Statistical challenges in observational cohort studies

Hof, M.H.P.

Publication date
2015
Document Version
Final published version

4 A mixture model for the analysis of data derived from record linkage

Combining information from two data sources depends on finding records that belong to the same individual (matches). Sometimes unique identifiers per individual are not available and we have to rely on partially identifying variables that are registered in both data sources. A risk of relying on these variables is that some records from both datasets are wrongly linked to each other, which introduces bias in further regression analyses. In this chapter, we propose a mixture model where we treat the indicator whether records belong to the same individual as missing. Each pair of records from both datasets contributes independently to a pairwise pseudo-likelihood, which we maximize with an expectation-maximization algorithm. Each part of the pseudo-likelihood is parameterized by the appropriate (parametric) density function. Moreover, some structures of the data allow for simplifying assumptions which makes the pseudo-likelihood considerably easier to parameterize. Because the optimization requires a product over all combinations of records from both datasets, we suggest a procedure that summarizes information from highly unlikely matches. With simulations we showed that the new approach produces accurate estimates in different linkage scenarios. Moreover, the estimator remained accurate in scenarios where previously proposed analysis approaches give biased results. We applied the method to estimation of the association between pregnancy duration of the first and second born children from the same mother from a register without mother identifier.

Introduction

Record linkage of data sources is an effective way to answer research questions without the need for new data collection. Records that belong to the same individual (matches) are separated from records that do not belong to the same individuals (non-matches) and inference is made on the matches [65, 66, 67]. However, a problem is that there is sometimes no unique identifier for matches. Then the record linkage strategy depends on partially identifying variables registered in the data sources, often referred to as linkage variables.

Most analyses of data derived from record linkage are secondary; the analyst only has access to a linked dataset, containing the record pairs that are declared as match. In this situation, the analyst is not the one that performed the record linkage procedure. A problem that arises in this set of declared links is the presence of false matches, caused by the partially identifying nature of the linkage variables [69]. In most linkage problems, some quantitative information is available about the quality of the linked dataset. To accurately estimate the relation between an outcome y from one database and a vector of covariates $\mathbf{x}$ from the other database, the model needs to correct for the fact that not all identified matches are true matches.

Different approaches have been proposed to reduce the influence of wrong matches present in the linked dataset. Scheuren and Winkler developed a method that requires that only the outcome y is derived from a linkage process and the covariates $\mathbf{x}$ are known for each individual. From all pairs of covariates and outcome $(\mathbf{x}, b)$, there is one true pair $(\mathbf{x}, y)$ and all the other pairs consist of wrong values. The outcome b with the highest posterior probability of a match is assigned to the true value y [70]. This method has been generalized by Lahiri and Larsen [30]. Their method uses the posterior probability that a pair $(\mathbf{x}, b)$ is a match and gives unbiased estimates when the outcome variable and covariates are located in separate datasets, both datasets have similar number of records, and all records from both datasets refer to the same population.

Kim and Chambers and Chipperfield et al. developed methods that reduce the influence of wrongly linked matches after a record linkage procedure has been performed [85, 72, 31]. To each $\mathbf{x}$ in a particular dataset, the most likely y value is linked from the other dataset. Each pair $(\mathbf{x}, y)$ has the probability of being correctly identified as a match, which is obtained from an external source. Chipperfield et al. recommended clerical review of a subset of the linked records to obtain this probability.

Hof and Zwinderman suggested a weighted least squares (WLS) approach [86]. When the assumptions made by Lahiri and Larsen hold, this WLS estimator has similar performance to the Lahiri and Larsen estimator. However, it is not necessary that the outcome variable and all covariates are located in separate datasets. Goldstein et al. proposed yet another method based on the fact that there are pairs $(\mathbf{x}, y)$ in the linked dataset that we suspect to be non-matches. From these pairs, y is removed and treated as a missing observation [87].

Because all these estimators have been developed for linked datasets, they require some arbitrary decision on whether to declare a record pair as match or not. Moreover, the accuracy of the analysis depends on the information that is available about the quality of the linked dataset. In this chapter, we develop a new method based on the situation in which the analyst has access to the original datasets based on the record linkage theory by Fellegi and Sunter [29] and the WLS approach. This new estimator can also be used when the analyst has only access to a linked dataset, given that the individual that performs the linkage procedure returns all pairs with their corresponding probability of being a match. The new estimator is potentially less restrictive than previous estimators. With simulations, we show that the new estimator gives accurate results in a number of linkage scenarios. We applied the new estimator to real data, in which we determined the association between pregnancy duration of first and second born children.

Record linkage

Consider two datasets $\mathbf{A}$ and $\mathbf{B}$ containing respectively n and m records and both contain records from the same individuals. A vector of binary indicators $\mathbf{d} = (d_{11}, \ldots, d_{ij}, \ldots, d_{nm})^T$ describes whether a record i from $\mathbf{A}$ and record j from $\mathbf{B}$ belong to the same individual (match): $d_{ij} = 1$. In addition to matches, both datasets might contain records that do not have a match in the other dataset.

In this chapter, we assume that the data from all matches are independent observations and can be analysed with a generalized linear model. The outcome $\mathbf{x}$ and covariates y of these matches are spread over the two datasets and we observe the vector $\mathbf{y} = (y_1, \ldots, y_j, \ldots, y_m)^T$ in $\mathbf{B}$ and the matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n)^T$ in $\mathbf{A}$, where $\mathbf{x}_i$ is the vector of covariates for the i^{th} record. We want to estimate the conditional expectation $E[y_j|\mathbf{x}_i, d_{ij} = 1]$, i.e. the relation between an outcome

variable y_j and covariates $\mathbf{x}_i$ in matches. Note that the outcome y_j can be either continuous or categorical and the vector $\mathbf{x}_i$ might contain covariates of any type.

Because the matches are independent observations, each record from $\mathbf{A}$ and $\mathbf{B}$ has *either one or no match* in the other dataset. In our record linkage situation, the vector $\mathbf{d}$ is not observed (i.e. there is no unique identifier for matches or prior knowledge on likely matches). Therefore it is not possible to use conventional regression techniques to estimate $E[y_j|\mathbf{x}_i, d_{ij} = 1]$.

In addition to the outcome and covariates, both datasets contain the same partially identifying variables $\mathbf{R}$. Examples of such variables are home address, date of birth, or initials. Each record from $\mathbf{A}$ and $\mathbf{B}$ has a vector of linkage variables $\mathbf{r}_i = (r_{i1}, \ldots, r_{ik}, \ldots, r_{il})$. Record i from $\mathbf{A}$ can be written as the vector $(\mathbf{x}_i, \mathbf{r}_i)$, and record j from $\mathbf{B}$ as the vector $(y_j, \mathbf{r}_j)$. The complete set of all nm record pairs is the union of two disjoint sets of matches $\mathbf{M}$ and non-matches $\mathbf{U}$, where $\mathbf{M} = \{(a, b); d = 1, a \in \mathbf{A}, b \in \mathbf{B}\}$ and $\mathbf{U} = \{(a, b); d = 0, a \in \mathbf{A}, b \in \mathbf{B}\}$.

Our main interest is to estimate the conditional expectation $E[\mathbf{y}|\mathbf{X}, \mathbf{M}]$, i.e. the relation between the outcome variable $\mathbf{y}$ and covariates $\mathbf{X}$ in the set of matches $\mathbf{M}$. Because we do not observe $\mathbf{d}$, we need to rely on the vectors of partially identifying variables $\mathbf{r}_i$ from $\mathbf{A}$ and $\mathbf{r}_j$ from $\mathbf{B}$ to determine whether record pair ij belongs to the set $\mathbf{M}$ or the set $\mathbf{U}$. The comparison of $\mathbf{r}_i$ and $\mathbf{r}_j$ can be done in multiple ways but often a binary agreement/disagreement comparison is used. This results in the matrix $\mathbf{G} = (\mathbf{g}_{11}, \ldots, \mathbf{g}_{ij}, \ldots, \mathbf{g}_{nm})^T$ containing all nm comparison vectors $\mathbf{g}_{ij} = (g_{ij1}, \ldots, g_{ijk}, \ldots, g_{ijl})$, where

$$g_{ijk} = \begin{cases} 1 & \text{if } r_{ik} = r_{jk}, \\ 0 & \text{if } r_{ik} \neq r_{jk}. \end{cases}$$

Each vector $\mathbf{g}$ can be observed in matches $\mathbf{M}$ ($d = 1$) and non-matches $\mathbf{U}$ ($d = 0$), where we can write the likelihood of $\mathbf{g}_{ij}$ as [29, 76]

$$L(\mathbf{g}_{ij}) = p(d_{ij} = 1)L(\mathbf{g}_{ij}|d_{ij} = 1) + (1 - p(d_{ij} = 1))L(\mathbf{g}_{ij}|d_{ij} = 0), \tag{4.1}$$

where $L(\mathbf{g}_{ij}|d_{ij} = 1)$ and $L(\mathbf{g}_{ij}|d_{ij} = 0)$ are the conditional distributions of $\mathbf{g}_{ij}$ in $\mathbf{M}$ and $\mathbf{U}$, and $p(d_{ij} = 1)$ the prevalence of matches in the nm record pairs. Note that this likelihood often requires an impracticable number of parameters which need to be estimated. Because there are 2^l unique patterns for l linking variables, there are also 2^l conditional distributions $L(\mathbf{g}_{ij}|d_{ij} = 1)$ and $L(\mathbf{g}_{ij}|d_{ij} = 0)$.

To decrease the number of parameters, Fellegi and Sunter suggested to assume independence between the comparison outcomes of each linking variable given d_{ij} [29].

Maximizing the likelihood of (4.1) requires the evaluation of all nm record pairs, which is computationally expensive when n and m are large. The number of record pairs can be reduced with blocking, for which we use one or more linkage variables as blocking variables [74]. Only when the blocking variable agrees, the record pair is evaluated on the linkage variables. When the blocking variable has highly discriminative properties, the number of comparisons we have to evaluate can be reduced substantially. However, it is necessary that the blocking variable has been entered correctly in both datasets, otherwise it might be possible that we miss matches which causes loss of power in further analyses.

Other assumptions and extensions of the record linkage mixture model are, among others, the introduction of approximate string comparison [78, 79], the introduction of clerical review in the estimation [76], the addition of interactions among comparison fields [76, 80, 88], and the introduction of extra classes in the non-matches (e.g. potential non-match and likely non-match classes) [71, 76].

After $L(\mathbf{g}_{ij}|d_{ij}=1)$, $L(\mathbf{g}_{ij}|d_{ij}=0)$, and $p(d_{ij}=1)$ have been estimated for each record pair ij, the posterior probability of a match given $\mathbf{g}_{ij}$ is calculated as

$$p(d_{ij}=1|\mathbf{g}_{ij}) = \frac{p(d_{ij}=1)L(\mathbf{g}_{ij}|d_{ij}=1)}{L(\mathbf{g}_{ij})}, \tag{4.2}$$

which can be included in a subsequent regression analysis by using the Lahiri and Larsen estimator [30] or the WLS estimator [86]. A disadvantage of using $p(d_{ij}=1|\mathbf{g}_{ij})$ is that only the comparison vector $\mathbf{g}$ is used to determine whether record pair ij is a match.

To reduce the influence of non-matches in the analysis, Jaro proposed to assign record pairs to either $\mathbf{M}$ or $\mathbf{U}$ under constraint that for each record from $\mathbf{A}$ only one record from $\mathbf{B}$ is a match and vice versa. Given this constraint and the estimated posterior probabilities of a match, the best assignment for all record pairs is the one that gives the highest sum of estimated posterior probabilities of a match in $\mathbf{M}$ [89].

An alternative method to reduce the influence of non-matches in the analysis has been proposed by Scheuren and Winkler. With their method, the estimated posterior probabilities of a match are used to assign record pairs to the set of

matches **M** or to the set of non-matches **U**. A regression model is fitted to **M**, and the residuals are used to find potential falsely identified matches [71]. This procedure is repeated until the estimated regression parameters remain the same.

Although we can reduce the influence of non-matches with these additional steps, their solutions are often suboptimal because they either fail to remove the influence of non-matches completely or they also remove true matches from the analysis depending on arbitrary decisions made by the analyst. Therefore we propose a new method, that uses $\mathbf{g}_{ij}$ but also the outcome y_j and covariates $\mathbf{x}_i$ to determine the probability that a record pair is a match.

New estimator

Pseudo-likelihood

Each pair of records ij from **A** and **B** can be written as the observed values $(\mathbf{x}_i, y_j, \mathbf{g}_{ij})$. In this chapter, we will focus on the situation in which each record from **A** and **B** has *at most* one match in the other dataset. Correlated observations which are observed in hierarchical data or longitudinal data will not be considered in this chapter. The vector **d**, containing the matching status of all nm record pairs, is not observed. We could maximize the expectation of the full joint-log-likelihood $\log(L_f(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d}))$ with respect to the posterior probabilities of the vector **d** given the observed data

$$E_{\mathbf{d}}[\log[L_f(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d})]] = \sum_{\mathbf{h} \in \mathbf{H}} \log[L(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d} = \mathbf{h})] p(\mathbf{d} = \mathbf{h} | \mathbf{y}, \mathbf{X}, \mathbf{G}), \quad (4.3)$$

where the full-likelihood $L_f(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d})$ is given by (4.12) in the appendix at page 87, and **H** is the set containing all possible values of **d**. Under the assumption of one or no match for each record in **A** and **B**, the size of **H** is

$$\sum_{k=1}^{n} \binom{n}{k} \prod_{l=0}^{k-1} (m - l),$$

which is a very large, even for moderate n and m. Therefore maximizing the full joint-log-likelihood becomes too computationally expensive. To reduce the complexity of the likelihood, we propose to ignore the dependence between record pairs and define a pseudo-likelihood $L(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d})$ as

$$L(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d}) = \prod_{i=1}^{n} \prod_{j=1}^{m} L(y_j, \mathbf{x}_i, \mathbf{g}_{ij} | d_{ij} = 1) p(d_{ij} = 1) + L(y_j, \mathbf{x}_i, \mathbf{g}_{ij} | d_{ij} = 0) p(d_{ij} = 0), \tag{4.4}$$

where the main advantage between (4.3) and (4.4) is that we do not marginalize over the set $\mathbf{H}$, but over $d_{ij} = 1$ and $d_{ij} = 0$ for each record pair ij. An explicit demonstration of the validity of the pseudo-likelihood is given in the appendix. In the pseudo-likelihood, we require the joint distributions $L(y_j, \mathbf{x}_i, \mathbf{g}_{ij} | d_{ij} = \ell)$, where $\ell = 1, 0$. Because both distributions are difficult to specify, we simplify the pseudo-likelihood by using the following three assumptions:

1. $y_j \perp \mathbf{x}_i | d_{ij} = 0$; the set of non-matches $\mathbf{U}$ only contains noise implying there is no relation between y_j and $\mathbf{x}_i$.
2. $(y_j, \mathbf{x}_i) \perp \mathbf{g}_{ij} | d_{ij} = 1$; in the set of matches $\mathbf{M}$, no information is contained in $\mathbf{g}_{ij}$ with respect to the joint distribution of y_i and $\mathbf{x}_j$.
3. $(y_j, \mathbf{x}_i) \perp \mathbf{g}_{ij} | d_{ij} = 0$; in the set of non-matches $\mathbf{U}$, no information is contained in $\mathbf{g}_{ij}$ with respect to the joint distribution of y_j and $\mathbf{x}_i$.

Assumption 1 is based on the fact that the set of all nm records pairs is the Cartesian product $\mathbf{C} = \mathbf{A} \times \mathbf{B}$. In this set, we consider all possible combinations of $\mathbf{A}$ and $\mathbf{B}$, which means that $\mathbf{C}$ contains all possible pairs of $\mathbf{x}_i$ and y_j. Consequently, there is no correlation between $\mathbf{x}_i$ and y_j in $\mathbf{C}$. Since each record from $\mathbf{A}$ and $\mathbf{B}$ has maximally one match, the maximum size of $\mathbf{M}$ is $n^1 = min(n, m)$. Because $\mathbf{M}$ and $\mathbf{U}$ are disjoint sets, the size of the set of $\mathbf{U}$ is $nm - n^1$, which is a considerable proportion of $\mathbf{C}$. For instance, with $n = 500$ and $m = 500$, the size of $\mathbf{M}$ is maximum 500 (0.2%) record pairs whereas the size of $\mathbf{U}$ is minimally 249500 (99.8%). In many record linkage situations, n and m are large and the set $\mathbf{M}$ is an ignorable proportion of $\mathbf{C}$. Therefore we can assume that there is no relation between $\mathbf{x_i}$ and y_j in $\mathbf{U}$.

To illustrate when assumptions 2 and 3 hold, we have to re-write (4.4) as

$$L(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d}) = \prod_{i=1}^{n} \prod_{j=1}^{m} L(\mathbf{g}_{ij}|y_j, \mathbf{x}_i, d_{ij} = 1) L(y_j, \mathbf{x}_i|d_{ij} = 1) p(d_{ij} = 1) + L(\mathbf{g}_{ij}|y_j, \mathbf{x}_i, d_{ij} = 0) L(y_j, \mathbf{x}_i|d_{ij} = 0) p(d_{ij} = 0).$$

For each record pair ij, assumptions 2 and 3 require that $L(\mathbf{g}_{ij}|y_j, \mathbf{x}_i, d_{ij}) = L(\mathbf{g}_{ij}|d_{ij})$ in $\mathbf{M}$ and $\mathbf{U}$. With linkage variables that cannot perfectly distinguish matches from non-matches we could observe, for instance, the comparison vectors $(0, 1, \ldots, 1)$ and $(1, 0, \ldots, 1)$ in $\mathbf{M}$. In this situation, assumption 2 holds when in the set of matches $\mathbf{M}$ the probability of agreement of the k^{th} linkage variable can be written as $p(g_{ijk} = 1|d_{ij} = 1) = \lambda_k^{(A)} \times \lambda_k^{(B)}$, where $\lambda_k^{(A)}$ and $\lambda_k^{(B)}$ are the probabilities that linkage variable $\mathbf{R}_k$ has been entered correctly in respectively dataset $\mathbf{A}$ and dataset $\mathbf{B}$. In $\mathbf{M}$, $g_{ijk} \neq 1$ when either r_{ik} or r_{jk} has been entered wrongly. Assuming that entry errors are independent of any observed or unobserved variables, we have $(y_j, \mathbf{x}_i) \perp \mathbf{g}_{ij}|d_{ij} = 1$.

Assumption 3 holds when in $\mathbf{U}$, the probability $p(g_{ijk} = 1|d_{ij} = 0)$ *only* depends on the non-uniqueness in the k^{th} linkage variable. For instance, gender as linkage variable has two unique values. In two datasets containing respectively 50% males and 50% females, we expect that $p(g = 1|d = 0) = 0.5$. Generally, we assume that $p(\mathbf{g}_{ij} = z|d_{ij} = 0) = \psi_{\mathbf{g}^*}$, where $\psi_{\mathbf{g}^*}$ is the probability to observe $\mathbf{g}_{ij} = \mathbf{g}^*$. When there are no observed or unobserved variables related to the probability of observing $\mathbf{g}_{ij}$, we have $(y_j, \mathbf{x}_i) \perp \mathbf{g}_{ij}|d_{ij} = 0$.

Because we do not observe the vector $\mathbf{d}$, assumptions 2 and 3 cannot be tested with the available data. However, we could determine the matching status d_{ij} for a subset of record pairs with, for instance, clerical review. Both assumptions can be tested by estimating the correlation between $\mathbf{g}_{ij}$ and $(y_j, \mathbf{x}_i)$ in a set of identified matches and in a set of identified non-matches. Unfortunately, assumption 2 is hard to test because the set of matches $\mathbf{M}$ is relatively small compared to the set of non-matches $\mathbf{U}$. Finding a reasonable number of matches means that we have to evaluate a substantial number of record pairs. For instance, when we have $n = 500$ and $m = 1000$ and 500 matches, we would have to check 1000 record pairs to find one match. Therefore, to obtain a set of matches to test assumption 2, we require a model driven approach to select record pairs with a high probability of a match. By fitting the Fellegi and Sunter record linkage model to the data, we can use the posterior probability of a match given $\mathbf{g}_{ij}$ to determine whether we should review record pair ij. This could reduce the number of record pairs we

have to evaluate considerably. Note that, because there are a lot of non-matches, clerical review of a random set of record pairs can be used to test assumption 3.

When all three assumptions hold, the pseudo-likelihood can be written as

$$\begin{aligned} &L(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d}) = \\ &\prod_{i=1}^{n}\prod_{j=1}^{m} L(y_j|\mathbf{x}_i, d_{ij}=1)L(\mathbf{x}_i|d_{ij}=1)L(\mathbf{g}_{ij}|d_{ij}=1)p(d_{ij}=1) + \\ &\qquad L(y_j|d_{ij}=0)L(\mathbf{x}_i|d_{ij}=0)L(\mathbf{g}_{ij}|d_{ij}=0)p(d_{ij}=0). \end{aligned} \tag{4.5}$$

Maximizing the pseudo-likelihood

Our approach to maximize the pseudo-likelihood in (4.5) is based on a two step-procedure. In the first step we use the Fellegi and Sunter record linkage mixture model from (4.1) to estimate $L(\mathbf{g}_{ij}|d_{ij} = \ell)p(d_{ij} = \ell)$, where $\ell = 0, 1$. We use these estimates as fixed-constants in the second step, and thus ignore their uncertainty. To emphasize this point, we re-write (4.5) as

$$\begin{aligned} L(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d}) = \prod_{i=1}^{n}\prod_{j=1}^{m} &L(y_j|\mathbf{x}_i, d_{ij}=1)L(\mathbf{x}_i|d_{ij}=1)\ w_{ij1} + \\ &L(y_j|d_{ij}=0)L(\mathbf{x}_i|d_{ij}=0)\ w_{ij0}, \end{aligned} \tag{4.6}$$

where $w_{ij1} = L(\mathbf{g}_{ij}|d_{ij} = 1)p(d_{ij} = 1)$ and $w_{ij0} = L(\mathbf{g}_{ij}|d_{ij} = 0)p(d_{ij} = 0)$. Note that we could also estimate the linkage part of the likelihood simultaneously with the regression part. However, we choose the two step procedure because it separates the linkage and the regression problem. A pre-requisite for this approach is that all nm record pairs $(y_j, \mathbf{x}_i)$ with their w_{ij1} and w_{ij0} values are available to the analyst that performs the regression analysis in the second step.

In the second step, the pseudo-likelihood from (4.6) is maximized. Note that our primary interest is the association between the outcome $\mathbf{y}$ and the vector of covariates $\mathbf{X}$ in the set of matches $\mathbf{M}$ (i.e. $E[\mathbf{y}|\mathbf{X}, \mathbf{M}]$). However, we do need to parameterize the other parts of the mixture in (4.6). At this stage we do not specify particular conditional distributions for $L(y_j|\mathbf{x}_j, d_{ij} = 1)$ and $L(y_j|d_{ij} = 0)$ but we assume that they are parameterized by the vectors β and γ, respectively.

In (4.6), $L(\mathbf{x}_i|d_{ij}=1)$ and $L(\mathbf{x}_i|d_{ij}=0)$, concern the multivariate distributions of $\mathbf{X}$ in the sets of matches and non-matches. For now, $L(\mathbf{x}_i|d_{ij}=1)$ and $L(\mathbf{x}_i|d_{ij}=0)$ are parametrized by respectively the vectors ζ and η. Later in this section, we will give some situations in which we can ignore $L(\mathbf{x}_i|d_{ij}=1)$ and $L(\mathbf{x}_i|d_{ij}=0)$.

We suggest an EM-algorithm to maximize the pseudo-likelihood defined in (4.6), in which the expected value of the pseudo-log-likelihood of each record pair $(y_j, \mathbf{x}_i, \mathbf{g}_{ij}, d_{ij})$ is maximized over the posterior distribution of d_{ij} given $(y_j, \mathbf{x}_i, \mathbf{g}_{ij})$ and the vector of model-parameters $\theta = (\beta, \gamma, \zeta, \eta)$. We can write

$$
\begin{aligned}
&E_{\mathbf{d}}[\log[L(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d})]|\theta] = \sum_{i=1}^{n}\sum_{j=1}^{m} E_{d_{ij}}\left[\log[L(y_j, \mathbf{x}_i, \mathbf{g}_{ij}, d_{ij})|\theta]\right] = \\
&\sum_{i=1}^{n}\sum_{j=1}^{m}\{\log[L(y_j|\mathbf{x}_i, d_{ij}=1)] + \log[L(\mathbf{x}_i|d_{ij}=1)] + \log[w_{ij1}]\} \times \\
&\qquad p(d_{ij}=1|y_j, \mathbf{x}_i, \mathbf{g}_{ij}, \theta) + \\
&\qquad \{\log[L(y_j|d_{ij}=0)] + \log[L(\mathbf{x}_i|d_{ij}=0)] + \log[w_{ij0}]\} \times \\
&\qquad (1 - p(d_{ij}=1|y_j, \mathbf{x}_i, \mathbf{g}_{ij}, \theta)),
\end{aligned}
\tag{4.7}
$$

where $p(d_{ij}=1|y_j, \mathbf{x}_i, \mathbf{g}_{ij}, \theta)$ is given by (4.13) in the appendix on page 91. Let $\hat{\theta}^{(0)}$ be the vector of initial starting values. The EM-algorithm seeks to find the maximum of (4.7) by iteratively applying an expectation-step and a maximization step. In the expectation step of iteration t, the posterior probability $p(d_{ij}=1|y_j, \mathbf{x}_i, \mathbf{g}_{ij}, \hat{\theta}^{(t-1)})$ is calculated, which is the posterior probability of a match for record pair ij given the data and the value of the estimated model-parameters $\hat{\theta}^{(t-1)}$ from the previous iteration $t-1$. In the maximization step, (4.7) is maximized, with fixed probabilities $p(d_{ij}=1|y_j, \mathbf{x}_i, \mathbf{g}_{ij}, \hat{\theta}^{(t-1)})$. The EM-algorithm is described in detail in the appendix.

Depending on the type of outcome $\mathbf{y}$ in the data, we have to parametrize the conditional distributions $L(y_j|\mathbf{x}_i, d_{ij}=1)$ and $L(y_j|d_{ij}=1)$. For $L(\mathbf{x}_i|d_{ij}=1)$ and $L(\mathbf{x}_i|d_{ij}=0)$, the parametrization depends on the types of covariates $\mathbf{X}$ in the data. For instance, when $\mathbf{X}$ can be written as a set of disjoint categories, we can use a multinomial distribution to parametrize both distributions. A disadvantage of this approach is that we lose accuracy when we divide continuous variables into categories. Therefore, when all covariates $\mathbf{X}$ are continuous, we propose to write $L(\mathbf{x}_i|d_{ij}=1)$ and $L(\mathbf{x}_i|d_{ij}=0)$ in terms of univariate marginal distributions and

describe their dependency structure with a copula. This approach is described in more detail in the appendix.

In some situations we may assume that $L(\mathbf{x}_i|d_{ij}=1) = L(\mathbf{x}_i|d_{ij}=0) = L(\mathbf{x}_i)$ and the pseudo-likelihood from (4.6) can be re-written as

$$L(\mathbf{y},\mathbf{X},\mathbf{G},\mathbf{d}) = \prod_{i=1}^{n}\prod_{j=1}^{m} L(\mathbf{x}_i)[L(y_j|\mathbf{x}_i,d_{ij}=1)\ w_{ij1} + L(y_j|d_{ij}=0)\ w_{ij0}]. \tag{4.8}$$

Since we are not primarily interested in $L(\mathbf{x}_i)$, we can ignore this term in the pseudo-likelihood. Note that besides our proposed EM-algorithm, this likelihood can also be maximized with standard optimization algorithms. In general, $L(\mathbf{x}_i|d_{ij}=1) = L(\mathbf{x}_i|d_{ij}=0)$ when the records from dataset **A** that have a *match* in dataset **B** and the records from dataset **A** *without* a match in dataset **B** are sampled from the same population.

When there is perfect linkage and all records in **A** have one true match in **B**, $L(\mathbf{x}_i|d_{ij}=1) = L(\mathbf{x}_i|d_{ij}=0)$. Because each record i from **A** has one true match and $m-1$ true non-matches, we know that $\sum_{j=1}^{m} q_{ij1} = 1$ and $\sum_{j=1}^{m} q_{ij0} = m-1$. Therefore the weighted empirical cumulative distribution of the e^{th} covariate specified in (4.16) in the appendix is similar for matches and for non-matches. With "almost perfect" linkage variables and all records from **A** have a match in B, this equality is almost satisfied. Therefore we could ignore the information that is contained in **X** regarding whether a record pair is a match or a non-match.

Variance estimation

In situations where both datasets have a small number of records, we use a bootstrap procedure to capture both the uncertainty of the record linkage parameters and the regression parameters. Let **C** be the Cartesian product of **A** and **B**, containing all nm record pairs. Simple random sampling with replacement is used to take replicate samples from **C** and we fit the mixture model on these replicate samples. This procedure is repeated Z times, where we get the estimate $\tilde{\theta}^z$ from the EM-algorithm. Using the parameter estimates $\hat{\theta}$ obtained from maximizing the pseudo-likelihood with the original data, the variance of $\hat{\theta}$ is $\frac{1}{Z}\sum_{z=1}^{Z}(\tilde{\theta}^z - \hat{\theta})^2$.

Because bootstrapping is too computationally expensive for moderate to large datasets, we propose as an approximate variance estimator the inverse of the Hessian matrix of the pseudo-likelihood. Note that this variance estimate is only correct when we assume that $p(d_{ij} = 1|y_j, \mathbf{x}_i, \mathbf{g}_{ij}, \hat{\theta})$ obtained from the pseudo-likelihood is the same as the probability that we would have obtained with a full-likelihood approach. In the appendix we showed that with respect to matches, our pseudo-likelihood has the same objective function as a full-likelihood approach. The Hessian entries for the estimated parameters $\hat{\beta}$ in $L(y_j|\mathbf{x}_i, d_{ij} = 1; \hat{\beta})$ are

$$I(\beta) = -\frac{\partial^2 \log L}{\partial \beta^2} = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{x}_i^T \ (\mathbf{Q}_{ij} - \mathbf{P}_{ij})\, \mathbf{x}_i, \tag{4.9}$$

where $\mathbf{Q}_{ij}$ is a diagonal matrix with entries $s_{ij1}^2 q_{ij1}$ and $\mathbf{P}_{ij}$ also a diagonal matrix with entries $(y_j - E[y_j|\mathbf{x}_i])^2 q_{ij1}(1-q_{ij1})$. When $L(y_j|\mathbf{x}_i, d_{ij} = 1)$ is described by a linear regression model, s_{ij}^2 is the usual residual variance and when $L(y_j|\mathbf{x}_i, d_{ij} = 1)$ is a logistic regression model $s_{ij}^2 = p(y_j = 1|\mathbf{x}_i, d_{ij} = 1) \times (1 - p(y_j = 1|\mathbf{x}_i, d_{ij} = 1))$. For other generalized linear models s_{ij}^2 has similar well-known forms.

Note that with perfect linkage $\mathbf{P}_{ij} = 0$ and $\mathbf{Q}_{ij} = \sigma_{ij}^2$ and therefore (4.9) simplifies to the Hessian of a common generalized linear model.

Reducing the number of record pairs

In most record linkage situations the number of records in $\mathbf{A}$ (i.e. n) and $\mathbf{B}$ (i.e. m) are large, say tens of thousands each, which means that linking $\mathbf{A}$ and $\mathbf{B}$ produces a huge number of record pairs (i.e. nm pairs). Consequently, estimating the relationship between outcome $\mathbf{y}$ and covariates $\mathbf{X}$ becomes time-consuming.

To decrease the number of record pairs in the EM-algorithm while retaining the same accuracy, we consider the following procedure. Firstly, we estimate the probability of a match given the comparison vectors $v_{ij} = p(d_{ij} = 1|\mathbf{g}_{ij})$ with (4.2). Secondly, we use the fact that there are many record pairs with v_{ij} close to zero. These record pairs add almost no information about the relation between $\mathbf{y}$ and $\mathbf{X}$ in matches ($E[\mathbf{y}|\mathbf{X}, \mathbf{M}]$). We use the binary indicator b_{ij} to mark these record pairs, where $b_{ij} = 1$ if $v_{ij} > T$ and $b_{ij} = 0$ otherwise.

We set T low enough that only non-matches have $b_{ij} = 0$ and almost all matches have $b_{ij} = 1$. Note that not all matches have $b = 1$ when the linkage variables

contain a considerable amount of entry error. For instance, consider the extreme case in which record pair ij is a match but all linkage variables have been registered wrongly. Its comparison vector, $\mathbf{g}_{ij} = (0, 0, \ldots, 0)$, is associated with a really low probability of a match and we have to set $T = 0$ to include this match in the analysis. Consequently, all record pairs are included in the analysis and we have no data reduction. However, ignoring this match has small effects on the results since its contribution to the matches part of the pseudo-likelihood is really low.

For all record pairs ij where $b_{ij} = 0$, we fix the distributions $L(\mathbf{g}_{ij}|d_{ij} = 0) = 1$ and $L(\mathbf{g}_{ij}|d_{ij} = 1) = 0$. The distributions $L(\mathbf{g}_{ij}|d_{ij} = 0)$ and $L(\mathbf{g}_{ij}|d_{ij} = 1)$ for the record pairs with $b_{ij} = 1$ are left unchanged. When $\mathbf{y}$ is a discrete variable, we can summarize all record pairs ij with $b_{ij} = 0$ into a limited number of combinations of $\mathbf{y}$ and $\mathbf{X}$ and associated frequency counts. We then use the frequency of each unique combination as a weight in the pseudo-log-likelihood of the data. Simulations showed that our pseudo-likelihood based method gave accurate estimates with good coverage (data not shown).

Another method to reduce the number of record pairs is blocking [89]. We could assign a probability of a match of zero to all record pairs for which the blocking variables disagree. However, in our pseudo-likelihood this does not decrease the number of record pairs we have to evaluate. A record pair with disagreeing blocking variables still contributes to the non-matches part, thus we still have to evaluate it. Further research is needed to investigate efficient approaches to reduce the number of record pairs with blocking.

Simulation

Scenarios

To illustrate the performance of our new method, we performed simulations of different linkage scenarios and compared the results of our new method to those of other regression methods that deal with linked data. In the simulations, we were interested in the relation between a binary outcome variable located in dataset **B** and three continuous covariates located in dataset **A**. Furthermore, each record from **A** had either one or no match in dataset **B**.

The data for the simulations were generated with the following procedure. First we simulated the outcome and covariates of n_1 true matches, using the logistic model $P_1(y = 1|\mathbf{x})$ for the association between $\mathbf{y}$ and $\mathbf{X}$ and the multivariate distribution for $\mathbf{X}$ with density $\mathcal{N}(\mu^1, \mathbf{\Sigma}^1)$. From these n_1 records, the outcome $\mathbf{y}$ was assigned to dataset $\mathbf{B}$ and the covariates $\mathbf{X}$ were assigned to dataset $\mathbf{A}$. This resulted in two datasets in which each record had one true match in the other dataset. To simulate additional records in $\mathbf{A}$ and $\mathbf{B}$ that had no true match, we added extra records to the two datasets. For dataset $\mathbf{A}$, n_0 records without a true match were drawn from the multivariate normal distribution $\mathcal{N}(\mu^0, \mathbf{\Sigma}^0)$. For dataset $\mathbf{B}$, m_0 records without a match were drawn from a binomial distribution with $P_0(y = 1)$. After the data generation, dataset $\mathbf{A}$ had $n = n_1 + n_0$ records and dataset $\mathbf{B}$ had $m = n_1 + m_0$ records.

The second step in the data generating procedure was adding the linkage variables to the data. Because we were not particularly interested in the linkage problem itself, we created a relatively easy linkage problem in which the linkage variables were independently distributed. Four linkage variables were drawn from uniform distributions. After simulating the linkage variables, we assigned similar linkage variable values to the n_1 true matches in both datasets.

Three scenarios were simulated and their characteristics have been summarized in table 4.1. In all scenarios, there were $n_1 = 250$ records from dataset $\mathbf{A}$ and dataset $\mathbf{B}$ that formed a true match. In all scenarios, $\mathbf{B}$ consisted of $n_1 + m_0 = 1000$ records by adding 750 records to $\mathbf{B}$ without a match in $\mathbf{A}$. The linkage variables were chosen such that the expected number of false matches (i.e. record pairs that are no true match with the highest linkage probability) was 25. To reach this expected number of false matches in scenario 1, four linkage variables with 20, 10, 10, and 5 uniformly distributed unique values were simulated. For the other two scenarios, where the number of records in A was $n_1 + n_0 = 500$ instead of 250, we simulated four linkage variables with respectively 20, 20, 10, and 5 uniformly distributed unique values.

In scenario 1, all records from $\mathbf{A}$ had one true match in dataset $\mathbf{B}$ and there was no record without a match in $\mathbf{B}$. An additional $n_0 = 250$ records in $\mathbf{A}$ had no true match in $\mathbf{B}$ in scenario 2. However, the distribution of $\mathbf{X}$ in the 250 records in $\mathbf{A}$ without a match was similar to the distribution of $\mathbf{X}$ in the 250 records with a true match. The most complicated linkage problem was simulated in scenario 3. Similar to scenario 2, dataset $\mathbf{A}$ consisted of 250 records with a true match and 250 records without a true match. However, in scenario 3, the distribution of $\mathbf{X}$ in the records with a true match was different from the distribution of $\mathbf{X}$ in the

records without a true match. This makes it necessary to use the mixture model defined in equations 4.5 and 4.6 and will introduce bias if the simpler mixture model defined in (4.8) is used. All scenarios were simulated 1000 times.

Composition of datasets			
	Scenario 1	Scenario 2	Scenario 3
n_1	250	250	250
n_0	0	250	250
m_0	750	750	750
Unique values of linkage variables			
	Scenario 1	Scenario 2	Scenario 3
$(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \mathbf{R}_4)$	$(20, 10, 10, 5)$	$(20, 20, 10, 5)$	$(20,\ 20,\ 10,\ 5)$
Distributions of outcome and covariates			
	Scenario 1	Scenario 2	Scenario 3
$logit(P_1(y=1\|\mathbf{x}))$	$-1+0.6\mathbf{x}_1-0.8\mathbf{x}_2+0.2\mathbf{x}_3$	$-1+0.6\mathbf{x}_1-0.8\mathbf{x}_2+0.2\mathbf{x}_3$	$-1+0.6\mathbf{x}_1-0.8\mathbf{x}_2+0.2\mathbf{x}_3$
$logit(P_0(y=1))$	-3	-3	-3
μ^1	$(0,0,0)$	$(0,0,0)$	$(0,0,0)$
$\mathbf{\Sigma}^1(\mathbf{X})$	$\begin{pmatrix} 1.0 & -0.3 & 0.2 \\ -0.3 & 1.0 & -0.1 \\ 0.2 & -0.1 & 1.0 \end{pmatrix}$	$\begin{pmatrix} 1.0 & -0.3 & 0.2 \\ -0.3 & 1.0 & -0.1 \\ 0.2 & -0.1 & 1.0 \end{pmatrix}$	$\begin{pmatrix} 1.0 & -0.3 & 0.2 \\ -0.3 & 1.0 & -0.1 \\ 0.2 & -0.1 & 1.0 \end{pmatrix}$
μ^0	$(0,0,0)$	$(0,0,0)$	$(2,-2,2)$
$\mathbf{\Sigma}^0(\mathbf{X})$		$\begin{pmatrix} 1.0 & -0.3 & 0.2 \\ -0.3 & 1.0 & -0.1 \\ 0.2 & -0.1 & 1.0 \end{pmatrix}$	$\begin{pmatrix} 1.0 & -0.1 & 0.2 \\ -0.1 & 1.0 & 0.2 \\ 0.2 & 0.2 & 1.0 \end{pmatrix}$

Table 4.1: Characteristics of the simulated scenarios, where the covariates $\mathbf{X}$ from records with a match and the records without a match were drawn from respectively $\mathcal{N}(\mu^1, \mathbf{\Sigma}^1)$ and $\mathcal{N}(\mu^0, \mathbf{\Sigma}^0)$, where $\mathcal{N}(\mu, \mathbf{\Sigma})$ is the multivariate normal distribution with mean μ and covariance matrix $\mathbf{\Sigma}$.

Regression models

For all of the following approaches, we calculated the bias, mean squared error, and coverage of the 95% confidence interval.

Estimator$_a$ This estimator was based on (4.6), where both $L(y_j|\mathbf{x}_i, d_{ij} = 1)$ and $L(y_j|d_{ij} = 0)$ were estimated with logistic regression models. For the distributions $L(\mathbf{x}_i|d_{ij} = 1)$ and $L(\mathbf{x}_i|d_{ij} = 0)$, i.e. the multivariate distribution of $\mathbf{X}$ in respectively matches and non-matches, we fitted univariate cumulative distributions and a copula to the data (see equations 4.16, 4.17, 4.18a, 4.18b in the appendix). We chose to model the dependencies between the univariate cumulative distributions in both distributions with a Gaussian copula, because it allowed the pairwise dependencies between covariates to differ for each pair [90]. In addition, we fitted the models in which the dependencies between the covariates $\mathbf{X}$ were ignored. Thereby, the number of parameters was reduced by ignoring the dependencies and therefore these models were less demanding to fit to the data. More specifically, we fitted the following models to the data.

	$L(\mathbf{x}_i\|d_{ij} = 1)$	$L(\mathbf{x}_i\|d_{ij} = 0)$
Estimator$_{a1}$	copula	copula
Estimator$_{a2}$	copula	no copula
Estimator$_{a3}$	no copula	no copula

The estimator was fitted to the data five times with different random initial vectors $\hat{\theta}^{(0)}$. The initial starting values $\hat{\beta}^{(0)}$ and $\hat{\eta}^{(0)}$ were drawn from a normal distribution with mean 0 and standard deviation 2. For Estimator$_{a1}$ and Estimator$_{a2}$, the dependencies within the covariates $\mathbf{X}$ in the distributions $L(\mathbf{x}_i|d_{ij} = 1)$ and $L(\mathbf{x}_i|d_{ij} = 0)$ were parametrized by the dependency parameters $\hat{\rho}^{(1)}$ and $\hat{\rho}^{(0)}$. In the vector of initial starting values $\hat{\theta}^{(0)}$, all dependency parameters were set to zero. The variances of the estimated regression coefficients of interest $\hat{\beta}$, regarding the conditional distribution $L(y_j|\mathbf{x}_i, d_{ij} = 1; \hat{\beta})$, were approximated with the inverse of the Hessian matrix from (4.9).

Estimator$_b$: This estimator was based on (4.8), where we assumed that the distribution of $\mathbf{x}_i$ is the same in matches and non-matches. Similarly to Estimator$_a$, the estimator was fitted to the data five times with different random initial vectors $\hat{\theta}^{(0)}$, where the initial starting values $\hat{\beta}^{(0)}$ and $\hat{\eta}^{(0)}$ were drawn a normal distribution with mean 0 and standard deviation

2. The variance of the estimated regression coefficients of interest $\hat{\beta}$ were estimated with the inverse of the Hessian matrix from (4.9).

WLS: This estimator was based on the weighted least squares approach proposed by [86]. Although each record i from dataset **A** had at most one true match in dataset **B**, the vector of posterior probabilities for record i would frequently not sum up to one or zero because of falsely identified matches. Therefore the weights were truncated, depending on whether we suspected that a true match was found. A true match was associated with a high posterior probabilities of a match $v_{ij} = p(d_{ij} = 1|\mathbf{g}_{ij})$. Therefore, the vector should sum up to one if the vector v_i contained a high probability that was higher than an arbitrary threshold Z and to zero otherwise. In our situation, the threshold Z was chosen to be the highest posterior probability of a match in v and the truncated probabilities v_{ij}^{trunc} were

$$v_{ij}^{trunc} = \begin{cases} \frac{v_{ij}}{\sum_{j=1}^{n} v_{ij}} & \text{if } v_{ij} \geq max(v_i), \\ 0 & \text{if } v_{ij} < max(v_i). \end{cases} \tag{4.10}$$

The truncated probabilities v_{ij}^{trunc} were used in a weighted logistic regression analysis. Variance estimation was performed with a bootstrap procedure with 250 replicate samples.

Naive: A weighted logistic regression analysis was performed over all combinations of records from both datasets. The weights per combination of record i from dataset **A** and record j from dataset **B** were the posterior probabilities of a match $p(d_{ij} = 1|\mathbf{g}_{ij})$ derived from the linkage procedure (see (4.2)). Variances of the estimated regression coefficients were obtained by using the inverse of the Hessian matrix.

Deterministic: A logistic regression analysis was performed with only the combinations of records from dataset **A** and dataset **B** with the highest posterior probability of a match $p(d_{ij} = 1|\mathbf{g}_{ij})$. Variances of the estimated regression coefficients were obtained by using the inverse of the Hessian matrix.

Simulation results

The results of the simulated scenarios have been summarized in table 4.2. In scenario 1, Estimator$_{a1}$, Estimator$_{a2}$, Estimator$_{a3}$, Estimator$_b$, and the WLS approach were unbiased and had about 95% coverages of the 95% confidence interval. In scenario 1, Estimator$_b$ was accurate because $L(\mathbf{x}_i|d_{ij}=1) \approx L(\mathbf{x}_i|d_{ij}=0)$. The WLS method was unbiased because its main assumption held in scenario 1; all records from dataset **A** had one match in dataset **B**. Almost all non-matches with a high probability of a match given their comparison vectors ($p(d_{ij}=1|\mathbf{g}_{ij})$) were excluded from the analysis because of the truncation defined in (4.10) and therefore the analysis almost only included true matches. The deterministic approach gave slightly biased estimates. However, the coverage of the 95% confidence interval was accurate. The naive approach gave the highest bias, which was also reflected in the underestimation of the coverages of the 95% confidence interval of β_1 and β_2. Moreover, the regression coefficients β were biased to zero. The intercept α was slightly overestimated in the naive and deterministic approach (respectively $\hat{\alpha}=-1.16$ and $\hat{\alpha}=-1.09$, where we simulated $\alpha=-1$). The results from scenario 2 were similar to the results from the first scenario. Estimator$_{a1}$, Estimator$_{a2}$, Estimator$_{a3}$, Estimator$_b$ and the WLS approach were unbiased and had good coverages of the 95% confidence interval. Estimator$_b$ was unbiased because the distributions of $\mathbf{X}$ in records with a true match and without a true match was similar. Therefore similarly to scenario 1, the distribution of X in matches was almost similar to the distribution of $\mathbf{X}$ in non-matches. In addition, the deterministic approach was slightly biased and the naive method gave the highest bias.

Estimator$_{a2}$ was the only estimator in scenario 3 with almost no bias and good coverage of the 95% confidence interval. Estimator$_{a1}$ was slightly biased in this scenario with an underestimation of the coverages of the 95% confidence interval. The mean squared error of β_1 with Estimator$_{a1}$ was almost twice as high as the mean squared error with Estimator$_{a2}$. In about 5% of the simulated datasets Estimator$_{a1}$ gave completely wrong estimates ($\hat{\beta}_1 < 0$) because $L(\mathbf{x}_i|d_{ij}=0)$ converged to a degenerate distribution, which strongly affected the probability of a match given the data ($p(d_{ij}=1|y_j,\mathbf{x}_i,\mathbf{g}_{ij},\hat{\theta})$). Estimator$_{a3}$, in which we ignored the dependencies in the distribution of $\mathbf{X}$ in both matches and non-matches, gave biased estimates. In addition, the mean squared error of the estimates from Estimator$_{a3}$ were slightly higher than of Estimator$_{a1}$. All other estimators were highly downward biased to zero and underestimated the coverages of the 95% confidence interval.

Estimated ($E(\hat{\theta})$)

	Scenario 1				Scenario 2				Scenario 3			
	α	β_1	β_2	β_3	α	β_1	β_2	β_3	α	β_1	β_2	β_3
Simulated	*-1.000*	*0.600*	*-0.800*	*0.200*	*1.000*	*0.600*	*-0.800*	*0.200*	*1.000*	*0.600*	*-0.800*	*0.200*
Estimator$_{a1}$	-1.032	0.613	-0.819	0.206	-1.030	0.618	-0.816	0.207	-1.044	0.569	-0.824	0.171
Estimator$_{a2}$	-1.025	0.569	-0.781	0.175	-1.021	0.576	-0.777	0.177	-1.034	0.630	-0.819	0.199
Estimator$_{a3}$	-1.033	0.613	-0.820	0.207	-1.030	0.619	-0.816	0.207	-1.033	0.686	-0.871	0.250
Estimator$_{b}$	-1.034	0.610	-0.817	0.206	-1.031	0.617	-0.813	0.207	-1.005	0.412	-0.559	-0.026
WLS	-1.025	0.616	-0.834	0.212	-1.075	0.596	-0.787	0.199	-1.053	0.460	-0.610	0.044
Naive	-1.159	0.503	-0.675	0.170	-1.158	0.506	-0.668	0.170	-1.130	0.263	-0.360	-0.107
Deterministic	-1.090	0.559	-0.749	0.188	-1.087	0.564	-0.744	0.189	-1.065	0.370	-0.500	-0.028

Mean squared error

	Scenario 1				Scenario 2				Scenario 3			
	α	β_1	β_2	β_3	α	β_1	β_2	β_3	α	β_1	β_2	β_3
Estimator$_{a1}$	0.035	0.037	0.042	0.031	0.036	0.038	0.040	0.031	0.034	0.079	0.046	0.040
Estimator$_{a2}$	0.033	0.041	0.043	0.034	0.034	0.040	0.042	0.032	0.031	0.041	0.044	0.032
Estimator$_{a3}$	0.035	0.037	0.041	0.031	0.036	0.038	0.039	0.031	0.033	0.047	0.047	0.033
Estimator$_{b}$	0.035	0.036	0.040	0.031	0.036	0.037	0.039	0.030	0.025	0.070	0.098	0.078
WLS	0.035	0.035	0.042	0.032	0.038	0.032	0.033	0.027	0.028	0.051	0.071	0.048
Naive	0.052	0.034	0.043	0.022	0.054	0.033	0.043	0.021	0.039	0.135	0.217	0.110
Deterministic	0.036	0.030	0.033	0.025	0.037	0.031	0.033	0.025	0.026	0.078	0.115	0.071

Coverage of the 95% confidence interval

	Scenario 1				Scenario 2				Scenario 3			
	α	β_1	β_2	β_3	α	β_1	β_2	β_3	α	β_1	β_2	β_3
Estimator$_{a1}$	0.941	0.949	0.941	0.951	0.942	0.937	0.954	0.944	0.948	0.853	0.927	0.887
Estimator$_{a2}$	0.938	0.935	0.930	0.936	0.947	0.936	0.933	0.941	0.947	0.939	0.934	0.944
Estimator$_{a3}$	0.941	0.949	0.941	0.950	0.941	0.939	0.956	0.943	0.947	0.944	0.939	0.952
Estimator$_{b}$	0.943	0.954	0.944	0.952	0.942	0.946	0.956	0.946	0.961	0.747	0.644	0.669
WLS	0.955	0.974	0.953	0.958	0.940	0.958	0.963	0.950	0.955	0.844	0.801	0.860
Naive	0.836	0.911	0.869	0.965	0.828	0.908	0.865	0.962	0.877	0.356	0.177	0.348
Deterministic	0.916	0.944	0.932	0.949	0.913	0.933	0.926	0.948	0.948	0.657	0.473	0.634

Table 4.2: Bias, mean squared error and coverage for all seven estimators in the scenarios described in table 4.1.

Real data example

To illustrate our procedure, we estimated the impact of risk factors for a preterm second delivery with data from a prospective nationwide cohort. Data from this cohort is stored in the Perinatal Registry Netherlands (PRN) registry, which contains data on pregnancies, deliveries and (re)admissions until 28 days after birth. The coverage is about 96% of all births in the Netherlands [91]. We used pregnancy data from 2004 till 2008 which resulted in 393302 first successful deliveries and 312871 second successful deliveries.

The problem of the PRN registry is that the deliveries are individually registered and there is no unique identifier to determine which first and second deliveries are from the same mother. Therefore, we used seven partially identifying variables (see table 4.3) with binary agreement/disagreement values. For the k^{th} partially identifying variable $\mathbf{R}_k$, this resulted in the binary comparison outcome g_{ijk} for each first delivery i and each second delivery j. There was a large number of missing values in the partially identifying variables and we assumed that missing values did not contribute to the likelihood of the record linkage mixture model. Therefore we also defined the vector h_{ijk} which was 1 if $\mathbf{R}_k$ was available for both first delivery i and second delivery j and 0 otherwise. We excluded comparisons in which the second delivery happened before the first delivery. Therefore we defined a binary indicator z_{ij}, which was 1 if the second delivery j happened at least 22 weeks after the first delivery i and 0 otherwise. The likelihood of the comparison vector data $\mathbf{g}_{ij}$ was then defined as

$$\prod_{i=1}^{n}\prod_{j=1}^{m}\pi\prod_{k=1}^{7}z_{ij}\left[\mu_k^{g_{ijk}}(1-\mu_k)^{(1-g_{ijk})}\right]^{h_{ijk}}+ (1-\pi)\prod_{k=1}^{7}\left[\upsilon_k^{g_{ijk}}(1-\upsilon_k)^{(1-g_{ijk})}\right]^{h_{ijk}}, \tag{4.11}$$

where π is the relative frequency of true matches among all comparison vectors, μ_k the probability of agreement of the k^{th} linkage variable among matches, and υ_k the probability of agreement of the k^{th} variable among non-matches.

After estimating the parameters from the record linkage model (μ, υ, π), we restructured our data with the procedure proposed in Section 4 because our data was far too big to analyze directly ($393302 \times 312871 \approx 1.2 \times 10^{11}$ record pairs). Based on the results from our record linkage model (see table 4.4), we used a threshold

Record linkage variables			
First deliveries ($n = 393302$)		Second deliveries ($m = 312871$)	
$\mathbf{R}_1$	Place of residence (zipcode)	$\mathbf{R}_1$	Place of residence (zipcode)
$\mathbf{R}_2$	Year of birth mother	$\mathbf{R}_2$	Year of birth mother
$\mathbf{R}_3$	Month of birth mother	$\mathbf{R}_3$	Month of birth mother
$\mathbf{R}_4$	Day of birth mother	$\mathbf{R}_4$	Day of birth mother
$\mathbf{R}_5$	Year of birth child	$\mathbf{R}_5$	Year of birth previous child
$\mathbf{R}_6$	Month of birth child	$\mathbf{R}_6$	Month of previous child
$\mathbf{R}_7$	Day of birth child	$\mathbf{R}_7$	Day of previous child
Regression variables			
First deliveries ($n = 393302$)		Second deliveries ($m = 312871$)	
$\mathbf{X}_1$	Mother's age < 20 years (Yes/No)	Y	Pregnancy duration (in weeks)
$\mathbf{X}_2$	Mother's age ≥ 35 years (Yes/No)		
$\mathbf{X}_3$	Low Social economic status (Yes/No)		
$\mathbf{X}_4$	Pregnancy duration (in weeks)		

Table 4.3: Partially identifying variables and regression data available from the PRN registry.

T of 0.01 which gave us a manageable number of record pairs. For the record pair ij with a posterior probability lower than T, $L(\mathbf{g}_{ij}|d_{ij} = 1)$ and $L(\mathbf{g}_{ij}|d_{ij} = 0)$ were set to respectively zero and one. All record pairs with a probability higher than 0.01 were included in the analysis with their original weights.

			Comparison patterns						
$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{R}_3$	$\mathbf{R}_4$	$\mathbf{R}_5$	$\mathbf{R}_6$	$\mathbf{R}_7$	z	Frequency	$p(d_{ij} = 1\|\mathbf{g}_{ij})$
Missing	1	0	1	1	1	1	1	9015	0.01
1	0	1	0	1	1	1	1	255	0.01
1	0	0	1	1	1	1	1	507	0.01
Missing	0	1	1	1	1	1	1	24043	0.02
0	1	1	1	1	1	Missing	1	4	0.14
1	1	1	1	1	1	0	1	41	0.14
1	Missing	Missing	Missing	1	1	Missing	1	4	0.31
1	1	1	1	0	1	1	1	58	0.32
Missing	1	1	1	1	1	Missing	1	2	0.56
1	1	1	1	1	0	1	1	38	0.65
0	1	1	1	1	1	1	1	9893	0.75
1	Missing	Missing	Missing	1	1	1	1	41761	0.89
1	1	1	0	1	1	1	1	211	0.93
1	1	0	1	1	1	1	1	197	0.94
Missing	1	1	1	1	1	1	1	19181	0.96
1	0	1	1	1	1	1	1	846	0.97
1	1	1	Missing	1	1	1	1	3	1.00
1	1	1	1	1	1	1	1	31898	1.00

Table 4.4: Unique comparison vectors sorted on their posterior probability of a match $p(d_{ij} = 1|\mathbf{g}_{ij})$.

We used the following methods to estimate the relation between the outcome and covariates in matches ($E[\mathbf{y}|\mathbf{X},\mathbf{M}]$) in the restructured data

- Estimator$_a$: Both $L(\mathbf{x}_i|d_{ij}=1)$ and $L(\mathbf{x}_i|d_{ij}=0)$ follow multinomial distributions.
- Estimator$_b$
- WLS: Include all record pairs ij with $p(d_{ij}=1|\mathbf{g}_{ij})>0.9$.
- Naive
- Deterministic$_a$: Include all record pairs ij with $p(d_{ij}=1|\mathbf{g}_{ij})>0.90$.
- Deterministic$_b$: Include all record pairs ij with $p(d_{ij}=1|\mathbf{g}_{ij})>0.70$.

Detailed descriptions of the regression methods are given in section 4. In our data, pregnancy duration was always observed. We performed two types of regression analyses; an analysis with a continuous $\mathbf{y}$ (pregnancy duration of the second delivery) using normally distributed residuals for $\mathbf{y}$ and an analysis with a binary $\mathbf{y}$ (pregnancy duration of the second delivery $<$ 37 weeks or not).

In the WLS and Deterministic approaches, we relied on an arbitrary threshold to distinguish true matches from falsely identified matches. For this analysis the threshold was put on 0.9 for the WLS approach and the Deterministic$_a$ approach. With this threshold, we had 52336 combinations of first and second deliveries with about 817 (1.6%) false matches. In addition, we also fitted the Deterministic$_b$ approach to the data, where the threshold was put at 0.7 (103990 matches with about 7773 (7.5%) false matches).

We used the estimated distributions $L(\mathbf{x}_i|d_{ij}=1)$ and $L(\mathbf{x}_i|d_{ij}=0)$ from Estimator$_a$ to investigate whether both distributions were different. For continuous and binary $\mathbf{y}$, both distributions $L(\mathbf{x}_i|d_{ij}=1)$ and $L(\mathbf{x}_i|d_{ij}=0)$ were almost similar (data not shown) and therefore Estimator$_a$ and Estimator$_b$ should give approximately similar estimates.

The results from the different estimators have been summarized in table 4.5 for the continuous outcome. As expected, Estimator$_a$ and Estimator$_b$ gave approximately similar effects, signs, and significance of the covariates. Compared to Estimator$_a$, the WLS approach and the Deterministic$_a$ approach gave somewhat smaller estimated effects of the covariates. This was especially seen in the effect of mother's age $<$ 20 (-0.337, -0.263, and -0.269 with respectively Estimator$_a$, WLS, and Deterministic$_a$) and for the effect of pregnancy duration (0.250, 0.208,

and 0.210 with respectively Estimator$_a$, WLS, and Deterministic$_a$). The WLS approach seemed overoptimistic about the standard errors of the estimates, which were small compared to Estimator$_a$. The standard errors of the Deterministic$_a$ approach were comparable to the standard errors of Estimator$_a$.

With the Naive and the Deterministic$_b$ approach the estimated effects of mother's age ≥ 35, mother's age < 20, and low social economic status were comparable to the estimated effects from Estimator$_a$. Moreover, the estimated effect of pregnancy duration was slightly lower. The standard errors for all covariates in the Deterministic$_b$ approach were comparable to the standard errors obtained with Estimator$_a$. The Naive approach, however, gave standard errors which were about 10 times lower than Estimator$_a$.

For a binary outcome $\mathbf{y}$, the results were comparable to the situation with a continuous $\mathbf{y}$ (data not shown). The estimated regression coefficients from Estimator$_a$ and Estimator$_b$ were almost the same. The results from the WLS, Naive, and Deterministic$_b$ approaches were also comparable. The estimated effects with the Deterministic$_a$ approach were, with exception of low social economic status, smaller.

	Estimator$_a$				Estimator$_b$			
Coefficients	Est.	S.E.	Z-value	P-value	Est.	S.E.	Z-value	P-value
intercept	38.727	0.010	—	—	38.744	0.010	—	—
mother's age $<$ 20 years	-0.337	0.050	-6.78	$<$0.001	-0.330	0.056	-5.94	$<$0.001
mother's age $\geq$ 35 years	-0.111	0.022	-5.05	$<$0.001	-0.107	0.025	-4.32	$<$0.001
low social economic status	-0.070	0.016	-4.30	$<$0.001	-0.072	0.019	-3.87	$<$0.001
pregnancy duration (in weeks)	0.250	0.003	87.58	$<$0.001	0.241	0.003	79.23	$<$0.001
	WLS ($p > 0.90$)				Naive			
Coefficients	Est.	S.E.	Z-value	P-value	Est.	S.E.	Z-value	P-value
intercept	38.654	0.003	—	—	38.810	0.001	—	—
mother's age $<$ 20 years	-0.263	0.015	-17.36	$<$0.001	-0.312	0.005	-65.20	$<$0.001
mother's age $\geq$ 35 years	-0.080	0.006	-12.92	$<$0.001	-0.101	0.002	-47.35	$<$0.001
low social economic status	-0.086	0.005	-16.49	$<$0.001	-0.069	0.002	-44.12	$<$0.001
pregnancy duration (in weeks)	0.208	0.001	314.03	$<$0.001	0.215	$<$0.001	864.88	$<$0.001
	Deterministic$_a$ ($p > 0.90$)				Deterministic$_b$ ($p > 0.70$)			
Coefficients	Est.	S.E.	Z-value	P-value	Est.	S.E.	Z-value	P-value
intercept	38.647	0.013	—	—	38.816	0.008	—	—
mother's age $<$ 20 years	-0.269	0.076	-3.52	$<$0.001	-0.320	0.045	-7.15	$<$0.001
mother's age $\geq$ 35 years	-0.074	0.031	-2.37	0.018	-0.101	0.020	-4.94	$<$0.001
low social economic status	-0.086	0.026	-3.25	0.001	-0.071	0.015	-4.78	$<$0.001
pregnancy duration (in weeks)	0.210	0.003	62.75	$<$0.001	0.216	0.002	90.69	$<$0.001

Table 4.5: Results for a continuous outcome $\mathbf{y}$: estimated regression coefficients (Est.) with their standard error (S.E.), Z-value, and P-value for all six regression approaches.

Discussion

In this chapter, we proposed a flexible approach to analyze linked data. The EM-type of optimization that we use to maximize the pseudo-likelihood of the data, allows us to specify the most appropriate (conditional) density for each part of the pseudo-likelihood. Not only is the performance of our new estimator comparable to the current existing estimators when their restrictions on the data structure hold, it also produces accurate estimates in more complex situations. The results from the simulation showed that the variance approximation from section 4 based on the inverse of the Hessian matrix gives accurate variance estimations, which was reflected by the good coverages of the 95% confidence intervals.

In our simulations, we did not consider the situation in which some linkage variables did not agree in the set of matches. However, this problem was clearly present in the real data. For instance, consider the variable $\mathbf{R}_1$, which is the place of residence of the mother. Between pregnancies some mothers move to another address and therefore the place of residence of the first and second delivery is different. Ignoring these mothers led to considerably different results. For instance, there were substantial differences between the Deterministic$_a$ and Deterministic$_b$ approaches in our real data example (table 4.5, which shows that the choice of which combinations to include in the analysis has great impact on the estimated regression coefficients. A great advantage of our newly proposed estimator is that this arbitrary threshold is not necessary. Mothers who moved to another address were always included in the analysis, but the fact that the place of residence was not similar in both datasets lowered the contribution of this mother to the part of the likelihood regarding matches.

In this chapter we have assumed that all covariates were located in dataset **A** and the outcome in dataset **B**. However, this assumption is not necessary and the model can easily be extended to a situation in which covariates are located in both **A** and **B**.

Although our new estimator showed promising results, there are still some unresolved problems. We simplified the pseudo-likelihood by assuming that, given the match status, the comparison outcomes $\mathbf{g}_{ij}$ from the linkage problem are independent of the covariates $\mathbf{x}_i$ and the outcome y_j from the regression problem. Although this seems to be a valid assumption for most record linkage problems, there may be situations in which this assumption does not hold. Consider for instance the real data example, in which we used place of residence as a linkage variable. When some mothers change their place of residence in between pregnancies,

the zipcode registered for the first delivery does not match with the second delivery. When the moved mother has a considerably different outcome $\mathbf{y}$ and different covariates $\mathbf{X}$ than the mothers that still live on the same address, the independence assumption is not valid. While we did not investigate this phenomenon, we could reformulate the pseudo-likelihood in (4.5) and include $\mathbf{g}_{ij}$ in the parts concerning the conditional distributions $L(y_j|\mathbf{x}_i, d_{ij} = 1)$ and $L(y_j|d_{ij} = 0)$. In this pseudo-likelihood, we have the conditional distributions $L(y_j|\mathbf{x}_i, \mathbf{g}_{ij}, d_{ij} = 1)$ and $L(y_j|\mathbf{g}_{ij}, d_{ij} = 0)$, which are considerably harder to parameterize. More research is needed to determine the effects of correlation between the linkage problem and the regression problem. In addition, further research is needed to develop model selection procedures to characterize all parts of the pseudo-likelihood. Related to this problem, more research is needed to investigate the robustness of our model to miss-specification.

In conclusion, our new estimator gave accurate results in both simulations and real data. It gives high flexibility to a researcher who is challenged with the analysis of data derived from record linkage.

Appendix

Validity pseudo-likelihood

Our simulations show that the pairwise pseudo-likelihood approach gives almost unbiased parameter estimates and excellent coverage of the confidence intervals (see section 4). Here we will show that the pseudo-log-likelihood estimates are approximately equivalent to estimates derived from maximizing a full joint-log-likelihood. Throughout this section, we use the same assumptions as we have used in the manuscript; matches are independently distributed, there is no relation between the outcome y_j and covariates $\mathbf{x}_i$ in non-matches ($d_{ij} = 0$), no information is contained in $\mathbf{g}_{ij}$ with respect to both the distribution of y_j and of $\mathbf{x}_i$ ($y_j, \mathbf{x}_i \perp \mathbf{g}_{ij}|d_{ij}$) (see page 65).

Under the assumption of independently distributed records in the set of matches, the full joint-likelihood L_f of the all outcomes $\mathbf{y}$, covariates $\mathbf{X}$, matching comparison outcomes $\mathbf{G}$, and true matching status d equals

$$\begin{aligned} &L_f(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d}) = p(d)L_f(\mathbf{y}, \mathbf{X}, \mathbf{G}|d) = \\ &p(d) \times \prod_{i=1}^{n}\prod_{j=1}^{m}\left[L(y_j|\mathbf{x}_i)^{d_{ij}}L(\mathbf{x}_i)^{d_{ij}}\right] \times \prod_{i=1}^{n}\prod_{j=1}^{m}\left[L(\mathbf{g}_{ij})^{d_{ij}}\right] \times \\ &\prod_{i=1}^{n}\prod_{j=1}^{m}\left[L(\mathbf{g}_{ij})^{(1-d_{ij})}\right] \times \prod_{i=1}^{n}\left[L_{nm}(\mathbf{x}_i)^{(1-\sum_{j=1}^{m} d_{ij})}\right] \times \\ &\prod_{j=1}^{m}\left[L_{nm}(y_j)^{(1-\sum_{i=1}^{n} d_{ij})}\right], \end{aligned} \tag{4.12}$$

where d is the $n \times m$ vector of matching-indicator values, $p(d)$ is the probability of observing d, $L(y_j|\mathbf{x}_i)$ is the conditional distribution of y_j given $\mathbf{x}_i$ in the records that form a true match, $L(\mathbf{x}_i)$ is the distribution of the covariates $\mathbf{x}_i$ in records from dataset $\mathbf{A}$ that have a match in dataset $\mathbf{B}$, $L_{nm}(\mathbf{x}_i)$ is the distribution of $\mathbf{x}_i$ in records from $\mathbf{A}$ that do not have a match, and $L_{nm}(y_j)$ is the distribution of outcome y_j in records of $\mathbf{B}$ without a match in $\mathbf{A}$. Since we assume that records from $\mathbf{A}$ have either one match or no match in $\mathbf{B}$ and the other way around, the subvectors $\mathbf{d}_i = (d_{i1}, ..., d_{im})$ and $d_j = (d_{1j}, ..., d_{nj})$ contain only zero's (i.e. record i or j have no match) or contain only one value $d_{ij} = 1$ (i.e. record i or j have a match), which also implies that $\sum_{i=1}^{n} d_{ij}$ (and $\sum_{j=1}^{m} d_{ij}$) are zero for records in

A that do not have a match in **B** and are 1 for records in **A** that do have a match in **B** (and the other way around). The second line of the full joint-likelihood in (4.12) therefore pertains to the records in **A** and **B** that do form a match and the third line refers to the records in **A** that do not have a match in **B** and to the records in **B** that do not have a match in **A**.

Because the vector **d** which contains the matching status of each record pair is not observed, the parameters in (4.12) must be estimated by maximizing the expectation of the log of (4.12) over all possible values of **d**, i.e.

$$E_{\mathbf{d}}[\log[L_f(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d})]] = \sum_{\mathbf{h} \in \mathbf{H}} \log[L_f(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d} = \mathbf{h})] p(\mathbf{d} = \mathbf{h} | \mathbf{y}, \mathbf{X}, \mathbf{G}),$$

where **H** is the set of all possible values of **d**. A problem of using the full joint-likelihood is that we have to maximize $E_{\mathbf{d}}[\log[L(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d})]]$ over **H**, which is, even for moderate n and m, a very large set. Therefore, we propose to use the pairwise pseudo-likelihood

$$E_{\mathbf{d}}[\log[L(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d})]] = \sum_{\mathbf{h} \in \mathbf{H}} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \log[L(y_j, \mathbf{x}_i, \mathbf{g}_{ij}, d_{ij} = h_{ij})] \right] p(\mathbf{d} = \mathbf{h} | \mathbf{y}, \mathbf{X}, \mathbf{G}).$$

Now, define $\mathbf{H}_{ij}^1$ as the subset of **H** of all possible values of **d** with $d_{ij} = 1$ and $\mathbf{H}_{ij}^0$ the subset of all possible values of **d** with $d_{ij} = 0$. The marginalized probabilities $p(d_{ij} = 1 | y_j, \mathbf{x}_i, \mathbf{g}_{ij})$ and $p(d_{ij} = 0 | y_j, \mathbf{x}_i, \mathbf{g}_{ij})$ can be specified as

$$p(d_{ij} = 1 | y_j, \mathbf{x}_i, \mathbf{g}_{ij}) = \sum_{\mathbf{h}^1 \in \mathbf{H}_{ij}^1} p(\mathbf{d} = \mathbf{h}^1 | \mathbf{y}, \mathbf{X}, \mathbf{G}),$$
$$p(d_{ij} = 0 | y_j, \mathbf{x}_i, \mathbf{g}_{ij}) = \sum_{\mathbf{h}^0 \in \mathbf{H}_{ij}^0} p(\mathbf{d} = \mathbf{h}^0 | \mathbf{y}, \mathbf{X}, \mathbf{G}).$$

where $\mathbf{h}^1$ and $\mathbf{h}^0$ are possible values of **d** with $d_{ij} = 1$ and $d_{ij} = 0$, respectively. The expectation of the log of the joint pseudo-likelihood can be written as

$$
\begin{aligned}
&E_{\mathbf{d}}[\log[L(\mathbf{y},\mathbf{X},\mathbf{G},\mathbf{d})]] \\
&= \sum_{\mathbf{h}\in\mathbf{H}} \left[\sum_{i=1}^{n}\sum_{j=1}^{m} \log[L(y_j,\mathbf{x}_i,\mathbf{g}_{ij},d_{ij}=h_{ij})] \right] p(\mathbf{d}=\mathbf{h}|\mathbf{y},\mathbf{X},\mathbf{G}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \left[\log[L(y_j,\mathbf{x}_i,\mathbf{g}_{ij},d_{ij}=1)] \sum_{\mathbf{h}^1\in\mathbf{H}^1_{ij}} p(\mathbf{d}=\mathbf{h}^1|\mathbf{y},\mathbf{X},\mathbf{G}) \right] + \\
&\qquad \sum_{i=1}^{n}\sum_{j=1}^{m} \left[\log[L(y_j,\mathbf{x}_i,\mathbf{g}_{ij},d_{ij}=0)] \sum_{\mathbf{h}^0\in\mathbf{H}^0_{ij}} p(\mathbf{d}=\mathbf{h}^0|\mathbf{y},\mathbf{X},\mathbf{G}) \right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \log[L(y_j,\mathbf{x}_i,\mathbf{g}_{ij},d_{ij}=1)] p(d_{ij}=1|y_j,\mathbf{x}_i,\mathbf{g}_{ij}) + \\
&\qquad \sum_{i=1}^{n}\sum_{j=1}^{m} \log[L(y_j,\mathbf{x}_i,\mathbf{g}_{ij},d_{ij}=0)] p(d_{ij}=0|y_j,\mathbf{x}_i,\mathbf{g}_{ij}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \log[L(y_j,\mathbf{x}_i|d_{ij}=1) L(\mathbf{g}_{ij}|d_{ij}=1) L(d_{ij}=1)] p(d_{ij}=1|y_j,\mathbf{x}_i,\mathbf{g}_{ij}) + \\
&\qquad \sum_{i=1}^{n}\sum_{j=1}^{m} \log[L(y_j|d_{ij}=0) L(\mathbf{x}_i|d_{ij}=0) L(\mathbf{g}_{ij}|d_{ij}=0) L(d_{ij}=0)] \times \\
&\qquad\qquad p(d_{ij}=0|y_j,\mathbf{x}_i,\mathbf{g}_{ij}),
\end{aligned}
$$

which is the expected log pseudo-likelihood which we use for the parameter estimation (see (4.7)). Thus, instead of taking conditional expectations given all observed data, we can take conditional expectations of the considered record pair.

It is important to stress the fact that the distribution of the covariates x for records in $\mathbf{A}$ without a match in $\mathbf{B}$, $L_{nm}(\mathbf{x}_i)$, in the full joint-likelihood is different from the distribution of non-matching records, $L(\mathbf{x}_i|d_{ij}=0)$, in the joint pseudo-likelihood. The latter is a mixture of $\mathbf{X}$ in the records with a match in $\mathbf{B}$ ($L(\mathbf{x}_i)$) and the distribution of $\mathbf{X}$ in records without a match in $\mathbf{B}$ ($L_{nm}(\mathbf{x}_i)$), to which each record from $\mathbf{A}$ contributes $\sum_{j=1}^{m} p(d_{ij}=0|y_j,\mathbf{x}_i,\mathbf{g}_{ij})$ times. For y_j, we have a similar difference.

Our main interest concerns the estimation of the parameters in the conditional distribution $L(y_j|\mathbf{x}_i,d_{ij}=1)$, for which it is easy to show that its estimation

equation in the M-step of the EM-algorithm is the same in the full-likelihood and pseudo-likelihood. Similarly to the main chapter, let $L(y_j|\mathbf{x}_i, d_{ij} = 1)$ be parameterized by the vector β. Taking derivatives with respect to β in respectively the full-likelihood and the pseudo-likelihood, we get

$$\begin{aligned}
&\frac{\partial E_{\mathbf{d}}[\log[L_f(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d})]]}{\partial \beta} \\
&= \sum_{\mathbf{h} \in \mathbf{H}} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\partial \log[L(y_j|\mathbf{x}_i, d_{ij} = 1; \beta)]}{\partial \beta} \right] p(\mathbf{d} = \mathbf{h}|\mathbf{y}, \mathbf{X}, \mathbf{G}) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{\partial \log[L(y_j|\mathbf{x}_i, d_{ij} = 1; \beta)]}{\partial \beta} p(d_{ij} = 1|y_j, \mathbf{x}_i, \mathbf{g}_{ij}) \right],
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial E_{\mathbf{d}}[\log[L(\mathbf{y}, \mathbf{X}, \mathbf{G}, \mathbf{d})]]}{\partial \beta} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{\partial \log[L(y_j|\mathbf{x}_i, d_{ij} = 1; \beta)]}{\partial \beta} p(d_{ij} = 1|y_j, \mathbf{x}_i, \mathbf{g}_{ij}) \right].
\end{aligned}$$

Although both derivatives are similar, the probability $p(d_{ij} = 1|y_j, \mathbf{x}_i, \mathbf{g}_{ij})$ calculated in the E-step of the EM-algorithm is different. However, the probabilities $p(d_{ij} = 1|y_j, \mathbf{x}_i, \mathbf{g}_{ij})$ from the pseudo-likelihood are close to the real probabilities.

EM-algorithm

The EM-algorithm seeks to find the maximum of (4.7) by iterating between an expectation step and a maximization step. In iteration t, we can write the model parameters as $\theta^t = (\beta^t, \gamma^t, \zeta^t, \eta^t)$. Before we start the EM-algorithm, we initialize $\hat{\theta}^{(0)}$ to some random values. Note that with certain starting vector $\hat{\theta}^{(0)}$, the algorithm converges to a local maximum. Therefore, we recommend to fit the EM-algorithm with several different random initial starting vectors $\hat{\theta}^{(0)}$. From all the converged solutions obtained with different initial starting vectors, we chose the vector with the highest maximum pseudo-likelihood from (4.6) to be the best estimate $\hat{\theta}$.

The expectation step of the t^{th} iteration, the posterior
log-odds$(d_{ij} = 1|y_j, \mathbf{x}_i, \mathbf{g}_{ij}, \hat{\theta}^{(t-1)})$ for record pair ij is calculated as

$$\text{log-odds}(d_{ij} = 1|y_j, \mathbf{x}_i, \mathbf{g}_{ij}; \hat{\theta}^{(t-1)}) = \log\left[\frac{L(y_j|\mathbf{x}_i, d_{ij} = 1; \hat{\beta}^{(t-1)})}{L(y_j|d_{ij} = 0; \hat{\gamma}^{(t-1)})}\right] + \log\left[\frac{L(\mathbf{x}_i|d_{ij} = 1; \hat{\zeta}^{(t-1)})}{L(\mathbf{x}_i|d_{ij} = 0; \hat{\eta}^{(t-1)})}\right] + \log\left[\frac{w_{ij1}}{w_{ij0}}\right], \tag{4.13}$$

where $\hat{\theta}^{(t-1)}$ is the parameter estimate from the previous iteration $t-1$. In the maximization step, the parameter vector $\hat{\theta}^{(t)}$ is estimated by maximizing

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{m} E_{\mathbf{d}}\left[\log[L(y_j, \mathbf{x}_i, \mathbf{g}_{ij}, d_{ij}; \hat{\theta}^{(t)})|\hat{\theta}^{(t-1)}]\right] = \\
&\sum_{i=1}^{n}\sum_{j=1}^{m}\{\log[L(y_j|\mathbf{x}_i, d_{ij} = 1; \hat{\beta}^{(t)})] + \log[L(\mathbf{x}_i|d_{ij} = 1; \hat{\zeta}^{(t)})] + \\
&\qquad \log[w_{ij1}]\}\ q_{ij1}\ + \{\log[L(y_j|d_{ij} = 0; \hat{\gamma}^{(t)})] + \\
&\qquad \log[L(\mathbf{x}_i|d_{ij} = 0; \hat{\eta}^{(t)})] + \log[w_{ij0}]\}\ q_{ij0},
\end{aligned} \tag{4.14}$$

where $q_{ij1} = p(d_{ij} = 1|y_j, \mathbf{x}_i, \mathbf{g}_{ij}, \hat{\theta}^{(t-1)})$ and $q_{ij0} = 1 - q_{ij1}$ are fixed. The parameter vector $\hat{\theta}^{(t)} = (\hat{\beta}^{(t)}, \hat{\gamma}^{(t)}, \hat{\zeta}^{(t)}, \hat{\eta}^{(t)})$ can be obtained by separately maximizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \log[L(y_j|\mathbf{x}_i, d_{ij} = 1; \hat{\beta}^{(t)})]\ q_{ij1}, \quad \text{and}$$
$$\sum_{i=1}^{n}\sum_{j=1}^{m} \log[L(y_j|d_{ij} = 0; \hat{\gamma}^{(t)})]\ q_{ij0},$$

which can be done with standard weighted regression procedures. Parameters of the distributions of $\mathbf{X}$ for matches and non-matches can be estimated by separately maximizing

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{m}\log[L(\mathbf{x}_i|d_{ij}=1;\hat{\zeta}^{(t)})]\ q_{ij1}, \qquad \text{and}\\ &\sum_{i=1}^{n}\sum_{j=1}^{m}\log[L(\mathbf{x}_i|d_{ij}=0;\hat{\eta}^{(t)})]\ q_{ij0}.\end{aligned} \tag{4.15}$$

The complexity of $L(\mathbf{x}_i|d_{ij}=1;\zeta)$ and $L(\mathbf{x}_i|d_{ij}=0;\eta)$ depends on the types of covariates in the data. When $\mathbf{X}$ can be written as a set of disjoint categories, we can use a multinomial distribution to parametrize both distributions. When we have both continuous and categorical variables, we could categorize the continuous variables. However, we lose accuracy depending on the number of categories we use to split the continuous variables.

To avoid this loss of accuracy when we only have continuous variables $\mathbf{X}$, we suggest to write both $L(\mathbf{x}_i|d_{ij}=1;\zeta)$ and $L(\mathbf{x}_i|d_{ij}=0;\eta)$ in terms of univariate marginal distributions and a copula which describes the dependency structure [92]. With this specification, we have $\zeta=\rho^{(1)}$ and $\eta=\rho^{(0)}$, where $\rho^{(\ell)}$ is the vector of dependency parameters for respectively matches ($\ell=1$) and non-matches ($\ell=0$).

The univariate cumulative distribution of each variable $\mathbf{X}_e$, where $(e=1,\ldots,p)$, given the match status $d=\ell$ can be estimated with weighted uni-variate empirical functions $F_{s,d}(z)$ defined as

$$F_{e,\ell}(z)=\frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{m}q_{ij\ell}}\sum_{i=1}^{n}\sum_{j=1}^{m}q_{ij\ell}\{x_{ie}\le z\}, \tag{4.16}$$

where $q_{ij\ell}$ is the posterior probability of a match ($\ell=1$) or non-match ($\ell=0$). The p-dimensional distribution of $L(\mathbf{x}_i|d=\ell;\rho^{(\ell)})$ can be written as

$$\begin{aligned}&L(x_{i1},x_{i2},\ldots,x_{ip}|d_{ij}=\ell,\rho^{(\ell)})=\\ &\qquad\left[c(F_{1,\ell}(x_{i1}),F_{2,\ell}(x_{i2}),\ldots,F_{p,\ell}(x_{ip});\rho^{(\ell)})\prod_{e=1}^{p}f_{e,\ell}(x_{ie})\right]^{q_{ij\ell}},\end{aligned} \tag{4.17}$$

where $f_{e,\ell}(x_{ie})$ is the univariate density function of the e^{th} variable, $c(u_1,u_2,\ldots,u_p;\rho^{(\ell)})=\frac{\partial C(u_1,u_2,\ldots,u_p;\rho^{(\ell)})}{\partial u_1\partial u_2\ldots\partial u_p}$ is the density of the p-dimensional parametric copula $C(u_1,u_2,\ldots,u_p;\rho^{(\ell)})$ with the vector of dependency parameters $\rho^{(\ell)}$.

To obtain the univariate density function $f_{e,\ell}(x_{ie})$ we use monotonic increasing polynomial splines bounded between zero and one to approximate $F_{e,\ell}(x_{ie})$ [93]. The first derivative of this spline is $f_{e,\ell}(x_{ie})$. When we write $L(\mathbf{x}_i|d_{ij} = \ell; \rho^{(\ell)})$ as (4.17), we can decompose the distributions from (4.15) into

$$\sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij\ell} \ \log\left[\prod_{e=1}^{p} f_{e,\ell}(x_{ie})\right], \qquad \text{and} \tag{4.18a}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij\ell} \ \log\left[c(F_{1,\ell}(x_{i1}), F_{2,\ell}(x_{i2}), \ldots, F_{p,\ell}(x_{ip}); \hat{\rho}^{(\ell)})\right], \tag{4.18b}$$

and both parts can be optimized separately. Firstly the spline functions describing the univariate cumulative densities from (4.18a) are estimated. Secondly the dependency parameters for matches ($\hat{\rho}^{(1)}$) and non-matches ($\hat{\rho}^{(0)}$) are estimated by maximizing (4.18b). The estimation of the dependency parameters $\hat{\rho}^{(1)}$ and $\hat{\rho}^{(0)}$ is demanding, especially in higher dimensions. To decrease the number of parameters in the model, we could assume independence between $\mathbf{X}$ which means that we do not have to model the dependency parameters. The dependency structure can also be ignored in matches or non-matches only. Robustness of our model, ignoring the dependency structure, is evaluated in our simulation described in section 4. There are many alternatives to estimate $L(x_i|d_{ij} = \ell)$ such as kernel estimation methods [94, 95] or mixture models [96, 97]. Whether these alternatives are more convenient or efficient than copula models is beyond the subject of this chapter.