Supplemental Materials for "Infinite Plaid Models for Infinite Bi-clustering"

Katsuhiko Ishiguro, Issei Sato*, Naonori Ueda, Masahiro Nakano, Akisato Kimura
NTT Communication Science Laboratories, Kyoto, Japan
The University of Tokyo, Tokyo, Japan

Abstract

This supplemental material provides several details and omitted results from the main manuscript.

Inference procedure

In this section, we present our inference procedure for the proposed Infinite Plaid models.

Here are the overview of the inference procedure development.

1. We focus on the easiness of derivations and implementations, than the computational efficiency and the mixing speed of the posterior.
2. We do not try marginalization of observation function parameters θ and ϕ as opposed to the original Bayesian Plaid model (Caldas and Kaski 2008).
3. For the inference of hidden variables $\mathbf{Z} = \{\mathbf{Z}_1\mathbf{Z}_2\}$, we took the similar approach of (Meeds et al. 2007): simple Gibbs samplers for the existing sub-matrix factors, and Metropolis-Hastings sampler for accepting the new sub-matrix proposal. In addition, we used split-merge MH moves for drastic searches over sub-matrix factors, based on (Jain and Neal 2004).
4. Given $\mathbf{Z}$, we can compute the exact posteriors for the parameters θ and ϕ.
5. For hyperparameters, we implemented the hyper-prior based posterior sampling following (Hoff 2005).

Sampling Z

We rely on a combination of a simple Gibbs sampler and Metropolis-Hastings samplers for sampling new $\mathbf{Z}$. We explain how to sample an instance $z_{1,i,k}$ from its posterior, since the procedure is completely symmetric for other i, k and $\mathbf{Z}_2$. Throughout this section, we use K as the number of currently instantiated factors (sub-matrices). The Beta variables λ never appear in the inference implementations since we can marginalize λ out.

Sampling for existing sub-matrices For existing factors $k \in \{1, 2, \ldots, K\}$, we can sample $z_{1,i,k}$ in a standard IBP manner. First, the posterior of $z_{1,i,k}$ for our Gibbs sampler is formulated as follows:

$$p\left(z_{1,i,k}|\mathbf{Z}_1^{\backslash(i,k)}, \mathbf{Z}_2, \boldsymbol{X}, \theta, \phi\right) \propto p\left(z_{1,i,k}|\mathbf{Z}_1^{\backslash(i,k)}\right) p\left(\boldsymbol{X}^{(1,i)}|z_{1,i,k}, \mathbf{Z}_1^{\backslash(i,k)}, \mathbf{Z}_2, \theta, \phi\right) . \tag{1}$$

$\mathbf{Z}_1^{\backslash(i,k)}$ is $\mathbf{Z}_1$ excluding $z_{1,i,k}$, $\boldsymbol{X}^{(1,i)} = \left\{x_{i,j}\right\}, j = 1, 2, \ldots, N_2$. Since $z_{1,i,k}$ is either 1 or 0, we want to compute the ratio of $\frac{p(z=1)}{p(z=0)}$.

The of the first part of the r.h.s. of Eq. (1) is the prior part which derives from IBP. Its form is well known (Griffiths and Ghahramani 2011; Doshi-Velez et al. 2009). The ratio between $z = 0$ and $z = 1$ is:

$$\frac{m_{1,k} - z_{1,i,k}}{N_1 - m_{1,k} + z_{1,i,k}} , \tag{2}$$

where $m_{1,k} = \sum_i \mathbb{I}(z_{1,i,k} = 1)$.

The second part of the r.h.s. of Eq. (1) is the likelihood part. With straightforward but cumbersome computations, we have the following equation for the ratio:

$$\prod_j \exp\left[-\frac{\tau_0}{2}\left\{z_{2,j,k}^2\theta_k^2 - 2z_{2,j,k}\theta_k y_{i,j,k}\right\}\right] , \tag{3}$$

where

$$y_{i,j,k} = x_{i,j} - \phi - \sum_{l\backslash k} z_{1,i,l}z_{2,j,l}\theta_l .$$

Then the probability of $z_{1,i,k} = 1$ is:

$$\frac{Eq.\ (2) \times Eq.\ (3)}{1.0 + Eq.\ (2) \times Eq.\ (3)} .$$

MH move for new sub-matrices A problem of our Two-way IBP is that we need to align indices and the number of instantiated sub-matrices K between two domains $\mathbf{Z}_1$ and $\mathbf{Z}_2$. Thus we need a special treatment for sampling new factors for $z_{1,i,k}$ as it effects on the cardinality of $\mathbf{Z}_2$.

Our MH sampling scheme follows (Meeds et al. 2007). First, we sample a number of new sub-matrices to be added to the model following the standard IBP definition:

$$k_{\text{new}} \sim \text{Poisson}\left(\frac{\alpha_1}{K}\right) .$$

We denote the new factors by $l = 1, 2, \ldots, k_{\text{new}}$. Now the proposal of $z_{1,i}$ has additional assignments on new sub-matrices: $z^*_{1,i,l} = 1$. At the same time, we draw k_{new} times for new θ from its prior:

$$\theta^*_l \sim \text{Normal}\left(\mu^\theta, (\tau^\theta)^{-1}\right) .$$

Different from (Meeds et al. 2007), the Two-way IBP requires additional draws of $z_{2,j,l}$ for all j in the second domain. We extend $\boldsymbol{Z}_2$ by $N_2 \times k_{\text{new}}$ empty entries, and put assignments by the following prior:

$$z^*_{2,j,l} \sim \text{Bernoulli}\left(\frac{\alpha_2}{\alpha_2 + N_2}\right) .$$

Finally we accept the proposal by the following acceptance rate (Meeds et al. 2007):

$$\min\left(1, \frac{p\left(\boldsymbol{X}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, z^*_{1,i,l}, z^*_{2,j,l}, \theta^*_l, k_{\text{new}}, \theta, \phi, \tau_0\right)}{p\left(\boldsymbol{X}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \theta, \phi, \tau_0\right)}\right)$$

Split-merge MH moves of sub-matrices To speed up mixing of $\boldsymbol{Z}$, we also implement the split-merge MH sampler for sub-matrices. Its implementation is straightforward: we follow the idea of (Meeds et al. 2007) that rely on the paper about MH for Dirichlet Process (Jain and Neal 2004).

Sampling θ and ϕ

Given the current Gibbs samples of the hidden assignment variables $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, sampling of remaining parameters θ and ϕ are straightforward. Posteriors of θ_k and ϕ are computed as follows:

$$p\left(\theta_k|\boldsymbol{X}, \boldsymbol{Z}, \theta_{\backslash k}, \phi\right) = \text{Normal}\left(\frac{\tau^\theta\mu^\theta + M_k\tau_0\bar{y}_k}{\tau^\theta + M_k\tau_0}, \left(\tau^\theta + M_k\tau_0\right)^{-1}\right) ,$$

$$p\left(\phi|\boldsymbol{X}, \boldsymbol{Z}, \theta, \phi\right) = \text{Normal}\left(\frac{\tau^\phi\mu^\phi + N_1N_2\tau_0\bar{y}_\phi}{\tau^\phi + N_1N_2\tau_0}, \left(\tau^\phi + N_1N_2\tau_0\right)^{-1}\right) .$$

In the above equations,

$$\bar{y}_k = \frac{1}{M_k}\sum_{i,j}\mathbb{I}\left(z_{1,i,k} = 1\right)\mathbb{I}\left(z_{2,j,k} = 1\right)y_{i,j,k} ,$$

$$\bar{y}_\phi = \frac{1}{N_1N_2}\sum_{i,j}y_{i,j,\phi} , y_{i,j,\phi} = x_{i,j} - \sum_k z_{1,i,k}z_{2,j,k}\theta_k ,$$

$$M_k = \sum_i\sum_j\mathbb{I}\left(z_{1,i,k} = 1\right)\mathbb{I}\left(z_{2,j,k} = 1\right) .$$

Sampling hyperparameters

We can further infer the hyperparameter values of the Infinite Plaid models. There are a few ways to estimate hyperparameters. One is to directly optimize hyperparameters (possibly by numerical gradients) to maximize the marginalized log likelihoods. Another is to perform posterior samplings by assuming "flat" hyper priors on those hyperparameters.

In our implementations, we took the second approach. We refer (Hoff 2005) for an efficient and easy implementations of hyper prior-based estimations.

Experiment details

Likelihoods as a evaluation measure

We first considered the test data log likelihood as a primal quantitative measure without ground-truth information. Given the full observation matrix, we randomly keep a small portion (10%) of the matrix entries as test data, and compute the likelihoods. However, in preliminary experiments we found the log likelihood does not effectively present the goodness of the extracted sub-matrices computed by NMI (based on the ground truth). One possible reason is that the likelihood is computed over all matrix entries including the majority "non-interesting" ones that are not proactively modeled. Therefore we do not adopt the test data log likelihood for evaluation.

Fig. 1 presents the test data log likelihoods on synthetic data sets. There was not so much differences in likelihoods between two models. We found the same impressions on real-world data sets.

Inference

All the latent variables of the Bayesian Plaid model are inferred by collapsed Gibbs samplers, similar to the proposed Infinite Plaid model. For hyperparameters, we adopted two strategies for the both models: (i) no updates or (ii) infer them simultaneously by hyper-prior sampling.

Synthetic data experiments

Data The first data (synth1) has $K = 3$ non-overlapping sub-matrices. All parameters for the second data (synth2) is the same with the first one, excepting slight overlaps between sub-matrices. All θ_k for synth1 and synth2 are set to the same value. The third and fourth (synth3, synth4) datasets have $K = 4$ overlapping sub-matrices. Instead of larger complexities (K), sub-matrices in the synth3 and synth4 have different θ_ks from each other.

Initialization. Before conducting inference, we first initialize values of $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$, θ, and ϕ for both models. For the Bayesian Plaid models, we choose a specific K and initialize the model parameters according to the generative model. For the Infinite Plaid models, we may use the original generative model. However, in this paper, we want to show robustness of the Infinite Plaid models under incorrect assumptions of K. Thus, we dare to initialize the values of $\boldsymbol{Z}$ of Infinite Plaid models with the Bayes Plaid models with fixed K. We expect Infinite Plaid models can find a better K through inference, while K-fixing Bayesian Plaid models suffer from incorrect K. After inferences, we evaluated two models using the quantitative measures discussed earlier.

Hyperparameters. Two models have the four sets of hyperparameters. First one is the hyperparameter set for λ ($\boldsymbol{Z}$) prior: $a^\lambda_{1,2}, b^\lambda_{1,2}$ for Bayesian Plaid models, and $\alpha_{1,2}$ for Infinite Plaid Models. Second is the hyperparameter set for θ prior: μ^θ, τ^θ. Third is the hyperparameter set for ϕ prior: μ^ϕ, τ^ϕ. And finally we have τ_0 for the observation function. To focus on the impact of the initial sub-matrix numbers, we initialized all the hyperparameters of the models to the true values i.e. hyperparameters used in synthesizing the

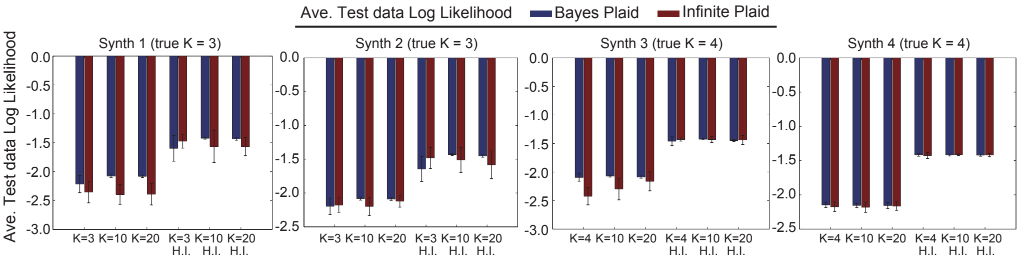


Figure 1: Averaged test data log likelihoods on synth4 data set. All log likelihoods are averaged over Gibbs iterations to obtain posterior averages. H.I. indicates that the hyperparameter inference is employed. Averages and standard deviations of 20 runs are presented.

Table 1: The averages and the standard deviations of the inferred final K by the Infinite Plaid models.

	$K^{\text{init}} = K^{\text{true}}$	$K^{\text{init}} = 5$	$K^{\text{init}} = 10$
Synth 1 ($K^{\text{true}} = 3$)	2.86 ± 0.25	5.23 ± 2.65	7.45 ± 4.63
Synth 2 ($K^{\text{true}} = 3$)	2.87 ± 0.44	4.16 ± 2.08	7.56 ± 4.11
Synth 3 ($K^{\text{true}} = 4$)	4.32 ± 0.39	4.22 ± 0.42	4.53 ± 0.85
Synth 4 ($K^{\text{true}} = 4$)	4.00 ± 0.25	4.77 ± 1.03	4.96 ± 1.36

data (excepting α for Infinite Plaid models). These hyperparameters are either (i) fixed during inferences or (ii) inferred simultaneously. In the manuscript, we only presented the results of (ii): all hyperparameters are inferred simultaneously.

Number of bi-clusters found. Table 1 presents the averages and the standard deviations of the final K inferred by the Infinite Plaid models on synthetic data sets. The numbers are in good accordance with the NMI measures. In general, the inference of K naturally becomes harder with the larger K^{init}. $K^{\text{init}} = 10$ for Synth1, 2 seems very difficult, judging from the averages. Larger standard deviations imply that the inference are trapped with local solutions with a variety of Ks.

Real-world data experiments

Pre-processing. For Enron datasets, we scaled the E-mail count values by $\log(1+x)$ where x is the actual E-mail count, to fit the Normal distribution. For Lastfm dataset, we remove all artists who have been tagged at most one word, and remove all tags that have been attached at most to one artist because we are interested in sub-matrices that group multiple objects.

Hyperparameters. It is fundamentally difficult to determine the best hyperparameters without the ground-truth information. In this experiment, we heuristically determined the initial values of hyperparameters. Collected real-world datasets are very sparse, which means that most of matrix entries are zero-valued. Thus we set the hyperparameters for ϕ (μ^{ϕ}, τ^{ϕ}) so that the distribution of ϕ concentrates near zero. We compute the averages and variance of non-zero entries, and use them as the hyperparameters for θ ($\mu^{\theta}, \tau^{\theta}$). For τ_0, we use the inverse of variances of the whole matrix data. These hyperparameters are either (i) fixed during inferences or (ii) inferred simultaneously. In the manuscript, we only presented the results of (ii): all hyperparameters are inferred simultaneously.

Result on Enron Aug. Data. Fig. 2 presents an example of sub-matrices from Enron Aug. data. The $k = 6$th submatrix (purple colored) is a cluster of VIP members. Note that the receivers consist of several Presidents, legal persons and the risk management head. We may imagine that some fatal problems are reported to these important persons to settle them. Another interesting sub-matrix is the $k = 3$rd (green colored) sub-matrix. This sub-matrix contains only one sender (1st domain object), who is the founder of Enron. He sends e-mails to many employees including many VIPs of group companies. (Ishiguro et al. 2010; Ishiguro, Ueda, and Sawada 2012) also found a similar partition, concluding that this specific relations suit well to the fact that *"the founder actually made an announcement to calm down the public"* ((Ishiguro et al. 2010)) concerning the resign of the Enron CEO at that time.

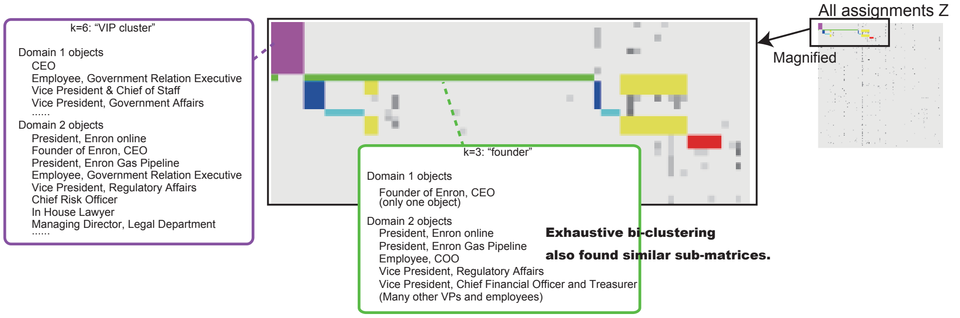
Result on Enron Nov. Data. Fig. 3 presents an example of sub-matrix extraction from Enron Nov. data. The $k = 1$st sub-matrix (red colored) sub-matrix is a VIP + legal expert community similar to the 8th sub-matrix at Enron Oct. data. Interestingly, there are a more bi-directional connections between executives than other months, implying more frequent contacts among VIPs at one month before the bankruptcy. The $k = 2$nd sub-matrix (orange colored) represents a small and tightly connected community. Unfortunately, we cannot examine details of the membership because included objects lacks demographic information, but we point out that the same community (with same members) appeared in all four datasets. One particular sub-matrix that was never found in other dataset is the $k = 8$th sub-matrix (sky-blue colored). This is a small community of traders and cash analysts.

Demonstration code sources

A Matlab implementation of the Baysian Plaid models and the Infinite Plaid models with simpler observation models is published in the GitHub: https://github.com/k-ishiguro/InfinitePlaidModels

References

Caldas, J., and Kaski, S. 2008. Bayesian Biclustering with the Plaid Model. In *Proceedings of the IEEE Interna-*



k=6: “VIP cluster”

Domain 1 objects
CEO
Employee, Government Relation Executive
Vice President & Chief of Staff
Vice President, Government Affairs
......

Domain 2 objects
President, Enron online
Founder of Enron, CEO
President, Enron Gas Pipeline
Employee, Government Relation Executive
Vice President, Regulatory Affairs
Chief Risk Officer
In House Lawyer
Managing Director, Legal Department
......

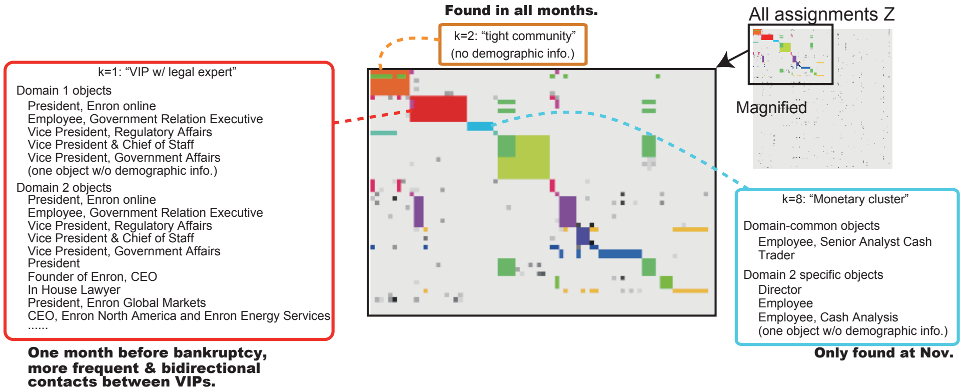
k=3: “founder”

Domain 1 objects
Founder of Enron, CEO
(only one object)

Domain 2 objects
President, Enron online
President, Enron Gas Pipeline
Employee, COO
Vice President, Regulatory Affairs
Vice President, Chief Financial Officer and Treasurer
(Many other VPs and employees)

Exhaustive bi-clustering also found similar sub-matrices.

Figure 2: Results on Enron Aug. data. Visualizations and descriptions of some sub-matrices.



Found in all months.

k=2: “tight community”
(no demographic info.)

k=1: “VIP w/ legal expert”

Domain 1 objects
President, Enron online
Employee, Government Relation Executive
Vice President, Regulatory Affairs
Vice President & Chief of Staff
Vice President, Government Affairs
(one object w/o demographic info.)

Domain 2 objects
President, Enron online
Employee, Government Relation Executive
Vice President, Regulatory Affairs
Vice President & Chief of Staff
Vice President, Government Affairs
President
Founder of Enron, CEO
In House Lawyer
President, Enron Global Markets
CEO, Enron North America and Enron Energy Services
......

One month before bankruptcy, more frequent & bidirectional contacts between VIPs.

k=8: “Monetary cluster”

Domain-common objects
Employee, Senior Analyst Cash Trader

Domain 2 specific objects
Director
Employee
Employee, Cash Analysis
(one object w/o demographic info.)

Only found at Nov.

Figure 3: Results on Enron Nov. data. Visualizations and descriptions of some sub-matrices.

tional Workshop on Machine Leaning for Signal Proceesing (MLSP).

Doshi-Velez, F.; Miller, K. T.; Van Gael, J.; and Teh, Y. W. 2009. Variational Inference for the Indian Buffet Process. In *Proceeding of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Griffiths, T. L., and Ghahramani, Z. 2011. The Indian Buffet Process : An Introduction and Review. *Journal of Machine Learning Research* 12:1185–1224.

Hoff, P. D. 2005. Subset clustering of binary sequences, with an application to genomic abnormality data. *Biometrics* 61(4):1027–1036.

Ishiguro, K.; Iwata, T.; Ueda, N.; and Tenenbaum, J. 2010. Dynamic Infinite Relational Model for Time-varying Relational Data Analysis. In Lafferty, J.; Williams, C. K. I.; Shawe-Taylor, J.; Zemel, R. S.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 23 (Proceedings of NIPS)*.

Ishiguro, K.; Ueda, N.; and Sawada, H. 2012. Subset Infinite Relational Models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume XX, 547–555.

Jain, S., and Neal, R. M. 2004. A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. *Journal of Computational and Graphical Statistics* 13:158–182.

Meeds, E. W.; Ghahramani, Z.; Neal, R. M.; and Roweis, S. 2007. Modeling Dyadic Data with Binary Latent Factors. In *Advances in Neural Information Processing Systems 19 (NIPS)*, 977–984.