Biometrika (2018), **105**, 2, *pp.* 431–446
Printed in Great Britain
doi: 10.1093/biomet/asy003
Advance Access publication 19 March 2018

Theoretical limits of microclustering for record linkage

By J. E. JOHNDROW

Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, California 94305, U.S.A.
johndrow@stanford.edu

K. LUM

Human Rights Data Analysis Group, San Francisco, California 94110, U.S.A.
kl@hrdag.org

and D. B. DUNSON

Department of Statistical Science, Duke University, Box 90251, Durham, North Carolina 27708, U.S.A.
dunson@duke.edu

Summary

There has been substantial recent interest in record linkage, where one attempts to group the records pertaining to the same entities from one or more large databases that lack unique identifiers. This can be viewed as a type of microclustering, with few observations per cluster and a very large number of clusters. We show that the problem is fundamentally hard from a theoretical perspective and, even in idealized cases, accurate entity resolution is effectively impossible unless the number of entities is small relative to the number of records and/or the separation between records from different entities is extremely large. These results suggest conservatism in interpretation of the results of record linkage, support collection of additional data to more accurately disambiguate the entities, and motivate a focus on coarser inference. For example, results from a simulation study suggest that sometimes one may obtain accurate results for population size estimation even when fine-scale entity resolution is inaccurate.

Some key words: Closed population estimation; Clustering; Entity resolution; Microclustering; Record linkage; Small clusters.

1. Introduction

Record linkage refers to the problem of assigning records to unique entities based on observed characteristics. One example, which is the motivating problem for this work, arises in human rights research (Lum et al., 2013; Sadinle & Fienberg, 2013; Sadinle, 2014), where there is interest in recording deaths or other human rights violations attributable to a conflict, such as the ongoing conflict in Syria. In this setting, the data are incomplete records of violations, which usually consist of a name, a date of death, and a place of death. In the turbulent atmosphere accompanying a conflict, often multiple organizations record information on deaths with little communication or standardization of recording practices. Because these data are usually gathered from oral recollections of survivors, measurement errors are common. The result is multiple

databases consisting of noisy observations on features of the deceased that in some cases would not uniquely identify the individual even in the absence of noise. There are two distinct inferential goals when applying record linkage in this setting: identification of specific victims and estimation of the total number of casualties in the conflict. These two objectives are shared by other common application areas. For example, in fraud detection, entity resolution itself is the objective, whereas in social science applications, coarser inferences such as correlations between linked variables or estimated regression coefficients (Lahiri & Larsen, 2005) are of primary interest; see D'Orazio et al. (2006) for specific examples.

A variety of methods for record linkage have been proposed (Winkler, 2006; Christen, 2012), though much of the literature has focused on the theoretical framework of Fellegi & Sunter (1969). In this set-up, every pair of records from two databases is compared using a discrepancy function of record features and classified as either a match, a nonmatch, or possibly a match. The goal is to design a decision rule that minimizes the number of possible matches for fixed match and nonmatch error rates. The necessity of performing pairwise comparisons leads to a combinatorial explosion, and a related literature has focused on the construction of blocking rules to limit the number of comparisons performed (Jaro, 1989, 1995; Al-Lawati et al., 2005; Bilenko et al., 2006; Michelson & Knoblock, 2006).

An alternative and more recent approach is to perform entity resolution through clustering, where the goal is to recover the entities from one or more noisy observations on each entity (Steorts et al., 2014, 2015; Steorts, 2015; Zanella et al., 2016). In this framework, entities and clusters are equivalent. Model-based or likelihood-based methods of this sort can be equated with mixture modelling, where the number of mixture components is large and the number of observations per component is very small. Historically, the focus in mixture modelling has been on regularization that penalizes large numbers of clusters, in order to obtain a more parsimonious representation of the data-generating process. Recognizing that this type of regularization is inappropriate for most record linkage problems, Miller et al. (2015) defined the concept of microclustering, where the cluster sizes grow at a sublinear rate with the number of observations. They proposed a Bayes nonparametric approach to clustering in this setting that takes advantage of a novel random partition process that has the microclustering property. This is applied to multinomial mixtures in Zanella et al. (2016).

While microclustering is appropriate for most record linkage problems, there is a lack of literature on performance guarantees and other theoretical properties of entity resolution procedures. Because microclustering methods favour sublinear growth in cluster sizes, the number of parameters of these models can grow at the same rate as the number of observations, so basic asymptotic properties such as central limit theorems, strong laws and consistency will not hold. For example, in the human rights applications that motivated Miller et al. (2015), the number of unique records per entity is thought to be very small, generally less than 10, while the number of unique entities is thought to be in the thousands or hundreds of thousands. As such, it is critical to consider the finite-sample performance of microclustering in cases where the number of records per cluster is a tiny fraction of the sample size, and to obtain theoretical upper bounds on how accurate cluster-based entity resolution can possibly be when the microclustering condition holds.

Working with simple mixture models where some of the parameters are known, we characterize the exact distributions of quantities related to entity resolution. Achievable performance is shown to be a function of entity separation and the noise level. Using these results, we provide minimal conditions for accuracy in entity resolution to be bounded away from zero asymptotically as the number of records grows. We also provide an information-theoretic bound on the best possible performance in the case where some of the entities cannot be uniquely identified from noiseless observations of the available features. These results are supported by several simulation studies.

Our problem is related to the extensive literature on mixture identifiability (Teicher, 1961, 1963; Yakowitz & Spragins, 1968; Holzmann et al., 2006) and estimation of the number of components (Day, 1969; Richardson & Green, 1997; Lo et al., 2001; Tibshirani et al., 2001), as well as the voluminous literature on clustering (see Hastie et al., 2009, Ch. 3 and works cited therein), with the important distinction that we focus on microclusters, mixtures with many components and few observations per component, and we are interested primarily in entity resolution, not in estimation of the parameters of the mixture.

Our results initially present a very dim view of entity resolution by microclustering; indeed, it appears that the full problem is unsolvable without further information except under very strong conditions. However, in many cases interest is focused on certain summary statistics of the linked records, which may be relatively insensitive to errors in entity resolution. Motivated by the human rights application mentioned above, we consider the case where the ultimate goal of entity resolution is to recover the total number of entities in the population. This corresponds to the total number of casualties in the conflict, the coarser inferential goal mentioned previously. A variety of methods exist for this problem, which is referred to as closed population estimation, and generally use as data a relatively small contingency table that characterizes the number of unique records appearing in every possible combination of the databases (Wolter, 1986; Zaslavsky & Wolfgang, 1993; Griffin, 2014). In a simulation study, we show that relatively accurate estimation of the total population size is possible even when entity resolution is inaccurate. The success of population estimation in this admittedly limited simulation study suggests further investigation of whether low-dimensional summaries are in general recoverable from linked databases even when the error rate in entity resolution is high.

2. Main results

2·1. *Preliminaries*

We work primarily with Gaussian mixtures of the form

$$L(y \mid \nu, \{\mu_k, \Sigma_k\}_{k=1,\dots,K}) = \sum_{k=1}^{K} \nu_k \phi(y; \mu_k, \Sigma_k), \tag{1}$$

where $\nu \in \mathcal{S}^{K-1}$ is an element of the $(K-1)$-dimensional probability simplex, $\mu_k, y \in \mathbb{R}^p$, K is a positive integer, Σ_k is a $p \times p$ positive-definite matrix, and $\phi(y; \mu, \Sigma) = |2\pi\Sigma|^{-1/2} \exp\{-(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)/2\}$ is the Gaussian density function. In (1), y are observed entity-specific features that we will use to perform record linkage. In our motivating application, typical features are name, time/date of death, and place of death. It is natural to treat time and place as continuous variables, and it is common to embed name into an abstract continuous space by way of a metric on text, such as Jaccard similarity or Levenshtein distance. As such, (1) provides a reasonable default mixture in our setting.

The mixture (1) differs from the mixture considered in Zanella et al. (2016), which is similar to that in Dunson & Xing (2009), a nonparametric Bayesian model for multivariate categorical data. Our rationale for using Gaussian mixtures comes from the results of Johndrow et al. (2017) and Fienberg et al. (2009), which make clear that the maximum number of unique mixture components in the model of Dunson & Xing (2009) is strictly less than d^p, where d is the number of distinct levels of the categorical variables. Thus, it is impossible to resolve more than d^p entities on the basis of p categorical measurements, motivating our focus on the case of continuous features, which does not suffer from this fundamental limitation.

Table 1. *Example of data for name problem*

Name (y_i)	Identifier (q_i)
John Smith	1
John Smith	2
Jane Wang	8
Jane Wang	9
Anna Rodriguez	11
Anna Rodriguez	14

In providing an upper bound on performance in entity resolution, we focus on a case that favours good performance; in particular, we consider the task of correctly determining which mixture component generated each $y_i \sim L(y \mid \nu, \{\mu_k, \Sigma_k\}_{k=1,\ldots,K})$ $(i = 1, \ldots, N)$, assuming that (1) is known. We focus on the estimator

$$\hat{k}(y) = \arg\max_k \log \phi(y; \mu_k, \Sigma_k) = \arg\max_k \log \phi_k(y). \tag{2}$$

We will assign y to the mixture component that maximizes the likelihood; this is the Bayes rule classifier with equal prior weight on each component. This estimator allows many-to-one matches. In what follows, we will study a series of cases where the set of unknown parameters in the model is gradually expanded, which provides a set of theoretically tractable finite-sample bounds on the best-case performance of clustering-based approaches to entity resolution. Although we focus on Gaussian mixtures for simplicity, many of the results apply equally to mixtures of any kernels that are functions of a metric on $\mathbb{R}^d$, and we point out extensions where appropriate.

2·2. *An information-theoretic bound*

We first consider multiple true entities with identical values of the entity-specific parameters (μ_k, Σ_k). Suppose that we observe two complete enumerations of a population, each containing a nearly mutually exclusive set of covariates about each individual. We assume that these two lists contain only one field in common. For example, suppose one list contains each individual's name and date of birth and the other contains each individual's name, location of death, and date of death. The goal is to match each individual on the first list to the correct individual on the second list to produce a complete dataset consisting of name, date of birth, date of death, and location of death for each individual in the population.

In locations with low entropy in the name distribution, as is the case in Syria, this list is likely to be composed of many individuals sharing exactly the same first and last name. In this section, we illustrate the limitations in performance of record linkage when multiple entities have identical values of (μ_k, Σ_k) and the data are observed without noise. In the context of (1), this corresponds to the limit as the maximum eigenvalue of Σ_k approaches zero, resulting in a mixture of delta measures. For simplicity, we focus on the case where the features are names, with an obvious parallel to the case where features are vectors in $\mathbb{R}^p$ and multiple entities have identical true values of the feature vector.

Suppose that we observe a list of names y_i for $i = 1, \ldots, N = K$, where y_i takes $M < N$ unique values. Let $N_m = \sum_i \mathbb{1}(y_i = \mu_m^*)$ for $m = 1, \ldots, M$, where $\{\mu_m^* : m = 1, \ldots, M\}$ is the set of unique values of μ_k; N_m is the number of times the name μ_m^* appears in the database. Let $q_i \in \{1, \ldots, K\}$ denote an unobserved identifier of the component that generated y_i. For example, the full data could look like Table 1 and we only observe the name column.

The goal is to assign the correct identifier to each record or, equivalently, to determine from which component each record was generated. This is related to the problem of relinking two

paired variables when the ordering of the variables has been independently permuted, as outlined in DeGroot & Goel (1980) and references therein. We consider the case where it is known that there is exactly one record corresponding to each person, and use a random allocation procedure. When multiple true entities have identical values of μ_k, the estimator in (2) does not give a unique solution, since $\mathrm{pr}(y \mid \mu_k) = 1$ if $y = \mu_k$ and 0 otherwise, so the likelihood has identical values for all k such that $\mu_k = y$.

Let $\mathcal{I}_m = \{i : y_i = \mu_m^*\}$ be the set of all records with name μ_m^*, and let $\mathcal{I}_m^* = \{k : \mu_k = \mu_m^*\}$ be the set of all components with mean μ_m^*; this is the set of values q_i can potentially take for each $i \in \mathcal{I}_m$. The procedure used is to randomly assign records $i \in \mathcal{I}_m$ to a permutation of the elements of $\mathcal{I}_m^*$ such that each record is assigned to exactly one of the mixture components that could have generated it. After making this assignment, the true value of q_i is revealed and the number of correct assignments enumerated. Clearly, there are $N_m!$ ways to assign each individual with the same name to an element of $\mathcal{I}_m^*$, where $|\mathcal{I}_m^*| = N_m$, and only one of these assignments will be exactly right. Let Z_m be the number of correct assignments with name μ_m^*, and let $Z = \sum_{m=1}^{M} Z_m$. Then the probability of assigning every record i to its true q_i is $\mathrm{pr}(Z = N) = \prod_{m=1}^{M} (N_m!)^{-1}$. On the log scale this turns out to be very intuitive since, by Stirling's approximation,

$$\begin{aligned}\log\{\mathrm{pr}(Z = N)\} &= -\sum_{m=1}^{M} \log N_m! = -\sum_{m=1}^{M} N_m \log N_m - N_m + O(\log N_m) \\ &= -H_Y + N - \sum_{m=1}^{M} O(\log N_m),\end{aligned}$$

where H_Y is the entropy of the name distribution. Moreover, the distribution of Z_m can be described by the probability mass function

$$\mathrm{pr}(Z_m = z) = \frac{1}{N_m!} \binom{N_m}{z} \{!(N_m - z)\} \tag{3}$$

where, for an integer n, $!n$ is the number of derangements of the integers $1, \ldots, n$, i.e., the number of ways to rearrange the sequence $1, \ldots, n$ such that none of the elements of the sequence are in their original locations. We have the relation

$$!n = \left\lfloor \frac{n!}{\mathrm{e}} + \frac{1}{2} \right\rfloor = \mathrm{round}(n!/\mathrm{e}) \tag{4}$$

for $n \geqslant 1$, where $\lfloor \cdot \rfloor$ is the floor function; also, $!0 \equiv 1$.

We now consider the expectation of Z_m. It is straightforward to compute upper and lower bounds; proofs are deferred to the Appendix.

Remark 1. The expectation of Z_m satisfies

$$\frac{\Gamma(N_m, 1)}{\Gamma(N_m)} \leqslant E(Z_m) \leqslant \frac{\Gamma(N_m, 1)}{\Gamma(N_m)} + \frac{2^{N_m - 1}}{(N_m - 1)!},$$

where $\Gamma(N_m, 1) = \int_1^\infty t^{N_m - 1} \exp(-t)\, \mathrm{d}t$ is the incomplete gamma function.

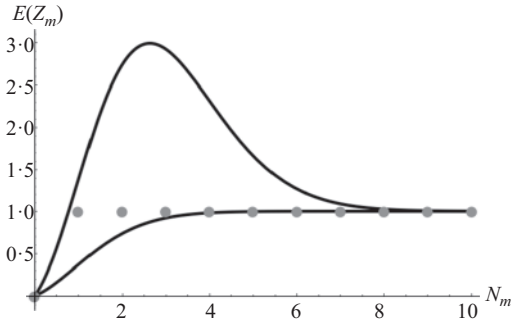

Fig. 1. Upper and lower bounds on $E(Z_m)$ (lines) and the exact value of $E(Z_m)$ (points) for $N_m \in \{0, \ldots, 10\}$.

The difference between the upper and lower bounds is less than 0·001 when $N_m = 11$, so for large N_m the lower bound is very accurate. Figure 1 shows the upper and lower bounds as well as the exact value of $E(Z_m)$, which can quickly be computed exactly for $N_m \leqslant 10$ and is identically 1 in all cases. From this it is clear that taking $E(Z_m) = 1$ for all N_m is at least a very accurate approximation, and is probably the exact value of the expectation. Assuming it is exact, we have $E(Z) = M$, and the expected proportion of correct assignments is MN^{-1}.

We give a concentration inequality for the proportion of correct assignments, Z/N. We have $N^{-1}Z = N^{-1}\sum_{m=1}^{M} Z_m$. As the Z_m are independent and $E(Z) = M$, by Hoeffding's inequality we have

$$\mathrm{pr}(|Z/N - M/N| > t) \leqslant 2\exp\left\{\frac{-2t^2}{\sum_{m=1}^{M}(N_m/N)^2}\right\} = 2\exp\left(\frac{-2N^2t^2}{\sum_{m=1}^{M}N_m^2}\right).$$

We obtained data from the U.S. Census Bureau on the frequency of all surnames and given names in the U.S. population. Assuming independent selection of first and last names in the overall population, we estimate $E(Z/N) = 0{\cdot}28$ for entity resolution of the U.S. population on the basis of only first name and last name. Dependence between first and last names will tend to decrease this expectation. We have $N^2(\sum_m N_m^2)^{-1} > 6 \times 10^5$, so

$$\mathrm{pr}(|Z/N - M/N| > t) \leqslant 2\exp(-1{\cdot}2 \times 10^6 t^2);$$

for example, the probability that $Z/N > 0{\cdot}29$ is less than 10^{-51}. Hence, in the United States names example, the distribution is highly concentrated around its expectation and there is an extremely low probability of getting even one third or more of the assignments correct.

For additional context, we also computed $E(Z/N)$ for two states. For the least populated state, Wyoming, we estimate $E(Z/N) = 0{\cdot}89$, while for the most populated state, California, we estimate $E(Z/N) = 0{\cdot}45$. We also compute $E(Z/N)$ for the entire United States, assuming that in addition to first and last names we also observe the last four digits of each person's social security number. We assume that these digits are assigned uniformly at random from integers between 0000 and 9999 independently of first and last name. Adding this extra information to first name and last name for the U.S. population gives $E(Z/N) = 0{\cdot}57$. Thus, in each case a substantial proportion of errors is likely. These examples illustrate the fact that in many entity resolution problems, the best possible performance is substantially less than perfect accuracy due to redundancy in the true values of the entity features. This provides an upper bound on the performance achievable when features are observed with noise.

2·3. *Analysis of noisy observations when mixture parameters are known*

Having established the limitations resulting from redundancy of the true entity features, we now analyse the effect of noise in the setting where all true entity features are distinct. We begin with a highly simplified case. Suppose we observe a data sequence $y_1, \ldots, y_N$ and that each observation originates from the mixture in (1) with $N = \lambda K$ for $\lambda \in \mathbb{N}$, $\nu_k = N^{-1}$, $\mu_k \in [a, b]$, and $\Sigma_k = \sigma^2$ for all k. Although the results are general, we have in mind situations in which λ is some small positive integer and most entities have on the order of λ records in the data, the typical situation in our motivating human rights applications.

Assume that the parameters $\{\mu_k\}_{1 \leqslant k \leqslant K}$, σ^2 and ν are known. On observing y, we use estimator (2) of the mixture component it originated from. Let k_0 be the true value of k. Then, letting $\mathrm{pr}_{\phi_{k_0}}\{A(y)\}$ denote the probability of event $A(y)$ if y is drawn from component k_0 of (1),

$$\mathrm{pr}\{\hat{k}(y) = k_0\} = \mathrm{pr}_{\phi_{k_0}}\left\{(y - \mu_{k_0})^2 = \bigwedge_{k=1}^{K} (y - \mu_k)^2\right\},$$

where $\bigwedge_{k=1}^{K} y_k$ denotes the minimum of the collection $\{y_k,\ k = 1, \ldots, K\}$. We make the simplifying assumption that the μ_k are equally spaced, so that $|\mu_k - \mu_{k+1}| = \delta_K = (K-1)^{-1}|b-a| = (K-1)^{-1}\ell$ for all k. Then, letting $\Phi(\cdot)$ denote the standard normal distribution function,

$$\begin{aligned}\mathrm{pr}\{\hat{k}(y) = k_0\} &= \mathrm{pr}_{\phi_{k_0}}\left\{(y - \mu_{k_0})^2 < \frac{\delta_K^2}{2}\right\} = \mathrm{pr}_{\phi_{k_0}}\left(\frac{|y - \mu_{k_0}|}{\sigma} < \frac{\delta_K}{2\sigma}\right) \\ &= \Phi\left(\frac{\delta_K}{2\sigma}\right) - \Phi\left(\frac{-\delta_K}{2\sigma}\right) = 2\Phi\left(\frac{\delta_K}{2\sigma}\right) - 1, \quad k_0 \neq 1, K. \end{aligned} \tag{5}$$

For $k_0 = 1$ or $k_0 = K$, the expression is $\Phi\{\delta_K/(2\sigma)\}$. When K is large, the effect of using (5) for all k is negligible, so to simplify exposition we will do so. A condition like that in (5) would hold for any mixtures where the component densities are a function of a metric on $\mathbb{R}$, with $\Phi(\cdot)$ replaced by a different distribution function. This includes many of the kernel functions commonly used in machine learning, as well as other common densities such as the t density.

With Z being the number of correct classifications, we have the following result for the Gaussian mixture.

Remark 2 (Infeasibility result for microclustering). Suppose $\mu_k \in \mathbb{R}$ are equally spaced and restricted to a compact set, so that $\delta_K = (K-1)^{-1}\ell$. Then

$$\mathrm{pr}\left(\left|\frac{Z}{N} - \left[2\Phi\left\{\frac{\ell}{2(K-1)\sigma}\right\} - 1\right]\right| > t\right) < 2\exp(-2t^2\lambda K)$$

and

$$\lim_{K\to\infty} \mathrm{pr}(Z = 0) = \lim_{N\to\infty}\left[2 - 2\Phi\left\{\frac{\ell}{2(N/\lambda - 1)\sigma}\right\}\right]^N = \exp\left[-\ell\lambda/\{(2\pi)^{1/2}\sigma\}\right].$$

Therefore, in large populations, the proportion of correct assignments, $N^{-1}Z$, is highly concentrated around its expectation given by (5), which will be very near zero when $K^{-1} \ll \ell/(2\sigma)$. Evidently, $Z \to 0$ almost surely and the probability of zero correct assignments is bounded away

from zero unless $\lim_{K\to\infty} \ell/(\sigma K) > 0$, which requires $\ell/\sigma = \Omega(K)$, where $f(K) = \Omega\{g(K)\}$ means that there exist constants C and $K_0 < \infty$ such that $f(K) > Cg(K)$ for all $K > K_0$. In other words, either the width of the set containing the means must grow at a rate of at least K, or the observation noise must go to zero at least as fast as K^{-1}. We refer to the condition $\ell/\sigma = \Omega(K)$ as infinite separation, as it effectively requires that the entities be infinitely far apart relative to the noise level in the limit. Practically, this means that for entity resolution via microclustering, measurements on entity-specific features must get more precise as the number of entities increases. Given that this regime applies when all the parameters of the mixture are known, Remark 2 suggests that the full problem of entity resolution by clustering is practically impossible in most cases. Estimates of these parameters would have standard error of the order $K^{1/2}N^{-1/2}$. Therefore, when $N = \lambda K$, which is the case in most record linkage applications, standard errors are constant in the number of observations, and uncertainty in parameters remains even asymptotically, so the result in Remark 2 understates the futility of the problem.

2·4. *The effect of dimension*

We now consider the case where the dimension p_K grows with K, and show that when the parameters of the mixture are known, infinite separation can be achieved when the means reside on a compact set and observation noise does not decay to zero as $K \to \infty$. Consider the mixture in (1) with $\mu_k \in \mathbb{R}^{p_K}$ and $\Sigma_k = \sigma^2 I_{p_K}$ for all k. Assume that the means are restricted to the Euclidean unit ball $\mathbb{B}(p_K)$ in $\mathbb{R}^{p_K}$ and they are arranged so that $\|\mu_j - \mu_k\| \geqslant \delta_K$ for every $j \neq k$, where $\|\cdot\|$ is the Euclidean norm. The maximum number of means that can fit inside $\mathbb{B}(p_K)$ while satisfying this separation condition is the δ_K-packing number $\mathcal{M}\{\delta_K; \mathbb{B}(p_K), \|\cdot\|\}$, which is related to the δ_K-covering number $\mathcal{N}\{\delta_K; \mathbb{B}(p_K), \|\cdot\|\}$ by the inequality

$$\mathcal{N}\{\delta_K; \mathbb{B}(p_K), \|\cdot\|\} \leqslant \mathcal{M}\{\delta_K; \mathbb{B}(p_K), \|\cdot\|\} \leqslant \mathcal{N}\{\delta_K/2; \mathbb{B}(p_K), \|\cdot\|\}. \tag{6}$$

The covering number of the unit ball satisfies

$$p_K \log \frac{1}{\delta_K} \leqslant \log \mathcal{N}\{\delta_K; \mathbb{B}(p_K), \|\cdot\|\} \leqslant p_K \log\left(1 + \frac{2}{\delta_K}\right). \tag{7}$$

If we have K points inside $\mathbb{B}$ that are δ_K-separated, then at most $K = \mathcal{M}\{\delta_K; \mathbb{B}(p_K), \|\cdot\|\}$, so combining (6) and (7) gives

$$K^{-1/p_K} < \delta_K < 4(K^{1/p_K} - 1)^{-1}. \tag{8}$$

The maximum likelihood estimator (2) satisfies

$$\mathrm{pr}\{\hat{k}(y) = k_0\} = \mathrm{pr}_{\phi_{k_0}}\left\{\frac{(y-\mu_{k_0})^{\mathrm{T}}(y-\mu_{k_0})}{\sigma^2} < \frac{\delta_K^2}{2\sigma^2}\right\} = \mathrm{pr}\left(\chi^2_{p_K} < \frac{\delta_K^2}{2\sigma^2}\right),$$

where $\chi^2_{p_K}$ is chi-squared with p_K degrees of freedom. Appealing to the central limit theorem,

$$\mathrm{pr}\left(\chi^2_{p_K} < \frac{\delta_K^2}{2\sigma^2}\right) \to \Phi\left\{\frac{\delta_K^2/(2\sigma^2) - p_K}{(2p_K)^{1/2}}\right\},$$

so $\delta_K^2/\sigma^2 \geqslant Cp_K$ for all large K with $0 < C < \infty$ is a necessary and sufficient condition for $\text{pr}\{\hat{k}(y) = k_0\}$ to converge to a nonzero constant as $K, p_K \to \infty$. Combining this with (8), we obtain that $\delta_K^2/\sigma^2 \geqslant Cp_K$ implies $\sigma^2 \leqslant 16(Cp_K)^{-1}(K^{1/p_K} - 1)^{-2}$. Thus, it is even possible to have σ^2 bounded away from zero if p_K grows fast enough with K. For example, if $p_K = (\log K)^2$, then $\lim_{K\to\infty} 16(Cp_K)^{-1}(K^{1/p_K} - 1)^{-2} = 16C^{-1}$. Of course, having $p_K \to \infty$ in the case where the mixture parameters are unknown means that for each mixture component we must estimate a growing number of parameters, and $N = \Omega(Kp_K)$ is a necessary condition for consistency. Therefore $N = \lambda K$ will not be sufficient, and we must have the number of records per entity growing faster than p_K, which cannot occur in the microclustering setting. The practical ramification is that, if we ignore the need to estimate the parameters of each component, one way to combat the failure of entity resolution as the number of entities increases is to attempt to increase the number of variables collected per record on each entity.

2·5. *The case where means are unknown: Bayesian mixtures*

We now consider the case where the mixture component means are unknown. Suppose that N observations are generated from the mixture given in (1) with σ^2 and ν known. Consider a Bayesian analysis with independent priors $\mu_k \sim N(0, \tau^2)$. The calculations leading to the following results can be found in the Appendix and Supplementary Material.

Let $\gamma_1, \ldots, \gamma_N$ for $\gamma_i \in \{1, \ldots, K\}$ be a configuration of the N observations into K classes, and let $N_k = \sum_i \mathbb{1}(\gamma_i = k)$. Let $\mathcal{G} = \{1, \ldots, K\}^N$ be the set of all possible configurations, with $|\mathcal{G}| = K^N$. The marginal likelihood of the configuration, integrating out the means, is

$$L(y, \gamma \mid \nu, \tau^2, \sigma^2) = \prod_{k=1}^{K} \frac{\nu_k^{N_k}\sigma}{(2\pi\sigma^2)^{N_k/2}(N_k\tau^2 + \sigma^2)^{1/2}} \exp\left\{\frac{\tau^2(N_k\bar{y}_k)^2}{2\sigma^2(N_k\tau^2 + \sigma^2)} - \frac{N_k\overline{y_k^2}}{2\sigma^2}\right\}, \quad (9)$$

where $\overline{y_k^2} = N_k^{-1}\sum_{i:\,\gamma_i=k} y_i^2$ and $\bar{y}_k = N_k^{-1}\sum_{i:\,\gamma_i=k} y_i$; so the posterior probability of the configuration is

$$p(\gamma \mid y, \nu, \tau^2, \sigma^2) = \frac{L(y, \gamma \mid \nu, \tau^2, \sigma^2)}{\sum_{\gamma^* \in \mathcal{G}} L(y, \gamma^* \mid \nu, \tau^2, \sigma^2)} = \frac{1}{1 + \sum_{\gamma^* \neq \gamma} \text{BF}(\gamma^*, \gamma)},$$

where the Bayes factor is $\text{BF}(\gamma^*, \gamma) \equiv L(y, \gamma^* \mid \nu, \tau^2, \sigma^2)/L(y, \gamma \mid \nu, \tau^2, \sigma^2)$. Consider the case where γ^* consists of all singleton clusters while γ consists of $N-2$ singleton clusters, one empty cluster, and a single cluster with two observations. There are N such elements of $\mathcal{G}$. The Bayes factor is

$$\text{BF}(\gamma^*, \gamma) = \frac{\nu_j}{\nu_k}\frac{\sigma(2\tau^2 + \sigma^2)^{1/2}}{(\tau^2 + \sigma^2)} \exp\left[\frac{\tau^2}{2\sigma^2}\left\{\frac{y_i^2 + y_{i'}^2}{\tau^2 + \sigma^2} - \frac{(y_i + y_{i'})^2}{2\tau^2 + \sigma^2}\right\}\right],$$

where i and i' are the indices of the two observations that are allocated to the same cluster in γ and different clusters in γ^*, j is the cluster that contains y_i in configuration γ^* and is empty

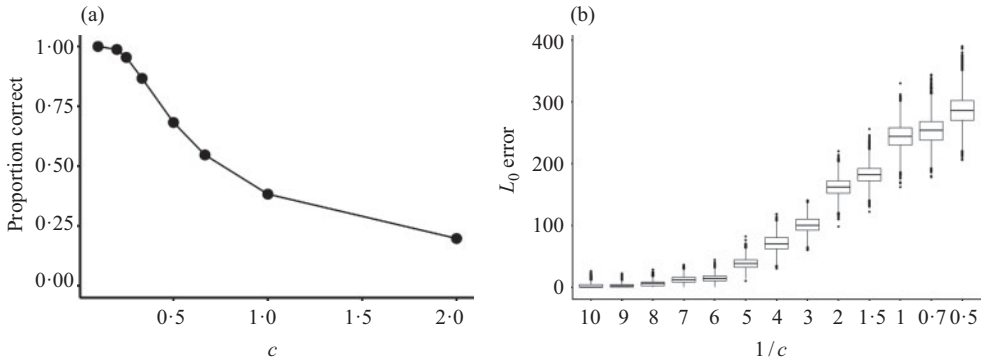

Fig. 2. Performance of Bayes mixtures in entity resolution: (a) proportion of entities correctly assigned using maximum likelihood assignment when parameters are known; (b) boxplot of Markov chain Monte Carlo samples of $\|A - A^{(0)}\|_0$ for Bayes mixtures with unknown means versus c^{-1}.

in configuration γ, and k is the index of the cluster that contains observation $y_{i'}$ in configuration γ^* and contains both y_i and $y_{i'}$ in configuration γ. Suppose that the truth is configuration γ^*, with $N = K$ distinct entities. Integrating the Bayes factor over the data distribution, we obtain

$$\int \mathrm{BF}(\gamma^*, \gamma)\phi(y_i; \mu_i, \sigma)\phi(y_{i'}; \mu_{i'}, \sigma) = \frac{\nu_j}{\nu_k} \frac{2\tau^2 + \sigma^2}{\{2(\sigma^2 + \tau^2)^2 - \sigma^4\}^{1/2}} \times \exp\left[-\frac{1}{4}\tau^2\left\{-\frac{(\mu_i - \mu_{i'})^2}{\sigma^4} + \frac{(\mu_i + \mu_{i'})^2}{2(\sigma^2 + \tau^2)^2 - \sigma^4}\right\}\right]. \tag{10}$$

From this it is clear that when $N = K$, as $\|\mu_i - \mu_{i'}\| \to 0$, the expectation of the Bayes factor converges to a constant, and a necessary condition for $E\{\mathrm{BF}(\gamma^*, \gamma)\} \to \infty$ is $\|\mu_i - \mu_{i'}\| \to \infty$. Therefore, when the μ_i are confined to a compact set, Bayes factors for infinitely many incorrect configurations will converge to constants in expectation as $K \to \infty$, since $K \to \infty$ implies $\|\mu_i - \mu_{i'}\| \to 0$ for infinitely many pairs i, i'. It follows that the posterior will not even be consistent for entity resolution, and will fail to concentrate on any finite set of configurations asymptotically.

3. EMPIRICAL ANALYSIS OF ENTITY RESOLUTION BY MICROCLUSTERING

We show through simulation studies that the infeasibility results are borne out empirically. We first consider the case where there are $K = 5000$ entities and we observe data $y_i \sim N(\mu_i, \sigma^2)$ for $i = 1, \ldots, N$ with $N = K$. The common variance parameter σ^2 is cK^{-1}, and c is varied between 0·1 and 2 across the simulations. In every case $\mu_i = i/K$, so the means are equally spaced on the unit interval. Entity resolution is performed using the estimator in (2).

The results are shown in Fig. 2(a). As expected, the proportion correctly assigned decreases with c. Entity resolution is nearly perfect for $c = 0{\cdot}1$, but begins to decline noticeably around $c = 0{\cdot}25$, which is intuitive since at that value, half the distance between the true means, the threshold at which misassignment occurs using the maximum likelihood estimate is twice the standard deviation. For $c = 2/3$, approximately half of the observations are correctly assigned. When $c = 2$, the proportion correctly assigned is about 0·2.

We perform a second simulation in which we conduct entity resolution without knowledge of the true means. We simulated $N = 100$ observations from

$$\gamma_i \sim \text{Categorical}\{(1/K, \ldots, 1/K)\}, \quad y_i \mid \gamma_i \sim N(\gamma_i/K, \sigma^2)$$

with $\sigma^2 = cK^{-1}$, where c varied between 0·1 and 2 across the simulations. We then performed posterior computation by collapsed Gibbs sampling for the Bayesian mixture model with known component weights and component variances described in § 2·5. We used identical priors $\mu_k \sim N(0, 9)$ on the means for each component. For each Markov chain Monte Carlo sample, we computed an adjacency matrix A for the 100 observations, where $A_{ij} = 1$ if observations i and j are assigned to the same component and $A_{ij} = 0$ otherwise. We then computed the L_0 distance between the sampled A and the true adjacency matrix $A^{(0)}$, defined as $\|A - A^{(0)}\|_0 = \sum_{i=1}^N \sum_{j=1}^N \mathbb{1}(A_{ij} \neq A_{ij}^{(0)})$, for each Markov chain Monte Carlo sample. Perfect entity resolution corresponds to $\|A - A^{(0)}\|_0 = 0$, while the value of $\|A - A^{(0)}\|_0$ can conceivably be as large as $100^2 - 100$, which occurs when A is a matrix of ones and $A^{(0)}$ is the identity. Figure 2(b) shows boxplots of the approximate posterior distribution of $\|A - A^{(0)}\|_0$ as a function of c^{-1}. As expected, performance in entity resolution degrades as c increases, with the error rate increasing sharply near the value $c = 0{\cdot}25$.

4. Population size estimation when entity resolution is poor

4·1. *Overview of population size estimation*

Estimation of the number of unique entities when some entities may not appear in any database is referred to as population size estimation and is the ultimate objective of entity resolution in our motivating human rights setting. In this section we give a positive empirical result for this inference problem. We construct a simulation in which it is possible to accurately estimate the number of unique entities from a clustering assignment even when the proportion of records correctly assigned to clusters is small.

We first describe the population size estimation problem and its relationship to entity resolution. Our observed data consist of noisy observations y_i of entity characteristics and an integer $d_i \in \{1, \ldots, T\}$ such that $d_i = j$ indicates that record i appeared in database j, and we aim to estimate K, the number of unique entities. The typical approach uses a two-stage procedure. First, we perform entity resolution on the observed data. The linked records are summarized as a 2^T contingency table that records the estimated number of individuals appearing in every possible combination of the T databases. Specifically, for every $x \in \{0, 1\}^T$, let $\hat{n}(x)$ be the estimated count of the number of entities that appeared in databases $\{j : x_j \neq 0\}$. For example, the entry $\hat{n}(011)$ in the case of three databases gives the estimated count of the number of entities that appear in the second and third databases but not the first. Performing entity resolution gives us $\hat{n}(x)$ for every x except $x = 00 \cdots 0$. In the following, we use $n(0)$ as shorthand for $n(00 \cdots 0)$. One then uses a second-stage population estimation procedure to estimate $\hat{n}(0)$, resulting in $\hat{K} = \sum_{x \in \{0,1\}^T} \hat{n}(x)$.

4·2. *Simulation set-up*

To simulate observations (y_i, d_i), we use the following procedure. We first generate a collection of T database-specific observation probabilities $\pi_1, \pi_2, \ldots, \pi_T$ from $\pi_j \sim \text{Be}(a, b)$. These are population-level probabilities that any given entity will appear in database j. We then use Algorithm 1 to generate data.

Algorithm 1. Generation of synthetic databases.

Set $i = 0$
For $k = 1, \ldots, K$
  For $j = 1, \ldots, T$
    Sample $x_j \sim \text{Ber}(\pi_j)$
    If $x_j = 1$
      $i \leftarrow i + 1$
      Sample $y_i \sim N(k/K, \sigma^2)$
      $d_i = j$
Output $N_{\text{obs}} = i, y_i, d_i$ for $i = 1, \ldots, N_{\text{obs}}$

This results in T synthetic databases which do not contain entries for any of the entities for which the sampled value of x in Algorithm 1 was the zero vector. These are the unobserved entities that are estimated in the second stage of the procedure, and their true count is $n(0)$. In general, we choose a and b in the beta distribution to make $n(0) \approx 0{\cdot}25K$. This is consistent with real population estimation problems encountered in the human rights field and makes the problem relatively challenging compared to, say, the choice $a = b = 1$, which results in much smaller proportions of unobserved entities.

4·3. *Inference procedure*

We perform inference using the following two-stage procedure. For the observed records y_i ($i = 1, \ldots, N_{\text{obs}}$), we first calculate an estimate $\hat{k}$ of the cluster assignments using (2). Let $\hat{x}_k \in \{0,1\}^T$ denote a binary vector with a 1 in element j if entity k is estimated to appear in database j and with zero entries otherwise, for $j = 1, \ldots, T$. For any $x \in \{0,1\}^T$, define $\hat{n}(x) = \sum_k \mathbb{1}(\hat{x}_k = x)$, giving an estimate of the list intersection counts $n(x)$ for all $x \neq 0$. Then, in the second stage, we estimate the number of unobserved entities $n(0)$ using a standard estimator implemented in the R (R Development Core Team, 2018) package Rcapture. We then define $\hat{K} = \sum_{x \in \{0,1\}^T} \hat{n}(x)$, the sum of the estimated number of entities appearing in every possible combination of the databases, including those that appear in no databases. We perform this inference process for 250 replicate, independent simulations for several values of σ^2.

To assess performance, we consider four metrics: (i) the mean proportion of records assigned to their correct entity/cluster; (ii) mean coverage of 95% confidence intervals for K, which is an output of Rcapture; (iii) accuracy of point estimates for the total number of entities K, as measured by

$$1 - \frac{\text{RMSE}(\hat{K})}{K} = 1 - \frac{1}{K}\left\{\frac{1}{r}\sum_{r=1}^{250}(\hat{K}_r - K_r)^2\right\}^{1/2},$$

where r indexes simulation replicate; and (iv) accuracy in estimation of $n(x)$, $x \neq 0$, as measured by $1 - \{1 - R^2(\hat{n}, n)\}$, where R^2 is the squared correlation of $\hat{n}(x)$ with $n(x)$ taken over the entries in $n(x)$ with $x \neq 0$ and the 250 replicate simulations.

The results are presented in Fig. 3 for a series of simulations with $\sigma^2 = cK^{-1}$ for values of c between 0·1 and 2 and $K = 5000$ in each case. As expected, as c increases, accuracy in entity resolution decreases markedly. On the other hand, coverage of 95% confidence intervals for K and the root mean squared error for estimation of K by $\hat{K}$ are insensitive to the value of c. Thus, at least in this example, population estimation on the basis of linked records is not sensitive to the

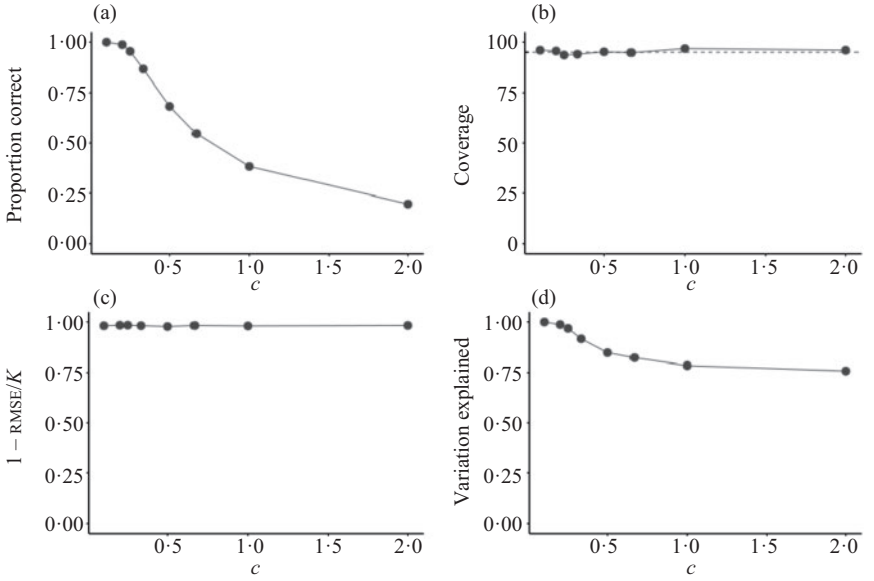

Fig. 3. Plots of simulation results as a function of c for population estimation after entity resolution as described in the text: (a) mean proportion of records correctly assigned; (b) mean coverage of 95% confidence intervals for K; (c) $1 - K^{-1}\,\mathrm{RMSE}(\hat{K})$; (d) $1 - \{1 - R^2(\hat{n}, n)\}^{1/2}$.

accuracy of entity resolution. This is particularly interesting, since estimation of $n(x)$ by $\hat{n}(x)$ for $x \neq 0$ is sensitive to the value of c, as shown in Fig. 3(d). In other words, poor entity resolution results in poorer estimates of the individual cells $n(x)$, $x \in \{0, 1\}^T$, of the contingency table, but their sum K is still estimated accurately.

5. Discussion

This work exposes a fundamental problem with entity resolution via clustering, even in idealized cases, such as when the true data-generating model is known. Empirically, it appears that some functionals of the linked records may be reliably estimated even if entity resolution performance is poor. Understanding which classes of functionals we can estimate and under what conditions is an important area for future research. Another interesting direction is to consider ways of checking whether extensive errors in entity resolution are likely to have occurred after performing model-based clustering by comparing component-specific variance with the separation between the cluster centres.

Acknowledgement

This work was inspired by research conducted at the Human Rights Data Analysis Group. The authors gratefully acknowledge funding support for this work from the Human Rights Data Analysis Group and the U.S. National Institutes of Health.

Supplementary material

Supplementary material available at *Biometrika* online includes a Mathematica notebook with computation of the expression in (10).

APPENDIX

Proof of Remark 1

From (3) and (4) we have

$$E(Z_m) \geqslant \sum_{j=0}^{N_m} \frac{j}{N_m!}\binom{N_m}{j}\frac{(N_m-j)!}{\mathrm{e}} \geqslant \sum_{j=0}^{N_m} \frac{j}{j!(N_m-j)!}\frac{(N_m-j)!}{\mathrm{e}} \geqslant \frac{1}{\mathrm{e}}\sum_{j=1}^{N_m}\frac{1}{(j-1)!} = \frac{\Gamma(N_m,1)}{\Gamma(N_m)},$$

where $\Gamma(N_m,1) = \int_1^\infty t^{N_m-1}\exp(-t)\,\mathrm{d}t$ is the incomplete gamma function. The corresponding upper bound is

$$\begin{aligned} E(Z_m) &\leqslant \sum_{j=0}^{N_m} \frac{j}{N_m!}\binom{N_m}{j}\left\{\frac{(N_m-j)!}{\mathrm{e}}+1\right\} \\ &\leqslant \sum_{j=0}^{N_m} \frac{j}{N_m!}\binom{N_m}{j}\left\{\frac{(N_m-j)!}{\mathrm{e}}+1\right\} \leqslant \frac{\Gamma(N_m,1)}{\Gamma(N_m)} + \frac{2^{N_m-1}}{(N_m-1)!}. \end{aligned}$$

Proof of Remark 2

If $Z \sim \mathrm{Bi}[N, 2\Phi\{\delta_K/(2\sigma)\}-1]$ then $\mathrm{pr}(Z=0) = [2-2\Phi\{\delta_K/(2\sigma)\}]^N$. Clearly, if the μ_k are equally spaced and restricted to be on a compact set of width ℓ, then $\delta_K = \ell/(K-1) = \ell(N/\lambda-1)^{-1}$ for $\ell < \infty$. Since

$$\lim_{K\to\infty}\mathrm{pr}(Z=0) = \lim_{N\to\infty}\left[2-2\Phi\left\{\frac{\ell}{2(N/\lambda-1\sigma)}\right\}\right]^N = \exp\{-\ell\lambda/(2\pi\sigma^2)^{1/2}\},$$

we obtain the second assertion. The first statement is obtained by an application of Hoeffding's inequality.

Gaussian mixture marginal likelihoods

We do the calculation that gives rise to (9). Since each μ_k is assigned an independent prior, we have

$$L(y,\gamma \mid \nu,\tau^2,\sigma^2) = p(y \mid \gamma,\sigma^2,\tau^2)p(\gamma \mid \nu) = p(\gamma \mid \nu)\prod_{k=1}^K p(y^{(k)} \mid \sigma^2,\tau^2),$$

where $y^{(k)} = (y_i)_{i:\gamma_i=k}$ are the observations in class k. The terms $p(y^{(k)} \mid \sigma^2,\tau^2)$ are marginal likelihoods of the data class k in the conjugate Gaussian model with unknown mean, with

$$p(y^{(k)} \mid \sigma^2,\tau^2) = \frac{\sigma}{(N_k\tau^2+\sigma^2)^{1/2}}\exp\left\{\frac{\tau^2(N_k\bar{y}_k)^2}{2\sigma^2(N_k\tau^2+\sigma^2)} - \frac{N_k\overline{y_k^2}}{2\sigma^2}\right\}$$

and $p(\gamma \mid \nu) = \prod_{i=1}^N\prod_{k=1}^K \nu_k^{\gamma_i}(1-\nu_k)^{1-\gamma_i} = \prod_{k=1}^K \nu_k^{N_k}$, where $\bar{y}_k$ and $\overline{y_k^2}$ are defined in the main text.

Bayes factors

The Bayes factor for comparing all singleton clusters γ^* to $N-2$ singleton clusters, one empty cluster, and one cluster with two observations γ is

$$\begin{aligned} \mathrm{BF}(\gamma^*,\gamma) &= \frac{\nu_k\nu_j}{\nu_k^2}\frac{(2\tau^2+\sigma^2)^{1/2}(\sigma^2)^{1/2}}{\tau^2+\sigma^2}\exp\left\{\frac{\tau^2(y_i^2+y_{i'}^2)}{2\sigma^2(\tau^2+\sigma^2)} - \frac{\tau^2(y_i+y_{i'})^2}{2\sigma^2(2\tau^2+\sigma^2)}\right\} \\ &= \frac{\nu_j}{\nu_k}\frac{\sigma(2\tau^2+\sigma^2)^{1/2}}{\tau^2+\sigma^2}\exp\left[\frac{\tau^2}{2\sigma^2}\left\{\frac{y_i^2+y_{i'}^2}{\tau^2+\sigma^2} - \frac{(y_i+y_{i'})^2}{2\tau^2+\sigma^2}\right\}\right], \end{aligned}$$

where the notation i, i' and k, j is defined in the main text.

Expectation of the Bayes factor

This expression can be obtained by repeatedly completing the square. The calculation is simple but tedious and was performed in Mathematica. A Mathematica notebook is provided in the Supplementary Material.

References

Al-Lawati, A., Lee, D. & McDaniel, P. (2005). Blocking-aware private record linkage. In *Proceedings of the 2nd International Workshop on Information Quality in Information Systems*. Association for Computing Machinery, pp. 59–68.

Bilenko, M., Kamath, B. & Mooney, R. J. (2006). Adaptive blocking: Learning to scale up record linkage. In *Sixth International Conference on Data Mining (ICDM'06)*. IEEE, pp. 87–96.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin: Springer.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56**, 463–74.

DeGroot, M. H. & Goel, P. K. (1980). Estimation of the correlation coefficient from a broken random sample. *Ann. Statist.* **8**, 264–78.

D'Orazio, M., Di Zio, M. & Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Chichester: Wiley.

Dunson, D. B. & Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Am. Statist. Assoc.* **104**, 1042–51.

Fellegi, I. P. & Sunter, A. B. (1969). A theory for record linkage. *J. Am. Statist. Assoc.* **64**, 1183–210.

Fienberg, S. E., Rinaldo, A. & Zhou, Y. (2009). Maximum likelihood estimation in latent class models for contingency table data. In *Algebraic and Geometric Methods in Statistics*, P. Gibilisco, E. Riccomagno, M. P. Rogantin & H. P. Wynn, eds., ch. 2. Cambridge: Cambridge University Press, pp. 27–63.

Griffin, R. A. (2014). Potential uses of administrative records for triple system modeling for estimation of census coverage error in 2020. *J. Offic. Statist.* **30**, 177–89.

Hastie, T. J., Tibshirani, R. J. & Friedman, J. H. (2009). Unsupervised learning. In *The Elements of Statistical Learning*. New York: Springer, pp. 485–585.

Holzmann, H., Munk, A. & Gneiting, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scand. J. Statist.* **33**, 753–63.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *J. Am. Statist. Assoc.* **84**, 414–20.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statist. Med.* **14**, 491–8.

Johndrow, J. E., Bhattacharya, A. & Dunson, D. B. (2017). Tensor decompositions and sparse log-linear models. *Ann. Statist.* **45**, 1–38.

Lahiri, P. & Larsen, M. D. (2005). Regression analysis with linked data. *J. Am. Statist. Assoc.* **100**, 222–30.

Lo, Y., Mendell, N. R. & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika* **88**, 767–78.

Lum, K., Price, M. E. & Banks, D. (2013). Applications of multiple systems estimation in human rights research. *Am. Statistician* **67**, 191–200.

Michelson, M. & Knoblock, C. A. (2006). Learning blocking schemes for record linkage. In *Proceedings of the National Conference on Artificial Intelligence*, vol. 21. Association for the Advancement of Artificial Intelligence, pp. 440–5.

Miller, J., Betancourt, B., Zaidi, A., Wallach, H. & Steorts, R. C. (2015). Microclustering: When the cluster sizes grow sublinearly with the size of the data set. *arXiv:* 1512.00792.

R Development Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with Discussion). *J. R. Statist. Soc.* B **59**, 731–92.

Sadinle, M. (2014). Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *Ann. Appl. Statist.* **8**, 2404–34.

Sadinle, M. & Fienberg, S. E. (2013). A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems. *J. Am. Statist. Assoc.* **108**, 385–97.

Steorts, R., Hall, R. & Fienberg, S. E. (2014). SMERED: A Bayesian approach to graphical record linkage and de-duplication. In *Artificial Intelligence and Statistics*. pp. 922–30.

Steorts, R. C. (2015). Entity resolution with empirically motivated priors. *Bayesian Anal.* **10**, 849–75.

Steorts, R. C., Hall, R. & Fienberg, S. E. (2015). A Bayesian approach to graphical record linkage and de-duplication. *J. Am. Statist. Assoc.* **111**, 1660–72.

Teicher, H. (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32**, 244–8.

Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* **34**, 1265–9.

TIBSHIRANI, R. J., WALTHER, G. & HASTIE, T. J. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Statist. Soc.* B **63**, 411–23.

WINKLER, W. E. (2006). Overview of record linkage and current research directions. *Research Report Series* (*Statistics #2006-2*). Washington, DC: U.S. Bureau of the Census.

WOLTER, K. M. (1986). Some coverage error models for census data. *J. Am. Statist. Assoc.* **81**, 337–46.

YAKOWITZ, S. J. & SPRAGINS, J. D. (1968). On the identifiability of finite mixtures. *Ann. Math. Statist.* **39**, 209–14.

ZANELLA, G., BETANCOURT, B., WALLACH, H., MILLER, J., ZAIDI, A. & STEORTS, R. C. (2016). Flexible models for microclustering with application to entity resolution. *arXiv:* 1610.09780.

ZASLAVSKY, A. M. & WOLFGANG, G. S. (1993). Triple-system modeling of census, post-enumeration survey, and administrative-list data. *J. Bus. Econ. Statist.* **11**, 279–88.

[*Received on* 23 *March* 2017. *Editorial decision on* 4 *December* 2017]