# BAYESIAN RECORD LINKAGE

RACHEL ANDERSON

## 1. Introduction

Ask any economist who works with survey or administrative data how they handle imperfect matches when performing a one-to-one merge, and you will get a different answer. Some pick the matches that they believe are most likely to be correct. Some estimate the same model using multiple configurations of matched data. Others avoid the issue entirely, by dropping all matches that are flagged during the automated merge processes in R or Stata. There is no one-size-fits-all solution for data merging, nor a formal theory about how subjective decisions in the merging process may introduce bias or uncertainty into estimation and inference in economic models.

The process of joining records from multiple data sources that describe the same entity appears in many fields, such as statistics, computer science, operations research, and epidemiology, under many names, such as record linkage, data linkage, entity resolution, instance identification, de-duplication, merge/purge processing, and reconciliation. In econometrics, the process is commonly referred to as "data combination", although literature notably lacking, despite the increasing availability of large administrative datasets (Ridder and Moffitt, 2007).

There are now many solutions to the record linkage problem, and yet little understanding of how these techniques may introduce sample selection bias, and how researchers should handle these issues in practice. Examining Bayesian record linkage tools is a promising place to start, as they provide a coherent framework for incorporating uncertainty at every point of the analysis.

In this paper, I review and implement the Bayesian record linkage procedures from Larsen and Rubin (2001) and Sadinle (2017) to obtain a posterior distribution over the bipartite matching/sets of matches vs. non-matches. I use simulated data (and maybe a real data set if I have time). I analyze adjustments for what makes a better match and compare them to traditional techniques.

Then, with these posteriors, I perform a basic regression analysis using methods from Lahiri and Larsen (2005) and Tancredi and Liseo (2015) Tancredi and Liseo (2011) about how to propagate uncertainty from the matches into a basic regression analysis.

I (1) review the two methods (2) implement both, and compare results (3) write a research plan – how I will continue the analysis

## 2. Setup

There are many ways to approach this (REFERNECES), I use the setup considered in Fellegi and Sunter.

Consider two datafiles $X_1$ and $X_2$ that record information from two overlapping sets of individuals or entities. The files originate from two record-generating processes that may induce errors and missing values, so that identifying which individuals are represented in both $X_1$ and $X_2$ is nontrivial. I assume that individuals appear at most once in each datafile, so that the goal of record linkage is to identify which records in files $X_1$ and $X_2$ refer to the same entities[1].

[1] If not, one would perform first de-duplication, another topic

Suppose that files $X_1$ and $X_2$ contain $n_1$ and $n_2$ records, respectively, and without loss of generality that $n_1 \geq n_2$. Denote also the number of entities represented in both files as $n_M$, so that $n_2 \geq n_M \geq 0$.

According to the probabilistic record linkage framework of Fellegi and Sunter (1969), the set of ordered record pairs $X_1 \times X_2$ is the union of two disjoint sets, *matches* ($M$) and *non-matches* ($U$):

$$M = \{(i,j) : i \in X_1, j \in X_2, i = j\}$$
$$U = \{(i,j) : i \in X_1, j \in X_2, i \neq j\}$$

A record pair $(i,j) \in X_1 \times X_2$ is evaluated according to $L$ different comparison criteria, which arise from comparing various data fields for $i$ and $j$, such as first name, last name, address, or social security number, if $i$ and $j$ represent people. These comparisons are represented by a *comparison vector*,

$$\gamma_{\mathbf{ij}} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^\ell, \ldots, \gamma_{ij}^L)$$

The comparison criteria may be binary, as in "$i$ and $j$ have the same birthday", or factor variables that account for partial agreement between strings (see Winkler, 1990, for details). The models presented herein use only binary comparison vectors, however they can be extended to allow for partial agreement using the framework from Sadinle (2017).

The probability of observing a particular configuration of $\gamma_{\mathbf{ij}}$ can be modeled as arising from the mixture distribution:

$$P(\gamma_{\mathbf{ij}}) = P(\gamma_{\mathbf{ij}}|M)p_M + P(\gamma_{\mathbf{ij}}|U)p_U \tag{1}$$

where $P(\gamma_{ij}|M)$ and $P(\gamma_{ij}|U)$ are the probabilities of observing the pattern $\gamma_{ij}$ conditional on the record pair $(i,j)$ belonging to $M$ or $U$, respectively. The proportions $p_M$ and $p_U = 1 - p_M$ are the marginal probabilities of observing a matched or unmatched pair.

Applying Bayes' Rule, we obtain the probability of $(i,j) \in M$ conditional on observing $\gamma_{\mathbf{ij}}$:

$$P(M|\gamma_{\mathbf{ij}}) = \frac{p_M P(\gamma_{\mathbf{ij}}|M)}{P(\gamma_{\mathbf{ij}})} \tag{2}$$

So that if we can estimate the quantities in (1), we can estimate the probability that any two records refer to the same entity.

Once these probabilities are estimated, it is possible to perform the analysis: weighting match pairs, assigning matches via linear program that forces one-to-one assignment, or whatever else the researcher desires.

This paper focuses on methods for matching that best capture uncertainty about matches. Future work will focus on how this introduces other biases.

## 3. Simplifying assumptions

Let $\mathbf{\Gamma} \equiv \{\gamma_{\mathbf{ij}} : (i,j) \in X_1 \times X_2\}$ denote the set of comparison vectors for all records pairs $(i,j) \in X_1 \times X_2$.

3.1. **Blocking.** Note that $\mathbf{\Gamma}$ contains potentially $n_1 \times n_2$ elements, so that calculating $\mathbf{\Gamma}$ may be computationally expensive when $X_1$ or $X_2$ is large. In practice, researchers partition $X_1 \times X_2$ into "blocks", where only records belonging to the same block are attempted to be linked, and records belonging to different blocks are assumed to be nonmatches. For example, postal codes and household membership are often used to define blocks in census applications, however it is important that the blocking variables are recorded without error (Herzog et al., 2007).

This paper assumes that no blocking is used; or, alternatively, that records are already divided into blocks that can be analyzed independently using the methods outlined below.

3.2. **Conditional independence.** In principle, we can model,

$$\gamma_{\mathbf{ij}} \mid M \sim \text{Dirichlet}(\delta_{\mathbf{M}})$$
$$\gamma_{\mathbf{ij}} \mid U \sim \text{Dirichlet}(\delta_{\mathbf{U}})$$

However, there are $2^L - 1$ possible configurations of each $\gamma_{\mathbf{ij}}$, so that $\delta_{\mathbf{M}}$ and $\delta_{\mathbf{U}}$ may be very high-dimensional if we want to allow assymetric probabilities to vary across different categories, which makes sense in many applications. It may be possible to perform factor analysis or something... some have done this

A common assumption in the literature is to define comparison fields $\ell$ so that $\gamma_{ij}^{\ell}$ are independent across $\ell$ conditional on match status (i.e. if errors in name and address are not correlated conditional on match status This implies:

$$P(\gamma_{ij}|C) = \prod_{\ell=1}^{L} P(\gamma_{ij}^{\ell}|C)^{\gamma_{ij}^{\ell}}(1 - Pr(\gamma_{ij}^{\ell}|C))^{1-\gamma_{ij}^{\ell}} \qquad C \in \{M, U\} \tag{3}$$

Hence the number of parameters used to describe each mixture class is reduced to $L$

## 4. Mixture Model Approach to Record Linkage

Since membership to $M$ or $U$ is not actually observed, a convenient way of simultaneously estimating $p_M, p_U$ and classifying $(i, j)$ into matches and non-matches is via mixture modeling, with mixture distributions $P(\gamma_{\mathbf{ij}}|M)$ and $P(\gamma_{\mathbf{ij}}|U)$.

4.1. **Estimation.** For convenience, denote $p_{M\ell} = P(\gamma_{ij}^{\ell}|M)$ and $p_{U\ell} = P(\gamma_{ij}^{\ell}|U)$. Assuming conditional independence across $\ell$ (and global parameters that do not vary by block, if using blocked records), a convenient prior distribution is the product of independent Beta distributions,

$$p_M \sim \text{Beta}(\alpha_M, \beta_M)$$
$$p_{M\ell} \sim \text{Beta}(\alpha_{M\ell}, \beta_{M\ell}),\ \ell = 1, \ldots, L$$
$$p_{U\ell} \sim \text{Beta}(\alpha_{U\ell}, \beta_{U\ell}),\ \ell = 1, \ldots, L$$

For $i = 1, \ldots, n_1 \in X_1$, $j = 1, \ldots, n_2 \in X_2$ define

$$I_{ij} = \begin{cases} 1, & \text{if records } i \in X_1 \text{ and } j \in X_2 \text{ refer to the same entity;} \\ 0, & \text{otherwise} \end{cases}$$

If the match indicators $\mathbf{I}$ were known, the posterior distributions of $(p_M, \mathbf{p_{M\ell}}, \mathbf{p_{U\ell}})$ would be:

$$p_M|I \sim \text{Beta}\left(\alpha_M + \sum_{(i,j)} I_{ij},\ \beta_M + \sum_{(i,j)}(1 - I_{ij})\right) \tag{4}$$

$$p_{M\ell}|I \sim \text{Beta}\left(\alpha_{M\ell} + \sum_{(i,j)} I_{ij}\gamma_{ij}^{\ell},\ \beta_{M\ell} + \sum_{(i,j)} I_{ij}(1 - \gamma_{ij}^{\ell})\right), \quad \ell = 1, \ldots, L \tag{5}$$

$$p_{U\ell}|I \sim \text{Beta}\left(\alpha_{U\ell} + \sum_{(i,j)}(1 - I_{ij})\gamma_{ij}^{\ell},\ \beta_{U\ell} + \sum_{(i,j)}(1 - I_{ij})(1 - \gamma_{ij}^{\ell})\right), \quad \ell = 1, \ldots, L \tag{6}$$

The posterior distribution of parameters can therefore be simulated from alternating conditional distribution draws via the Gibbs Sampling scheme in Larsen (2005). See Appendix A.1 for details.

4.2. **Interpreting the results.** For a candidate pair $(i, j)$, the posterior probability of a match is $P(I_{ij} = 1|p_M, p_{M\ell}, p_{U\ell}) = \frac{1}{K}\sum_k I_{ij}^{(k)}$. Options for designating matches are to (1) designate all candidate pairs exceeding a cutoff as matches, or (2) use a linear program to enforce one-to-one matching.

4.3. **Limitations.**

(1) There is no guarantee that the clusters will correspond to matches and non-matches. This is a more general criticism about using mixture models, which suffer from this identification problem. In practice, 3-component mixtures tend to work better, even though theoretically we would like two. Winkler (2002) mentioned conditions for the mixture model to give good results: the proportion of matches should be greater than 5%, the classes of matches and non-matches should be well-separated, typographical errors must be relatively low, there must be redundant fields that overcome errors in other fields, among others.

(2) Many-to-one matches can still happen unless the linear sum assignment is used. Even if mixture model is fitted with the one-to-one constraint, the FS decision rule alone may lead to many-to-many assignments. The linkage decision for the pair $(i, j)$ not only depends on $\gamma_{ij}$ but on the other pairs.

Larsen (2012) notes that it is also possible to specify a prior distribution over the whole probability vector associated with the set of comparison vectors $\gamma$ as two Dirichlet distributions:

$$Pr(\gamma|M) \sim \text{Dirichlet}(\delta_M)$$
$$Pr(\gamma|U) \sim \text{Dirichlet}(\delta_U)$$

4.4. **Enforcing one-to-one assignment.** The FS decision rule does not enforce the maximum one-to-one assignment that is desirable in many economic applications. In practice, the optimal

assignment of record pairs is obtained by solving the linear sum assignment problem:

$$\max_{\Delta} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} w_{ij} \Delta_{ij}$$
$$\text{subject to } \Delta_{ij} \in \{0,1\}; \sum_{i=1}^{n_1} \Delta_{ij} \leq 1,\ j = 1, \ldots, n_2$$
$$\text{and } \sum_{j=1}^{n_2} \Delta_{ij} \leq 1,\ i = 1, \ldots, n_1$$

where the constraints ensure that $\Delta$ represents a bipartite matching. The output of this step is a bipartite matching that maximizes the sum of the weights $w_{ij}$ among matched pairs, and the pairs that are not matched. Sadinle (2017) shows that this can be thought of as the MLE under the assumption that the comparison vectors are conditionally independent given the bipartite matching.

## 5. Beta Record Linkage (Sadinle, 2017)

The main disadvantage of the mixture model approach is that it does not enforce one-to-one matching. Intuitively, if a one-to-one matching is desired, then the matching status of $(i, j)$ should affect the matching status of $(i, j')$. Yet the mixture model approach does not allow for this.

This issue was noted by Larsen (2005) and Sadinle (2017). Formally, the parameter of interest is a bipartite matching, which can be represented compactly as a *matching labeling* $Z = (Z_1, Z_2, \ldots, Z_{n_2})$ for the records in file $X_2$ such that

$$Z_j = \begin{cases} i, & \text{if records } i \in X_1 \text{ and } j \in X_2 \text{ refer to the same entity;} \\ n+1, & \text{if record } j \in X_2 \text{ does not have a match in } X_1 \end{cases}$$

### 5.1. Beta Prior for Bipartite Matchings $Z$.

Following Larsen (2005) and Sadinle (2017).

Just as in the mixture model approach, the prior probability that $j \in X_2$ is:

$$I(Z_j \leq n_1) \overset{i.i.d}{\sim} \text{Bernoulli}(p_M)$$

where $p_M$ represents the proporiton of matches expected a priori as a fraction of the smallest file $X_2$. Same as before, the hyperprior for $p_M$ is:

$$p_M \sim \text{Beta}(\alpha_M, \beta_M)$$

The prior on $p_M$ implies $n_{12}(Z) = \sum_{j=1}^{n_2} I(Z_j \leq n_1)$, the number of matches according to matching labeling $Z$ is distributed as:

$$n_{12}(Z) \sim \text{Beta-Binomial}(n_2, \alpha_M, \beta_M)$$

after marginalizing over $p_M$.

Conditioning on $\{I(Z_j \leq n_1)\}_{j=1}^{n_2}$, all possible bipartite matchings are taken to be equally likely, so

$$Pr(Z \mid n_{12}) = \left( \frac{n_1!}{(n_1 - n_{12})!} \right)^{-1}$$

These conditions imply the joint prior over $Z$:

$$Pr(Z \mid \alpha_M, \beta_M) = \frac{(n_1 - n_{12}(Z))!}{n_1!} \frac{\text{Beta}(n_{12}(Z) + \alpha_M,\ n_2 - n_{12}(Z) + \beta_M)}{\text{Beta}(\alpha_M, \beta_M)}$$

5.2. **Gibbs sampler for bipartite matching (Larsen, 2005).**

(1) Pick an initial values of $p_M, p_{M\ell}, p_{U\ell}$ and a valid configuration of $Z$. Repeat the following until convergence:

   (a) Draw $p_M$ from

   $$p_M \mid Z \sim \text{Beta}(\alpha_M + n_M(Z),\ \beta_M + n_2 - n_M(Z))$$

   Note this is same as before.

   (b) Draw $p_{M\ell}$ and $p_{U\ell}$ from their conditional distributions (same as before).

   (c) Use Metropolis-Hastings algorithm to draw values of $Z$ and $n_M(Z)$ from their full conditional distributions.

(2) Stop when converged.

The only difference is step 3. We no longer draw matches according to $n_1 \times n_2$ independent Bernoulli random variables. The procedure used in this paper is exactly that from the appendix of Larsen (2005). It is quite long, so see Appendix for details.

## 6. Data and Simulations

I wrote Python scripts that implement Gibbs sampling for the mixture model- and bipartite matching-based approaches to record linkage. I test them on fake data with different parameters and compare the results. Documentation is available in `README.md`.

6.1. **Data.** For each combination of $(L, n_1, n_2, p_M, p_{M\ell}, p_{U\ell})$, in Table 1, I randomly selected $n_M = p_M n_2$ records from $X_2$ and $n_M$ records from $X_1$ and assigned them as matches. For these record pairs, I generated $\boldsymbol{\gamma_{ij}}$ according to

$$\gamma_{ij}^{\ell} \mid M \overset{i.i.d.}{\sim} \text{Bernoulli}(p_{M\ell}), \quad \ell = 1, \dots L$$

For the remaining $(n_1 n_2 - n_M)$ record pairs, I generated $\boldsymbol{\gamma_{ij}}$ from

$$\gamma_{ij}^{\ell} \mid U \overset{i.i.d.}{\sim} \text{Bernoulli}(p_{U\ell}), \quad \ell = 1, \dots L$$

For the results presented here, I test $p_{M\ell} = 0.9$ and $p_{U\ell} = 0.2$ for all $\ell$, however my programs also allow flexible choices.

6.2. **Priors.** I use a flat prior for all parameters ($\alpha_M = \beta_M = \alpha_{M\ell} = \beta_{M\ell} = \alpha_{U\ell} = \beta_{U\ell} = 1$). However, these are also flexible.

TABLE 1. Parameter combinations and time to perform 100,000 iterations (in seconds)

| $L$ | $n_1$ | $n_2$ | $p_M$ | $p_{M\ell}$ | $p_{U\ell}$ | MM | BPM |
|---|---|---|---|---|---|---|---|
| 2 | 10 | 10 | 0.2 | 0.9 | 0.2 | 237 | 4174 |
| 3 | 10 | 10 | 0.2 | 0.9 | 0.2 | 292 | 4315 |
| 4 | 10 | 10 | 0.2 | 0.9 | 0.2 | 329 | 4361 |
| 5 | 10 | 10 | 0.2 | 0.9 | 0.2 | 379 | 4417 |
| 2 | 20 | 20 | 0.5 | 0.9 | 0.2 | 900 | 14967 |
| 3 | 20 | 20 | 0.5 | 0.9 | 0.2 | 1274 | 18643 |
| 4 | 20 | 20 | 0.5 | 0.9 | 0.2 | 1490 | 18618 |
| 5 | 20 | 20 | 0.5 | 0.9 | 0.2 | 1469 | 15470 |
| 2 | 10 | 10 | 0.9 | 0.9 | 0.2 | 242 | 4246 |
| 3 | 10 | 10 | 0.9 | 0.9 | 0.2 | 289 | 4255 |
| 4 | 10 | 10 | 0.9 | 0.9 | 0.2 | 334 | 4327 |
| 5 | 10 | 10 | 0.9 | 0.9 | 0.2 | 382 | 4249 |

## 7. RESULTS

For each dataset, I perform Gibbs sampling according to the Mixture Modeling and Beta Record linkage approaches. Unless otherwise specified I use a burn-in of 5,000. Here is what I find,

### 7.1. **Mixture Model.**

7.1.1. *Convergence.* To evaluate convergence for the mixture models, I look at the trace plots of $(p_M, p_{M\ell}, p_{U\ell})$. A common pattern across all chains is that $p_M$ converges to values near 0 or 1 within a few initial draws, and then stays close to that value for all remaining draws. Additionally, one of $p_{M\ell}$ or $p_\ell$ oscillates between 0 and 1 for all draws, while the other oscillates between 0.15 and 0.35. A–n example of this is Figure 1.

Aside from the strange pattern of $p_M$, my posterior density estimates are robust to different burn-in amounts, which suggests that all parameters are either stuck or, in some sense, converged. It may be possible to rewrite the sampler to avoid this issue, however there are deeper issues about the model such that it did not merit the effort.[2]

7.1.2. *Identifiability.* For each parameter combination tested, the posterior densities and trace plots for $(p_{M1}, \ldots, p_{ML})$ and $(p_{U1}, \ldots, p_{UL})$ appear to be about the same across $\ell$. This accurately reflects the data generating process, since $p_{M\ell}$ and $p_{U\ell}$ are fixed across $\ell$. Across parameter combinations, however, the posterior densities for $p_{M\ell}$ and $p_{U\ell}$ flip. For example, in Figure 2, the posterior density for $p_{M\ell}$ is approximately Uniform(0,1), while $p_{U\ell}$ is approximately normally distributed with mean 0.22. Yet for $L = 2$ and $L = 3$ (and all other parameters unchanged), the roles of $M$ and $U$ flip (Figure 3)

This example illustrates that identifiability is an issue. First, there is the issue of label degeneracy, which is the fact that the estimated "M" mixture may actually be the "U" mixture if I do not impose additional restrictions. A deeper issue is that in each of my datasets, there are between 2-10 matched record pairs, out of 100-400 possible links.

[2]I tried fixing this in earlier versions of the code, by adding 1 to the denominator when sampling from $P(\mathbf{I}|(p_M, p_{M\ell}, p_{U\ell}))$ if $p_M$ was converging to 1 too quickly.

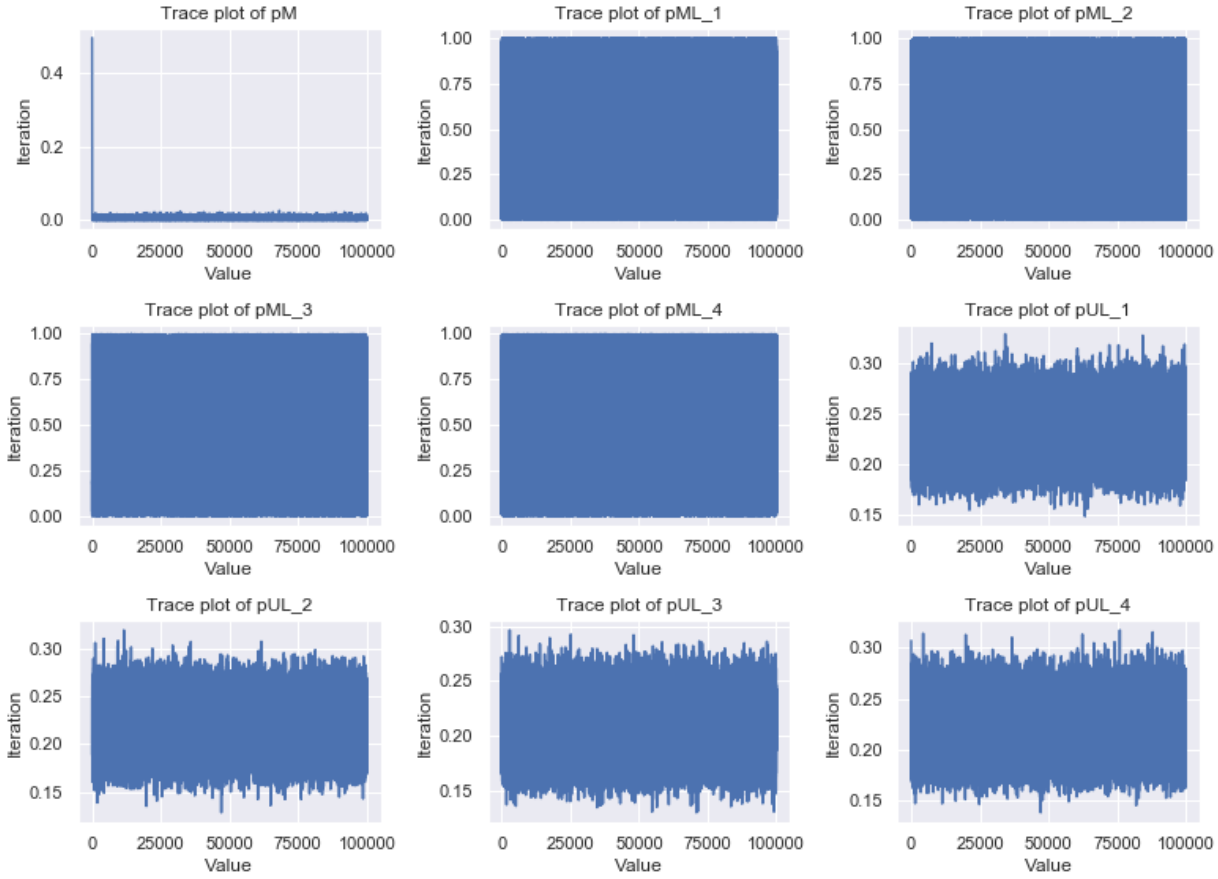


FIGURE 1. Trace plots for 100,000 draws of $(p_M, p_{M\ell}, p_{U\ell})$ with true values of $(L, n_1, n_2, p_M, p_{M\ell}, p_{U\ell}) = (4, 20, 20, 0.5, 0.9, 0.2)$ and no burn-in

7.1.3. *Matching and matching weights.*

7.2. **Beta Record Linkage.** Next steps might include writing a function that calculates the effective sample size, similar to the one from the `coda` package in `R`, or that uses the implements the appropriate convergence criteria from Brooks and Gelman (1998) and Gelman and Rubin (1992).

7.3. **How correct?** Comments - Mixture Model

(1) ($n_M = 2$ case, mixture model): For $L = 2$, $p_M \to 1$ and $p_{M\ell}$ look approximately normal (but low probability), $p_{U\ell}$ is uniform. For $L > 2$, $p_M \to 0$, $p_{U\ell}$ are approx normal and low, $p_{M\ell}$ like uniform. This holds for all $L$. Adding more comparison info doesn't help because not enough matches. Seems to mix well.

(2) ($n_M = 2$ case, bpm): $L = 4$ case: llh bell shaped – maybe getting stuck? Unclear. need to look at $Z's$. $p_M$ is flat, $p_M\ell$ look like chi sq., pUL look like normal.

(3) in the $n_M = 2$

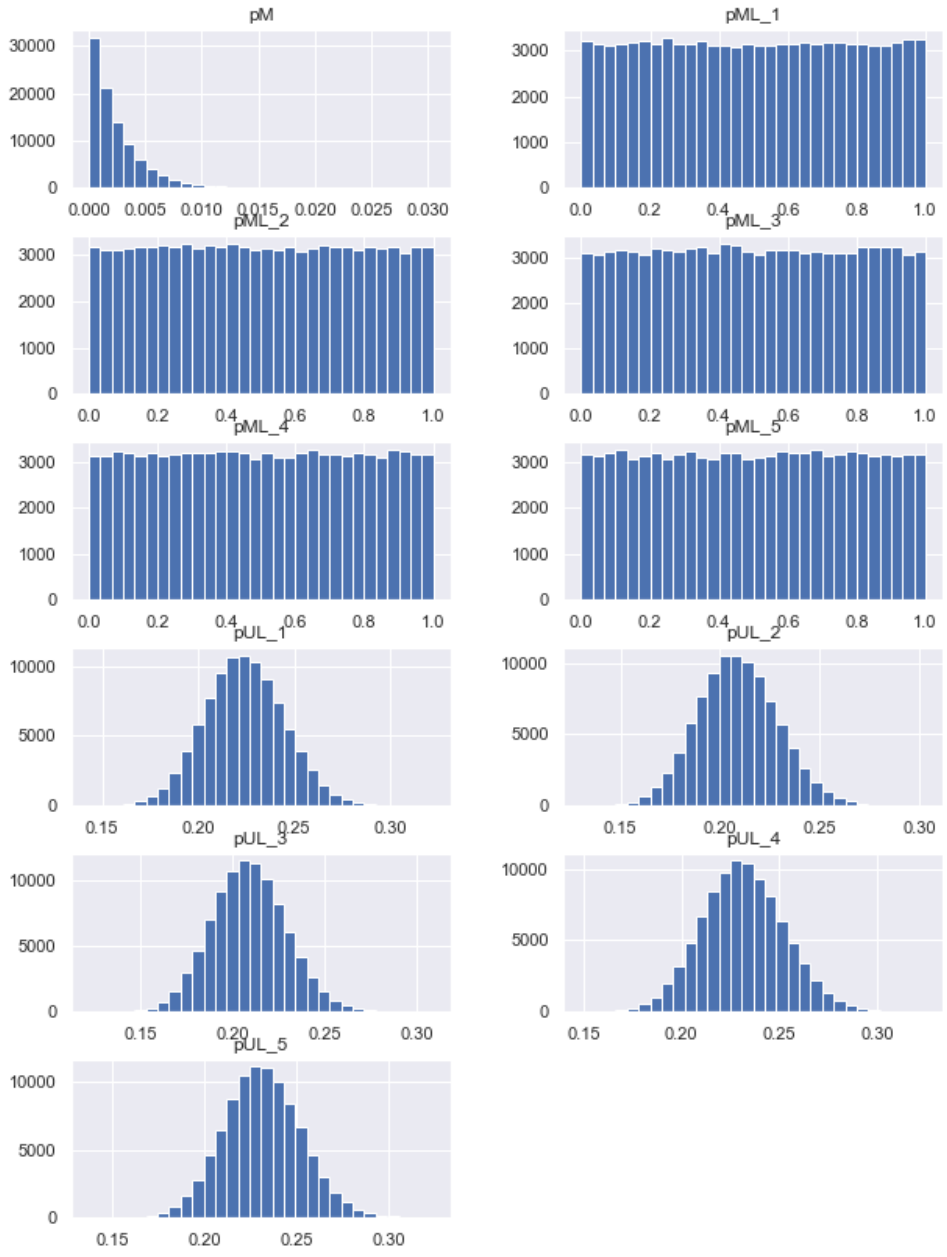


FIGURE 2. Histogram of 100,000 draws (burn-in = 5,000) from mixture model for all parameters with true parameter values $(L, n_1, n_2, p_M, p_{M\ell}, p_{U\ell}) = (5, 20, 20, 0.5, 0.9, 0.2)$ and independent, flat priors

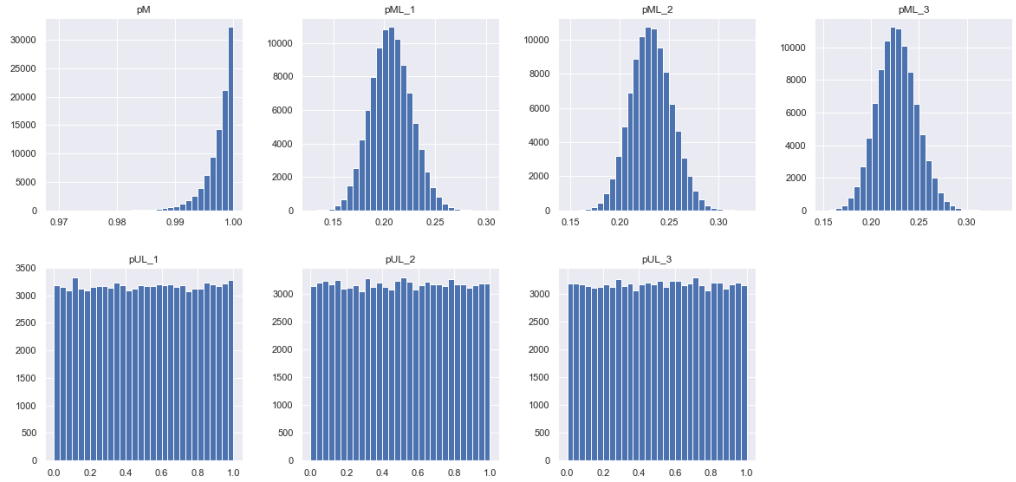


FIGURE 3. Histogram of 100,000 draws (burn-in = 5,000) from mixture model for all parameters with true parameter values $(L, n_1, n_2, p_M, p_{M\ell}, p_{U\ell}) = (3, 20, 20, 0.5, 0.9, 0.2)$ and independent, flat priors

BPM: Patterns in $(p_M, p_{M\ell}, p_{U\ell})$ are persistent across all parameter combinations in bipartite matchings.

## 8. Discussion

8.1. **How to improve the results by making more prior restrictions.** It is expected that the probability of agreeing on an individual field of comparison is higher for matches than for nonmatches:

$$Pr\left(\gamma_{ij}^{\ell} = 1 |\ (i,j) \in M\right) \geq Pr\left(\gamma_{ij}^{\ell} = 1 |\ (i,j) \in U\right)$$

8.2. **Results.** Some plots will go here.

This method does not perform well, even when I impose strong priors and set initial parameter values equal to the truth. First, the posterior for $p_M$ is heavily skewed toward 1 when I sample $I$ using the formula in Step 1. This corresponds with high posterior probabilities of $I(a, b)$, which may imply large false positive matching rates if threshold is set too low.

This issue reflects the fact that updates to $(p_M, p_{M\ell}, p_{U\ell})$ depend on assignments of $I(a, b)$. The $sample_I$ function is too quick to assign matches. This may result from the fact that I use two clusters, and that once $p_{U\ell}$ probabilities get set low, the chain cannot recover. I test this issue by adding 1 to the denominator of the Bernoulli parameter in Step 1:

$$p \equiv Pr(\ I(a,b)^{(k+1)} = 1 \mid \gamma(a,b)) = Pr(\ M \mid \gamma(a,b)) = \frac{p_M^{(k)} Pr(\ \gamma(a,b) \mid M)}{Pr(\gamma(a,b)) + 1}$$

This change prevents $p_M$ from converging to 1 ¡span style="color:blue"¿(but I need to write more tests) ¡/span¿

My results are extremely sensitive to a choice of prior! Choosing the prior will be important to explore.

Could I sample from the joint distribution of $(p_M, p_{M\ell}, p_{U\ell})$ $|I$? Could I model as Dirichlet?

Ultimately it is not worth the time and energy trying to fix this broken method so now I focus on the bipartite matching, which will fix many of these issues.

## 9. Dicussion/Literature

Other option is to use training data to obtain the weights. Another is to apply mixture models to the comparison vectors directly. This latter method is favorable because in many contexts training data is not available, or creating a sufficiently large, representative training data set is costly.

Referee for Belin (93): "every gain which is achieved by a superior record linkage procedure must be justified by the cost of implementing that procedure".

The record linkage literature can be divided into two categories, broadly. The first are those who develop methods to perform matches that are in some way "optimal" – defined by reducing false match rates, or in terms of speed. The second, where this work will eventually fall, is studying how to incorporate uncertainty from the matching process into the subsequent analysis. Rarely are we interested in the match itself, but the analysis produced using matched data.

9.1. **Methods people.** Larsen (2005) shows how to use a hierarchical Bayesian model to allow for matching probability of agreeing on fields of information to vary by block (but share a hyperprior $(\log(\alpha_s/\alpha_s + \beta_s) \sim \mathcal{N})$. He also shows one-to-one restrictions. Now $n_{12}$, the number of matches is:

$$n_m \sim Binomial(n_2, p_m)$$

where $p_m \sim Beta(\alpha_m, \beta_m)$ as before. The prior distribution for the set of matches, is uniform over the space of possible matching configurations. Suggests doing linear sum assignment procedure at each step.

9.2. **Experimental.** Belin (1993): Performance of record-linkage procedure can depend on a number of factors, including:

(1) Choice of matching variables

(2) Choice of blocking variables

(3) Assignment of weights to agreement or disagreement on various matching variables

(4) Handling of close but not exact agreement between matching variables

(5) Handling of missing data in one or both of a pair of records

(6) Algorithm for assigning candidate matches

(7) Choice of cutoff weight above which record pairs will be declared matched

(8) The site or setting from which data are obtained

In the Bayesian, case we don't need to declare anyone matched, could calculate the probability of "model" (match) space explored via Geweke and then use this to calculate proper weights on reviewed matches?

9.3. **bias people.** Scheuren and Winkler (1993) and Lahiri and Larsen (2005) propose bias-corrected estimators of coefficients in a linear regression model given data from a probabilistically linked file.

Chipperfield et al. (2011) consider the analysis of linked binary variables.

Building on Chambers (2009), Kim and Chambers (2012a, 2012b) (referred to as KC hereafter) investigate the analysis of linked data using a more general set of models fitted using estimating equations.

Kim and Chambers (2012b) review recent development in inference for regression parameters using linked data.

Chipperfield and Chambers (2015) describes a bootstrap approach to inference using estimating equations that is based on probabilistically linked data where the linked data file is created under the 1-1 constraint. They say this of the previous papers:

> Linkage models form the key feature of all of the above approaches. The linkage model describes the probability that a record on one file is linked to each of the records on another file. For a linkage model to be useful, it must properly take into account how records were linked. SW and LL do not allow for 1-1 linkage, where every record on one file is linked to a distinct and different record on the other, or for linkage in multiple passes or stages, both of which are commonly used in probabilistic record linkage. In theory, KC allows for 1-1 linkage, but imposes strong constraints on the linkage model in order to do so. KC also requires a clerical sample to estimate the parameters of the linkage model, something which is not always available in practice and which itself can be subject to measurement errors.

In a Bayesian approach none of this necessary maybe.

## 10. Bipartite matching

Bayesian approaches of Fortini et al. (2001) and Larsen (2002, 2005, 2010) improve the mixture model implementation by properly treating the parameter of interest as a bipartite matching, which relaxes the assumption that record pairs' matching status is independent of one another. The setup is as follows:

Two sets: $X = \{x_1, \ldots, x_{n1}\}$ and $Y = \{y_1, \ldots, y_{n2}\}$. The goal is to find an assignment of items so that every item in $X$ is matched to exactly one item in $Y$ and no two items share the same match. An assignment corresponds to a permutation $\pi$ where $\pi$ is a one-to-one mapping (check?) $\{1, \ldots, n_1\} \to \{1, \ldots, n_2\}$ mapping each item in $X$ to its match in $Y$. We define $\pi(i) = j$ to denote the index of a match $y_{\pi(i)} = y_j$ for an item $x_i$, and $\pi^{-1}(j) = i$ to denote the reverse (if it exists).

Uncertainty over assignments expressed as:

$$P(\pi|\theta) = \frac{1}{Z(\theta)} \exp(-E(\pi, \theta))$$

## 11. Priors

It is expected that the probability of agreeing on an individual field of comparison is higher for matches than for non-matches is $P(\gamma_k(a, b) = 1|(a, b) \in M) > P(\gamma_k(a, b) = 1|(a, b) \in U)$. And the probability of a match $p_M$ is less than or equal to the minimum file size divided by the number of possible pairs.

## 12. Datasets

The techniques are tested first using a synthetic data generator developed by Christen and Pudjijono (2009) and Christen and Vatsalan (2013), and then with two real datasets from Enamorado (2018) and Enamorado, Fifield and Imai (2018).

(1) In the first empirical application I merge two datasets on local-level candidates for the 2012 and 2016 municipal elections in Brazil. Each dataset contains more than 450,000 observations with a perfectly-recorded unique identifier, the Brazilian individual taxpayer registration identification number (called the *Cadastro de Pessoas Físicas*). All other common identifiers are manually entered into the database, so they may contain errors.

(2) In the second application, I merge the 2016 American National Election Study(ANES) with a nationwide voter file containing over 160 million voter records.

## 13. Simulations

Description of the fake data.

13.1. **Sadinle (2017) Stuff.** What this section will include:

(1) Have plots for posterior of all M and U probabilities as produced by Sadinle procedure. Put on same graph?

(2) Posterior probability of matches? Try doing geweke?

(3) Plot of likelihood to show convergence

Files will look like

(1) For each data set, generates chain of bipartite matches Z, parameters for M and U probabilities, and writes these to files

(2) File that analyzes $Z$ draws and gives posterior probability for pct correct, pct match, etc. across draws. Can also look at individual matches (will be useful for comparing the two methods)

(3) File that produces posterior distributions for $m$ and $u$ probabilities.

(4) Plot the likelihood to examine convergence properties.

## References

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics 7*(4), 434?455.

Fortini, M., B. Liseo, A. Nuccitelli, and M. Scanu (2001). On bayesian record linkage. *Research in Official Statistics 4*(1), 185–198.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457?472.

Herzog, T. N., F. J. Scheuren, and W. E. Winkler (2007). *Data Quality and Record Linkage Techniques* (1st ed.). Springer Publishing Company, Incorporated.

Lahiri, P. and M. D. Larsen (2005). Regression analysis with linked data. *Journal of the American Statistical Association 100*(469), 222?230.

Larsen, M. (2005, 10). Hierarchical bayesian record linkage theory.

Larsen, M. D. and D. B. Rubin (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association 96*(453), 32?41.

Ridder, G. and R. Moffitt (2007). Chapter 75 the econometrics of data combination. *Handbook of Econometrics*, 5469?5547.

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

Tancredi, A. and B. Liseo (2011). A hierarchical bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics 5*(2B), 1553?1585.

Tancredi, A. and B. Liseo (2015). Regression analysis with linked data: problems and possible solutions. *Statistica 75*(1), 19–35.

Winkler, W. (1990, 01). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*.

## Appendix A. Appendix

### A.1. **Gibbs Sampler (Larsen, 2005).**

(1) Specify parameters for the prior distributions. Choose initial values of $\left(p_M^{(0)}, p_{M\ell}^{(0)}, p_{U\ell}^{(0)}\right)$ for $\ell = 1, \dots L$.

(2) Repeat the following steps numerous times until the distribution of draws has converged to the posterior distribution of interest:

(a) Using the current values of $\left(p_M^{(k)}, p_{M\ell}^{(k)}, p_{U\ell}^{(k)}\right)$, draw $I_{ij}^{(k+1)}$ for each $(i, j)$ candidate pair as an independent draw from a Bernoulli distribution with parameter

$$Pr\left(I_{ij}^{(k+1)} = 1 \middle| \gamma_{ij}\right) = Pr\left(M|\gamma_{ij}\right) = \frac{p_M^{(k)} Pr\left(\gamma_{ij}|M, p_{M\ell}^{(k)}\right)}{Pr\left(\gamma_{ij}\middle|p_M^{(k)}, p_{M\ell}^{(k)}, p_{U\ell}^{(k)}\right)} \tag{7}$$

where

$$Pr\left(\gamma_{ij}|M, p_{M\ell}^{(k)}\right) = \prod_{\ell=1}^{L} \left(p_{M\ell}^{(k)}\right)^{\gamma_{ij}^{\ell}} \left(1 - p_{M\ell}^{(k)}\right)^{1-\gamma_{ij}^{\ell}}$$

and the denominator is calculated according to (1) above.

(b) Draw a value of $p_M^{(k+1)}$ from (4).

(c) Draw values of $\{p_{M\ell}^{(k+1)}\}_{\ell=1}^{L}$ independently from (5).

(d) Draw values of $\{p_{U\ell}^{(k+1)}\}_{\ell=1}^{L}$ independently from (6).

(3) Stop once the algorithm has converged. Criteria for Convergence ARE X Y Z.

### A.2. **Gibbs sampler for bipartite matching.**

I implement the incremental method for modifying $n_M$ and $Z$ via the Metropolis-Hastings steps outlined in the appendix of Larsen (2005). As Larsen notes, there are various Gibbs and Metropolis-Hastings sampling procedures that could generate draws from the target distribution, however most are computationally infeasible. The following procedure is designed to cover the space of possible configurations and to produce higher probabilities of change across iterations.

The full conditional distribution of $(n_m, Z)$ is:

$$Pr\left(n_m, Z|\gamma, p_{M\ell}, p_{U\ell}, p_M, \alpha, \beta\right) \propto Pr(n_m|p_M) Pr(Z|n_M) Pr\left(\gamma|Z, \{p_M\ell, p_{U\ell}\}_{\ell=1}^{L}, p_M, \alpha, \beta\right) \tag{8}$$

This is used to calculate the jump probabilities $P(n_M, Z)$ in the moves below. For all types of moves, there are more clever ways to select which pairs to add, drop, or switch, however these are computationally expensive. Future steps may include optimizing these steps, although ACCEPTANCE PROBABILITY IS PRETTY GOOD

**Move 1:** $n_M^* = n_M - 1$

A pair $(i, j)$ is picked at random from the set of matched record pairs according to the current configuration of $Z$ with equal probabilities. The probability of picking pair $(i, j)$ is $(n_M)^{-1}$.

The inverse move is to add the deleted pair of records to the set of designated matches. If a non-matching pair is selected at random with equal probabilities, the probability of selecting the dropped

match is $((n_1 - n_M + 1)(n_2 - n_M + 1))^{-1}$. Hence the acceptance probability for the Metropolis-Hastings algorithm is:

$$\min\left\{1, \frac{Pr\left(n_M^*, Z^* | \text{ current parameter values}\right) n_M}{Pr\left(n_M, Z | \text{ current parameter values}\right)(n_1 - n_M + 1)(n_2 - n_M + 1)}\right\}$$

**Move 2:** $n_M^* = n_M + 1$

A pair $(i, j)$ is selected at random from the set of non-matches according to the current configuration of $Z$ with probability $((n_1 - n_M)(n_2 - n_M))^{-1}$.

The inverse move is to delete a pair of records from the set of designated matches with equal probability $n_M^{-1}$ for each pair. This implies the acceptance probability for the Metropolis-Hastings algorithm:

$$\min\left\{1, \frac{Pr\left(n_M^*, Z^* | \text{ current parameter values}\right)(n_1 - n_M)(n_2 - n_M)}{Pr\left(n_M, Z | \text{ current parameter values}\right)(n_M + 1)}\right\}$$

**Move 3:** $n_M^* = n_M$**, but** $Z$ **changes**

There are three variations of this move to consider.

**Move 3.1: Two matches switch pairings**

Two matched pairs $(i, j)$ and $(k, l)$ are selected at random from the set of matched pairs according to the current configuration of $Z$. Then the new pairs $(i, l)$ and $(k, j)$ are assigned as matches, and the old pairs $(i, j)$ and $(k, l)$ are assigned non-match status.

The reverse move is to undo the switch, so that the acceptance probability of the M-H algorithm is

$$\min\left\{1, \frac{Pr(\gamma_{il}|M)Pr(\gamma_{kj}|M)Pr(\gamma_{ij}|U)Pr(\gamma_{kl}|U)}{Pr(\gamma_{ij}|M)Pr(\gamma_{kl}|M)Pr(\gamma_{il}|U)Pr(\gamma_{kj}|U)}\right\}$$

**Move 3.2: A matched pair replaces one of its matching records with a non-matching record**

A matched pair $(i, j)$ is chosen with uniform probability $(n_M)^{-1}$ from the set of matches. With probability $1/2$, $i \in X_1$ is assigned a new match at random; otherwise, $j \in X_2$ is assigned a new match at random.

The new match is selected at random from the corresponding set of unmatched records in $X_1$ or $X_2$, according to whether $i$ or $j$ is dropped. If the record $i \in X_1$ is replaced through random selection with $k \in X_1$ then the M-H acceptance probability is

$$\min\left\{1, \frac{Pr(\gamma_{kj}|M)Pr(\gamma_{ij}|U)}{Pr(\gamma_{ij}|M)Pr(\gamma_{kj}|U)}\right\}$$

If a record $j \in X_2$ is replaced through random selection with $l \in X_2$, then the M-H acceptance probability is

$$\min\left\{1, \frac{Pr(\gamma_{il}|M)Pr(\gamma_{ij}|U)}{Pr(\gamma_{ij}|M)Pr(\gamma_{il}|U)}\right\}$$

**Move 3.3 A matched pair is deleted and two unmatched records are paired**

A matched pair $(i, j)$ is selected at random from the set of matched records according to current configuration $Z$ with equal probability. An unmatched pair $(k, l)$ is selected at random from the set of unmatched candidate pairs with equal probability. The acceptance probability for the M-H algorithm is

$$\min\left\{1, \frac{Pr(\gamma_{kl}|M)Pr(\gamma_{ij}|U)n_M}{Pr(\gamma_{ij}|M)Pr(\gamma_{kl}|U)(n_1 - n_M)(n_2 - n_M)}\right\}$$

A.3. **Notes on computation/implementation.** Following the advice of (Gelman ref), all probabilities for the algorithm above are computed first in logs and later exponentiated to calculate the jump probability to avoid numerical overflow or underflow.

Unless otherwise noted, simulations were conducted on Adroit computing, with this machine.

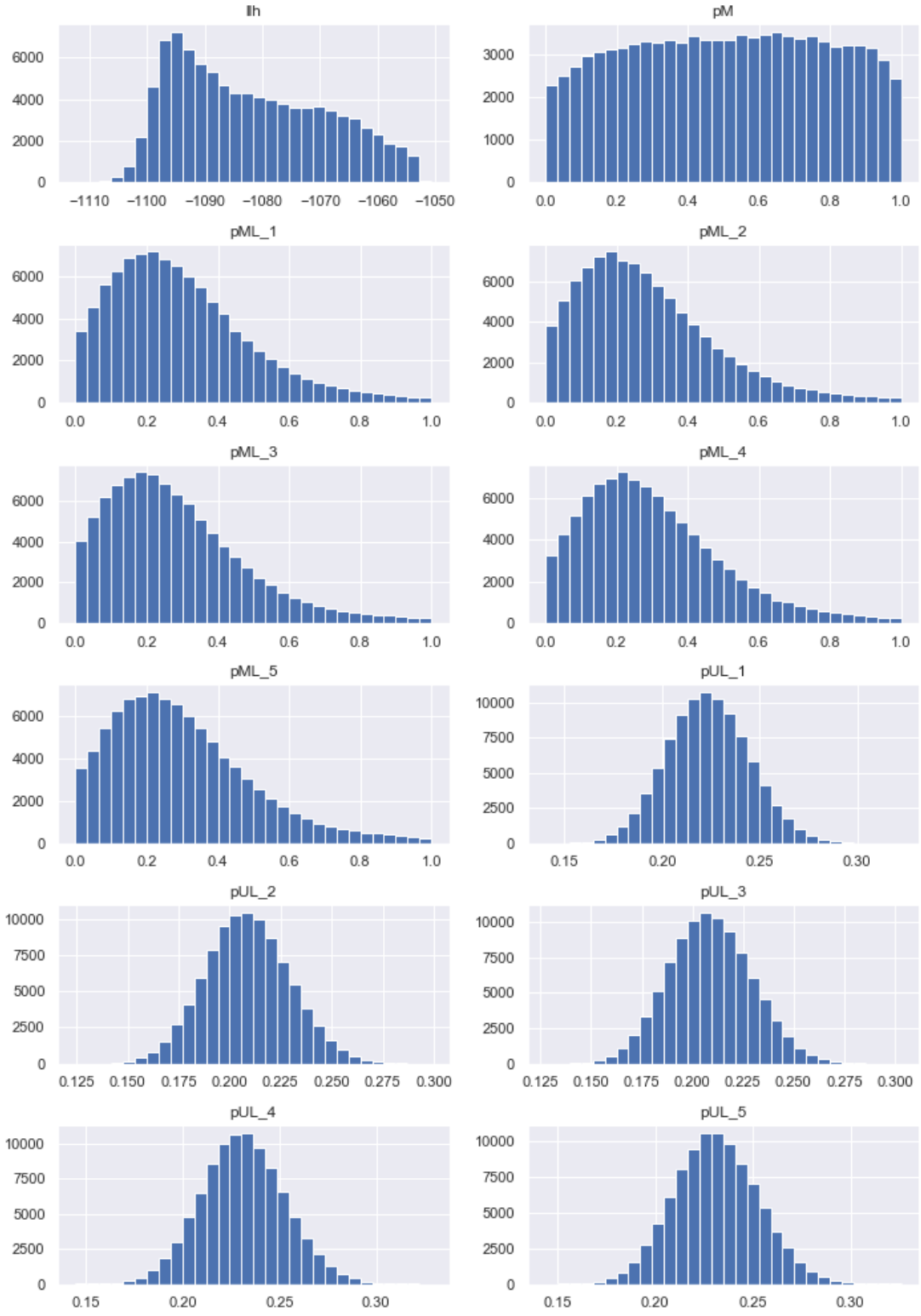


FIGURE 4. Histogram of 100,000 draws (burn-in = 5,000) from Beta Record Linkage model for all parameters and likelihood of $Z$ with true parameter values $(L, n_1, n_2, p_M, p_{M\ell}, p_{U\ell}) = (5, 20, 20, 0.5, 0.9, 0.2)$ and independent, flat priors