Generalized Bayesian Record Linkage and Regression with Exact Error Propagation

Rebecca C. Steorts[1](✉), Andrea Tancredi[2], and Brunero Liseo[2]

[1] Department of Statistical Science, affiliated faculty, Computer Science, Biostatistics and Bioinformatics, the information initiative at Duke (iiD), and the Social Science Research Institute (SSRI) Duke University; Principal Researcher, Center for Statistical Research Methodology, Duke University and U.S. Census Bureau, Durham, USA
beka@stat.duke.edu

[2] Department of Methods and Models for Economics, Territory and Finance, La Sapienza, Rome, Italy
{andrea.tancredi,brunero.liseo}@uniroma.it

Abstract. Record linkage (de-duplication or entity resolution) is the process of merging noisy databases to remove duplicate entities. While record linkage removes duplicate entities from such databases, *the downstream task* is any inferential, predictive, or post-linkage task on the linked data. One goal of the downstream task is obtaining a larger reference data set, allowing one to perform more accurate statistical analyses. In addition, there is inherent record linkage uncertainty passed to the downstream task. Motivated by the above, we propose a generalized Bayesian record linkage method and consider multiple regression analysis as the downstream task. Records are linked via a random partition model, which allows for a wide class to be considered. In addition, we jointly model the record linkage and downstream task, which allows one to account for the record linkage uncertainty exactly. Moreover, one is able to generate a feedback propagation mechanism of the information from the proposed Bayesian record linkage model into the downstream task. This feedback effect is essential to eliminate potential biases that can jeopardize resulting downstream task. We apply our methodology to multiple linear regression, and illustrate empirically that the "feedback effect" is able to improve the performance of record linkage.

1 Introduction

Record linkage (de-duplication or entity resolution) is the process of merging noisy databases to remove duplicate entities. While record linkage removes duplicate entities from such databases, *the downstream task* is any inferential, predictive or post-linkage task on the linked data. In this paper, we propose a joint model for the record linkage and the downstream task of linear regression. Our proposed model can link records over an arbitrary number of databases

J. Domingo-Ferrer and F. Montes (Eds.): PSD 2018, LNCS 11126, pp. 297–313, 2018.
https://doi.org/10.1007/978-3-319-99771-1_20

(lists or files). We assume there is duplication within each database, known as "duplicate detection." Our record linkage model can be expressed as a random partition model, which leads to a large family of distributions. Next, we jointly model the record linkage task and the downstream task (linear regression), which allows for the exact propagation of the record linkage uncertainty into the downstream task. Crucially, this generates a feedback propagation mechanism from the proposed Bayesian record linkage model into the downstream task of linear regression. This feedback effect is essential to eliminate potential biases that can jeopardize resulting inference in the downstream task. We apply our methodology to multiple linear regression, and illustrate empirically that the "feedback effect" is able to improve performance of record linkage.

1.1 Prior Work

Our work builds off [14,16–18], which all proposed Bayesian record linkage models well suited for categorical data. [18] modeled the fully observed records through the "hit-and-miss" measurement error model [2]. One natural way to handle record linkage uncertainty is via a joint model of the record linkage and downstream task. [10] introduced a record linkage model for continuous data based on a multivariate normal model with measurement error. Turning to just record linkage tasks, [16,17] were the first to perform simultaneous record linkage and de-duplication on multiple files by using the fully observed records, creating a scalable record linkage algorithm. In similar work, de-duplication in a single database framework was tackled from a Bayesian perspective in [14] by using the information provided by the comparison data.

Related work regarding the record linkage and downstream task has been considered under specific assumptions. [9] assumed that the two databases represent a permutation of the same database of units and proposed an estimator (LL) of the regression coefficients which is unbiased, conditionally on the matching probabilities provided by the record linkage task. [7] extended this approach to handle more complex and realistic linkage scenarios and logistic regression problems. Generalizations of the LL estimator have been also provided by [8] using estimating equations. In addition, [4] proposed to consider the probabilities of being a match—provided by the record linkage algorithm—as an ingredient to be used within a multiple imputation scenario. Finally, [5] proposed a Bayesian method that jointly models the record linkage and the association between the overlapping features in two different databases. The authors consider somewhat simpler situation where the number of records to match in the two databases is relatively small and relies upon a specific blocking criteria. In addition, one potential limitation of the approach is the assumption of specific matching pattern. For each single block of comparisons, all cases in the smaller database will certainly appear in the other databases. We refer to [6] for details.

Section 2 introduces our Bayesian record linkage model, providing extensions to priors on random partitions. Section 3 generalizes our record linkage methodology to the downstream task of linear regression. Section 4 provides experiments for the record linkage task on synthetic data. We then provide three experiments on the joint record linkage and downstream task of linear regression on synthetic data. Section 5 provides a discussion and extensions to future work.

2 Bayesian Record Linkage and Priors on Partitions

In this section, we introduce notation used through the paper, our Bayesian record linkage model, and an alternative and more intuitive construction for the prior on co-referent records, known as the linkage structure λ.

2.1 Notation

Assume L databases (lists, data sets, or files) $F_1, F_2 \dots, F_L$ that consist of either qualitative and/or categorical records, which are noisy due to the data collection process. Each record corresponds to an underlying latent entity (statistical unit) of partially overlapping samples (or populations). In addition, assume all databases have p overlapping features (fields). Assume L sets of records are collected from a given population of size N_{pop} where $1 \leq N_{\text{pop}} \leq \infty$ in the same framework as [15, 17]. As such, assign a label j' ($j' = 1, \dots, N_{\text{pop}}$) to each member of the population. Next, let $\tilde{v}_{j'} = (\tilde{v}_{j'1}, \dots, \tilde{v}_{j'p})$ be the vector of the p categorical overlapping features for the population individual j'. Finally, denote the entire set of population records by $\tilde{v} = (\tilde{v}_1, \dots, \tilde{v}_{N_{pop}})$.

2.2 Bayesian Record Linkage Model

Assume the set of population records $\tilde{v}$ is generated independently, for $j' = 1, \dots, N_{\text{pop}}$, from a vector of independent categorical variables $\tilde{V} = (\tilde{V}_1, \dots, \tilde{V}_\ell, \dots, \tilde{V}_p)$ such that $\tilde{V}_l \in \{v_{\ell\,1}, \dots, v_{\ell\,M_\ell}\}$ and

$$P\left(\tilde{V}_\ell = v_{\ell\,s}\right) = \theta_{\ell\,v_{\ell\,s}} \quad s = 1, \dots, M_\ell, \tag{1}$$

where M_ℓ is the number of categorical values for the ℓth feature. At the sample level, assume that one does not observe the "true" population values due to measurement errors. Thus, the observed records, which is a database of size N_i, $i = 1, \dots, L$, consists of distorted versions of subsets of the vectors $\tilde{v}_{j'}$. Let $v_{ij} = (v_{ij1}, \dots, v_{ijp})$ denote the observed values for the j-th record of the i-th database, where $i = 1, \dots, L$ and $j = 1, \dots, N_i$. Denote the observed records (across the L databases) by $v = (v_{11}, \dots, v_{1N_1}, \dots, v_{L1}, \dots v_{LN_L})$. Next, let the set of latent indicator variables $\lambda_{ij} \in \{1, \dots, N_{\text{pop}}\}$ denote the unknown co-reference (matching) pattern between the observed records v and the population records $\tilde{v}$, where $\lambda_{ij} = j'$ indicates that the population record j' generated the observed record v_{ij}.[1] In general, let $\lambda = (\lambda_{11}, \dots, \lambda_{1N_1}, \dots, \lambda_{L1}, \dots, \lambda_{LN_L})$ denote the linkage structure.

[1] The relation $\lambda_{ij_1} = \lambda_{ij_2}$, with $j_1 \neq j_2$, implies that records j_1 and j_2 of the i-th database represents are co-referent to the same population record. This is an instance of duplicate-detection within the same database. When $\lambda_{i_1j_1} = \lambda_{i_2j_2}$, with $i_1 \neq i_2$, one has the usual record linkage framework with the same individual appearing in two different databases.

Next, we formalize the distortion mechanism when the population records are observed in the L databases using the *hit-and-miss model* [2]. Let $V_{ij\ell}$ be the random variable that generates observed record v_{ijl}. Assume that $V_{ijl} \in \{v_{l\,1}, \ldots, v_{l\,M_\ell}\}$, that is, $V_{ij\ell}$ has the same support of $\tilde{V}_\ell$. Let $\delta_{a,b} = 1$ if $a = b$ and $\delta_{a,b} = 0$ if $a \neq b$, which implies that

$$P(V_{ij\ell} = v_{\ell\,s} \mid \lambda_{ij}, \tilde{v}, \alpha_\ell) = (1 - \alpha_\ell)\delta_{\tilde{v}_{\lambda_{ij}\ell}, v_{\ell\,s}} + \alpha_\ell \theta_{\ell\,v_{\ell\,s}} \quad s = 1, \ldots, M_\ell \qquad (2)$$

for $i = 1, \ldots, L$; $j = 1, \ldots, N_i$; $\ell = 1, \ldots, p$, where $\alpha_\ell \in [0, 1]$ represents the distortion probability for the ℓ-th overlapping feature. Here, the true population value is observed with probability $1 - \alpha_\ell$, and a different value is drawn from the random variable $\tilde{V}_\ell$ generating the population values with probability α_ℓ. Finally, assuming the conditional independence among all the overlapping features given their respective unobserved population counterparts, one obtains

$$p(v \mid \tilde{v}, \lambda, \alpha) = \prod_{i=1}^{L} \prod_{j=1}^{N_i} \prod_{\ell=1}^{p} P(v_{ijl} \mid \tilde{v}, \lambda, \alpha) = \prod_{i=1}^{L} \prod_{j=1}^{N_i} \prod_{\ell=1}^{p} [(1 - \alpha_\ell)\delta_{\tilde{v}_{\lambda_{ij}\ell}, v_{ij\ell}} + \alpha_\ell \theta_{\ell\,v_{ij\ell}}]. \quad (3)$$

We assume that the distortion probabilities are exchangeable, that is

$$\alpha_\ell \overset{iid}{\sim} \text{Beta}(f, g),\ \ell = 1, \ldots, p,$$

and we assume the probabilities $\theta_{\ell\,1} \ldots \theta_{\ell\,M_\ell}$ are considered known and equal to the corresponding population frequencies. The model summarized by Eqs. (1) and (3) can be viewed as a latent variable model where the unobserved population records $\tilde{v}$ generate the observed records v and $\alpha = (\alpha_1 \ldots, \alpha_p)$ can be viewed as the unknown model distortion parameter.

Remark: A convenient property of the hit-miss model is that one can integrate out the unknown population values $\tilde{v}$ to directly obtain the distribution $p(v|\alpha, \lambda)$. The resulting marginal distribution $p(v|\alpha, \lambda)$ is the product of within-cluster distributions. To improve mixing, we use a Metropolis within Gibbs algorithm to simulate from the joint posterior $p(\lambda, \alpha|v)$ (See Appendix A).

2.3 The Prior Distribution for $\boldsymbol{\lambda}$

In this section, we propose a more intuitive and subjective construction of a prior distribution on λ. Let z denote the random partition of the observed records determined by λ and let $\mathcal{P}$ denote the set containing all the possible partitions of the N observed records. The distribution on the sample labels λ induces a distribution on $\mathcal{P}$. Furthermore, matches and duplicates are completely specified given the knowledge of the random partition z, which is invariant with respect to the labelings of the partition blocks. Given this construction, one can directly focus on the partition distribution of the observed records without linking the labels distribution to a sample design and to a population size N_{pop}, see for example, [14]. One can effectively consider the distribution of λ as a prior distribution for the latent linkage structure and concentrate only on its probabilistic properties.

Both the interpretations of the role of λ (either as a consequence of the sampling design or a model represented by partitions) may provide useful insights for a correct choice of its prior distribution. One difficult and related question in the record linkage literature has been the subjective specification on the space of partitions. A simple, alternative prior for the number of distinct entities $k(z)$ can be obtained looking at the following allocation rule for the record labels which is based on a generalization of the Chinese Restaurant Process, namely the Pitman-Yor process (PYP) ([3,13]). (See Appendix B for details).

3 The Downstream Task of Linear Regression

In this section, we propose record linkage methodology for the downstream task of linear regression. Consider the model $\tilde{Y} = \sum_{l=1}^{p} \tilde{X}_l \beta_l + \epsilon$ for the population units, where the goal is to estimate the regression coefficients $\beta = (\beta_1, \dots, \beta_p)^t$. We observe Y and $X = (X_1, \dots, X_p)$, where X represents a noisy measurement of the true covariates $\tilde{X} = (\tilde{X}_1, \dots \tilde{X}_p)$ and Y is a random copy of the corresponding population variable $\tilde{Y}$.

To better illustrate our approach, we consider two scenarios. In the first scenario—*the complete regression scenario*—each database reports a set of overlapping features, the response variable, and the covariates. Let y_{ij} and $x_{ij} = (x_{ij1} \dots, x_{ijp})$ denote the observed values for the j-th unit of the i-th database, where $i = 1, \dots, L$ and $j = 1, \dots, N_i$. In addition, let (y, x) denote the entire set of regression data observed across the L databases. In the *complete scenario*, there is not a bias problem concerning the estimation of the β coefficients. In the second scenario—*the broken regression scenario*—we assume that the overlapping features are observed in each database, the response variable is observed in only the first database, and specific subsets of covariates are observed in the other databases. In this situation, let (y, x) denote the observation y_{1j}, where $j = 1, \dots, N_1$ and x_{ij}, where $i = 2, \dots, L$ and $j = 1, \dots, N_i$. Note that x_{ij} represents only a fixed subset of the values $x_{ij1} \dots x_{ijp}$ for $j = 1, \dots, N_i$. Here, there is a bias issue regarding estimating the β coefficients.[2]

3.1 Simple Linear Regression

In this section, we consider linear regression and the two scenarios mentioned above with a single covariate X. First, consider the *complete regression scenario.* Let $\tilde{X}_{j'}$ be the true value of observation X corresponding to the records of cluster $C_{j'}$. Now consider a cluster $C_{j'} = \{(i, j)\}$ with one record. Given the true value of $\tilde{X}_{j'} = \tilde{x}_{j'}$ and membership to cluster $C_{j'}$, we assume that the response variable Y_{ij} follows a standard normal regression model with covariate $\tilde{x}_{j'}$, where the observed value for the covariate X_{ij} is normal with mean $\tilde{x}_{j'}$ and Y_{ij} and X_{ij} are independent. That is,

[2] In both scenarios, we assume that the covariates have zero mean and the regression model does not have an intercept.

$$\begin{bmatrix} Y_{ij} \\ X_{ij} \end{bmatrix} \mid \tilde{X}_{j'} = \tilde{x}_{j'} \sim N_2 \left[\begin{pmatrix} \beta & 0 \\ 0 & 1 \end{pmatrix} \begin{bmatrix} \tilde{x}_{j'} \\ \tilde{x}_{j'} \end{bmatrix}, \begin{pmatrix} \sigma^2_{y|\tilde{x}} & 0 \\ 0 & \sigma^2_{x|\tilde{x}} \end{pmatrix} \right]. \tag{4}$$

We assume that $\tilde{X}_{j'} \sim N(0, \sigma^2_{\tilde{x}})$, which allows one to integrate $X_{j'}$ via Eq. 4. In fact, setting $Z_{ij} = (Y_{ij}, X_{ij})'$, one can easily show that conditionally on the event $\{(i,j) \in C_{j'}\}$, it follows that

$$Z_{ij} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2_{\tilde{x}} \begin{pmatrix} \beta^2 & \beta \\ \beta & 1 \end{pmatrix} + \begin{pmatrix} \sigma^2_{y|\tilde{x}} & 0 \\ 0 & \sigma^2_{x|\tilde{x}} \end{pmatrix} \right]. \tag{5}$$

For ease of notation, let I_n denote the $n \times n$ identity matrix, 0_n denote the n-vector of zero; 1_n denote a vector of all 1's, and $J_n = 1_n 1'_n$. Next, set

$$B = \begin{pmatrix} \beta^2 & \beta \\ \beta & 1 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma^2_{y|\tilde{x}} & 0 \\ 0 & \sigma^2_{x|\tilde{x}} \end{pmatrix}.$$

Consider a cluster $C_{j'} = \{(i_1, j_1), (i_2, j_2)\}$ with two records. The two pairs $Z_{i_1 j_1}$ and $Z_{i_2 j_2}$ are random vectors, both depending on the same "true" value $\tilde{X}_{j'}$. Let $\otimes$ be the Kronecker product. Conditionally on $\tilde{X}_{j'} = \tilde{x}_{j'}$ and on the cluster membership, we replicate the model for a cluster with one record by assuming that $Z_{i_1 j_1}$ and $Z_{i_2 j_2}$ are two independent bivariate normal random variables with joint distribution

$$N_4 \left[\left(I_2 \otimes \begin{pmatrix} \beta & 0 \\ 0 & 1 \end{pmatrix} \right) (1_4 \tilde{x}_{j'}), I_2 \otimes \Sigma \right]. \tag{6}$$

Then the marginal distribution of $(Z_{i_1 j_1}, Z_{i_2 j_2})'$ is

$$\begin{pmatrix} Z_{i_1 j_1} \\ Z_{i_2 j_2} \end{pmatrix} \sim N_4 \big(0_4, I_2 \otimes \Sigma + \sigma^2_{\tilde{x}} J_2 \otimes B \big).$$

This argument can be extended to any cluster size. When $\text{card}(C_{j'}) = n$, the marginal distribution of $Z = (Z_{i_1 j_1}, \ldots, Z_{i_n j_n})$ is again multivariate normal: $Z \sim N_{2n} \left(0_{2n}, I_n \otimes \Sigma + \sigma^2_{\tilde{x}} J_n \otimes B \right)$.

Next, consider the *broken regression scenario*. In this case, when some information is missing—either the covariate in the first database or the response variable in some of the other databases—one can easily marginalize over the missing variables by using standard properties of multivariate normal distribution. Let $(y,x)_{C'_j} = ((y_{ij}, x_{ij}) : \lambda_{ij} = j')$ denote the set of regression observations, which conditionally on λ, correspond to the j'-th population unit. For example, for a cluster $C_{j'} = \{(1,j)\}$ with one record in the first database, we denote this as $(y,x)_{C'_j} = y_{1j}$. Using the marginal density of Y_{ij} in Eq. 5, we can write the likelihood, conditional on λ, as $p((y,x)_{C'_j} | \lambda, \beta, \sigma^2_{y|\tilde{x}}, \sigma^2_{x|\tilde{x}})$. Similarily, suppose $C_{j'} = \{(i,j)\}$ with $i > 1$, then $(y,x)_{C'_j} = x_{ij}$ and the likelihood is given by marginal density of X_{ij}. Next, consider a cluster $C_{j'} = \{(1, j_1), (i_2, j_2)\}$ with a record in the first database and the other record in a different database, i.e. $i_2 > 1$. It follows that $(y, x)_{C'_j} = (y_{1j_1}, x_{i_2j_2})$ and the corresponding likelihood is found by marginalizing over the missing values X_{1j_1}, Y_{2j_2} in Eq. 6, where we obtain the joint density in Eq. 5. Finally, it follows that the likelihood function (as a function of $\lambda, \beta, \sigma^2_{y|\tilde{x}}, \sigma^2_{x|\tilde{x}}$) for both the complete and broken regression scenarios can be generally written as $p(y, x|\lambda, \beta, \sigma^2_{x|\tilde{x}}, \sigma^2_{y|\tilde{x}}) = \prod_{j'=1}^{N_{pop}} p((y, x)_{C'_j}|\beta, \sigma^2_{x|\tilde{x}}, \sigma^2_{y|\tilde{x}})$.[3]

In order to handle the record linkage and downstream regression task simultaneously, we assume conditional independence on λ between the overlapping features in the record linkage model and the set of variables in the downstream task of linear regression. Assuming conditional independence, we find

$$\begin{aligned} p(\lambda, \beta, \alpha, \sigma^2_{y|\tilde{x}}, \sigma^2_{x|\tilde{x}}|v, x, y) \propto\, & p(v|\lambda, \alpha)p(y, x|\lambda, \beta, \sigma^2_{y|\tilde{x}}, \sigma^2_{x|\tilde{x}}) \\ & \times p(\lambda)p(\alpha)p(\beta, \sigma^2_{y|\tilde{x}}, \sigma^2_{x|\tilde{x}}). \end{aligned} \tag{7}$$

The first factor is related to the record linkage process, and second factor is related to the downstream task of linear regression, and the other factors represent the prior distributions. We assume independent diffuse priors for $\beta, \sigma^2_{y|\tilde{x}}, \sigma^2_{x|\tilde{x}}$. To update the appropriate regression parameters $\beta, \sigma^2_{y|\tilde{x}}, \sigma^2_{x|\tilde{x}}$, we use the Metropolis-Hastings algorithm in Appendix A. Using the factorization of the posterior in Eq. (7), the proposed method can be generalized to any statistical model.

3.2 Multiple Linear Regression

We extend the downstream task to that of multiple regression, first considering the complete regression scenario. Let $C_{j'}$ denote a cluster of size n, $Y_{C_{j'}}$ denote a vector with n observations of the response variable in this cluster, and $X_{C_{j'}}$ denote the $n \times p$ matrix with the values of the p covariates observed in the cluster units. Let $[YX]_{C_{j'}}$ denote the vector of $n(p+1)$ elements with the n rows of the matrix $(Y_{C_{j'}}, X_{C_{j'}})$ vertically stacked and let $\tilde{X}_{j'}$ denote the vector containing the true values of the p covariates. Equation 4 can be generalized assuming that

$$[YX]_{C_{j'}} \mid \tilde{X}_{j'} \sim N_{n(p+1)} \left[\left(I_{n\times n} \otimes \begin{pmatrix} \beta^t & 0^t_p \\ 0_{p\times p} & I_{p\times p} \end{pmatrix} \right) \left(1_{2n} \otimes \tilde{X} \right), I_{n\times n} \otimes \begin{pmatrix} \sigma^2_{y|\tilde{x}} & 0 \\ 0 & \Sigma_{x|\tilde{x}} \end{pmatrix} \right],$$

where

$$1_{2n} \otimes \tilde{X} \sim N_{2np} \left(0_{2np}, (1_n 1^t_n) \otimes \begin{pmatrix} \Sigma_{\tilde{x}} & \Sigma_{\tilde{x}} \\ \Sigma_{\tilde{x}} & \Sigma_{\tilde{x}} \end{pmatrix} \right).$$

This way the marginal distribution of $[YX]_{C_{j'}}$ is $n(p+1)$-variate normal with zero mean and covariance matrix

$$\left(I_{n\times n} \otimes \begin{pmatrix} \beta^t & 0^t_p \\ 0_{p\times p} & I_{p\times p} \end{pmatrix} \right) \left((1_n 1^t_n) \otimes \begin{pmatrix} \Sigma_{\tilde{x}} & \Sigma_{\tilde{x}} \\ \Sigma_{\tilde{x}} & \Sigma_{\tilde{x}} \end{pmatrix} \right) \left(I_{n\times n} \otimes \begin{pmatrix} \beta^t & 0^t_p \\ 0_{p\times p} & I_{p\times p} \end{pmatrix} \right)^t + \left(I_{n\times n} \otimes \begin{pmatrix} \sigma^2_{y|\tilde{x}} & 0 \\ 0 & \Sigma_{x|\tilde{x}} \end{pmatrix} \right),$$

[3] We assume that population units j' that do not have an observed cluster size contribute to the likelihood with a factor equal to 1.

which simplifies into

$$(1_n 1_n^t) \otimes \begin{pmatrix} \beta^t \Sigma_{\tilde{x}} \beta & \beta^t \Sigma_{\tilde{x}} \\ \Sigma_{\tilde{x}} \beta^t & \Sigma_{\tilde{x}} \end{pmatrix} + I_{n \times n} \otimes \begin{pmatrix} \sigma^2_{y|\tilde{x}} & 0 \\ 0 & \Sigma_{x|\tilde{x}} \end{pmatrix}.$$

The likelihood provided by the multiple regression model is the product of the factors $p([YX]_{C_{j'}} = [y,x]_{C'_j} | \beta, \sigma^2 y|\tilde{x}, \Sigma_{x|\tilde{x}})$ for the observed clusters. The same considerations from linear regression regarding modeling the prior and the computational aspects apply to multiple linear regression. Note the major difference is in the marginalization pattern in the broken regression scenario. In fact, for a cluster joining records across more than one database, we may need to integrate out the covariate values missing in the databases that share a cluster.

4 Experiments

To investigate the performance of our proposed methodology we consider the `RLdata500` data set from the `RecordLinkage` package in `R`. This synthetic data set consists of 500 records, each comprising first and last name and full date of birth. We modify this data set to consider two databases, where each database contains 250 records, respectively, with duplicates in and across the two databases. To consider the case without duplicate detection, we modify the original `RLdata500` such that it has no duplicate records within each of the two databases. Without duplicate detection is a special case of our general methodology (see Appendix C). We provide experiments for both record linkage and the downstream task.

4.1 Record Linkage with and Without Duplicate-Detection

We provide two record linkage experiments—one with duplicate detection and one without duplicate detection. In Figs. 1 and 2, we report the prior and the posterior for $k(z)$ and the performance of the record linkage procedure measured in terms of the posteriors of the false negative rates (FNR) and the false discovery rates (FDR). (For a review of FNR and FDR, see [1,15]).

Figure 1 (with duplicate detection) illustrates that the resulting posteriors of $k(z)$ appears robust to the choices of θ and σ (first row). We observe similar behavior for the posteriors of FNR and FDR (second and third rows). Figure 2 illustrates that as we vary the PYP parameters, the posterior of T is weakly dependent on their values. The two database framework without duplicate detection leads, a posteriori, to similar FNR (second row) and lower FDR (third row) compared to the previous case. (See Appendix D for the PYP parameter settings).

4.2 Regression Experiments

We consider three regression experiments on the `RLdata500` data set. In Experiment I, we consider the complete regression scenario in a single database framework with duplicate detection. In Experiment II, we consider the broken regression scenario with record linkage and duplicate detection. In Experiment III, we

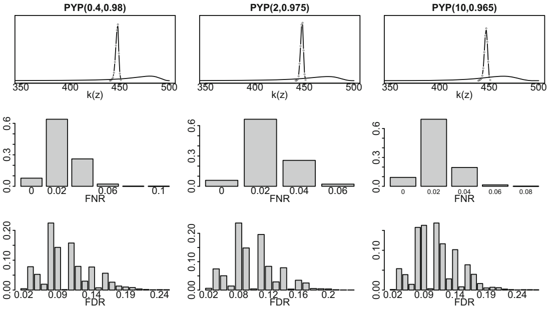

Fig. 1. Prior and posteriors for $k(z)$ (first row), FNR posteriors (second row), FDR posteriors (third row) for the `RLdata500` data set.

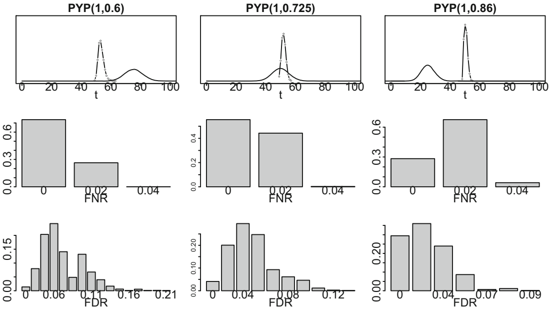

Fig. 2. Prior and posteriors for t (first row), FNR posteriors (second row), FDR posteriors (third row) for the `RLdata500` data set assuming a two database record linkage framework without duplicate-detection.

consider the broken multiple regression scenario in a two database framework without duplicate detection. (See Appendix E for details).

Figure 3 gives the results of Experiment I. The posteriors of $(\beta,\ \sigma_{y|\tilde{x}},\ \sigma_{x|\tilde{x}})$ from our joint modeling approach (first row, solid lines) do not show remarkable differences when compared to their true counterpart (first row, dotted lines), which were obtained by fitting the regression model conditional on the true value of λ. The similarity between the posteriors is mainly due to the large concentration of λ around the true pattern of duplications. The mode of the posterior of the number of distinct entities is exactly the true value (450), where the FNR and FDR are considerably smaller with respect to case without the y and x columns. Hence, the effect of considering the information provided by the regression model has improved the record linkage process.

Figure 4 gives the results of Experiment II. The posteriors (first row, solid lines) of $(\beta, \sigma_{y|\tilde{x}}, \sigma_{x|\tilde{x}})$ are similar to the corresponding true posteriors (first row, dashed lines). We report the posteriors obtained by fixing λ equal to the point estimate provided by the hit-and-miss model applied to the categorical variables alone (first row, dotted lines). The posteriors of β and $\sigma^2_{y|\tilde{x}}$ obtained with the

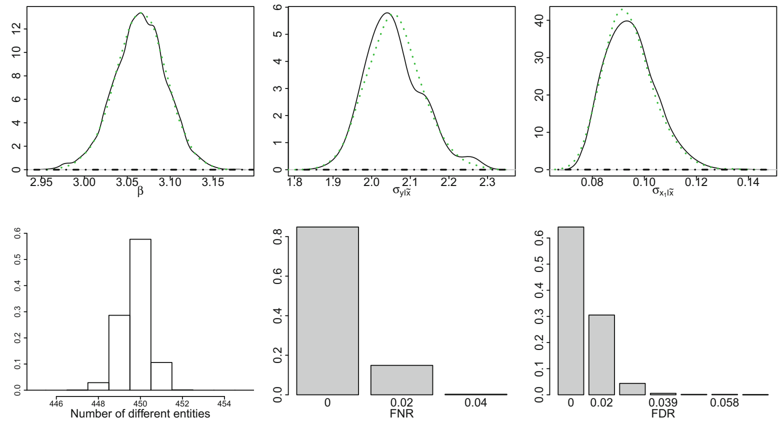

Fig. 3. Experiment I. Upper panels: prior (dotdash lines) and posterior of $\beta, \sigma_{y|\tilde{x}}, \sigma_{x|\tilde{x}}$ with the joint record linkage and regression model (solid lines) and the true linkage structure (dotted lines). Lower panels: posterior for $k(z)$, FNR and FDR.

plug-in approach are strongly biased for the presence of false matches which, on the other hand, are not affecting the posterior of $\sigma_{x|\tilde{x}}$. This distribution depends on the 13 duplicated entities with two copies of x which are correctly accounted for in the plug-in approach. To better illustrate the causes of the distortion in the estimation of the regression parameters, the right panel on the top row shows all the (x, y) pairs resulting from the plug-in approach. The solid black circles represent the true matches, and the empty red circles represent the false matches, with independent y and x values. We report the corresponding regression lines, where the three false matches are lowering the β estimate and increasing the $\sigma_{y|\tilde{x}}$ estimate. Further analysis reveals that the posterior for $k(z)$ (second row) with the integrated hit-miss and regression model is less concentrated with respect to the first experiment but it is more concentrated with respect to the single hit-miss model. We reduce the FDR, leaving the FNR almost unchanged. We coin this the *feedback effect* of the regression from the downstream task. For example, if we consider a false link, the posterior probability of being a match will typically be down-weighted by the low likelihood arising from the regression part of the model. Hence, in addition to centering the estimates of the regression coefficient β, the joint regression-hit miss model improves record linkage performance.

Figure 5 gives the results of the Experiment III. The joint model gives posteriors similar to the true ones while the plug-in approach gives biased estimates and larger variability (first row, left upper panels). The presence of false matches in the plug-in approach gives a positive bias in estimating the variance $\sigma_{y|\tilde{x}}$ and affects the posterior of the measurement error parameters (first row, right upper panels). The posteriors of $\sigma_{x_1|\tilde{x}}$ and $\sigma_{x_2|\tilde{x}}$ (not reported) both with the joint model and the true λ are essentially equal to the prior, while the plug-in posterior is concentrated on larger values. Under such conditions, even with the true linkage structure, we do not have any useful information for estimating the mea-

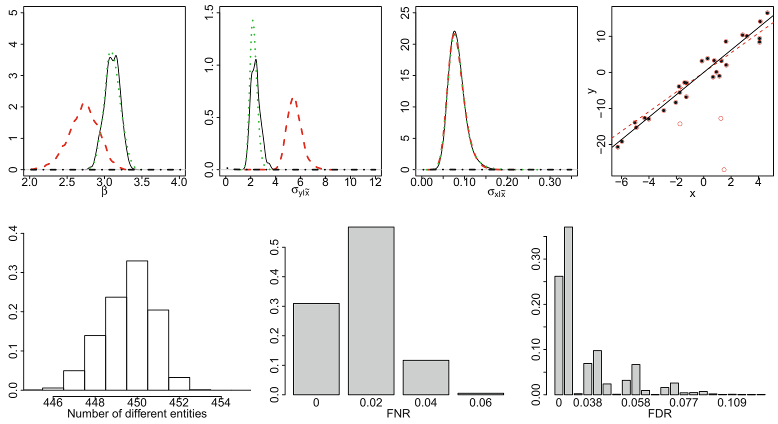

Fig. 4. Experiment II. Left upper panels: prior (dotdash lines) and posterior of $\beta, \sigma_{y|\tilde{x}}, \sigma_{x|\tilde{x}}$ with the joint record linkage and regression model (solid lines), the true linkage structure (dotted lines) and the plug-in approach (dashed lines). Right upper panel: estimated regression line and (x, y) pairs with the joint model (solid line and full circles) and the plug-in approach (dashed line and empty circles) Lower panels: posterior for $k(z)$, FNR and FDR.

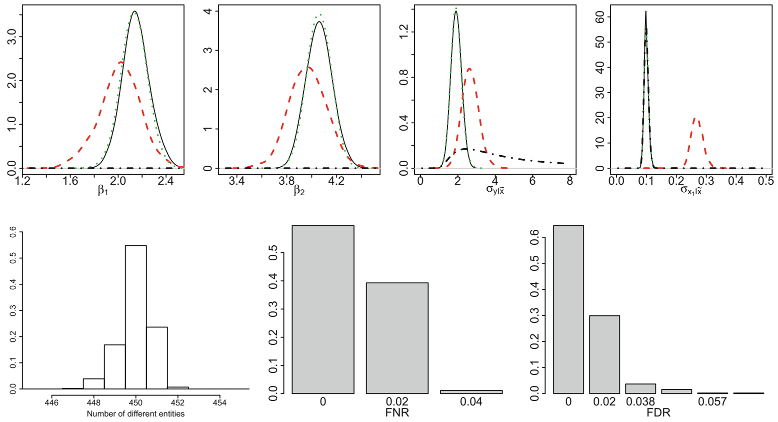

Fig. 5. Experiment III. Same caption as Fig. 3.

surement error variances due to the lack of duplicated x values. Thus, while the joint model correctly does not contrast the information provided by the prior, the presence of false matches creates (y, x) pairs that could be also explained by a larger measurement error of the covariates. We observe that the joint modeling of the record linkage and regression data improves the matching process as noted by the higher concentration of $k(z)$ (second row, left lower panel) around the true value of 450 and the lower FNRs and FDRs (second row, right lower panels) with respect to results obtained with the hit-and-miss model only.

5 Discussion

We have made three major contributions in this paper. First, we have proposed a Bayesian record linkage model investigating the role that prior partition models may have on the matching process. Second, we have proposed a generalized framework for record linkage and regression that accounts for the record linkage error exactly. Using our methodology, one is able to generate a feedback mechanism of the information provided by the working statistical model on the record linkage process. This feedback mechanism is essential to eliminate potential biases that can jeopardize the resulting post-linkage inference. Third, we illustrate our record linkage and multiple regression methodology on many experiments involving a synthetic data set, where improvements are gained in terms of standard record linkage evaluation metrics.

Acknowledgments. Steorts was supported by NSF-1652431 and NSF-1534412. Tancredi and Liseo were supported by Ministero dell' Istruzione dell' Universita e della Ricerca, Italia PRIN 2015.

Appendix

A Metropolis Algorithm

We provide our Metropolis-within-Gibbs algorithm that allows direct simulation from the joint posterior. Let $\lambda_{(-ij)}$ be the vector λ with the element λ_{ij} removed and let $C_{j'} \setminus (ij)$ be the set of all the observed records with record (ij) removed, which conditionally on λ, refer to the population individual j' The full conditional distribution of λ_{ij} is

$$\begin{aligned} p(\lambda_{ij} = q|\lambda_{(-ij)}, \alpha, v) &\propto \prod_{j'=1}^{N_{pop}} p(V_{C_{j'}} = v_{C_{j'}}|\alpha, \lambda)\, p(\lambda_{ij} = q|\lambda_{(-ij)}) \\ &\propto \prod_{j'=1}^{N_{pop}} \frac{p(V_{C_{j'}} = v_{C_{j'}}|\alpha, \lambda)}{p(V_{C_{j'\setminus(ij)}} = v_{C_{j'}\setminus(ij)}|\alpha, \lambda)} p(\lambda_{ij} = q|\lambda_{(-ij)}) \qquad (8) \\ &\propto \frac{p(V_{C_q} = v_{C_q}|\alpha, \lambda)}{p(V_{C_q\setminus(ij)} = v_{C_q\setminus(ij)}|\alpha, \lambda)} p(\lambda_{ij} = q|\lambda_{(-ij)}), \qquad (9) \end{aligned}$$

where $q = 1, \ldots N_{pop}$. In Eq. 8, set $\lambda_{ij} = q$, which implies that $C_{j'} = C_{j'\setminus ij}$ $\forall j' \neq q$. It follows that $\dfrac{p(v_{C_{j'}}|\alpha, \lambda)}{p(v_{C_{j'}\setminus ij}|\alpha, \lambda)} = 1 \quad \forall j' \neq q$. When the population entity q represents an already existing cluster given $\lambda_{-(ij)}$, the above ratio can also be written as

$$\frac{p(V_{C_q} = v_{C_q}|\alpha, \lambda)}{p(V_{C_q\setminus(ij)} = v_{C_q\setminus(ij)}|\alpha, \lambda)}$$
$$= \prod_{l=1}^{p} \left[\alpha_l \theta_{l\,v_{ijl}} + (1-\alpha_l) \frac{\prod_{(i_h,j_h)\in C_q\setminus(ij)} \left((1-\alpha_l)\delta_{v_{i_h j_h l}, v_{ijl}} + \alpha_l \theta_{l v_{i_h j_h l}} \right)}{p(V_{C_q\setminus(ij)\,l} = v_{C_q\setminus(ij)\,l}|\alpha, \lambda)} \right].$$

When the label q identifies a new cluster, the following simplification is possible:

$$\frac{p(V_{C_q} = v_{C_q}|\alpha, \lambda)}{p(V_{C_q\setminus(ij)} = v_{C_q\setminus(ij)}|\alpha, \lambda)} = \prod_{l=1}^{p} \theta_{l, v_{ijl}}.$$

Note that the posterior $p(\lambda, \alpha|v)$ is invariant with respect to the cluster labels and that we are only interested in the cluster composition. Thus, we can avoid simulating the entire population label distribution, and instead set $q \in \{1, \dots, N\}$ (since there can be at most N clusters) and update λ_{ij} with the following:

$$q(\lambda_{ij} = q) = \begin{cases} \frac{p(V_{C_q} = v_{C_q}|\alpha,\lambda)}{p(V_{C_q\setminus(ij)} = v_{C_q\setminus(ij)}|\alpha,\lambda)} p(\lambda_{ij} = q|\lambda_{(-ij)}) & \text{if } q \text{ labels an observed cluster} \\ \prod_{j=1}^{p} \theta_{l,v_{ijl}} p(\lambda_{ij} = new|\lambda_{(-ij)})/(N - k_{(-ij)}) & \text{if } q \text{ labels a new cluster} \end{cases} \quad (10)$$

for $i = 1, \dots, L, j = 1, \dots, N_i$, where $k_{(-ij)}$ is the number of clusters without the label λ_{ij}. This way of updating the cluster assignment is standard when the CRP is used for a prior on the cluster assignments. In addition, the marginal likelihood of the cluster observations is known or can be easily calculated using a recursive formula, see for example [11,12].

To adapt the algorithm to the two different prior distribution of λ, note that, when q labels an observed cluster, the use of a uniform prior on λ implies that

$$p(\lambda_{ij} = q|\lambda_{-(ij)}) = 1/N_{pop} \quad \text{and} \quad p(\lambda_{ij} = new|\lambda_{-(ij)}) = (N_{pop} - k_{-(ij)})/N_{pop}.$$

With the PYP prior, the above mentioned probabilities are, respectively,

$$p(\lambda_{ij} = q|\lambda_{-(ij)}) = (n_q - \sigma)/(N - 1 + \vartheta) \quad p(\lambda_{ij} = new|\lambda_{-(ij)}) = (k_{-(ij)}\sigma + \vartheta)/(N - 1 + \vartheta)$$

where n_q here denotes the size of the cluster C_q without the entity λ_{ij}. Finally, when a uniform prior on the partition space is considered, one has

$$p(\lambda_{ij} = q|\lambda_{-(ij)}) \propto 1/(N_{pop})_{k(-ij)} \quad \text{and} \quad p(\lambda_{ij} = new|\lambda_{-(ij)}) \propto (N_{pop} - k_{-(ij)})/(N_{pop})_{k_{(-ij)+1}}.$$

Finally, full conditional distributions of the components of α have a computationally manageable form using a recursive formula. In fact, assuming a standard Beta prior on each α_l, one obtains

$$p(\alpha_l|\lambda, v, \alpha_{-l}) \propto \prod_{j'=1}^{N} p(V_{C_{j'}l} = v_{C_{j'}l}|\alpha_l)\alpha_l^{p-1}(1-\alpha_l)^{q-1},$$

and a straightforward Metropolis step can be easily implemented.

B Construction of PYP Priors

We now briefly describe adapting the PYP prior to our L database framework. Assume the first j records of the i-th database and all the records of the first $i-1$ databases are classified into $k_{i,j}$ clusters identified by the population labels $j'_1, \ldots, j'_{k_{i,j}}$ with sizes $n_1, n_2, \ldots, n_{k_{i,j}}$ respectively. Also, let $N_{i,j} = \sum_{l=1}^{i-1} N_l + j$ denote the total number of these records. Next, the label of the record $\lambda_{i,j+1}$ identifies a new cluster with probability

$$\mathrm{P}\left(\lambda_{i,j+1} = \text{``new''} | \lambda_{1,1}, \ldots, \lambda_{i,j}\right) = \frac{k_{i,j}\sigma + \vartheta}{N_{i,j} + \vartheta},$$

where (ϑ, σ) are two parameters whose admissible values are $\sigma \in [0, 1)$ with $\vartheta > -\sigma$ or $\sigma < 0$ with $\theta = m|\sigma|$ for some positive integer m. Moreover, $\lambda_{i,j+1}$ will assume an already observed label j'_g identifying a cluster with size n_g with probability

$$\mathrm{P}\left(\lambda_{i,j+1} = j'_g | \lambda_{1,1}, \ldots, \lambda_{i_1,j_1}\right) = \frac{n_g - \sigma}{N_{i,j} + \vartheta} \quad g = 1, \ldots, k_{i,j}.$$

The above updating rule induces a prior on the set of the possible partitions of all the N records which can be written as [13]

$$\mathrm{P}\left(z(\lambda) = z\right) = \frac{(\vartheta + \sigma)_{k-1,\sigma}}{(\vartheta + 1)_{N-1,1}} \prod_{g=1}^{k} (1 - \sigma)_{n_g - 1, 1},$$

where $\{n_1, \ldots, n_k\}$ are the cluster sizes of the partition z and $x_{r,s} = x(x + s) \cdots (x + (r-1)s)$. It can also be proved [13] that, under this prior, the expected value of $k(z)$ is

$$E(k(z)) = \sum_{i=1}^{N} \frac{(\vartheta + \sigma)_{(i-1)\uparrow}}{(\vartheta + 1)_{(i-1)\uparrow}} = \frac{\vartheta}{\sigma} \left[\frac{(\vartheta + \sigma)_{N\uparrow}}{\vartheta_{N\uparrow}} - 1 \right]$$

and the variance is

$$Var(k(z)) = \frac{\vartheta(\vartheta + \sigma)}{\sigma^2} \frac{(\vartheta + 2\sigma)_{N\uparrow}}{\vartheta_{N\uparrow}} - \frac{\vartheta^2}{\sigma^2} \left(\frac{(\vartheta + \sigma)_{N\uparrow}}{\vartheta_{N\uparrow}} \right)^2 - \frac{\vartheta}{\sigma} \frac{(\vartheta + \sigma)_{N\uparrow}}{\vartheta_{N\uparrow}}$$

with $x_{s\uparrow} = \Gamma(x + s)/\Gamma(x)$. For more details, we refer to [19].

The above equations can be used for prior elicitation by fixing ϑ and σ in order to have $E(k(z))$ equal to a rough prior guess for the number of clusters and a specific amount of prior variability for $k(z)$. Moreover, in evaluating the asymptotic properties, [13] observes that as $N \to \infty$, $E(k(z))$ becomes infinite for non negative values of σ; on the other hand, if σ is negative, $k(z)$ is equal almost surely to m which thus takes the role of the size N_{pop} in a finite population framework.

C Record Linkage Without Duplicate-Detection

We now consider record linkage of two databases without duplicate-detection. To consider this case, we simply modify the prior distribution on the λ's such that $\lambda_{ij_1} \neq \lambda_{ij_2}\ \forall j_1 \neq j_2$ and for $i = 1, 2$. In this case, clusters consist of at most two elements so that the distribution of the observed records v, conditional on λ and α, can be calculated analytically without exploiting the recursive formula.

If a uniform prior on the label space is assumed, the above conditioning is equivalent to assuming that the two databases are two simple random samples with replacement from a population of N_{pop} units. This is the same situation described in [18], where N_{pop} is assumed unknown. Assume that T denotes the number of common units between the two databases; then $k(z)$ is equal to $N_1 + N_2 - T$, where T follows a hypergeometric distribution

$$P(T=t) = \frac{\binom{N_1}{t}\binom{N_{pop}-N_1}{N_2-t}}{\binom{N_{pop}}{N_2}}, \quad \max\{0, N_1+N_2-N_{pop}\} \leq t \leq \min\{N_1, N_2\}.$$

From a computational perspective, the conditioning of the uniform prior does not imply substantial changes. In fact if a PYP prior is assumed, the standard record linkage framework can be tackled by imposing that $\lambda_{1j} = j$ for $j = 1, \dots, N_1$ and that the units of the second database may only join a cluster composed by a single unit of the first database or create a new cluster, that is

$$p(\lambda_{2\,j+1} = q|\lambda_{11}\dots,\lambda_{2\,j}) = \begin{cases} \frac{1-\sigma}{k_{2j}-j(1-\sigma+\vartheta)} & \text{if } q \leq N_1 \text{ and } n_q = 1 \\ 0 & \text{if } q \leq N_1 \text{ and } n_q = 2 \\ 0 & \text{if } q > N_1 \text{ and } n_q = 1, \end{cases} \qquad j = 0, 1\dots, N_2 - 1$$

and

$$p(\lambda_{2\,j+1} = new|\lambda_{11}\dots,\lambda_{2\,j}) = \frac{k_{2j}\sigma + \vartheta}{k_{2j} - j(1-\sigma)+\vartheta} \quad j = 0, 1\dots, N_2 - 1$$

where $k_{20} = N_1$ and k_{2j} is the number of distinct elements considering the first database and the first j elements of the second database. Finally, notice that

$$p(\lambda_{21},\dots,\lambda_{2N_2}|\lambda_{11},\dots,\lambda_{1N_1}) = \frac{(1-\sigma)^{N-k_{2N_2}}\prod_{l=N_1+1}^{k_{2N_2}}(\sigma(l-1)+\vartheta)}{\prod_{l=1}^{N_2}(k_{2\,l-1}-(l-1)(1-\sigma)+\vartheta)}\frac{(N-k_{2N_2})!}{N!}.$$

This implies that the λ's are no longer exchangeable. This problem, although interesting from a theoretical perspective, does not cause computational issues.

The conditional prior probabilities for the Gibbs step updating of λ_{2j} to be used from Eq. (10) are

$$p(\lambda_{2j} = q|\lambda_{-(2j)}) \propto \begin{cases} (1-\sigma) & \text{if } q \leq N_1 \text{ and } n_q = 1 \\ 0 & \text{if } q \leq N_1 \text{ and } n_q = 2 \\ 0 & \text{if } q > N_1 \text{ and } n_q = 1, \end{cases}$$

and

$$p(\lambda_{2j} = new|\lambda_{-(2j)}) \propto (k_{-(2j)}\sigma + \vartheta)\prod_{l=j}^{N_2-1}\left[\frac{k_{2l} - l(1-\sigma)+\vartheta}{k_{2l}+1-l(1-\sigma)+\vartheta}\right].$$

D Record Linkage Experiment

We provide the parameter settings for the record linkage experiments. For the case with duplicate detection, we considered the effect of the PYP prior for λ with $(\theta, \sigma) = (0.4, 0.98)$, $(2, 0.975)$, $(10, 0.965)$. These prior distributions have a common prior mean of $k(z)$ almost equal to 450; however, their respective variance are quite different. For the case of no duplicate detection, we consider the effect of the constrained PYP prior for λ with $(\theta, \sigma) = (1, 0.6)$, $(1, 0.725)$, and $(1, 0.86)$. These values of the hyper-parameters (θ, σ) produce prior means for the number of matches equal to 75, 50 and 25.

E Regression Experiments

We elaborate on our three regression experiments. In the first experiment, we modify the data set by adding two columns with the pairs y and x generated from the model in Sect. 3.1, conditional on the true λ structure. For clusters with two records we simulate a single true value $\tilde{x}$ of the covariate from a normal distribution with zero mean and variance equal to $\sigma^2_{\tilde{x}} = 9$. Then, conditionally on $\tilde{x}$, we generate two independent draws x from a normal distribution with mean $\tilde{x}$ and variance $\sigma^2_{x|\tilde{x}} = 0.01$ and two corresponding independent draws y from a normal distribution with mean $\beta\tilde{x}$ with $\beta = 3$ and variance $\sigma^2_{y|\tilde{x}} = 4$. Instead, the records without duplication are augmented with a single pair (y, x) that is generated conditionally on a single value $\tilde{x}$ following the same model of the duplicated records.

In the second experiment, we use the modified RL500 data set that consists of two databases. We then remove y from the second databases and x from the first database. Given the 50 entities with duplication, 28 belong to both the databases reporting the y variable on the first database and the x variable on the second database. Moreover, 9 entities only belong to first database with 2 duplicate records of y, and 13 entities only belong to the second database with 2 copies of x. In addition, we assume the same priors as in the first experiment.

In the third experiment, we modify the RL500 data set by generating data from a regression model with two covariates, where we assume $\beta_1 = 2$ and $\beta_2 = 4$, $\sigma^2_{y|\tilde{x}} = 4$ and a diagonal covariance matrix $\Sigma_{x|\tilde{x}}$ with elements $\sigma^2_{x_1|\tilde{x}} = \sigma^2_{x_2|\tilde{x}} = 0.01$. We then split this data set into two databases of size 250, and then remove y from the second database and remove the two covariates from the first database. To mimic the case of record linkage without duplicate detection we arrange the two databases so that they share 50 entities without duplications within each databases.

References

1. Christen, P.: Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31164-2
2. Copas, J., Hilton, F.: Record linkage: statistical models for matching computer records. J. R. Stat. Soc. A **153**, 287–320 (1990)
3. De Blasi, P., Favaro, S., Lijoi, A., Mena, R., Prunster, I., Ruggiero, M.: Are gibbs-type priors the most natural generalization of the dirichlet process? IEEE Trans. Pattern Anal. Mach. Intell. **37**(2), 803–821 (2015)
4. Goldstein, H., Harron, K., Wade, A.: The analysis of record-linked data using multiple imputation with data value priors. Stat. Med. **31**, 3481–3493 (2012)
5. Gutman, R., Afendulis, C.C., Zaslavsky, A.M.: A Bayesian procedure for file linking to analyze end-of-life medical costs. J. Am. Stat. Assoc. **108**, 34–47 (2013)
6. Gutman, R., Sammartino, C., Green, T., Montague, B.: Error adjustments for file linking methods using encrypted unique client identifier (eUCI) with application to recently released prisoners who are HIV+. Stat. Med. **35**, 115–129 (2016)
7. Hof, M., Zwinderman, A.: Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. Stat. Med. **31**, 4231–4242 (2012)
8. Kim, G., Chambers, R.: Regression analysis under incomplete linkage. Comput. Stat. Data Anal. **56**, 2756–2770 (2012)
9. Lahiri, P., Larsen, M.D.: Regression analysis with linked data. J. Am. Stat. Assoc. **100**, 222–230 (2005)
10. Liseo, B., Tancredi, A.: Bayesian estimation of population size via linkage of multivariate normal data sets. J. Off. Stat. **27**, 491–505 (2011)
11. MacEachern, S.N.: Estimating normal means with a conjugate style Dirichlet process prior. Commun. Stat.-Simul. Comput. **23**, 727–741 (1994)
12. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**, 249–265 (2000)
13. Pitman, J.: Combinatiorial Stochastic Processes. Ecole d'Eté de Probabilités de Saint-Flour XXXII. LNM, vol. 1875. Springer, Berlin (2006). https://doi.org/10.1007/b11601500
14. Sadinle, M.: Detecting duplicates in a homicide registry using a Bayesian partitioning approach. Ann. Appl. Stat. **8**, 2404–2434 (2014)
15. Steorts, R.C.: Entity resolution with empirically motivated priors. Bayesian Anal. **10**, 849–875 (2015)
16. Steorts, R.C., Hall, R., Fienberg, S.E.: SMERED: a Bayesian approach to graphical record linkage and de-duplication. J. Mach. Learn. Res. **33**, 922–930 (2014)
17. Steorts, R.C., Hall, R., Fienberg, S.E.: A Bayesian approach to graphical record linkage and de-duplication. J. Am. Stat. Soc. (2016)
18. Tancredi, A., Liseo, B.: A hierarchical Bayesian approach to record linkage and population size problems. Ann. Appl. Stat. **5**, 1553–1585 (2011)
19. Yamato, H., Shibuya, M.: Moments of some statistics of pitman sampling formula. Bull. Inf. Cybern. **32**, 1–10 (2000)