)\, dy_j\, dx_i},$$

which converges to 0 as $\alpha_N \to 0$. Thus, we obtain that $\sup_{\theta\in\Theta} \|B_{N1} - \pi E^*[h(X)\rho(\widetilde{Y}, X; \theta)]\| \to 0$.

Finally, consider $\sup_{\theta\in\Theta} \|B_{N2}\|$. This norm is bounded from above by the sum of

$$\sup_{y_j, x_i}\big|o_{yx}(1)\big| \cdot \int \sup_{\theta\in\Theta}\big\|h(x_i)\rho(y_j, x_i; \theta)\big\| f_Y(y_j) f_X(x_i)\, dy_j\, dx_i$$
$$\times \frac{C_{N1}}{C_{N1} + \int\int D_{N1}(y_j, x_i) f_Y(y_j) f_X(x_i)\, dy_j\, dx_i}$$

and

$$\sup_{y_j, x_i}\big|o_{yx}(1)\big| \cdot \sup_{|z_i^x| > \frac{1}{\alpha_N}} \sup_{|z_j^y| > \frac{1}{\alpha_N} - \alpha_N} \sup_{y_j, x_i} \big|o_{yz^y xz^x}(1)\big|$$
$$\times \frac{C_{N1}\int \sup_{\theta\in\Theta}\big\|h(x_i)\rho(y_j, x_i; \theta)\big\| f_Y(y_j) f_X(x_i)\, dy_j\, dx_i}{C_{N1} + \int\int D_{N1}(y_j, x_i) f_Y(y_j) f_X(x_i)\, dy_j\, dx_i},$$

and, hence, $\sup_{\theta\in\Theta}\|B_{N2}\| \to 0$ as $\alpha_N \to 0$.

To summarize our results so far, we shown that

$$\sup_{\theta\in\Theta}\left\|g^N(\theta) - g_\pi(\theta)\right\| \le \sup_{\theta\in\Theta}\|A_{N2}\| + \sup_{\theta\in\Theta}\left\|B_{N1} - \pi E^*\left[h(X)\rho(\widetilde{Y}, X; \theta)\right]\right\| + \sup_{\theta\in\Theta}\|B_{N2}\|$$

and, thus, $\sup_{\theta\in\Theta}\|g^N(\theta) - g_\pi(\theta)\| \to 0$ as $\alpha_N \to 0$. This implies that

$$\sup_{\theta\in\Theta}\left|Q_N\left(\theta, \pi^N\right) - Q(\theta, \pi)\right| \to 0. \tag{S7}$$

Now fix $\varepsilon > 0$. Let us show that for large enough N^x, N^y, $Q(\theta^N, \pi) < Q(\widetilde{\theta}, \pi) + \varepsilon$. Indeed, (S7) implies that when N^x and N^y are large enough, $Q(\theta^N, \pi) < Q_N(\theta^N, \pi^N) + \varepsilon/3$. Also, $Q_N(\theta^N, \pi^N) < Q_N(\widetilde{\theta}, \pi^N) + \varepsilon/3$ because θ^N is an arg min of $Q_N(\theta^N, \pi^N)$. Finally, (S7) implies that when N^x and N^y are large enough, $Q_N(\widetilde{\theta}, \pi^N) < Q(\widetilde{\theta}, \pi) + \varepsilon/3$.

Let S be any open neighborhood of $\widetilde{\theta}$ and let S^c be its complement in $\mathbb{R}^l$. From the compactness of Θ and the continuity of $\rho(\cdot, \cdot, \cdot)$ in θ, we conclude that $\min_{S^c\cap\Theta} Q(\theta, \pi)$ is attained. The fact that $\widetilde{\theta}$ is the unique minimizer of $Q(\theta, \pi)$ gives that $\min_{S^c\cap\Theta} Q(\theta, \pi) > Q(\widetilde{\theta}, \pi)$. Denote $\varepsilon = \min_{S^c\cap\Theta} Q(\theta, \pi) - Q(\widetilde{\theta}, \pi)$. As we showed above, for this ε we have, when N^x and N^y are large enough, that

$$Q\left(\theta^N, \pi\right) < Q(\widetilde{\theta}, \pi) + \varepsilon = \min_{S^c\cap\Theta} Q(\theta, \pi),$$

which for large enough N^x and N^y gives $\theta^N \in S$. Since S can be chosen arbitrarily small, this means that $\theta^N \to \widetilde{\theta}$.

References

Chaudhuri, S., K. Ganjam, V. Ganti, and R. Motwani (2003), "Robust and efficient fuzzy match for online data cleaning." In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, 313–324, ACM. [2]

Fellegi, I. and A. Sunter (1969), "A theory for record linkage." *Journal of the American Statistical Association*, 1183–1210. [1]

Gusfield, D. (1997), *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press. [2]

Jaro, M. (1989), "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida." *Journal of the American Statistical Association*, 414–420. [2]

Newcombe, H., J. Kennedy, S. Axford, and A. James (1959), "Automatic linkage of vital and health records." *Science*, 130, 954–959. [1]

Ridder, G. and R. Moffitt (2007), "The econometrics of data combination." *Handbook of Econometrics*, 6, 5469–5547. [2]

Salton, G. and D. Harman (2003), *Information Retrieval*. John Wiley and Sons Ltd. [2]

Winkler, W. (1999), "The state of record linkage and current research problems." In *Statistical Research Division*, US Census Bureau, Citeseer. [2]

Co-editor Rosa L. Matzkin handled this manuscript.

Manuscript received 20 April, 2015; final version accepted 26 January, 2017; available online 12 May, 2017.