

# Supplementary Material for “Efficient and Scalable Bipartite Matching with Fast Beta Linkage (fabl)”

Brian Kunderinger<sup>\*</sup>, Jerome P. Reiter<sup>\*</sup> and Rebecca C. Steorts<sup>†</sup>

## Supplement A: Derivations of Full Conditionals

We provide detailed derivations of the full-conditionals provided in Section 3. In our derivations below, we assume that the new update to the Gibbs sampler is  $s + 1$  and the current state of the sampler is  $s$ . As to not overload notation, we suppress the state of the sampler until presenting the final update of the full conditional that appears in Section 3.

The  $\mathbf{m}$  and  $\mathbf{u}$  parameters are updated through standard multinomial-Dirichlet distributions. For a particular  $m_{fl}$ , we have

$$\mathcal{L}(m_{fl}|\gamma, \alpha_{fl}, \mathbf{u}, \mathbf{Z}, \pi) \propto \prod_{i=1}^{n_A} \prod_{j=1}^{n_B} m_{fl}^{I(Z_j=i)I(\gamma_{ij}^f=l)I_{obs}(\gamma_{ij}^f)} m_{fl}^{\alpha_{fl}-1} = m_{fl}^{\alpha_{fl}(\mathbf{Z})-1},$$

where  $\alpha_{fl}(\mathbf{Z}) = \alpha_{fl} + \sum_{i,j} I_{obs}(\gamma_{ij}^f)I(\gamma_{ij}^f = l)I(Z_j = i)$ . Analogous procedures lead to the posterior distribution  $\mathcal{L}(u_{fl}|\gamma, \beta_{fl}, \mathbf{m}, \mathbf{Z}, \pi) \propto u_{fl}^{\beta_{fl}(\mathbf{Z})-1}$ , where  $\beta_{fl}(\mathbf{Z}) = \beta_{fl} + \sum_{i,j} I_{obs}(\gamma_{ij}^f)I(\gamma_{ij}^f = l)I(Z_j \neq i)$ . Thus for the vectors of parameters  $\mathbf{m}_f$  and  $\mathbf{u}_f$ , we have

$$\begin{aligned} \mathbf{m}_f^{(s+1)}|\mathbf{Z}^{(s)}, \gamma &\sim \text{Dirichlet}(\alpha_{f1}(\mathbf{Z}^{(s)}), \dots, \alpha_{fL_f}(\mathbf{Z}^{(s)})), \\ \mathbf{u}_f^{(s+1)}|\mathbf{Z}^{(s)}, \gamma &\sim \text{Dirichlet}(\beta_{f1}(\mathbf{Z}^{(s)}), \dots, \beta_{fL_f}(\mathbf{Z}^{(s)})). \end{aligned}$$

Since  $\pi$  encodes the rate of matching across the two data files, the full conditional  $p(\pi|\gamma, \mathbf{Z}, \mathbf{m}, \mathbf{u}, \alpha_\pi, \beta_\pi)$  depends only on the number of links  $n_{AB}(\mathbf{Z}) = \sum_{i=1}^{n_B} I(Z_j < n_A + j)$  encoded by  $\mathbf{Z}$  and hyperparameters. We have the full conditional

$$\begin{aligned} p(\pi|\mathbf{Z}, \alpha_\pi, \beta_\pi) &\propto p(\mathbf{Z}|\pi)p(\pi) \\ &\propto \pi^{n_{AB}(\mathbf{Z})}(1-\pi)^{n_B-n_{AB}(\mathbf{Z})}\pi^{\alpha_\pi-1}(1-\pi)^{\beta_\pi-1} \\ &\propto \pi^{n_{AB}(\mathbf{Z})+\alpha_\pi-1}(1-\pi)^{n_A-n_{AB}(\mathbf{Z})+\beta_\pi-1}. \end{aligned}$$

<sup>\*</sup>Department of Statistical Science, Duke University, P.O. Box 90251, Durham, NC 27708, USA [brian.kunderinger@duke.edu](mailto:brian.kunderinger@duke.edu), [jreiter@duke.edu](mailto:jreiter@duke.edu)

<sup>†</sup>Departments of Statistical Science and Computer Science, Duke University, P.O. Box 90251, Durham, NC 27708, USA [beka@stat.duke.edu](mailto:beka@stat.duke.edu)

## 2 Efficient and Scalable Bipartite Matching with Fast Beta Linkage (fabl)

Thus,  $\pi^{(s+1)} | \mathbf{Z}^{(s+1)}, \alpha_\pi, \beta_\pi$  has a  $\text{Beta}(n_{AB}(\mathbf{Z}^{(s)}) + \alpha_\pi, n_B - n_{AB}(\mathbf{Z}^{(s)}) + \beta_\pi)$  distribution.

To express the full conditional for  $\mathbf{Z}$ , we consider the cases where  $Z_j$  does not have a link in  $A$  and where  $Z_j$  does have a link in  $A$ . Because we sample each  $Z_j$  independently of all other  $Z_{j'}$  (for  $j \neq j'$ ), we need only the full conditional for each individual  $Z_j$ . Let  $\Gamma_{.j}$  denote the matrix of  $n_A$  comparison vectors relating to record  $B_j$ . Following the observation of Wortman (2019), when  $B_j$  does not link to any record in  $A$ , the contribution to the likelihood is simply a product of  $u$  parameters, which we will call  $c_j$ :

$$p(\Gamma_{.j} | \mathbf{m}, \mathbf{u}, \pi, Z_j = n_A + j) = \prod_{i=1}^{n_A} \prod_{f=1}^F \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f=l)I_{obs}(\gamma_{ij}^f)} = c_j. \quad (1)$$

When  $Z_j = q$  for some  $q \leq n_A$ , we have

$$p(\Gamma_{.j} | \mathbf{m}, \mathbf{u}, \pi, Z_j = q) = \prod_{f=1}^F \prod_{l=1}^{L_f} m_{fl}^{I(\gamma_{qj}^f=l)I_{obs}(\gamma_{qj}^f)} \prod_{i \neq q} \prod_{f=1}^F \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f=l)I_{obs}(\gamma_{ij}^f)}. \quad (2)$$

We multiply and divide by the  $u$  parameters for the matching record pair to obtain

$$p(\Gamma_{.j} | \mathbf{m}, \mathbf{u}, \pi, Z_j = q) \propto \prod_{f=1}^F \prod_{l=1}^{L_f} \left( \frac{m_{fl}}{u_{fl}} \right)^{I(\gamma_{qj}^f=l)I_{obs}(\gamma_{qj}^f)} \prod_{i=1}^{n_A} \prod_{f=1}^F \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f=l)I_{obs}(\gamma_{ij}^f)} \quad (3)$$

$$= w_{qj} c_j. \quad (4)$$

We can divide the result of each case by  $c_j$  to get

$$p(\Gamma_{.j} | \mathbf{m}, \mathbf{u}, \pi, Z_j) \propto \begin{cases} w_{qj}, & q \leq n_A; \\ 1, & q = n_A + j. \end{cases} \quad (5)$$

Lastly, we multiply the likelihood by the fast beta prior in (5a) to obtain the full conditional

$$p\left(Z_j^{(s+1)} = q | \gamma, \mathbf{m}^{(s+1)}, \mathbf{u}^{(s+1)}, \pi^{(s+1)}\right) \propto \begin{cases} \frac{\pi^{(s+1)}}{n_A} w_{qj}^{(s+1)}, & q \leq n_A; \\ 1 - \pi^{(s+1)}, & q = n_A + j. \end{cases} \quad (6)$$

Observe that the full conditional in (6) does not depend on  $\mathbf{Z}^{(s)}$ , implying that

$$p(\mathbf{Z}^{(s+1)} | \gamma, \mathbf{m}^{(s+1)}, \mathbf{u}^{(s+1)}, \pi) \propto \prod_q \left( \prod_j p\left(Z_j^{(s+1)} = q | \gamma, \mathbf{m}^{(s+1)}, \mathbf{u}^{(s+1)}, \pi\right) \right). \quad (7)$$

Therefore, we update the full conditional for  $\mathbf{Z}^{(s+1)}$  in parallel. This contrasts the Gibbs sampler of Sadinle (2017), where the update to the matching matrix is sequential and accomplished using a partially collapsed Gibbs sampler, where one integrates out  $\pi$ . We do not opt for this approach as it would lead to similar sequential updates, which are computationally slow.

## 22 Supplement B: Bayes Estimate

We calculate a Bayes estimate  $\hat{Z}$  for the linkage parameter  $Z$  by assigning different positive losses to different types of errors, and minimizing posterior expected loss. We adopt the loss function proposed in [Sadinle \(2017\)](#) in which  $\hat{Z}_j \in \{1, \dots, n_A, n_A + j, R\}$ , with  $R$  representing the option to leave the matching undetermined by the model. Specifically, we have

$$L(\hat{Z}_j, Z_j) = \begin{cases} 0, & \text{if } Z_j = \hat{Z}_j; \\ \theta_R, & \text{if } \hat{Z}_j = R; \\ \theta_{10}, & \text{if } Z_j \leq 1, \hat{Z}_j = n_A + j; \\ \theta_{01}, & \text{if } Z_j = n_A + j, \hat{Z}_j \leq n_A; \\ \theta_{11}, & \text{if } Z_j \leq n_A, \hat{Z}_j \leq n_A, Z_j \neq \hat{Z}_j. \end{cases}$$

23 Here,  $\theta_R$  is the loss from not making a decision on the linkage status,  $\theta_{10}$  is the loss  
24 from a false nonmatch,  $\theta_{01}$  is the loss from a false match, and  $\theta_{11}$  is the loss from the  
25 special case of a false match in which the record has a true match other than the one  
26 estimated by the model.

In general, we set  $(\theta_{10}, \theta_{01}, \theta_{11}, \theta_R) = (1, 1, 2, \infty)$  inducing the decision rule

$$\hat{Z}_j = \begin{cases} i, & \text{if } p(Z_j = i | \gamma) > \frac{1}{2}; \\ 0, & \text{otherwise.} \end{cases}$$

27 Since **fabl** does not strictly enforce one-to-one matching, it is possible for this Bayes  
28 estimate to link multiple records in  $B$  to one record in  $A$ . In the event that we have two  
29 records  $B_j$  and  $B_{j'}$  such that both  $p(\hat{Z}_j = i | \gamma) > \frac{1}{2}$  and  $p(\hat{Z}_{j'} = i | \gamma) > \frac{1}{2}$ , we accept the  
30 match with the higher posterior probability, and declare the other to have no match.  
31 Since each  $Z_j$  is independent, this is equivalent to minimizing the expected loss subject  
32 to the constraint that  $\hat{Z}_j \neq \hat{Z}_{j'}$  for all  $j \neq j'$ . A similar approach appears in the most  
33 probable maximal matching sets used by [Steorts et al. \(2016\)](#) to match records to latent  
34 entities.

35 When we seek a partial estimate of the linkage structure, leaving a portion of record  
36 pairs to be classified manually in clerical review, we adopt losses  $(\theta_{10}, \theta_{01}, \theta_{11}, \theta_R) =$   
37  $(1, 1, 2, .1)$ . For a more in-depth explanation of this function and the induced Bayes  
38 estimate, see [Sadinle \(2017\)](#).

## 39 Supplement C: Accuracy under Partial Estimates

40 In this section, we repeat the simulation study in Section 5.2 of the main text, allowing  
41 for clerical review rather than forcing all records to have or not have links. Specifically,  
42 by leaving  $\theta_{10} = \theta_{01} = 1$  and  $\theta_{11} = 2$ , but setting  $\theta_R = 0.1$ , we allow the model  
43 to decline to decide a match for certain records, with nonassignment being 10% as  
44 costly as a false match. In this context, we are no longer focused on finding all true  
45 matches, but rather protecting against false matches. Thus, instead of recall, we use the

#### 4 Efficient and Scalable Bipartite Matching with Fast Beta Linkage (*fabl*)

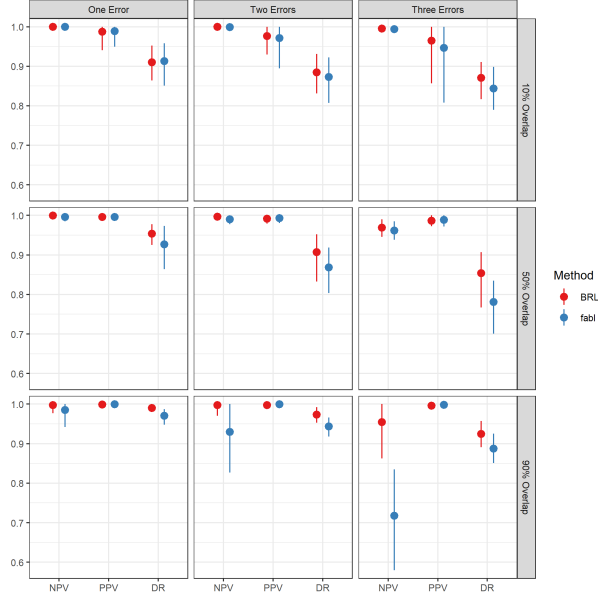


Figure 1: Negative predictive value (NPV), positive predictive value (PPV), and decision rate (DR) on data files in the simulation in Appendix C. We see poorer performance for **fabl** only in situations with high overlap.

negative predictive value (NPV), defined as the proportion of non-links that are actual nonmatches. Mathematically,  $NPV = \sum_{j=1}^{n_B} I(\hat{Z}_j = Z_j = n_A + j) / \sum_{j=1}^{n_B} I(\hat{Z}_j = n_A + j)$ . We continue to use the precision, which is renamed the positive predictive value (PPV) in this context. Lastly, we also examine the rejection rate (RR), or how often the model declines to make a linkage decision, defined as  $RR = \sum_{j=1}^{n_B} I(\hat{Z}_j = R) / n_B$ . To convey this information alongside NPV and PPV, for which values close to 1 indicate strong performance, we report the decision rate (DR), defined as  $DR = 1 - RR$ .

In Figure 1, we see that **fabl** maintains equivalently strong PPV as BRL across all linkage settings. However, with high amounts of error, and thus fewer accurate and discerning fields of information, the rejection rate under **fabl** rises, leading to a decrease in NPV. Since **fabl** does not remove previously matched records from consideration for a new record, posterior probabilities of matches at times can be split across more records; in contrast, BRL is able to maintain higher confidence in matches in this setting. If one wishes to use partial estimates, **fabl** will possibly leave more linkages for the modeler to match by hand than would be left under BRL, but the decisions made by each method should have nearly equal accuracy.

## Supplement D: Traceplots for Simulation Study

Figures 2, 3, and 4 are traceplots for one of the 900 linkage tasks that comprise the simulation in Section 5.2. It is set up with one error across the linkage fields and 50 duplicates across files. Traceplots across other settings exhibit similar behavior. Note that traceplots for  $u$  parameters show very little variation because the overwhelming majority of record pairs are nonmatching.

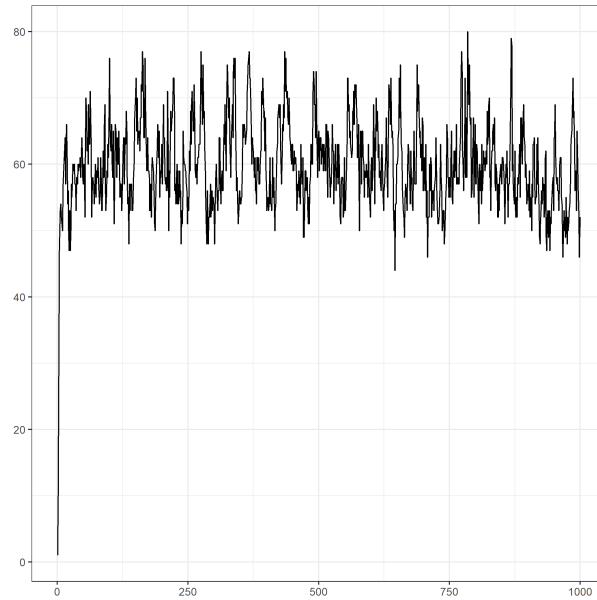


Figure 2: Representative traceplot of overlap between files from simulation study in Section 5.2.

## 6 Efficient and Scalable Bipartite Matching with Fast Beta Linkage (fabl)

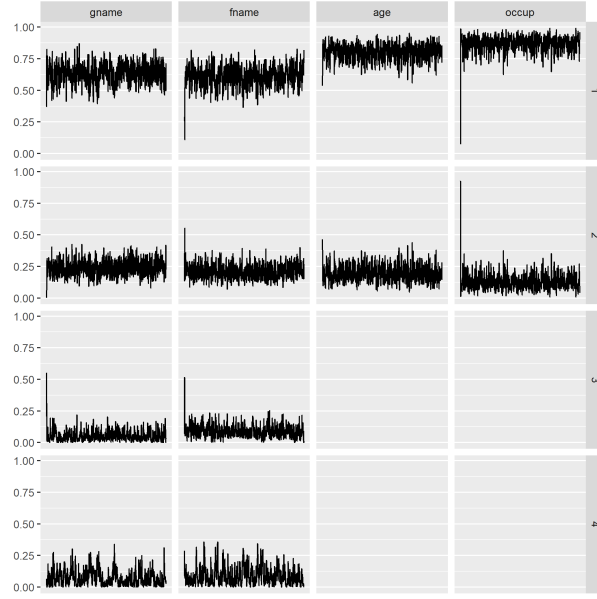


Figure 3: Representative traceplot of  $m$  parameter from simulation study in Section 5.2.

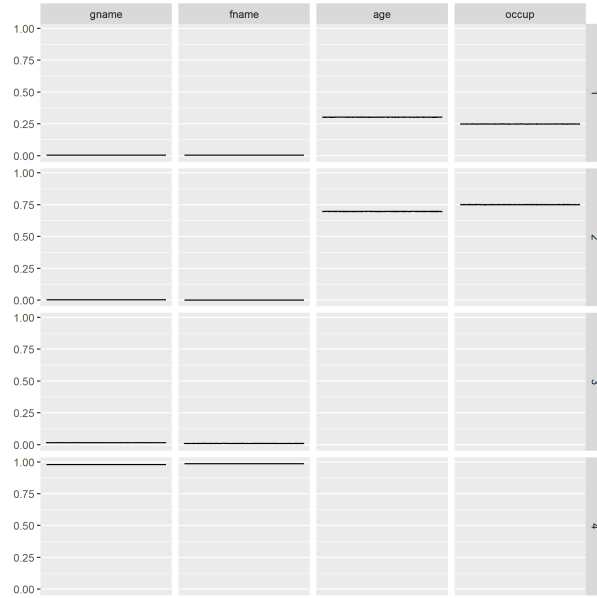


Figure 4: Representative traceplot of  $u$  parameters from simulation study in Section 5.2.

	$\mathbf{m}$			$\mathbf{u}$		
	Agree	Partial	Disagree	Agree	Partial	Disagree
Feature 1	$\frac{9}{10}$	$\frac{9}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{3}{100}$	$\frac{96}{100}$
Feature 2	$\frac{9}{10}$	$\frac{9}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{3}{100}$	$\frac{96}{100}$
Feature 3	$\frac{9}{10}$	$\frac{9}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{3}{100}$	$\frac{96}{100}$
Feature 4	$\frac{9}{10}$	$\frac{9}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{3}{100}$	$\frac{96}{100}$

Table 1: Probabilities used for  $\mathbf{m}$  and  $\mathbf{u}$  distributions in simulation study in Appendix E.

## Supplement E: Additional Speed Simulation Study

Figures 5 and 6 illustrate that different constructions of the comparison vectors lead to similar speed gains. We replicate the speed study of Section 5.1 under different settings. Here, we use four fields of comparison, each with three possible levels of agreement, resulting in  $3^4 = 81$  possible patterns. The  $\mathbf{m}$  and  $\mathbf{u}$  parameters for this simulation are shown Table 1.

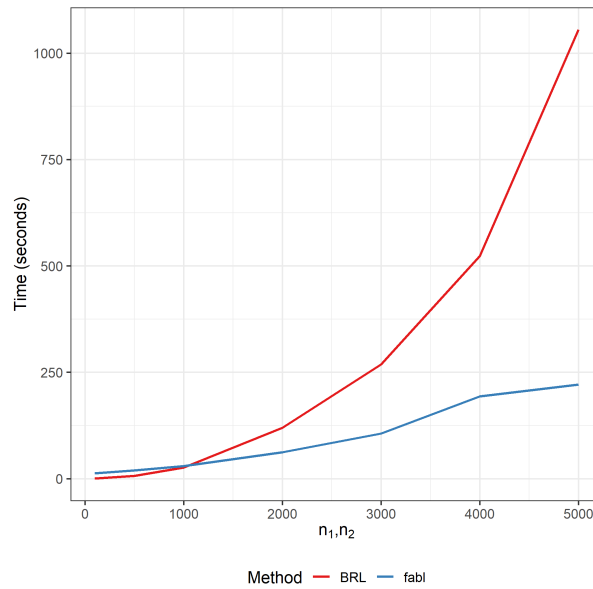


Figure 5: Run-time for BRL and `fabl` to run 1000 Gibbs iterations in simulation study in Appendix E, including hashing step for `fabl`, for increasing values of both  $n_A$  and  $n_B$ . We see near quadratic growth in run-time for BRL, and near linear growth for `fabl`.

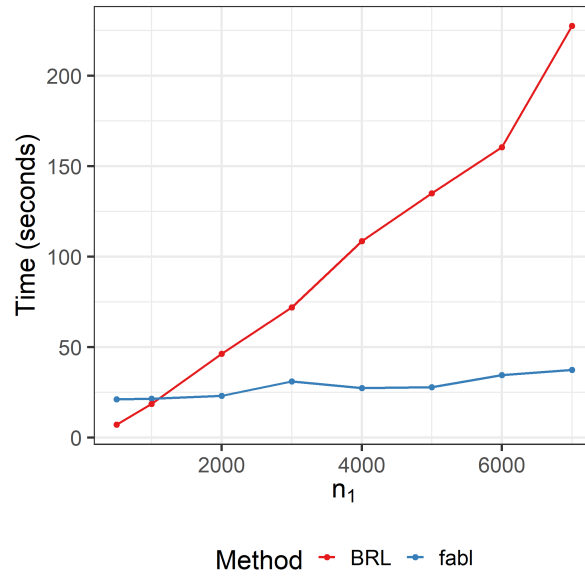


Figure 6: Run-time for BRL and `fabl` to run 1000 Gibbs iterations in simulation study in Appendix E, including hashing step for `fabl`, with increasing  $n_A$  and  $n_B$  fixed at 500. We see linear growth in run-time for BRL, and near constant run-time for `fabl`.



74 **Supplement F: Traceplots for El Salvador Case Study**

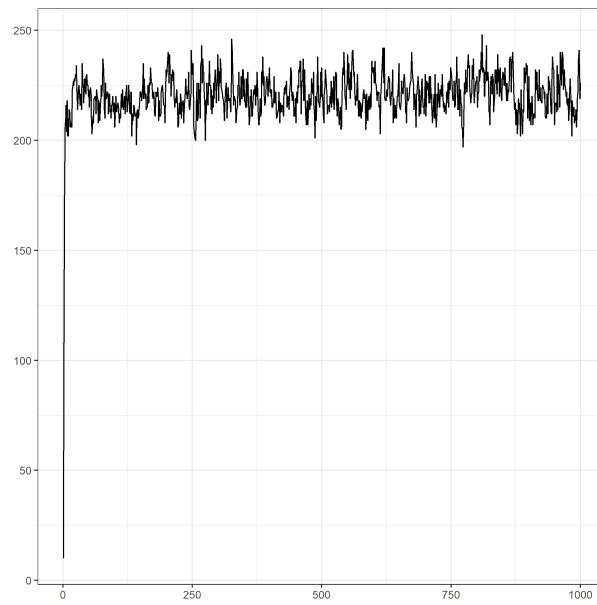


Figure 7: Traceplot for number of matches found across data files in El Salvador case study.

10 *Efficient and Scalable Bipartite Matching with Fast Beta Linkage (fabl)*

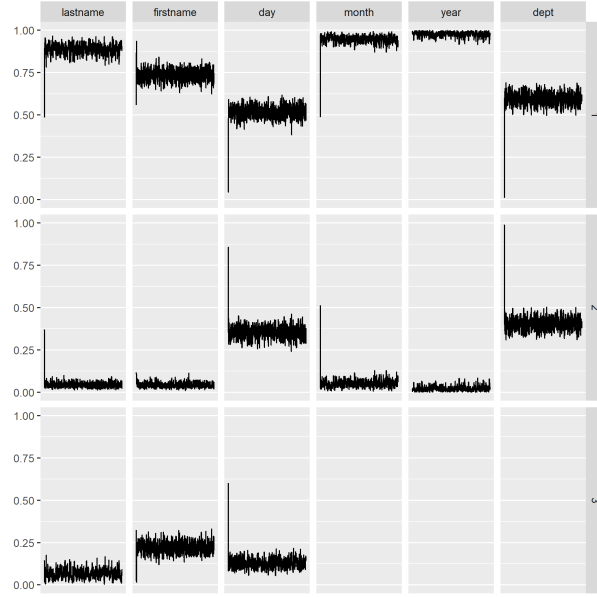


Figure 8: Traceplot for  $m$  parameter in El Salvador case study.

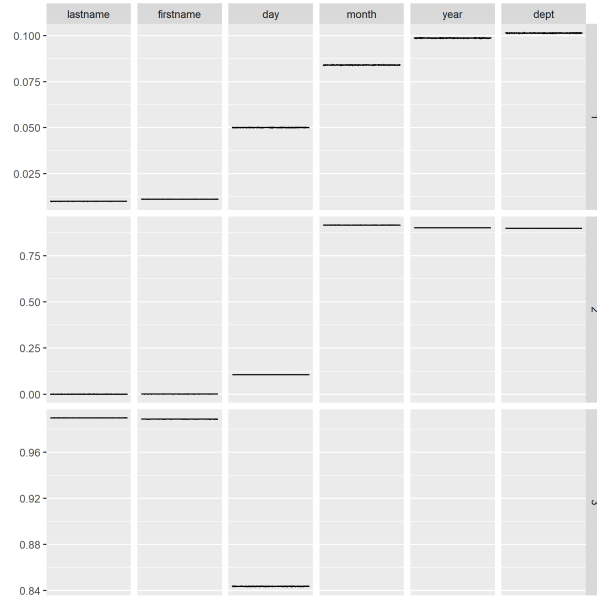


Figure 9: Traceplot for  $u$  parameter in El Salvador case study.