STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 123–143, DOI 10.21307/stattrans-2020-034
Received – 31.01.2020; accepted – 30.06.2020

High dimensional, robust, unsupervised record linkage

Sabyasachi Bera[1], **Snigdhansu Chatterjee**[2]

ABSTRACT

We develop a technique for record linkage on high dimensional data, where the two datasets may not have any common variable, and there may be no training set available. Our methodology is based on sparse, high dimensional principal components. Since large and high dimensional datasets are often prone to outliers and aberrant observations, we propose a technique for estimating robust, high dimensional principal components. We present theoretical results validating the robust, high dimensional principal component estimation steps, and justifying their use for record linkage. Some numeric results and remarks are also presented.

Key words: record linkage, principal components, high dimensional, robust.

1. Introduction

In recent times, owing to rapid advancement of a variety of technological resources and services, and increasingly digitally connected environment, numerous kinds of datasets are available. For example, for a given community of individual's, there may be very high dimensional data available on each individual's (i) online shopping patterns, as well as on their (ii) social media presence and usage. It may of interest to businesses to understand their customers better based on their social and cultural backgrounds, consequently it is of interest to link an online shopper's profile with their social media data. Owing to privacy rights of individuals and confidentiality concerns, identifying information may not be available to the statistician linking the records.

This paper is primarily on a methodology for linking high dimensional datasets of above type. Many existing approaches for entity resolution and record linkage are applicable only on low dimensional datasets, and where the datasets have shared features or variables. We do not require the two datasets to have a common set of features and in fact, present our discussion for the case where the datasets have no common variable.

In this context, we also develop a mathematical framework for the topic of record linkage, for better understanding and tractability of the theoretical properties of such linking algorithms. Parts of the existing literature on record linkage and entity resolution are based on *ad hoc* principles, and we hope to address some foundational challenges in this topic.

[1]University of Minnesota. USA. E-mail: berax008@umn.edu. ORCID: https://orcid.org/0000-0002-9053-4094

[2]University of Minnesota. USA. E-mail: chatt019@umn.edu. ORCID: https://orcid.org/0000-0002-7986-0470.

Additionally, it is routinely observed that high dimensional datasets can contain outliers or aberrant observations. Consequently a major aspect of our proposed methodology is to develop *robust* techniques for data linkage. Our proposal borrows recent studies on high-dimensional principal components and extends them to the case of robust principal components.

Another aspect of this paper is that we propose a computationally much simpler and easily implementable method for linking records than the available Bayesian approaches such as the ones given in LISEO and TANCREDI (2013). Other machine learning approaches involving graphs and networks that are sometimes adopted for entity resolution, also require heavy machinery computing. It is not clear if such extremely computation intensive methodology is either necessary, or whether there is a principled statistical foundation to such methodology. Apropos of this, the computational burden from our proposal is significantly lower.

This paper advocates a principled approach. Our approach is broadly as follows: we implement a robust, sparse, high-dimensional principal component analysis (PCA often hereafter) on both datasets, and consolidate the information about each observation (that is, each row of both matrices) into a low-dimensional vector (p_0 in the notations of this paper). Then, we compute correlations between these p_0 dimensional vectors from the two matrices, with the understanding that an existing linkage will show up as a highly correlated entry. The threshold for the correlation is based on the training set in the current paper, but our principle is workable even when there is no training set available. Owing to the facts that (a) our proposal requires no common features or variables common to both datasets, and (b) we do not require a training dataset, we call our proposal *unsupervised* record linkage.

In order to ensure clarity of our presentation and to keep the technicalities at a reasonable level, in this paper we only present results on unsupervised record linkage where all the variables are continuous in nature. In particular, commonly used variables for record linkage, like name, date of birth, address, do not satisfy our technical assumptions. In practice, we may use a traditional method using the nominal and ordinal variables to do a preliminary subsetting of potential linkages, after which the unsupervised record linkage method may be used on the continuous variables. Also, it is possible in some cases to use continuous variables as underlying latent variables governing the behavior of a categorical random variable. These directions of research will be part of our future work.

The rest of this paper is organized as follows: In Section 2 we present a brief and necessarily incomplete state of the literature on record linkage, in order to clarify how our contribution differs from the advancements on this topic thus far. Then, in Section 3 we discuss notations, the conceptual framework of the datasets that we propose to link, and the linking model. Following that, in Section 4 we present our statistical model and technical arguments. In particular, Section 4.1 contains the record linking algorithm, and Section 4.2 contains the theoretical framework and justifications for our statistical model and algorithmic steps. Then, in Section 5 we present some numeric results based on simulation studies, along with additional comments on practical implementation of our proposed methodology. A final Section 6 collects our concluding remarks.

2. A Broad Overview of Record Linkage

Record linkage, also often referred to entity resolution, de-duplication or co-reference is a widely used technique for identifying records referring to the same entity across different databases. Although this is a trivial task when unique error-free identifiers of the entities are recorded in the data files, in general it need to be solved in the absence of unique identifiers using other information that the sources have in common on the entities.

The seminal article by FELLEGI and SUNTER (1969) presented the first mathematical model for this topic, based on earlier work by NEWCOMBE and KENNEDY (1962) for one-to-one entity resolution across two databases in terms of Neyman-Pearson hypothesis testing.

In this brief review, we focus primarily on *bipartite record linkage*, where the key assumption is that each entity is recorded *at-most* once in each files. Most of the literature (including our set-up) on record linkage falls in this scenario. This assumption implies a maximum one-to-one restriction in the linkage, that is, a record from one file can be linked with maximum one record from the other file.

The main principle of bipartite record linkage may be described as follows: Consider two data files $Y_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$ that record information from two overlapping sets of individuals or entities. These data files contain n_ℓ records respectively (without loss of generality we assume $n_1 \leq n_2$) for $\ell = 1,2$ with n_0 being the number of entities simultaneously recorded in both files, hence $0 \leq n_0 \leq n_1$.

In the bipartite record linkage context, we can think of the records from files Y_1 and Y_2 as two disjoint sets of nodes, where an edge between two records represents them referring to the same entity, which we also call being co-referent or being a match. Formally, this match can be encoded into a matrix $\triangle_{n_1 \times n_2}$ as follows:

$$\triangle_{ij} = \begin{cases} 1 & \text{if records } i \in Y_1 \text{ and } j \in Y_2 \text{ represent the same entity} \\ 0 & \text{otherwise} \end{cases}$$

The characteristics of a bipartite matching imply that at-most one entry in each column and each row of $\triangle$ can be equal to 1. The goal of bipartite record linkage is to estimate $\triangle$ using the information contained in Y_1 and Y_2.

The set of ordered record pairs $Y_1 \times Y_2$ can be thought as the union of the set of matches $M = \{(i,j) : i \in Y_1,\ j \in Y_2,\ \triangle_{ij} = 1\}$ and the set of non-matches $U = \{(i,j) : i \in Y_1,\ j \in Y_2,\ \triangle_{ij} = 0\}$. Thus, the problem of estimating $\triangle$ from Y_1 and Y_2 can be seen as identifying the sets M and U. When record pairs are estimated to be matches they are called links and when estimated to be non-matches they are called non-links.

2.1. The Fellegi–Sunter Approach of Record Linkage

The key idea of the Fellegi-Sunter approach is as follows: Comparison vectors γ_{ij} are obtained for each record pair (i,j) in $Y_1 \times Y_2$ with the goal of finding evidence of whether they represent matches or not. These vectors can be written as $\gamma_{ij} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^f, \ldots, \gamma_{ij}^F)$, where F denotes the number of criteria used to compare the records. Traditionally, these

F criteria correspond to one comparison for each variable that the data files have in common.

Let $S_f(i,j)$ denote a similarity measure computed from field f of records i and j. The range of S_f can be divided into L_f+1 intervals $I_{f0}, I_{f1}, \ldots, I_{fL_f}$, which represent different disagreement levels. In this construction, the interval I_{f0} represents the highest level of agreement, which includes total agreement, and the last interval I_{fL_f} represents the highest level of disagreement.

In their paper, FELLEGI and SUNTER (1969), the authors propose to the log-likelihood ratios

$$w_{ij} = \log \frac{\mathbb{P}[\gamma_{ij}|\triangle_{ij}=1]}{\mathbb{P}[\gamma_{ij}|\triangle_{ij}=0]}$$

as weights to estimate which record pairs are matches. The expression for w_{ij} assumes that γ_{ij} is a realization of a random vector, say, G_{ij} whose distribution depends on the matching status $\triangle_{ij}$ of the record pair. Similar to the Neyman-Pearson hypothesis testing, if this ratio is large we favor the hypothesis of the pair being a match.

When $\mathbb{P}[\gamma_{ij}|\triangle_{ij}=1]$ and $\mathbb{P}[\gamma_{ij}|\triangle_{ij}=0]$ are known, the procedure orders the possible values of γ_{ij} by their weights w_{ij} in non-increasing order, indexing by the subscript h, and determines two values, $h^{'}$ and $h^{''}$, such that $\sum_{h\leq h'-1}\mathbb{P}[\gamma_{ij}|\triangle_{ij}=0] < \mu \leq \sum_{h\leq h'}\mathbb{P}[\gamma_{ij}|\triangle_{ij}=0]$ and $\sum_{h\geq h''}\mathbb{P}[\gamma_{ij}|\triangle_{ij}=1] \geq \lambda > \sum_{h\geq h''+1}\mathbb{P}[\gamma_{ij}|\triangle_{ij}=1]$, where $\mu = \mathbb{P}[\text{assign } (i,j) \text{ as link}|\triangle_{ij}=0]$ and $\lambda = \mathbb{P}[\text{assign } (i,j) \text{ as non-link}|\triangle_{ij}=1)$ are two admissible "type 1" and "type 2" error levels.

Finally, the record pairs are classified into 3 groups:

1. Those with $h \leq h^{'}-1$ are declared links

2. Those with $h \geq h^{''}+1$ are non-links and

3. Those with configurations between $h^{'}$ and $h^{''}$ require clerical review.

Fellegi and Sunter showed that this decision rule is optimal in the sense that for fixed values of μ and λ it minimizes the probability of sending a pair to clerical review.

However, in practice, $\mathbb{P}[\gamma_{ij}|\triangle_{ij}=1]$ and $\mathbb{P}[\gamma_{ij}|\triangle_{ij}=0]$ are not known, and have to be estimated from Y_1 and Y_2. So, JARO (1989); LARSEN and RUBIN (2001) proposed to model the comparison data using mixture models of the type

$$\begin{aligned} G_{ij}|\triangle_{ij}=1 &\overset{iid}{\sim} M(m) \\ G_{ij}|\triangle_{ij}=0 &\overset{iid}{\sim} U(u), \\ \triangle_{ij} &\overset{iid}{\sim} \text{Bernoulli}(\theta) \end{aligned}$$

for comparison variables G_{ij}, some distributions $M(m)$ and $U(u)$, and $\theta \in (0,1)$. Estimation of M and U is usually done by EM-type algorithms.

2.2. Machine Learning/Classification Approach

In-general, record linkage task becomes quickly infeasible with size (n_ℓ) as well as the dimension (p_ℓ) of data files. A common solution to this problem is to partition the data files into blocks (e.g. geography, or gender and year of birth) of records determined by information that is thought to be accurately recorded in both data files, and then solve the task only within blocks. See CHRISTEN (2011); STEORTS et al. (2014) for extensive surveys. Earlier development into blocking is presented in HERZOG et al. (2007), who also discuss the use of blocking to identify duplicate list entires and for matching records between two sample surveys.

Recently a common approach of tackling the record linkage problems has been to treat it as a traditional supervised or semi-supervised classification problem: we need to classify record pairs into matches and non-matches. If we have a sample of record pairs for which the true matching statuses are known, we can train a classifier on this sample using comparisons between the pairs of records as our predictors, and then predict the matching statuses of the remaining record pairs. See MARTINS (2011); TORVIK and SMALHEISER (2009); TREERATPITUK and GILES (2009); VENTURA et al. (2015) for some examples.

2.3. Bayesian Methods

Bayesian methods have a long history of use in record linkage models. A major advantage of Bayesian methods is their natural handling of uncertainty quantification for the resulting estimates. For a review of recent development in Bayesian methods, see LISEO and TANCREDI (2013). While some of the Bayesian work incorporates the record data only through pairwise similarity scores (SADINLE, 2017; SADINLE and FIENBERG, 2013), other works (STEORTS et al., 2016) directly model the actual record data which usually requires crafting specific models for each type of field, and therefore mostly deal with categorical information. However, recently STEORTS et al. (2015) has generalized Bayesian methods to incorporate string variables such as addresses, phone numbers, or dates.

In addition, SADINLE and FIENBERG (2013) has extended the Fellegi-Sunter approach to linking records across more than two databases. Also, SINGLA and DOMINGOS (2006); ENAMORADO et al. (2018) generalized the underlying mixture models (specially the i.i.d. assumptions) in the Fellegi-Sunter approach.

3. Notations, the data and linking model

The main focus of this paper is to *link* observations from two datasets. Both datasets are matrices, with iid rows. However, the same observational units may have been used for both datasets. For example, the 17-th row of the first dataset and 47-th row of the second dataset may belong to the same individual. For a number of cases (n in the notation used below) we know the linkage, and thus can match and pair the information from both datasets. However, such linkage is not known for many other rows of both datasets. The main goal of a ***record linkage*** exercise is to establish such linkages.

It is also commonly understood that most observations from both datasets are linked, and only a handful of observations from either dataset do not have an entry on the other. For this paper, we assume that the datasets do not have any common variables. We also assume that both datasets considered here are high-dimensional.

3.1. Notations

Since the data, various parameters and latent variables will have multiple indices, we establish some notations first. The notation $a_\square$ denotes a vector, of dimension that will be determined by the context. All vectors are column vectors, and the notation $a_\square^T$ or a^T denotes the transpose, and $|a|$ denotes its Euclidean norm. The $n \times m$ matrix A has column vectors denoted by $A_{\square,1}, \ldots, A_{\square,m} \in \mathbb{R}^n$, and row vectors denoted by $A_{1,\square}, \ldots, A_{n,\square} \in \mathbb{R}^m$, thus

$$A = (A_{\square,1} : A_{\square,2} : \ldots : A_{\square,m})_{n\times m} = \begin{pmatrix} A_{1,\square} \\ A_{2,\square} \\ . \\ . \\ . \\ A_{n,\square} \end{pmatrix}_{n\times m}.$$

We index the datasets used in this paper by $\ell = 1,2$, and the notation Y_ℓ stands for the ℓ-th dataset with dimensions $n_\ell \times p_\ell$, consisting of n_ℓ independent observations of the p_ℓ features that are stacked as row vectors of the Y_ℓ matrix. The k-dimensional multivariate Normal distribution with mean $\mu \in \mathbb{R}^p$ and variance $\Sigma \in \mathbb{R}^{p\times p}$ will be denoted by $N_p(\mu, \Sigma)$. The notation $X_i \overset{i.i.d.}{=} \mathbb{F}$ denotes that the X_i's are independent, identically distributed according to $\mathbb{F}$.

3.2. The data and the linking framework

We consider two datasets for linkage, $Y_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$ for $\ell = 1,2$. Without loss of generality, $n_1 \le n_2$. In both datasets, each row represents an observation, and each column a feature. We assume, for mathematical simplicity, that there is no duplication of features in the two datasets. Within each matrix, the rows are independent. However, a pair of rows, one from each matrix, may have dependency.

For any positive integer k, let $\mathbb{N}_k = \{1,2,\ldots,k\}$, the set of positive integers or natural numbers up to and including k. For any finite set $\mathscr{S}$ (for example, $\mathbb{N}_k$), let $\sigma(\mathscr{S})$ be any permutation of the elements of $\mathscr{S}$.

Suppose that $n_0 \le n_1 (\le n_2)$ is the unknown number of *linked* observations between the datasets. Define $\tilde{n}_\ell = n_\ell - n_0$ for $\ell = 1,2$, denoting the number of unmatched observations from either dataset. Define $N = n_0 + \tilde{n}_1 + \tilde{n}_2$, this is the total number of observational units we consider, and we label the observational units with the index set $\mathbb{N}_N$. The i-th observation of the first dataset, $Y_{1,i,\square} \in \mathbb{R}^{p_1}$, corresponds to the unit whose index matches with the i-th element of $\sigma\big(\{1,\ldots,n_0+\tilde{n}_1\}\big)$, a random permutation of the index subset $\mathscr{S}_1 = \{1,\ldots,n_0+\tilde{n}_1\}$. The index subset of the second data is $\mathscr{S}_2 =$

$\{1,\ldots,n_0,n_0+\tilde{n}_1+1,\ldots,N\}$. The i-th observation of the second dataset, $Y_{2,i,\square}\in\mathbb{R}^{p_2}$, corresponds to the unit whose index matches with the i-th element of $\sigma\left(\mathscr{S}_2\right\}\big)$, a random permutation of the index subset $\mathscr{S}_2$. Note that $\mathscr{S}_1$ and $\mathscr{S}_2$ have exactly n_0 elements in common, reflecting the n_0 matched observational units for the two datasets.

In a few cases, the linkage between some observations units is known, and forms the training set. A training set is not needed for the present paper, but if indeed we have known and established linkages between, say $n\leq n_0$, observational units, without loss of generality we stack these known linkage cases as the first n rows of Y_ℓ, $\ell=1,2$.

We assume that the observations satisfy

$$Y_{\ell,i,\square}\overset{i.i.d.}{=}N_{p_\ell}\Big(0,\Sigma_\ell\Big),i=1,\ldots,n_\ell,\ \ \Sigma_\ell \text{ unknown},\ \ell=1,2.$$

We use the spectral representation

$$\begin{aligned}
&\Sigma_\ell=\Gamma_\ell\Lambda_\ell\Gamma_\ell^T,\ \text{where}\\
&\Lambda_\ell=diag\Big(\lambda_{\ell,1},\ldots,\lambda_{\ell,p_\ell}\Big),\ \text{with}\\
&\quad\lambda_{\ell,1}\geq\lambda_{\ell,2}\geq\ldots\geq\lambda_{\ell,p_0}\gg\lambda_{\ell,p_0+1}\geq\ldots\lambda_{\ell,p_\ell},\\
&\Gamma_\ell=\Big[\gamma_{\ell,\square,1}:\ldots:\gamma_{\ell,\square,p_\ell}\Big]\in\mathbb{R}^{p_\ell\times p_\ell}.
\end{aligned}$$

Thus, Λ_ℓ is the diagonal matrix of the eigenvalues of Σ_ℓ, and the columns of Γ_ℓ contain the corresponding eigenvectors. It is assumed that the first p_0 eigenvalues are considerably higher than the rest, and contain information relevant for linking records. Also for mathematical simplicity, we assume henceforth that $\lambda_{1,j}=\lambda_{2,j}$ for $j=1,\ldots,p_0$. That is, the top p_0 eigenvalues are the same. This is not a necessary assumption, but makes the presentation and technicalities of the developments presented below considerably simpler. We assume that the top p_0 eigenvectors of both $\Sigma_\ell,\ell=1,2$ are sparse, in that all but κ_ℓ of the entries in these eigenvectors are zero. This is a necessary assumption to obtain statistically consistent and computationally obtainable estimators of the principal components that we use in this paper, see WANG et al. (2016) for further details.

There are multiple record linkage contexts in which the above framework may be useful. First, traditional linkage techniques that rely on nominal and ordinal variables like names, addresses and so on often result in plausible subsets of observations from one dataset linked to each unit of the other dataset. At that stage, a further analysis based on the continuous variables as described here may be useful. Second, due to confidentiality and privacy considerations, datasets are often anonymized. In such cases, the model presented above may be extremely useful, either directly for modeling the reported continuous variables, or in conjunction with other continuous but latent variables. Third, our framework allows the scope of record linkage to extend beyond the traditional applications of linking sample surveys involving individuals or households, into linking data from multitude of sources, like social media, online shopping platforms, and electronic records of various kinds (LI et al., 2020; FATEMI et al., 2018). In many such contexts, the observed and often suitably anonymized data may be modeled us-

ing high-dimensional continuous (observed or latent) variables. With such an extended scope, record linkage may provide an increase in precision and accuracy of recommender systems (DRACHSLER et al., 2010; SHABTAI et al., 2013; SLOKOM, 2018), for providing online security and privacy (ZHU et al., 2016; SALAS, 2019), for transfer learning (RONG et al., 2012) and for distributed computing and related technical developments.

4. The statistical model

Without loss of generality and to considerably simplify the presentation below, we assume that the first n_0 rows of $Y_\ell, \ell = 1,2$ are linked. To relate the two datasets $Y_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$, $\ell = 1,2$, we define the following quantities:

$$\begin{pmatrix} Z_{1,i,j} \\ Z_{2,i,j} \end{pmatrix} \overset{i.i.d.}{=} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \text{ when } i = 1,\ldots,n_0 \text{ and } j = 1,\ldots,p_0,$$
$$Z_{\ell,i,j} \overset{i.i.d.}{=} N(0,1), \text{ for } \ell = 1,2, \text{ when } i > n_0 \text{ or } j > p_0.$$

Thus, $Z_{\ell,i,j}$ are all standard normal random variables, and for the case $i = 1,\ldots,n_0$ and $j = 1,\ldots,p_0$, the two random variables $Z_{1,i,j}$ and $Z_{2,i,j}$ share a correlation ρ between them. We arrange the $Z_{\ell,i,j}$ into two matrices of dimensions identical to those of our datasets Y_ℓ. Thus, $Z_{\ell,i,j}$ is the (i,j)-th element of the matrix $Z_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$, $\ell = 1,2$. It can be seen that each matrix Z_ℓ has iid $N(0,1)$ entries, but the top left corners of Z_1 and Z_2 are related.

We model the data as

$$Y_{\ell,i,\square} = \Gamma_\ell \Lambda_\ell^{1/2} Z_{\ell,i,\square}, \text{ where}$$
$$Z_{\ell,i,\square} \overset{i.i.d.}{=} N_{p_\ell}\left(0, \mathbb{I}_{p_\ell}\right), i = 1,\ldots,n_\ell$$

described above. In matrix terms, this then translates to

$$Y_\ell = Z_\ell \Lambda_\ell^{1/2} \Gamma_\ell^T, \text{ where}$$
$$Z_{\ell,i,j} \overset{i.i.d.}{=} N(0,1).$$

Note, however, that we do not imply with the above that the matrices $Z_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$ are independent of each other, and indeed they are not.

4.1. The record linking algorithm

Our proposed algorithm is as follows:

We now discuss in details the steps outlined above. First, the scaling $X_{\ell,i,\square} = Y_{\ell,i,\square}/|Y_{\ell,i,\square}|$ ensures that each $X_{\ell,i,\square}$ has unit norm, thus ensuring that outliers do not affect the PCA and subsequent computations. It is well known that PCA is very sensitive to outliers. The second step of the above algorithm is about computation of the high dimensional principal components using an established procedure. The third

Algorithm 1 Record Linking Algorithm

1. Scale each row of the two datasets by their respective norms, to get $X_{\ell,i,\square} = Y_{\ell,i,\square}/|Y_{\ell,i,\square}|$, for $\ell = 1,2$ and $i = 1,\ldots,n_l$. Collect these in the matrices $X_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$, $\ell = 1,2$.

2. Run the high dimensional sparse PCA algorithm due to WANG et al. (2016) on $X_\ell, \ell = 1,2$. This obtains the leading eigenvalue and eigenvector for these matrices. Project the data on the orthogonal space to this estimated eigenvector, and repeat the process to obtain the leading p_0 eigenvectors.

3. Obtain the coefficients $W_{\ell,i,\square} \in \mathbb{R}^{p_0}$ (given in (4.1) below) for each $i = 1,\ldots,n_\ell$, $\ell = 1,2$ from the projections of the observations on the top p_0 eigenvectors.

4. Obtain the correlations $C(i,\tilde{i})$ of $W_{1,i,\square}$ and $W_{2,\tilde{i},\square}$.

5. Arrange the correlations in descending order. Based on the values corresponding to the training set and a pre-set value for the maximum proportion of false positive matches, select a correlation threshold. If there are multiple matches above this threshold for any $i \in \{1,\ldots,n_1\}$ or $\tilde{i} \in \{1,\ldots,n_2\}$, the match with the higher correlation value is chosen. The proportion of false negatives is estimated from the number of training sample matches below the threshold.

through last steps of the algorithm are about using the principal component scores to obtain the record linkage. There can be considerable variation in the details in these steps depending on the context, we have presented one simple procedure.

The above algorithm used the training data only for the last step of setting a threshold for the correlations. We can easily formulate a variation, where a training set is not needed. Since we compute $n_1 n_2$ correlations of which only $\min(n_1,n_2)$ can possibly correspond to linked data, a threshold can be determined based a change-point in the correlation values, or on multiple matches. We illustrate this aspect in simulations reported later in this paper.

4.2. Theoretical properties

The justification for using the above algorithm rests on the fact that there is a clear separation of the correlation values between the linked data-pairs, as opposed to the correlation between the not-linked cases. We establish this fact in the following result:

Theorem 4.1. *Under the conditions of the model, the population correlation value for each linked pair of observations is ρ, and is zero for two observations that are not linked.*

Thus, there is a clear separation of the correlation values from the linked data-pairs from the rest. There would be sample variations, and the value of ρ is not known. Consequently, either a training set-based threshold or a change detection technique can be used to sort the true linkages from the rest.

Proof of Theorem 4.1: Let

$$\tilde{\Gamma}_\ell = \left[\gamma_{\ell,\square,1} : \ldots : \gamma_{\ell,\square,p_0}\right] \in \mathbb{R}^{p_\ell \times p_0},$$

the first p_0 columns of Γ_ℓ, for which the eigenvalues are considerably higher than the rest. We project the datapoints $Y_{\ell,i,\square}$ on the column space of $\tilde{\Gamma}_\ell$, for $i = 1, \ldots, n_\ell$. Note that all columns of $\tilde{\Gamma}_\ell$ are orthonormal (by construction, since these are estimators of successive eigenvectors, so they are orthogonal to each other and have unit norm). Given this, it is easy to see that the projection of $Y_{\ell,i,\square}$ is

$$\begin{aligned}\tilde{Y}_{\ell,i,\square} &= \sum_{j=1}^{p_0} \left\langle Y_{\ell,i,\square}, \gamma_{\ell,\square,j} \right\rangle \gamma_{\ell,\square,j} \\ &= \tilde{\Gamma}_\ell \tilde{\Gamma}_\ell^T Y_{\ell,i,\square} \\ &= \tilde{\Gamma}_\ell W_{\ell,i,\square}.\end{aligned}$$

The relevant information about the projections is carried in the low-dimensional ***weights*** $W_{\ell,i,\square} \in \mathbb{R}^{p_0}$, consequently we develop our analysis based on these below. Putting these weights as rows in a matrix, we have

$$\begin{aligned}W_\ell &= Y_\ell \tilde{\Gamma}_\ell \\ &= Z_\ell \Lambda_\ell^{1/2} \Gamma_\ell^T \tilde{\Gamma}_\ell \in \mathbb{R}^{n_\ell \times p_0}\end{aligned} \tag{4.1}$$

This last expression can be simplified further, since

$$\begin{aligned}\Gamma_\ell^T \tilde{\Gamma}_\ell &= \begin{pmatrix} \gamma_{\ell,\square,1}^T \\ \gamma_{\ell,\square,2}^T \\ . \\ . \\ . \\ \gamma_{\ell,\square,p_\ell}^T \end{pmatrix} \left[\gamma_{\ell,\square,1} : \ldots : \gamma_{\ell,\square,p_0}\right] \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ . & & & \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ . & & & \\ . & & & \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \mathbb{I}_{p_0} \\ \mathbf{0}_{(p_\ell - p_0) \times p_0} \end{pmatrix}.\end{aligned}$$

Thus we have

$$\begin{aligned}
&\Lambda_\ell^{1/2}\Gamma_\ell^T\tilde{\Gamma}_\ell \\
&\qquad = \begin{pmatrix}
\lambda_{\ell,1}^{1/2} & 0 & 0 & 0 \\
0 & \lambda_{\ell,2}^{1/2} & 0 & 0 \\
. & & & \\
0 & 0 & 0 & \lambda_{\ell,p_0}^{1/2} \\
0 & 0 & 0 & 0 \\
. & & & \\
. & & & \\
0 & 0 & 0 & 0
\end{pmatrix} \\
&\qquad = \begin{pmatrix} \tilde{\Lambda}_\ell^{1/2} \\ \mathbf{0}_{(p_\ell - p_0)\times p_0} \end{pmatrix} \in \mathbb{R}^{p_\ell \times p_0}.
\end{aligned}$$

Define

$$\tilde{Z}_\ell = \left[Z_{\ell,\square,1} : Z_{\ell,\square,2} : \ldots : Z_{\ell,\square,p_0} \right] \in \mathbb{R}^{n_\ell \times p_0}.$$

Consequently, we have

$$\begin{aligned}
W_\ell &= Y_\ell \tilde{\Gamma}_\ell \\
&= Z_\ell \Lambda_\ell^{1/2} \Gamma_\ell^T \tilde{\Gamma}_\ell \\
&= \left[\lambda_{\ell,1}^{1/2} Z_{\ell,\square,1} : \lambda_{\ell,2}^{1/2} Z_{\ell,\square,2} : \ldots : \lambda_{\ell,p_0}^{1/2} Z_{\ell,\square,p_0} \right] \\
&= \tilde{Z}_\ell \tilde{\Lambda}_\ell^{1/2} \in \mathbb{R}^{n_\ell \times p_0}.
\end{aligned}$$

We thus have, for any $i \in \{1, \ldots, n_0\}$

$$\begin{aligned}
\mathbb{E}W_{\ell,i,\square} &= 0, \\
\mathbb{V}W_{\ell,i,\square} &= \tilde{\Lambda}_\ell, \\
\mathbb{E}W_{1,i,\square}W_{2,i,\square}^T &= \tilde{\Lambda}_1 \mathbb{E}\tilde{Z}_{1,i,\square}\tilde{Z}_{2,i,\square}^T \tilde{\Lambda}_2, \\
\mathbb{E}W_{1,i,\square}^T W_{2,i,\square} &= \mathbb{E}\tilde{Z}_{1,i,\square}^T \tilde{\Lambda}_1 \tilde{\Lambda}_2 \tilde{Z}_{2,i,\square}^T = \sum_{j=1}^{p_0} \lambda_{1,j}^{1/2} \lambda_{1,j}^{1/2} \mathbb{E}\tilde{Z}_{1,i,j}\tilde{Z}_{2,i,j}.
\end{aligned}$$

Then it follows that

$$Cor\left(W_{1,i,\square}, W_{2,i,\square}\right) = \rho.$$

The algebra for the observations that are not linked is similar and omitted here. □

One important consideration for our framework is to ensure that the robustness procedure we implemented in the first step does not alter the eigenvector structure of the original data. That is, we need to ensure that the eigenvectors of Σ_ℓ match those of

the variance of X_ℓ, $\ell = 1,2$. This is ensured in the following result:

Theorem 4.2. *Suppose $X \in \mathbb{R}^p$ is a random vector with variance Σ_X, and let the variance of $U = X/|X|$ be denoted by Σ_U.*

(i) When X has an elliptically symmetric distribution and zero mean, the eigenvectors of Σ_X and Σ_U are identical.

(ii) If $\mathbb{E}U = 0 \in \mathbb{R}^p$, $\mathbb{E}U|X| = 0 \in \mathbb{R}^p$ and $\mathbb{E}|X|^2UU^T = \mathbb{E}|X|^2\mathbb{E}UU^T \in \mathbb{R}^{p\times p}$, then again the eigenvectors of Σ_X and Σ_U are identical. Moreover, if the eigenvalues of Σ_X are $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$, then the eigenvectors of Σ_U are $\frac{\lambda_1}{\sum_{i=1}^p \lambda_i} \geq \ldots \frac{\lambda_p}{\sum_{i=1}^p \lambda_i} > 0$.

Note that a p-dimensional random vector X is said to elliptically distributed if there exist a vector $\mu \in \mathbb{R}^p$, a positive semi-definite matrix $\Sigma \in \mathbb{R}^{p\times p}$ and a function $\phi : (0,\infty) \to \mathbb{R}$ such that the characteristic function of X is $\exp\{it^T\mu\}\phi(t^T\Sigma t)$ for $t \in \mathbb{R}^p$. See FANG et al. (1990) for several alternative and equivalent definitions of the elliptically contoured family, as well as for additional details. An example of elliptically contoured distribution is the multivariate Gaussian distribution, thus the framework adopted in this paper satisfies the first condition of Theorem 4.2. The second part of Theorem 4.2 is for general interest, in case an elliptic distributional assumption is not satisfied.

Proof of Theorem 4.2: First, consider the case where X has an elliptically symmetric distribution with mean zero. In such cases, we may write $X = R\Gamma\Lambda^{1/2}E$, where Γ is a rotation matrix, Λ is a diagonal matrix with positive elements, E is uniformly distributed on the unit sphere and R is a positive random variable that is independent of E. Then, we have $|X|^2 = R^2E^T\Lambda E$. Let $\tilde{E} = \frac{\Lambda^{1/2}E}{|\Lambda^{1/2}E|}$. Consequently, $U = X/|X| = \Gamma\tilde{E}$ is a function of E alone. Note that in the circularly symmetric case where $\Lambda = \lambda\mathbb{I}$, we now have $|X|$ independent of U, and the above conditions are trivially satisfied. For general Λ, note that

$$\mathbb{E}UU^T = \Gamma\mathbb{E}\tilde{E}\tilde{E}^T\Gamma^T,$$

and it can be easily shown that $\mathbb{E}\tilde{E}\tilde{E}^T$ is a diagonal matrix. Thus, $\Sigma_X = \Gamma\Lambda\Gamma^T$ and Σ_U have the same eigenvectors in this case. This proof is reminiscent of the arguments used in TASKINEN et al. (2012).

Under the assumptions of the second part, that is, $\mathbb{E}U = 0 \in \mathbb{R}^p$, $\mathbb{E}U|X| = 0 \in \mathbb{R}^p$ and $\mathbb{E}|X|^2UU^T = \mathbb{E}|X|^2\mathbb{E}UU^T \in \mathbb{R}^{p\times p}$, U and $|X|$ are uncorrelated, as is $|X|^2$ and UU^T. This immediately implies that $\mathbb{E}X = \mathbb{E}U|X| = 0$, thus we have $\Sigma_X = \mathbb{E}XX^T$ and $\Sigma_U = UU^T$. We also easily have

$$\begin{aligned}\Sigma_X &= \mathbb{E}XX^T \\ &= \mathbb{E}|X|^2UU^T \\ &= \mathbb{E}|X|^2\mathbb{E}UU^T.\end{aligned}$$

Thus, the eigenvectors of Σ_X and Σ_U are identical, since $\mathbb{E}|X|^2$ is a scalar.

Now, note that

$$\Sigma_U = \frac{\Sigma_X}{\mathbb{E}|X|^2} = \frac{\Sigma_X}{\mathbb{E}[\text{Trace } XX^T]} = \frac{\Sigma_X}{\text{Trace } \Sigma_X} = \frac{\Sigma_X}{\sum_{i=1}^p \lambda_i}.$$

□

The rest of this section is on the estimation of the high dimensional principal components. We present the results only for the first principal component. Our development here closely follows that of WANG et al. (2016), and we essentially make use of their theoretical machinery and algorithm for the rest of this paper. The results below are primarily designed to show that the technical conditions of WANG et al. (2016) hold for our case, and the eigenvector estimation algorithm they established also works for us. We omit many algebraic details, since they are similar to those of WANG et al. (2016).

Our first result is to show that U has a sub-Gaussian distribution. This is immediate, since U is bounded. We have multiple proofs of this result with sharp bounds on the constant σ^2, but present the simplest one here for clarity. Ensuring that U has a sub-Gaussian distribution facilitates the use of various known concentration inequality and other probabilistic results.

Lemma 4.1. $U \in$ *Sub-Gaussian*(2).

Proof of Lemma 4.1: We recall the definition of Sub-Gaussian distributions

$$X \in \text{Sub-Gaussian}(\sigma^2) \text{ if } \forall u \in \mathbb{R}^p,\ \mathbb{E}[e^{u^T X}] \leq e^{\frac{\sigma^2|u|^2}{2}}$$

Since $|U| = 1$, we have for $|u| \geq 1$

$$\begin{aligned} u^T U \leq |u|^2 (\text{C-S inequality}) \text{ for any } u \text{ on } \mathbb{R}^p \\ \Rightarrow e^{u^T U} \leq e^{|u|^2} \text{ for any } u \text{ on } \mathbb{R}^p \\ \Rightarrow U \in \text{Sub-Gaussian}(2). \end{aligned}$$

The case for $|u| < 1$ is more delicate, but can be handled with some routine algebra. We omit the details here. □

We now recall some definitions from WANG et al. (2016) for developing our next set of results. Since our framework is high dimensional, we need structural assumptions on the nature of the eigenvectors of Σ_U (or Σ_X), and the most common and convenient assumption here is one of sparsity. We define the sparse unit ball in p-dimensions having at most k non-zero entries as follows:

$$B_0(k) = \left\{ x \in \mathbb{R}^p : |x| = 1, \sum_{j=1}^p \mathscr{I}_{\{x_i \neq 0\} \leq k} \right\}.$$

Based on this and sample size n, for any $j \in \{1, \ldots, p\}$ and $C > 0$, a probability measure P is said the satisfy the *Restricted Covariance Condition* (RCC) with parameters p, n, j

and C, and written as $P \in RCC_p(n, j, C)$ if

$$P\{ \sup_{u \in B_0(l)} |\hat{V}(u) - V(u)| \geq C \max(\sqrt{\frac{j \log(p/\delta)}{n}}, \frac{j \log(p/\delta)}{n})\} \leq \delta$$

for all $\delta > 0$, where $V(u) = \mathbb{E} u^T \Sigma u$, $\hat{V}(u) = \mathbb{E} u^T \hat{\Sigma} u$ and $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n U_i U_i^T$ for $U_1, U_2, \ldots, U_n \overset{iid}{\sim} P$. We also define

$$RCC_p(C) = \bigcap_{l=1}^{p} \bigcap_{n=1}^{\infty} RCC_p(n, l, C).$$

Suppose, associated with a generic distribution $\mathbb{P}$ on $\mathbb{R}^p$, is the variance matrix Σ with the j-th eigenvalue and eigenvector respectively being λ_j and $\gamma_{\square, j}$, $j = 1, \ldots, p$. The results for the rest of this section are valid for the following class of probability measures. For $\theta > 0$, define

$$\mathscr{P}_p(n, k, \theta) = \Big\{ \mathbb{P} \in RCC_p(n, 2, 1) \cap RCC_p(n, 2k, 1) : \gamma_{\square, 1} \in B_0(k), \lambda_1 - \lambda_2 \geq \theta \Big\}.$$

Our next result states that U, after suitable scaling, has a distribution that satisfies the restricted covariance condition with appropriate selection of constants.

Lemma 4.2. *For the random variable $U \in \mathbb{R}^p$ with $p \geq 2$, assume that $\gamma_{\square, 1} \in B_0(k)$ and $\theta = \big(\lambda_1(\Sigma_X) - \lambda_2(\Sigma_X)\big) > 0$. Then $Z = \frac{U}{22} \in \mathscr{P}_p(n, k, \frac{\theta}{22\, Trace(\Sigma_X)})$.*

Proof of Lemma 4.2: Recall from Proposition 1 of WANG et al. (2016) that for every $\sigma > 0$,

$$\text{Sub-Gaussian}(\sigma^2) \subseteq RCC_p\big(16\sigma^2(1 + \frac{9}{\log(p)})\big).$$

Therefore using Lemma 4.1, we have that

$$\begin{aligned}
& U = \frac{X}{|X|} \in RCC_p\big(32\,(1 + \frac{9}{\log(p)})\big), \\
\Rightarrow & \frac{U}{\sqrt{32\,(1 + \frac{9}{\log(p)})}} \in RCC_p(1), \\
\Rightarrow & \frac{U}{22} \in RCC_p(1), \\
\Rightarrow & \frac{U}{22} \in \mathscr{P}_p(n, k, \frac{\theta}{22 \text{Trace}(\Sigma_X)}).
\end{aligned}$$

□

For a symmetric matrix $A \in \mathbb{R}^{p \times p}$, let us define

$$\hat{\gamma}^k_{\square, max}(A) = \text{sargmax}_{u \in B_0(k)} u^T A u$$

to be the k-sparse maximum eigenvector of A, where *sargmax* denotes the smallest element of the argmax in the lexicographic ordering. We use A as an argument of

$\hat{\gamma}^k_{\square,max}(\cdot)$ here to distinguish between the estimated eigenvectors from various matrices. Also, between 2 unit vectors u and v , we define the loss function

$$L(u,v) = \left(1-(u^T v)^2\right)^{1/2}.$$

Note that $\hat{\gamma}^k_{\square,max}(A)$ are all identical for $A = \hat{\Sigma}_U, \hat{\Sigma}_X, \hat{\Sigma}_Z$. Our next result is the main consistency and probabilistic guarantee result on the sample version of the k-sparse maximum eigen-vector. This result ensures in particular that under suitable conditions with $\log(p) = o(n)$, the sample k-sparse maximum eigenvector is consistent for the population maximum eigenvector.

Theorem 4.3. *For $2k\log(p) \le n$, the k-sparse empirical maximum eigen-vector, $\hat{\gamma}^k_{\square,max}(\hat{\Sigma}_U)$ satisfies*

$$\mathbb{E}L\left(\hat{\gamma}^k_{\square,max}(\hat{\Sigma}_U), \gamma_{\square,1}(\Sigma_U)\right) \le 44\sqrt{2}\left(1+\frac{1}{\log(p)}\right)\sqrt{\frac{k\log(p)}{n\theta^2}}\ \mathit{Trace}(\Sigma_X).$$

Proof of Theorem 4.3: We apply Theorem 2 of WANG et al. (2016) on Z and note that for $2k\log(p) \le n$, the k-sparse empirical maximum eigen-vector, $\hat{\gamma}^k_{\square,max}(\hat{\Sigma}_Z)$ satisfies

$$\mathbb{E}L\left(\hat{\gamma}^k_{\square,max}(\hat{\Sigma}_Z), \gamma_{\square,1}(\Sigma_Z)\right) \le 2\sqrt{44}\left(1+\frac{1}{\log(p)}\right)\sqrt{\frac{k\log(p)}{n\theta^2}}\ \mathrm{Trace}(\Sigma_X).$$

Proof follows by noting that

$$\mathbb{E}L\left(\hat{\gamma}^k_{\square,max}(\hat{\Sigma}_Z), \gamma_{\square,1}(\Sigma_Z)\right) = \mathbb{E}L\left(\hat{\gamma}^k_{\square,max}(\hat{\Sigma}_U), \gamma_{\square,1}(\Sigma_U)\right).$$

□

Let $U_i = \frac{X_i}{|X_i|}$ for $i = 1,2,\ldots,n$ where $X_i \overset{iid}{\sim} \mathbb{P}$, and we denote $\hat{\gamma}^{SDP}_{\square,1}(U)$ for the output of the SDP algorithm of WANG et al. (2016). with input $U = (U_1,\ldots,U_n)^T$, $\lambda = 4\sqrt{\frac{\log(p)}{n}}$ and $\varepsilon = \frac{\log(p)}{4n}$.

While Theorem 4.3 provided a general probabilistic guarantee on the error of the sample k-sparse maximum eigenvector, we need a similar result for the sparse maximum eigenvector that is obtained using the SDP algorithm. Note that the SDP algorithm allows for computation of a sparse maximum eigenvector in real time, and is thus both practical and is of theoretical relevance. The following result establishes a probabilistic guarantee and consistency for this version of the sparse empirical maximum eigenvector.

Theorem 4.4. *If $4\log(p) \le n \le k^2 p^2 \theta^{-2}\log(p)$, then*

$$\mathbb{E}L\left(\hat{\gamma}^{SDP}_{\square,1}(U), \gamma_{\square,1}(\Sigma_U)\right) \le (352\sqrt{2}+44)\ \mathit{Trace}\Sigma_X\sqrt{\frac{k^2\log(p)}{n\theta^2}}.$$

Proof of Theorem 4.4: We apply Theorem 5 of WANG et al. (2016) on Z and note

that

$$\mathbb{E}L\big(\hat{\gamma}_{\square,1}^{SDP}(Z), \gamma_{\square,1}(\Sigma_Z)\big) \leq 22(16\sqrt{2}+2)\sqrt{\frac{k^2\log(p)}{n\theta^2}}\,\mathrm{Trace}(\Sigma_X).$$

for $Z = (Z_1, \ldots, Z_n)^T$ where $Z_i = \frac{U_i}{22}$. The result is immediate.

□

In general, $\hat{\gamma}_{\square,1}^{SDP}(U)$ is not a sparse estimator. However, it turns out that a k-sparse version of $\hat{\gamma}_{\square,1}^{SDP}(U)$, that is, some $\hat{\gamma}_{\square,1}^{SDP,k}(U) \in B_0(k)$, may be obtained by setting all but the top k coordinates of $\hat{\gamma}_{\square,1}^{SDP}(U)$ in absolute value to zero and renormalizing the vector. In particular, $\hat{v}_0^{SDP}$ is computable in polynomial time and under the same condition as in Theorem 4.4.

5. Some Simulation Results

In this section, we present a simulation exercise to illustrate the performance of the proposed record linkage methodology, and also to illustrate some practical implementation steps.

To generate data, we followed the framework laid out at the start of Section 4. That is, we generated a set of independent bivariate Gaussian random variables with common correlation ρ, and several independent univariate standard Normal random variables, and used these to populate the two data matrices. We tested various choices of sample sizes, dimensions, correlation ρ. For brevity, we report the case where the two matrix datasets that we use are of dimensions $n_1 = 60, p_1 = 100$, and $n_2 = 70, p_2 = 120$ of independent rows each, and $\rho = 0.8$. The first $n_0 = 50$ entries of these two matrices are linked to each other. The rest (10 for the first matrix, 20 for the second matrix) are not linked. We use the first $n = 20$ observations for training in the version of the algorithm where a training set is used, thus leaving the last 30 linked data points for testing. We also demonstrate the performance of our method when no training dataset is available. We fix $p_0 = 10$ for this exercise, and repeated the entire simulation 100 times.

As practical steps, we found that using a sparse version of the estimated eigenvalues, as proposed in Theorem 4.4, considerably improves performance, owing to reduction of the effect of noise terms in the eventual linkage. Also, the estimated principal components for the two datasets may not have the same orientation and may not appear in the same order. Hence, when needed, a principled permutation and sign reversal of the estimated eigenvectors of the second dataset is done to improve linkage accuracy. While in theory the estimators of the eigenvalues are not required for the linkage steps, using those as weights improves linkage.

Over 100 replications of the simulation experiment, the correct linkage established on the test set by our proposed method was about 43.5% times, with a standard error of about 5.37%. When no training sample is used and instead a threshold for the correlation estimated from the data, the correct linkage percentage increases to about 56%, however, the standard error also increases to about 14.8%. The estimated threshold for correlation was at a lower value than the case with training data: a pattern that we noticed in

multiple simulations, which we will study further later. The linkage accuracy may seem low, however, we need to remember that this is a ***unsupervised*** framework, involving high dimensional datasets with no common variables, and minimal or no training data.

6. Conclusions and Future Work

Record linkage is in-general a difficult exercise. Bayesian models are often too complex for practical purposes and some Bayesian formulations fail to accommodate non-categorical variables. Supervised classification methods assume the existence of large, accurate sets of training data, which are often difficult and/or expensive to obtain. Also, in those methodologies, it is difficult to guarantee maximum one-to-one assignment constrain of bipartite record linkage. While some theoretical modifications have been developed to ensure an one-to-one assignment(SADINLE, 2017), typically some subsequent post-processing step is required to solve these inconsistencies.

In view of these difficulties, in this paper we propose a completely new approach towards record linkage. We do not require a common set of variables between the two datasets, we do not require a training set, and the dimensionality and sample sizes can both be large. Naturally, our methodology extends to cases where a common set of variables exist, we will elaborate on this in a future work. If a training set exists, we can make use of it, as illustrated in this paper. We have presented the case for the bipartite record linkage, but our model conceptually extends to other cases as well.

Some of the technical assumptions of this paper, like the two covariances matrices having the same set of leading eigenvalues, or the number leading eigenvalues p_0 being known, or the latent random variables that link the two datasets having the same correlation ρ, can be addressed with some additional work and methodological developments. The assumption of multivariate normality of the data is not critical: our proposal only depends on robust, high dimensional principal components, and these are available for data from many distributions with both discrete and continuous components. The assumption of sparsity in the leading eigenvectors is owing to the fact that for high dimensional modeling, some structural assumptions are needed since the sample size is not adequate to estimate all relevant unknown parameters. In any case, there are considerable challenges to estimating high dimensional principal components, see PAUL (2007).

The robust, high dimensional principal component we use is built on the work by WANG et al. (2016). The credit for both the theoretical framework and the algorithm goes to that work primarily. Our setup differs from WANG et al. (2016) in the detail that for robustness purposes we transform each observation to be on the unit sphere. One future work for us is to establish the theoretical results under weaker assumptions than WANG et al. (2016), or to show better theoretical properties.

We have used a simple method for linking observations in this paper, using correlations. A correlation-based linkage is not critical to our primary methodological steps. More complex and realistic measures of linkage will be studied in the future. The case where no training data is present needs further investigation, which will also be part of our future work. In absolute terms, our simulation results are not excellent; however,

we do not have a baseline for comparison since most other papers on record linkage do not use as general a framework as ours with (i) high dimensional data, (ii) no common set of features, and (iii) possibly no training set. Our framework may be termed *unsupervised learning of record linkage*, and in the unsupervised learning framework, our numeric results are perhaps acceptable. However, considerable fine tuning and experimentation with the algorithm for record linkage is needed. We have ensured that our high-dimensional, robust and potentially sparse principal component estimator is highly accurate, and some of our studies (not reported here) suggest that using a small number of common features dramatically increases linkage accuracy. A part of our future work is on including nominal and categorical variables for linkages in our framework, which will make our proposed approach more aligned with traditional record linkage techniques. In this context, we will also investigate how much additional gain results from using PCA in addition to available matching fields, compared to the traditional Felligi-Sunter method.

An important topic to consider in future from this paper is on statistical inference based on linked datasets. This is a non-trivial task, since the datasets are used multiple times in the process of linking, estimation of various quantities of interest, and then inference. The article HAN and LAHIRI (2019) provides review of the current state of the art in this direction of work. Some alternatives to fully Bayesian methods, for example regression analysis using linked data LAHIRI and LARSEN (2005); SCHEUREN and WINKLER (1997, 1993), have both computational efficiency and analytical tractability, which may make them attractive practical choices for applications. Comparisons with such alternatives is an additional future work.

An additional future work for us is to extend the methodology proposal here to multiple datasets. We will also work on real data examples, which has not been possible for this paper owing to data access limitations. It will be of interest to compare our unsupervised record linkage approach with more traditional record linking algorithms.

Acknowledgements

This research is partially supported by the US National Science Foundation (NSF) under grants # DMS-1622483, # DMS-1737918, # OAC-1939916 and #DMR-1939956.

References

CHRISTEN, P., (2011). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE transactions on knowledge and data engineering*, 24(9), pp. 1537–1555.

DRACHSLER, H., BOGERS, T., VUORIKARI, R., VERBERT, K., DUVAL, E., MANOUSELIS, N., BEHAM, G., LINDSTAEDT, S., STERN, H., FRIEDRICH, M., et al., (2010). Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. *Procedia Computer Science*, 1(2), pp. 2849–2858.

ENAMORADO, T., FIFIELD, B., and IMAI, K., (2018). Using a probabilistic model to assist merging of large-scale administrative records. *Available at SSRN 3214172*.

FANG, K.-T., KOTZ, S., and NG, K.-W., (1990). *Symmetric Multivariate and Related Distributions*. CRC Press.

FATEMI, B., KAZEMI, S. M., and POOLE, D., (2018). Record linkage to match customer names: A probabilistic approach. *arXiv preprint arXiv:1806.10928*.

FELLEGI, I. P. and SUNTER, A. B., (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), pp. 1183–1210.

HAN, Y. and LAHIRI, P., (2019). Statistical analysis with linked data. *International Statistical Review*, 87, pp. S139–S157.

HERZOG, T. N., SCHEUREN, F. J., and WINKLER, W. E., (2007). *Data quality and record linkage techniques*. Springer Science & Business Media.

JARO, M. A., (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406), pp. 414–420.

LAHIRI, P. and LARSEN, M. D., (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469), pp. 222–230.

LARSEN, M. D. and RUBIN, D. B., (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96(453), pp. 32–41.

LI, J., DOU, Z., ZHU, Y., ZUO, X., and WEN, J.-R., (2020). Deep cross-platform product matching in e-commerce. *Information Retrieval Journal*, 23(2), pp. 136–158.

LISEO, B. and TANCREDI, A., (2013). Some advances on Bayesian record linkage and inference for linked data. *URL http://www. ine. es/e/essnetdi_ws2011/ppts/Liseo_Tancredi. pdf*.

MARTINS, B., (2011). A supervised machine learning approach for duplicate detection over gazetteer records. In *International Conference on GeoSpatial Sematics*, pp. 34–51, Springer.

NEWCOMBE, H. B. and KENNEDY, J. M., (1962). Record linkage: making maximum use of the discriminating power of identifying information. *Communications of the ACM*, 5(11), pp. 563–566.

PAUL, D., (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4), pp. 1617–1642.

RONG, S., NIU, X., XIANG, E. W., WANG, H., YANG, Q., and YU, Y., (2012). A machine learning approach for instance matching based on similarity metrics. In *International Semantic Web Conference*, pp. 460–475, Springer.

SADINLE, M., (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518), pp. 600–612.

SADINLE, M. and FIENBERG, S. E., (2013). A generalized fellegi–sunter framework for multiple record linkage with application to homicide record systems. *Journal of the American Statistical Association*, 108(502), pp. 385–397.

SALAS, J., (2019). Sanitizing and measuring privacy of large sparse datasets for recommender systems. *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12.

SCHEUREN, F. and WINKLER, W. E., (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, pp. 39–58.

SCHEUREN, F. and WINKLER, W. E., (1997). Regression analysis of data files that are computer matched-ii. *Survey Methodology*, 23, pp. 157–165.

SHABTAI, A., ROKACH, L., and ELOVICI, Y., (2013). Occt: A one-class clustering tree for implementing one-to-many data linkage. *IEEE Transactions on Knowledge and Data Engineering*, 26(3), pp. 682–697.

SINGLA, P. and DOMINGOS, P., (2006). Entity resolution with markov logic. In *Sixth International Conference on Data Mining (ICDM'06)*, pp. 572–582, IEEE.

SLOKOM, M., (2018). Comparing recommender systems using synthetic data. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 548–552.

STEORTS, R. C. et al., (2015). Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10(4), pp. 849–875.

STEORTS, R. C., HALL, R., and FIENBERG, S. E., (2016). A bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516), pp. 1660–1672.

STEORTS, R. C., VENTURA, S. L., SADINLE, M., and FIENBERG, S. E., (2014). A comparison of blocking methods for record linkage. In *International Conference on Privacy in Statistical Databases*, pp. 253–268, Springer.

TASKINEN, S., KOCH, I., and OJA, H., (2012). Robustifying principal component analysis with spatial sign vectors. *Statistics & Probability Letters*, 82(4), pp. 765–774.

TORVIK, V. I. and SMALHEISER, N. R., (2009). Author name disambiguation in medline. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(3), pp. 1–29.

TREERATPITUK, P. and GILES, C. L., (2009). Disambiguating authors in academic publications using random forests. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pp. 39–48.

VENTURA, S. L., NUGENT, R., and FUCHS, E. R., (2015). Seeing the non-stars:(some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Research Policy*, 44(9), pp. 1672–1701.

WANG, T., BERTHET, Q., and SAMWORTH, R. J., (2016). Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5), pp. 1896–1930.

ZHU, J., ZHANG, S., SINGH, L., YANG, G. H., and SHERR, M., (2016). Generating risk reduction recommendations to decrease vulnerability of public online profiles. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 411–416, IEEE.