

Capturing Small Discrepancies in Record Linkage: A Maximum Entropy Framework with Continuous Similarity Measures^{*}

Adam Struzik ^{†, §}

Maciej Beręsewicz ^{‡, §}

August 2026

Abstract

Record linkage (RL) is the process of joining records residing in two datasets that are believed to be related to the same entity. In this paper, we extend the unsupervised maximum entropy classification (MEC) algorithm for record linkage, originally proposed by Lee et al. (2022), to allow for the use of continuous comparison functions. The classical approach relies on binary comparisons. Consequently, small discrepancies between key variable values are overlooked. To address this limitation, we propose incorporating continuous comparison functions by developing two approaches: parametric and nonparametric. The former is based on the hurdle Gamma distribution (Vo et al., 2023), whereas the latter utilizes the Kullback-Leibler Importance Estimation Procedure (KLIEP; see Sugiyama et al. (2008)). The simulation results show that the proposed parametric algorithm outperforms the approach of Lee et al. (2022) across all comparison measures, and the nonparametric one performs better than the original algorithm across the majority of measures.

^{*}Adam Struzik was responsible for developing and implementing the proposed algorithms, conducting the simulation study, and writing the paper, while Maciej Beręsewicz conceptualized the study and supervised the entire project. For the paper, Adam Struzik received an award in the Student Paper Competition organized by the American Statistical Association’s Statistical Computing and Graphics Sections.

[†]Adam Mickiewicz University in Poznań, Poland.

[‡]Poznań University of Economics and Business, Poznań, Poland.

[§]Statistical Office in Poznań, Poland.

1 Introduction

Many statistical analyses (e.g., related to population studies) rely on data collected from different sources. Given two datasets, a researcher must identify the records associated with the same entity. Ideally, data contain unique identifiers (keys). However, in practice, it is often not the case. *Record linkage* (RL) is the term describing the process of joining records that are believed to be related to the same entity (Lee et al., 2022). Performing this process requires using *key variables* that appear in both datasets.

In this setting, the problem is inherently probabilistic. The classical record linkage approach, proposed by Fellegi and Sunter (1969), is based on the likelihood ratio test. Its disadvantage is that some pairs need to be verified and classified by clerical review. Additionally, it assumes independence between measures of agreement of different record pairs, but in real situations, this assumption is rarely fulfilled. To address these limitations, Lee et al. (2022) proposed adapting the maximum entropy method for classification to record linkage. Supervised and unsupervised algorithms were considered. The unsupervised approach is particularly useful since training data are often not available. However, the proposed maximum entropy classification (MEC) algorithm relies on binary comparisons. Consequently, small discrepancies between key variable values are overlooked (e.g., Thomas and Tomas).

To remedy this issue, we propose the following:

1. The use of continuous comparison functions – for character variables, one may use the Jaro-Winkler similarity (Winkler, 1990), while for numerical variables, the Euclidean distance can be considered.
2. A modification of the MEC algorithm that incorporates continuous comparisons – we develop two approaches: parametric and nonparametric. The former is based on the hurdle Gamma distribution (Vo et al., 2023), whereas the latter utilizes the Kullback-Leibler Importance Estimation Procedure (KLIEP; see Sugiyama et al. (2008)).

The results of simulations demonstrate the following:

1. The proposed parametric algorithm outperforms the approach of Lee et al. (2022) across all comparison measures, and the nonparametric one performs better than the original algorithm across the majority of measures.
2. By incorporating continuous measures, the proposed methods effectively capture small discrepancies between records.

The paper is organized as follows. In Section 2, we introduce the notation and the MEC approach. In Section 3, we present the proposed continuous MEC algorithms. In Section 4, we discuss the simulation study and its results. Finally, in Section 5, we conclude the paper. Technical details of the proposed algorithms, as well as plots showing the distributions of the estimated number of matches in the conducted simulation study, are presented in the Appendix.

2 Maximum Entropy Classification for Record Linkage

Consider two datasets without duplicates: A and B . Let the bipartite comparison space $\Omega = A \times B$ consist of matches M and nonmatches U between the records in A and B ($\Omega = M \cup U$). Let $n = |\Omega|$ and $n_M = |M|$. The matching indicator is denoted as g_{ab} , i.e., $g_{ab} = \mathbb{I}((a, b) \in M)$. For any pair of records $(a, b) \in \Omega$, let $\gamma_{ab} = (\gamma_{ab}^1, \gamma_{ab}^2, \dots, \gamma_{ab}^K)'$ be the comparison vector based on a set of K key

variables, where prime denotes transposition. Let $m(\boldsymbol{\gamma})$ and $u(\boldsymbol{\gamma})$ be the conditional probability or density functions of $\boldsymbol{\gamma} = \boldsymbol{\gamma}_{ab}$ given $g_{ab} = 1$ and $g_{ab} = 0$, respectively. Additionally, let $S(M)$ and $S(U)$ be the supports of $\boldsymbol{\gamma}_{ab}$ over M and U , respectively.

Assuming $\mathcal{S}(M) \subseteq \mathcal{S}(U)$, let the probability or density ratio be defined as:

$$r(\boldsymbol{\gamma}) = \frac{m(\boldsymbol{\gamma})}{u(\boldsymbol{\gamma})}. \quad (1)$$

The MEC algorithm aims to minimize the following objective function with respect to r :

$$Q = \sum_{(a,b) \in M} \frac{u(\boldsymbol{\gamma}_{ab})}{n_M(\boldsymbol{\gamma}_{ab})} r(\boldsymbol{\gamma}_{ab}) - \frac{1}{n_M} \sum_{(a,b) \in M} \log r(\boldsymbol{\gamma}_{ab}), \quad (2)$$

where $n_M(\boldsymbol{\gamma}_{ab}) = \sum_{(i,j) \in M} \mathbb{I}(\boldsymbol{\gamma}_{ij} = \boldsymbol{\gamma}_{ab})$. A *classification set* for record linkage, denoted by $\hat{M}$, consists of record pairs from Ω such that any record from A or B appears in at most one record pair within $\hat{M}$ (Lee et al., 2022). Let the *entropy* of a classification set $\hat{M}$ be given by

$$D_{\hat{M}} = \frac{1}{|\hat{M}|} \sum_{(a,b) \in \hat{M}} \log r(\boldsymbol{\gamma}_{ab}). \quad (3)$$

Lee et al. (2022) proposed the following greedy method for constructing the classification set.

Definition 1 (MEC set). A MEC set of given size $n^* = |\hat{M}|$ is the first classification set that is of size n^* , obtained by deduplication in the descending order of $r(\boldsymbol{\gamma}_{ab})$ over Ω .

In this framework, it is possible that $(a, b') \notin \hat{M}$ even if $r(\boldsymbol{\gamma}_{ab'}) > r(\boldsymbol{\gamma}_{a'b'})$ for some $(a', b') \in \hat{M}$. This occurs if there exists $(a, b) \in \hat{M}$ with $r(\boldsymbol{\gamma}_{ab}) > r(\boldsymbol{\gamma}_{ab'})$.

A natural question concerns the construction of $\boldsymbol{\gamma}_{ab}$. Lee et al. (2022) proposed using K binary agreement indicators, defined as $\gamma_{ab}^k = 1$ if records $a \in A$ and $b \in B$ agree on the k -th variable, and $\gamma_{ab}^k = 0$ otherwise. Assuming conditional independence among the components of the comparison vector, we obtain

$$m(\boldsymbol{\gamma}; \boldsymbol{\theta}) = \prod_{k=1}^K \theta_k^{\gamma_{ab}^k} (1 - \theta_k)^{1 - \gamma_{ab}^k}, \quad u(\boldsymbol{\gamma}; \boldsymbol{\xi}) = \prod_{k=1}^K \xi_k^{\gamma_{ab}^k} (1 - \xi_k)^{1 - \gamma_{ab}^k}, \quad (4)$$

where $\theta_k = \mathbb{P}(\gamma_{ab}^k = 1 | g_{ab} = 1)$, $\xi_k = \mathbb{P}(\gamma_{ab}^k = 1 | g_{ab} = 0)$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_K)$, and $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_K)$.

The unsupervised MEC algorithm, proposed by Lee et al. (2022), is iterative. Initially, insofar as $\Omega = M \cup U$ is highly imbalanced, it estimates the parameters of $\boldsymbol{\xi}$ using all pairs from Ω . Subsequently, it iteratively estimates the parameters of $\boldsymbol{\theta}$ (based on the current set of matches), predicts the number of matches, and creates a new classification set (following the procedure described in Definition 1). To estimate n_M , the formula

$$n_M = \sum_{(a,b) \in \Omega} \hat{g}(\boldsymbol{\gamma}_{ab}) \quad (5)$$

is applied, where

$$\hat{g}(\boldsymbol{\gamma}_{ab}) = \min \left\{ \frac{|M| r(\boldsymbol{\gamma}_{ab})}{|M| (r(\boldsymbol{\gamma}_{ab}) - 1) + n}, 1 \right\}, \quad (6)$$

and $|M|$, $r(\boldsymbol{\gamma}_{ab})$ are based on estimates from the previous iteration.

The convergence of this procedure is proven in the Appendix of Lee et al. (2022). The primary advantage of this algorithm over the approach by Fellegi and Sunter (1969) is that it is fully automatic.

3 Algorithms Incorporating Continuous Comparisons

The algorithm presented in Section 2 does not capture small discrepancies between key variable values. We propose the use of continuous comparison functions to address the limitations of binary comparison methods. We consider any semi-metric. For example, we can use $1 - \text{Jaro-Winkler similarity}$ (Winkler, 1990) for character variables or the Euclidean distance for numerical variables.

A continuous comparison function can be modeled using a hurdle distribution, i.e., a distribution with a positive probability of taking the value 0 and a continuous component for positive values. We describe how to incorporate this idea into the MEC procedure using both parametric and nonparametric approaches.

3.1 Continuous Parametric Algorithm

Vo et al. (2023) considered record linkage using the hurdle Gamma distribution, which is defined as follows.

Definition 2 (hurdle Gamma distribution). The density function of a hurdle Gamma distribution is characterized by three parameters $p_0 \in (0, 1)$ and $\alpha, \beta > 0$ as follows:

$$f(x; p_0, \alpha, \beta) = p_0^{\mathbb{I}(x=0)} [(1 - p_0)v(x; \alpha, \beta)]^{\mathbb{I}(x>0)}, \quad (7)$$

where $x \geq 0$ and

$$v(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} \exp(-\beta x)}{\Gamma(\alpha)} \quad (8)$$

is the density function of the Gamma distribution.

Incorporating this distribution into the MEC procedure requires estimating its parameters. To estimate p_0 , we can use the proportion of zeros in a sample. The maximum likelihood estimation of the Gamma distribution parameters was described by Choi and Wette (1969).

Assume that the components of γ are conditionally independent (both in M and U). Let $\gamma^k|M$ and $\gamma^k|U$ follow hurdle Gamma distributions with parameters $p_{0M}^k, \alpha_M^k, \beta_M^k$ and $p_{0U}^k, \alpha_U^k, \beta_U^k$, respectively. Let θ and ξ denote the parameter vectors associated with M and U , respectively.

Assume a common support $\mathcal{S}(M) = \mathcal{S}(U)$. Let M_1 denote the maximal MEC set, consisting of record pairs with perfect and unique agreement on all key variables. An outline of the unsupervised continuous parametric MEC algorithm is provided below.

1. Set initial parameters $\theta^{(0)} = (p_{0M}^{1(0)}, \alpha_M^{1(0)}, \beta_M^{1(0)}, p_{0M}^{2(0)}, \alpha_M^{2(0)}, \beta_M^{2(0)}, \dots, p_{0M}^{K(0)}, \alpha_M^{K(0)}, \beta_M^{K(0)})$ and set $M^{(0)} = M_1$.
2. Utilizing all pairs from Ω , estimate p_{0U}^k using the proportion of zeros in a sample, and estimate α_U^k, β_U^k based on the formulas from Choi and Wette (1969).
3. Given $\hat{\xi} = (\hat{p}_{0U}^1, \hat{\alpha}_U^1, \hat{\beta}_U^1, \hat{p}_{0U}^2, \hat{\alpha}_U^2, \hat{\beta}_U^2, \dots, \hat{p}_{0U}^K, \hat{\alpha}_U^K, \hat{\beta}_U^K)$, calculate $n_M^{(0)}$ using Equation (5).
4. Using the current density ratio values and following the procedure described in Definition 1, find the MEC set $M^{(1)}$ such that $|M^{(1)}| = n_M^{(0)}$.
5. Perform the t -th iteration ($t \geq 1$) as follows.
 - (a) Update the estimates of p_{0M}^k, α_M^k , and β_M^k based on $M^{(t)}$.
 - (b) Calculate $n_M^{(t)}$.

- (c) Find the MEC set $M^{(t+1)}$.
6. Repeat Step 5 until $|n_M^{(t)} - n_M^{(t+1)}| < \delta$ or $\|\theta^{(t)} - \theta^{(t+1)}\| < \varepsilon$, where δ and ε are pre-defined thresholds (small positive values).

A detailed algorithm is presented in Appendix A. The theoretical convergence of this procedure follows the proof outlined in the Appendix of Lee et al. (2022).

3.2 Continuous Nonparametric Algorithm

Instead of assuming that γ follows the hurdle gamma distribution, we can assume that the density ratio r is given by

$$r(\gamma) = \prod_{k=1}^K \left\{ \left(\frac{p_{0M}^k}{p_{0U}^k} \right)^{\mathbb{I}(\gamma^k=0)} \left(\frac{1 - p_{0M}^k}{1 - p_{0U}^k} r_+^k(\gamma^k) \right)^{\mathbb{I}(\gamma^k>0)} \right\}, \quad (9)$$

where p_{0M}^k and p_{0U}^k are unknown parameters, and $r_+^k(\cdot)$ is an unknown function. We modify the algorithm presented in Section 3.1 by changing the density ratio estimation method.

In each iteration, the values of p_{0M}^k and p_{0U}^k are estimated based on the proportion of zeros in the current set of matches and nonmatches, respectively. To estimate the density ratio of the continuous components, i.e., $r_+^k(\cdot)$, we employ the Kullback-Leibler Importance Estimation Procedure (KLIEP), proposed by Sugiyama et al. (2008). This method allows for direct estimation of the density ratio between two continuous distributions. KLIEP models the ratio as a weighted sum of basis functions and formulates an objective function (based on the Kullback-Leibler divergence) to be maximized with respect to the weights. This formulation leads to a convex optimization problem, which is solved by iteratively performing gradient ascent and satisfying the constraint. The KLIEP method is implemented in the `densityratio` R package (Volker et al., 2024). The continuous nonparametric algorithm is detailed in Appendix B.

4 Simulation Study

Using the programming language R (R Core Team, 2025), we have implemented the approaches presented in Sections 3.1 and 3.2, as well as the binary MEC algorithm, proposed by Lee et al. (2022) and outlined in Section 2. To compare these three methods, we conduct a simulation study, analogous to Lee et al. (2022, Section 6, *Scenario-II (Informative)*). The only difference is that in the binary approach, we do not employ the Soundex algorithm (Copas and Hilton, 1990).

We use the CENSUS, CIS, and PRD datasets from McLeod et al. (2011). The considered key variables are: `PERNAME1` (forename), `PERNAME2` (surname), `SEX` (gender), `DOB_DAY` (day of birth), `DOB_MON` (month of birth), and `DOB_YEAR` (year of birth). The assumption of conditional independence of the components of γ (both in M and U) allows us to use different methods for different variables. Therefore, we consistently apply the binary approach for the variables `SEX`, `DOB_DAY`, `DOB_MON`, and `DOB_YEAR`, while for `PERNAME1` and `PERNAME2`, we employ different methods: binary, continuous parametric, and continuous nonparametric. The comparison function for the continuous algorithms is $1 - \text{Jaro-Winkler similarity}$.

For details regarding the sampling procedure, we refer to Lee et al. (2022). The simulation comprises three cases. The true number of matches is set to 400, 250, and 150, respectively. In each case, A and B contain 500 and 1000 records, respectively.

We consider two error measures: the *false link rate* (FLR) and the *missing match rate* (MMR), defined as:

$$\text{FLR} = \frac{1}{|\hat{M}|} \sum_{(a,b) \in \hat{M}} (1 - g_{ab}), \quad \text{MMR} = 1 - \frac{1}{n_M} \sum_{(a,b) \in \hat{M}} g_{ab}. \quad (10)$$

These exact values can be calculated because unique identifiers are available in the considered datasets. Additionally, we compute estimates of these errors (denoted as $\widehat{\text{FLR}}$ and $\widehat{\text{MMR}}$), following the procedure described by Lee et al. (2022). Finally, n_M denotes the true number of matches, while $\hat{n}_M$ represents its estimate, with the standard error calculated from the sample data given in parentheses.

The simulation results are presented in Table 1. The continuous methods outperform the binary approach regarding the estimation of n_M and the true MMR values (particularly when $n_M = 150$). The continuous parametric algorithm yields the lowest FLR, whereas the binary method exhibits slightly lower FLR values than the continuous nonparametric one. Regarding FLR estimation, the continuous parametric algorithm provides the most accurate results. The MMR estimation procedure, following Lee et al. (2022), is almost identical to that of FLR. Therefore, future research should consider improving this estimation method.

As expected, the use of continuous comparisons results in better n_M estimates than those obtained with the binary approach. The details regarding the distributions of the estimated number of matches are presented in Appendix C. The fact that the continuous parametric method outperforms the other considered methods in terms of error rates suggests that the hurdle Gamma distribution is suitable for modeling continuous comparisons based on $1 - \text{Jaro-Winkler}$ similarity.

Table 1: The number of matches and averages of its estimate, averages of error rates and their estimates, over 100 simulations.

n_M	Method	$\hat{n}_M$	FLR	MMR	$\widehat{\text{FLR}}$	$\widehat{\text{MMR}}$
400	Binary	380.5 (12.4)	0.0055	0.0540	0.0370	0.0370
	Continuous parametric	386.7 (7.3)	0.0032	0.0366	0.0114	0.0117
	Continuous nonparametric	386.1 (8.3)	0.0092	0.0437	0.0327	0.0327
250	Binary	237.0 (5.8)	0.0137	0.0652	0.0439	0.0438
	Continuous parametric	241.2 (4.9)	0.0063	0.0415	0.0122	0.0125
	Continuous nonparametric	241.6 (6.4)	0.0185	0.0518	0.0359	0.0361
150	Binary	140.9 (4.6)	0.0251	0.0843	0.0564	0.0561
	Continuous parametric	144.3 (3.8)	0.0111	0.0494	0.0124	0.0129
	Continuous nonparametric	144.1 (5.3)	0.0288	0.0676	0.0406	0.0406

5 Final Remarks

We have developed two unsupervised maximum entropy classification algorithms for record linkage, which allow for continuous comparisons of record pairs. One is based on the hurdle Gamma distribution, and the other utilizes the Kullback-Leibler Importance Estimation Procedure. Including continuous comparisons is essential to capture small discrepancies between records. Unlike standard binary comparison methods, our approaches prevent information loss caused by typos or minor errors. Notably, they outperform the binary MEC algorithm in terms of estimation of the number of matches.

A challenge that warrants further research concerns the large number of record comparisons required to perform the MEC algorithm. To mitigate this computational burden, *blocking* can be employed to place similar records into blocks. Another area that merits further investigation is improving the estimation of error rates, including estimation within blocks. Finally, a key direction for future work is developing a method to evaluate the standard error of the number of matches estimate.

Acknowledgments

The authors' work has been financed by the National Science Centre in Poland, OPUS 20 grant no. 2020/39/B/HS4/00941.

References

- Choi, S. C. and Wette, R. (1969). Maximum likelihood estimation of the parameters of the Gamma distribution and their bias. *Technometrics*, 11(4):683–690.
- Copas, J. and Hilton, F. (1990). Record linkage: statistical models for matching computer records. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 153(3):287–312.
- Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.
- Lee, D., Zhang, L.-C., and Kim, J. K. (2022). Maximum entropy classification for record linkage. *Survey Methodology*, 48(1). Catalogue No. 12-001-X, Paper available at <http://www.statcan.gc.ca/pub/12-001-x/2022001/article/00007-eng.htm>.
- McLeod, P., Heasman, D., and Forbes, I. (2011). Simulated data for the on the job training. Essnet DI. Accessed: 2025-09-08.
- R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Büna, P., and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746.
- Vo, T. H., Chauvet, G., Happe, A., Oger, E., Paquelet, S., and Garès, V. (2023). Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system. *Computational Statistics & Data Analysis*, 179(C).
- Volker, T. B., Gonzalez Poses, C., and van Kesteren, E.-J. (2024). densityratio: Distribution comparison through density ratio estimation.
- Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, pages 354–359. American Statistical Association.

Appendix

A Detailed Continuous Parametric Algorithm

Assume that the components of $\boldsymbol{\gamma}$ are conditionally independent (both in M and U). Let $\gamma^k|M$ and $\gamma^k|U$ follow hurdle Gamma distributions with parameters $p_{0M}^k, \alpha_M^k, \beta_M^k$ and $p_{0U}^k, \alpha_U^k, \beta_U^k$, respectively. Let $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ denote the parameter vectors associated with M and U , respectively. In this setting, we have

$$m(\boldsymbol{\gamma}; \boldsymbol{\theta}) = \prod_{k=1}^K f(\gamma^k; p_{0M}^k, \alpha_M^k, \beta_M^k), \quad u(\boldsymbol{\gamma}; \boldsymbol{\xi}) = \prod_{k=1}^K f(\gamma^k; p_{0U}^k, \alpha_U^k, \beta_U^k).$$

Assume a common support $\mathcal{S}(M) = \mathcal{S}(U)$. Let M_1 denote the maximal MEC set, consisting of record pairs with perfect and unique agreement on all key variables. The continuous parametric MEC algorithm for record linkage is detailed below.

1. Set initial parameters $\boldsymbol{\theta}^{(0)} = (p_{0M}^{1(0)}, \alpha_M^{1(0)}, \beta_M^{1(0)}, p_{0M}^{2(0)}, \alpha_M^{2(0)}, \beta_M^{2(0)}, \dots, p_{0M}^{K(0)}, \alpha_M^{K(0)}, \beta_M^{K(0)})$ and set $M^{(0)} = M_1$.

2. Calculate

$$\hat{p}_{0U}^k = \frac{1}{n} \sum_{(a,b) \in \Omega} \mathbb{I}(\gamma_{ab}^k = 0).$$

Denote

$$\bar{\gamma}_+^k = \frac{\sum_{(a,b) \in \Omega} \mathbb{I}(\gamma_{ab}^k > 0) \gamma_{ab}^k}{\sum_{(a,b) \in \Omega} \mathbb{I}(\gamma_{ab}^k > 0)}.$$

$\hat{\alpha}_U^k$ is given by the solution (found using iterative methods) of the following equation:

$$\sum_{(a,b) \in \Omega} \mathbb{I}(\gamma_{ab}^k > 0) \left[\log \gamma_{ab}^k - \log \bar{\gamma}_+^k - \psi(\alpha_U^k) + \log \alpha_U^k \right] = 0,$$

where

$$\psi(z) = \frac{\partial}{\partial z} \log \Gamma(z)$$

is the digamma function. Calculate

$$\hat{\beta}_U^k = \frac{\hat{\alpha}_U^k}{\bar{\gamma}_+^k}.$$

3. Given $\hat{\boldsymbol{\xi}} = (\hat{p}_{0U}^1, \hat{\alpha}_U^1, \hat{\beta}_U^1, \hat{p}_{0U}^2, \hat{\alpha}_U^2, \hat{\beta}_U^2, \dots, \hat{p}_{0U}^K, \hat{\alpha}_U^K, \hat{\beta}_U^K)$, calculate

$$r^{(0)}(\boldsymbol{\gamma}_{ab}) = \frac{m(\boldsymbol{\gamma}_{ab}; \boldsymbol{\theta}^{(0)})}{u(\boldsymbol{\gamma}_{ab}; \hat{\boldsymbol{\xi}})}$$

and

$$n_M^{(0)} = \sum_{(a,b) \in \Omega} \hat{g}^{(0)}(\boldsymbol{\gamma}_{ab}),$$

where

$$\hat{g}^{(0)}(\boldsymbol{\gamma}_{ab}) = \min \left\{ \frac{|M^{(0)}| r^{(0)}(\boldsymbol{\gamma}_{ab})}{|M^{(0)}| (r^{(0)}(\boldsymbol{\gamma}_{ab}) - 1) + n}, 1 \right\}.$$

4. Find the MEC set $M^{(1)}$ such that $|M^{(1)}| = n_M^{(0)}$ by deduplication in the descending order of $r^{(0)}(\gamma_{ab})$ over Ω .
5. For the t -th iteration, let $g_{ab}^{(t)} = \mathbb{I}((a, b) \in M^{(t)})$.

(a) Update

$$p_{0M}^{k(t)} = \frac{1}{|M^{(t)}|} \sum_{(a,b) \in \Omega} g_{ab}^{(t)} \mathbb{I}(\gamma_{ab}^k = 0).$$

Calculate

$$\bar{\gamma}_+^{k(t)} = \frac{\sum_{(a,b) \in \Omega} g_{ab}^{(t)} \mathbb{I}(\gamma_{ab}^k > 0) \gamma_{ab}^k}{\sum_{(a,b) \in \Omega} g_{ab}^{(t)} \mathbb{I}(\gamma_{ab}^k > 0)}$$

and obtain $\hat{\alpha}_M^{k(t)}$ as the solution of the following equation:

$$\sum_{(a,b) \in \Omega} g_{ab}^{(t)} \mathbb{I}(\gamma_{ab}^k > 0) \left[\log \gamma_{ab}^k - \log \bar{\gamma}_+^{k(t)} - \psi(\alpha_M^k) + \log \alpha_M^k \right] = 0.$$

Update

$$\hat{\beta}_M^{k(t)} = \frac{\hat{\alpha}_M^{k(t)}}{\bar{\gamma}_+^{k(t)}}.$$

(b) Calculate

$$n_M^{(t)} = \sum_{(a,b) \in \Omega} \hat{g}^{(t)}(\gamma_{ab}),$$

where

$$\hat{g}^{(t)}(\gamma_{ab}) = \min \left\{ \frac{|M^{(t)}| r^{(t)}(\gamma_{ab})}{|M^{(t)}| (r^{(t)}(\gamma_{ab}) - 1) + n}, 1 \right\}$$

and

$$r^{(t)}(\gamma_{ab}) = \frac{m(\gamma_{ab}; \boldsymbol{\theta}^{(t)})}{u(\gamma_{ab}; \hat{\boldsymbol{\xi}})}.$$

- (c) Find the MEC set $M^{(t+1)}$ such that $|M^{(t+1)}| = n_M^{(t)}$ by deduplication in the descending order of $r^{(t)}(\gamma_{ab})$ over Ω .
6. Repeat Step 5 until $|n_M^{(t)} - n_M^{(t+1)}| < \delta$ or $\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t+1)}\| < \varepsilon$, where δ and ε are pre-defined thresholds (small positive values).

The theoretical convergence of this procedure follows the proof outlined in the Appendix of [Lee et al. \(2022\)](#).

B Detailed Continuous Nonparametric Algorithm

Let the density ratio r be given by

$$r(\gamma) = \prod_{k=1}^K \left\{ \left(\frac{p_{0M}^k}{p_{0U}^k} \right)^{\mathbb{I}(\gamma^k=0)} \left(\frac{1 - p_{0M}^k}{1 - p_{0U}^k} r_+^k(\gamma^k) \right)^{\mathbb{I}(\gamma^k>0)} \right\},$$

where p_{0M}^k and p_{0U}^k are unknown parameters, and $r_+^k(\cdot)$ is an unknown function.

Assume a common support $\mathcal{S}(M) = \mathcal{S}(U)$. The continuous nonparametric MEC algorithm for record linkage is detailed below.

1. Set initial density ratio $r^{(0)}(\gamma_{ab})$ and set $M^{(0)} = M_1$.

2. Calculate

$$n_M^{(0)} = \sum_{(a,b) \in \Omega} \hat{g}^{(0)}(\gamma_{ab}),$$

where

$$\hat{g}^{(0)}(\gamma_{ab}) = \min \left\{ \frac{|M^{(0)}| r^{(0)}(\gamma_{ab})}{|M^{(0)}| (r^{(0)}(\gamma_{ab}) - 1) + n}, 1 \right\}.$$

3. Find the MEC set $M^{(1)}$ such that $|M^{(1)}| = n_M^{(0)}$ by deduplication in the descending order of $r^{(0)}(\gamma_{ab})$ over Ω .

4. For the t -th iteration, let $g_{ab}^{(t)} = \mathbb{I}((a,b) \in M^{(t)})$.

(a) Calculate

$$\hat{p}_{0M}^{k(t)} = \frac{1}{|M^{(t)}|} \sum_{(a,b) \in \Omega} g_{ab}^{(t)} \mathbb{I}(\gamma_{ab}^k = 0)$$

and

$$\hat{p}_{0U}^{k(t)} = \frac{1}{|\Omega \setminus M^{(t)}|} \sum_{(a,b) \in \Omega} (1 - g_{ab}^{(t)}) \mathbb{I}(\gamma_{ab}^k = 0).$$

For the samples $\{\gamma_{ab}^k : (a,b) \in M^{(t)}, \gamma_{ab}^k > 0\}$ and $\{\gamma_{ab}^k : (a,b) \in \Omega \setminus M^{(t)}, \gamma_{ab}^k > 0\}$, estimate the density ratio $r_+^{k(t)}(\cdot)$ using the KLIEP method.

(b) Update

$$r^{(t)}(\gamma_{ab}) = \prod_{k=1}^K \left\{ \left(\frac{\hat{p}_{0M}^{k(t)}}{\hat{p}_{0U}^{k(t)}} \right)^{\mathbb{I}(\gamma_{ab}^k = 0)} \left(\frac{1 - \hat{p}_{0M}^{k(t)}}{1 - \hat{p}_{0U}^{k(t)}} r_+^{k(t)}(\gamma_{ab}^k) \right)^{\mathbb{I}(\gamma_{ab}^k > 0)} \right\}.$$

Calculate

$$n_M^{(t)} = \sum_{(a,b) \in \Omega} \hat{g}^{(t)}(\gamma_{ab}),$$

where

$$\hat{g}^{(t)}(\gamma_{ab}) = \min \left\{ \frac{|M^{(t)}| r^{(t)}(\gamma_{ab})}{|M^{(t)}| (r^{(t)}(\gamma_{ab}) - 1) + n}, 1 \right\}.$$

(c) Find the MEC set $M^{(t+1)}$ such that $|M^{(t+1)}| = n_M^{(t)}$ by deduplication in the descending order of $r^{(t)}(\gamma_{ab})$ over Ω .

5. Repeat Step 4 until $|n_M^{(t)} - n_M^{(t+1)}| < \delta$, where δ is a pre-defined threshold (a small positive value).

The theoretical convergence of this procedure follows the proof outlined in the Appendix of [Lee et al. \(2022\)](#).

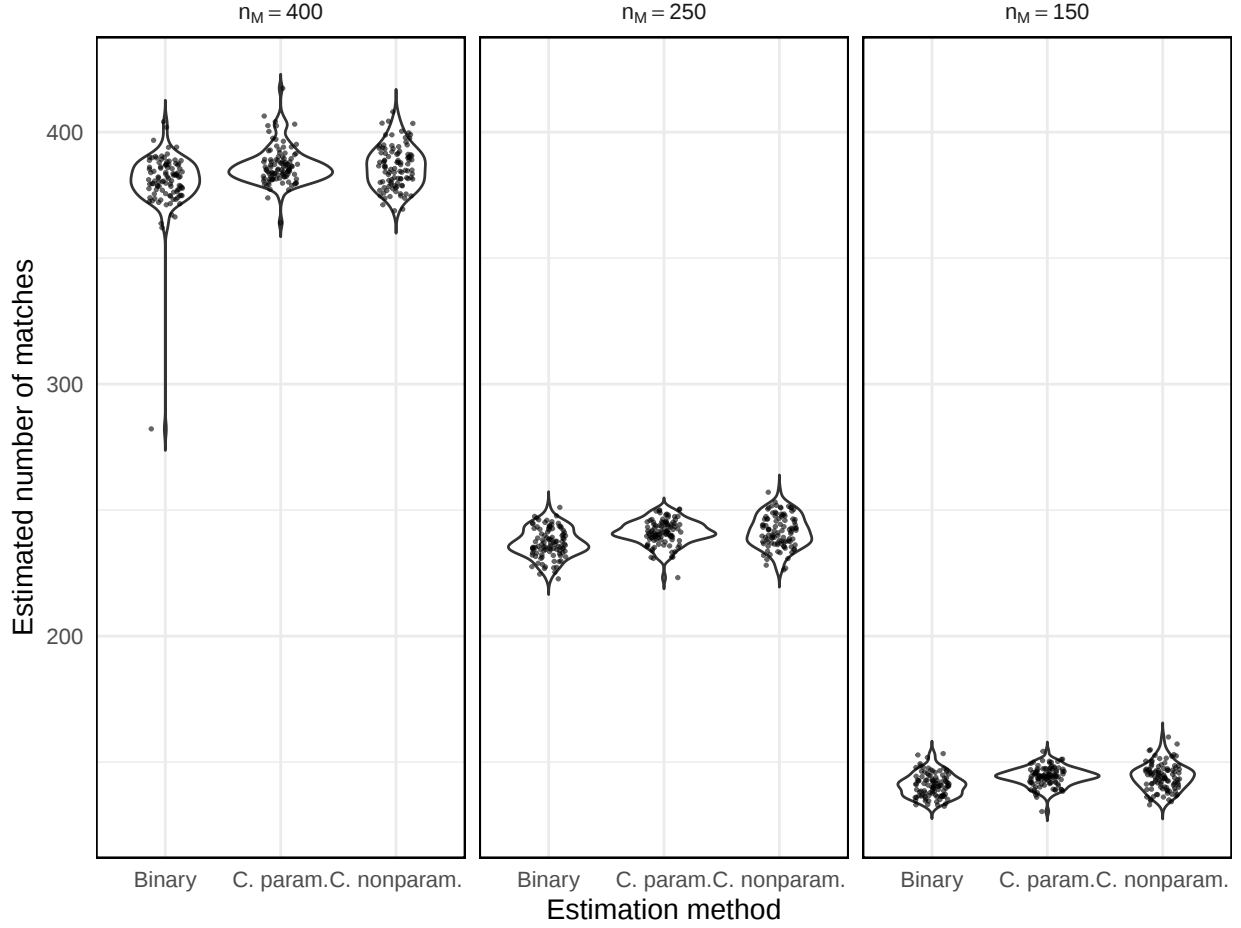


Figure 1: Violin plots of $\hat{n}_M$ obtained from 100 Monte Carlo simulations.

C Distributions of the Estimated Number of Matches

The violin plots of $\hat{n}_M$ obtained from 100 Monte Carlo simulations are presented in Figure 1. As also shown in Table 1, the continuous methods provide more stable estimates of n_M than the binary approach. When $n_M = 400$, the binary algorithm produces significant outliers, which is not the case with the continuous approaches. Across all considered n_M values, the continuous parametric method yields the lowest standard errors calculated from the sample data. Although in this simulation the continuous parametric method outperforms its nonparametric counterpart, there might be situations when the hurdle Gamma distribution is not suitable for modeling γ , and using the nonparametric approach is necessary.