Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Infinite Plaid Models for Infinite Bi-Clustering

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

Abstract

We propose a probabilistic model for non-exhaustive and overlapping (NEO) bi-clustering. Our goal is to extract a few sub-matrices from the given data matrix, where entries of a sub-matrix are characterized by a specific distribution or parameters. Existing NEO bi-clustering methods typically require the number of sub-matrices to be extracted, which is essentially difficult to fix a priori. In this paper, we extend the plaid model, known as one of the best NEO bi-clustering algorithms, to allow *infinite bi-clustering*; NEO bi-clustering without specifying the number of sub-matrices. Our model can represent infinite sub-matrices formally. We develop a MCMC inference without the finite truncation, which potentially addresses all possible numbers of sub-matrices. Experiments quantitatively and qualitatively verify the usefulness of the proposed model. The results reveal that our model can offer more precise and in-depth analysis of sub-matrices.

Introduction

In this paper, we are interested in bi-clustering for matrix data analysis. Given the data, the goal of bi-clustering is to extract a few (possibly overlapping) sub-matrices[1] from the matrix where we can characterize entries of a sub-matrix by a specific distribution or parameters. Bi-clustering is potentially applicable for many domains. Assume an mRNA expression data matrix, which consists of rows of *experimental conditions* and columns of *genes*. Given such data, biologists are interested in detecting pairs of *specific conditions* × *gene subsets* that have different expression levels compared to other expression entries. From a *user* × *product* purchase record, we can extract sub-matrices of *selected users* × *particular products* that sell very good. Successful extraction of such sub-matrices is the basis for efficient ad-targeting.

More specifically we study the *non-exhaustive* and *overlapping* (NEO) bi-clustering. Here we distinguish between the *exhaustive* and *non-exhaustive* bi-clustering.

Exhaustive bi-clustering (e.g. (Erosheva, Fienberg, and Lafferty 2004; Kemp et al. 2006; Roy and Teh 2009; Nakano et al. 2014)) is an extension of typical clustering (such as k-means) for matrices: the entire matrix is partitioned into many rectangle blocks, and all matrix entries are assigned to one (or more) of these blocks (Fig. 1). However, for knowledge discovery from the matrix, we are only interested in few sub-matrices that are essential or interpretable. For that purpose exhaustive bi-clustering techniques require post-processing in which all sub-matrices are examined to identify "informative" ones, often by painful manual effort.

On the contrary, the goal of non-exhaustive bi-clustering methods is to extract the most informative or significant sub-matrices, not partitioning all matrix entries. Thus the non-exhaustive bi-clustering is more similar to clique detection problems in network analysis: not all nodes in a network are extracted as clique members. Non-exhaustive bi-clustering would greatly reduce the man-hour costs of manual inspections and knowledge discovery, because it ignores non-informative matrix entries (Fig. 1).

Hereafter let us use "bi-clustering" to mean "NEO bi-clustering". There has been a plenty of bi-clustering researches over years, e.g. (Cheng and Church 2000; Lazzeroni and Owen 2002; Caldas and Kaski 2008; Shabalin et al. 2009; Fu and Banerjee 2009). However, there is one fatal problem that has not been solved yet: determining the number of sub-matrices to be extracted. Most existing bi-clustering methods assume that the number of sub-matrices is fixed a priori. This brings two essential difficulties. First, finding the appropriate number of sub-matrices is very difficult. Recall that determining "k" for classical k-means clustering is not trivial in general. And the same holds, perhaps more difficult (Gu and Liu 2008), for bi-clustering. Second, sub-optimal choices of the sub-matrix number inevitably degrade bi-clustering performances. For example, assume there are $K = 3$ sub-matrices incorporated in the given data matrix. Solving the bi-clustering by assuming $K = 2(< 3)$ never recover the true sub-matrices. If we solve with $K = 5(> 3)$, then the model seeks for two extra sub-matrices to fulfill the assumed K; e.g. splitting a correct sub-matrix into multiple smaller sub-matrices.

There are few works on this problem. One such work (Ben-Dor et al. 2003) suffers computational complexity equivalent to the cube to the number of columns; thus it is feasible only for matrices with very few columns. Gu and Liu (Gu and Liu 2008) adopted the model selection approach based on BIC. However model selection inevitably requires

[1]By sub-matrix, we mean a direct product of a subset of row indices and a subset of column indices.

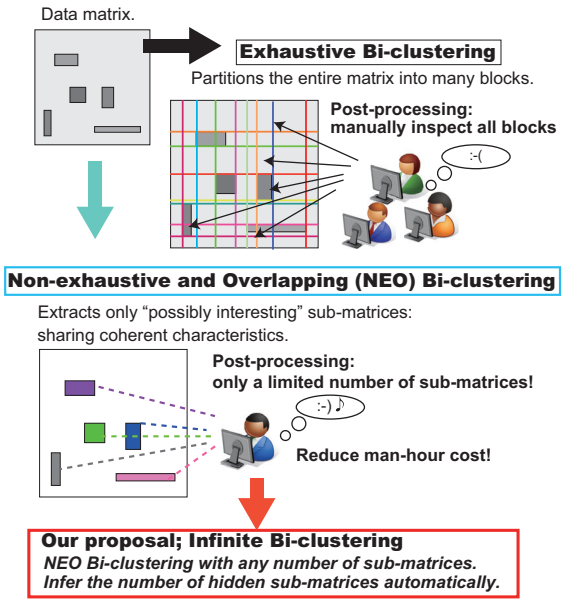

Figure 1: Exhaustive bi-clustering VS. NEO bi-clustering. We offer the *infinite* bi-clustering: NEO bi-clustering without knowing the exact number of distinctive sub-matrices. Row and column indices are not consecutive: thus we do not extract consecutive rectangles, but direct products of subsets of rows and columns.

multiple inference trials for different choices of model complexities (the number of sub-matrices, K). This consumes a lot of computation and time resources.

The main contribution of this paper is to propose a probabilistic model that allows **infinite bi-clustering**; NEO bi-clustering without specifying the number of sub-matrices. The proposed model is based on the plaid models (Lazzeroni and Owen 2002), which are known to be one of the best bi-clustering methods (Eren et al. 2013; Oghabian et al. 2014). The proposed *Infinite Plaid models* introduce a simple extension of the Indian Buffet Process (Griffiths and Ghahramani 2011) and can formulate bi-clustering patterns with infinitely many sub-matrices. We develop a MCMC inference that allows us to infer the appropriate number of sub-matrices for the given data automatically. The inference does not require the finite approximation of typical variational methods. Thus it can potentially address all possible numbers of sub-matrices, unlike the existing variational inference method for similar prior model (Ross et al. 2014). Experiment results show that the proposed model quantitatively outperforms the baseline finite bi-clustering method both for synthetic data and for real-world sparse datasets. We also qualitatively examine the extracted sub-matrices, and confirm that the Infinite Plaid models can offer in-depth sub-matrix analysis for several real-world datasets.

Background

Baseline: (simplified) Bayesian Plaid model

Bi-clustering has been studied intensively for years. A seminal paper (Cheng and Church 2000) has applied the bi-clustering technique for the analysis of gene-expression data. After that, many works have been developed (e.g. (Shabalin et al. 2009; Fu and Banerjee 2009)). Among them, the Plaid model (Lazzeroni and Owen 2002) is recognized as one of the best bi-clustering methods in several review studies (Eren et al. 2013; Oghabian et al. 2014). Bayesian models of the plaid model have been also proposed and reported effective (Gu and Liu 2008; Caldas and Kaski 2008).

We adopted a simplified version of Bayesian Plaid models (Caldas and Kaski 2008) as the baseline. The observed data is a matrix of $N_1 \times N_2$ continuous values, and K is the number of sub-matrices to model interesting parts. We define the simplified Bayesian Plaid model as follows:

$$\lambda_{1,k} \sim \mathrm{Beta}\left(a_1^\lambda, b_1^\lambda\right) \quad \lambda_{2,k} \sim \mathrm{Beta}\left(a_2^\lambda, b_2^\lambda\right), \tag{1}$$

$$z_{1,i,k} \sim \mathrm{Bernoulli}\left(\lambda_{1,k}\right), \quad z_{2,j,k} \sim \mathrm{Bernoulli}\left(\lambda_{2,k}\right), \tag{2}$$

$$\theta_k \sim \mathrm{Normal}\left(\mu^\theta, (\tau^\theta)^{-1}\right), \ \phi \sim \mathrm{Normal}\left(\mu^\phi, (\tau^\phi)^{-1}\right), \tag{3}$$

$$x_{i,j} \sim \mathrm{Normal}\left(\phi + \sum_k z_{1,i,k} z_{2,j,k} \theta_k, (\tau_0)^{-1}\right). \tag{4}$$

In the above equations, $k \in \{1, \ldots, K\}$ denotes sub-matrices, $i \in \{1, \ldots, N_1\}$ and $j \in \{1, \ldots, N_2\}$ denote objects in the first (row) and the second (column) domains, respectively. $\lambda_{1,k}$ and $\lambda_{2,k}$ (Eq. (1)) are the probabilities of assigning an object to the kth sub-matrix in the first and the second domain. $z_{1,i,k}$ and $z_{2,j,k}$ in Eq. (2) represent the sub-matrix (factor) memberships. If $z_{1,i,k} = 1(0)$ then the ith object of the first domain is (not) a member of the kth sub-matrix, and similar for $z_{2,j,k}$. θ_k and ϕ in Eq. (3) are the mean parameters for the kth sub-matrix and the "background" factor. Eq. (4) combines these to generate an observation. Note that this observation process is simplified from the original Plaid models. Throughout the paper, however, we technically focus on the modeling of $\boldsymbol{Z}$ thus we employ this simplified model.

As stated, existing bi-clustering methods including the Bayesian Plaid model require us to fix the number of sub-matrices, K, beforehand. It is very possible to choose a sub-optimal K in practice since it is innately difficult to find the best K by hand.

Indian Buffet Process (IBP)

Many researchers have studied sophisticated techniques for exhaustive bi-clustering. Especially, the Bayesian Nonparametrics (BNP) becomes a standard tool for the problem of an unknown number of sub-matrices in exhaustive bi-clustering (Kemp et al. 2006; Roy and Teh 2009; Nakano et al. 2014; Zhou et al. 2012; Ishiguro, Ueda, and Sawada 2012). Thus it is reasonable to consider the BNP approach for the (NEO) bi-clustering.

For that purpose, we would first consider an IBP (Griffiths and Ghahramani 2011), a BNP model for binary factor matrices. Assume the following Beta-Bernoulli process for N collections of $K = \infty$ feature factors $\boldsymbol{F} = \{F_i\}_{i=1,\ldots,N}$:

$B \sim \mathrm{BP}(\alpha, B_0)$, $F_i \sim \mathrm{BerP}(B)$. $\mathrm{BP}(\alpha, B_0)$ is a Beta process with a concentration parameter α and a base measure B_0. $B = \sum_{k=0}^{\infty} \lambda_k \delta_{\theta_k}$ is a collection of infinite pairs of $(\lambda_k, \theta_k), \lambda_k \in [0,1], \theta_k \in \Omega$. F_i is sampled from the Bernoulli process (BerP) as $F_i = \sum_{k=0}^{\infty} z_{i,k} \delta_{\theta_k}$ using a binary variable $z_{i,k} \in \{0,1\}$. IBP is defined as a generative process of $z_{i,k}$:

$$v_l \sim \mathrm{Beta}(\alpha, 1)\ ,\ \lambda_k = \prod_{l=1}^{k} v_l\ ,\ z_{i,k} \sim \mathrm{Bernoulli}(\lambda_k)\ .$$

$\boldsymbol{Z} = \{z_{i,k}\}$ acts like an infinite-K extension of $\boldsymbol{Z}_1$ or $\boldsymbol{Z}_2$ in Eq. (2). Assume that we observed N objects, and an object i may involve $K = \infty$ clusters. Then $z_{i,k}$ serves as an membership indicator of the object i to a cluster k.

Proposed model

Unfortunately, a vanilla IBP cannot achieve the infinite bi-clustering. This is because infinite bi-clustering requires to associate a sub-matrix parameter θ_k with *two* λs, which correspond to *two* binary variables $z_{1,i,k}, z_{2,j,k}$ (Eqs.(1,2)). However, the IBP ties the parameter θ_k with only *one* parameter as it is built on the pairs of (λ_k, θ_k) in BP.

For example, (Miller, Griffiths, and Jordan 2009) proposed to employ an IBP to factorize a square matrix. This model and its followers (Palla, Knowles, and Ghahramani 2012; Kim and Leskovec 2013), however, cannot perform bi-clustering on non-square matrix data, such as *user* $\times$ *item* matrix. (Whang, Rai, and Dhillon 2013) employed a product of independent IBPs to over-decompose a matrix into sub-matrices, and associated binary variables over sub-matrices to choose "extract" or "not-extract". This is not yet optimal because this model explicitly models the not-extract sub-matrices, results in unnecessary model complications.

Infinite Plaid models for infinite bi-clustering

Now we propose the *Infinite Plaid models*, an infinite bi-clustering model for general non-square matrices with the plaid factor observation models:

$$\boldsymbol{Z}_1, \boldsymbol{Z}_2 \mid \alpha_1, \alpha_2 \sim \mathrm{TIBP}(\alpha_1, \alpha_2) \tag{5}$$

$$\theta_k \sim \mathrm{Normal}\left(\mu^{\theta}, (\tau^{\theta})^{-1}\right)\ ,\ \phi \sim \mathrm{Normal}\left(\mu^{\phi}, (\tau^{\phi})^{-1}\right)\ , \tag{6}$$

$$x_{i,j} \sim \mathrm{Normal}\left(\phi + \sum_k z_{1,i,k} z_{2,j,k} \theta_k\ , (\tau_0)^{-1}\right)\ . \tag{7}$$

In Eq. (5), the Two-way IBP (TIBP), which will be explained later, is introduced to generate two binary matrices for sub-matrix memberships: $\boldsymbol{Z}_1 = \{z_{1,i,k}\} \in \{0,1\}^{N_1 \times \infty}$ and $\boldsymbol{Z}_2 = \{z_{2,j,k}\} \in \{0,1\}^{N_2 \times \infty}$. Different from the Bayesian Plaid models, the cardinality of sub-matrices may range to $K = \infty$; i.e. the Infinite Plaid models can formulate bi-cluster patterns with infinitely many sub-matrices. Remaining generative processes are the same as in the Bayesian Plaid models.

Now let us explain what is TIBP and why TIBP can remedy the problem of IBP. TIBP assumes *triplets* consists of *2 weights and an atom*, instead of pairs of *a weight and an atom* of IBP. Consider an infinite collection of triplets $B = (\lambda_{1,k}, \lambda_{2,k}, \theta_k)$, where $\lambda_{1,k}, \lambda_{2,k} \in [0,1]$ are the probability of activating a sub-matrix (factor) k in the 1st domain and the 2nd domain, respectively, and θ_k is a parameter drawn from the base measure B_0. Given B, we generate feature factors for 2-domain indices (i,j): $F_{(i,j)} = \sum_{k=1}^{\infty} z_{1,i,k} z_{2,j,k} \delta_{\theta_k}$ where two binary variables $z_{1,i,k}$ and $z_{2,j,k}$ represent sub-matrix(factor) responses.

More precisely, TIBP is defined as a generative process of such $z_{1,i,k}$ and $z_{2,j,k}$ that extends IBP:

$$v_{1,l} \sim \mathrm{Beta}(\alpha_1, 1)\ ,\ \lambda_{1,k} = \prod_{l=1}^{k} v_{1,l}\ ,\ z_{1,i,k} \sim \mathrm{Bernoulli}(\lambda_{1,k})\ ,$$

$$v_{2,l} \sim \mathrm{Beta}(\alpha_2, 1)\ ,\ \lambda_{2,k} = \prod_{l=1}^{k} v_{2,l}\ ,\ z_{2,j,k} \sim \mathrm{Bernoulli}(\lambda_{2,k})\ .$$

Repeating for all (i,j), we obtain $\boldsymbol{Z}_1 = \{z_{1,i,k}\} \in \{0,1\}^{N_1 \times \infty}$ and $\boldsymbol{Z}_2 = \{z_{2,j,k}\} \in \{0,1\}^{N_2 \times \infty}$. In shorthand we write: $\boldsymbol{Z}_1, \boldsymbol{Z}_2 \mid \alpha_1, \alpha_2, \sim \mathrm{TIBP}(\alpha_1, \alpha_2)$. Because of the triplets, a parameter atom θ_k and the corresponding index k are associated between $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ for $k \in \{1, 2, \ldots, \infty\}$. Therefore the binary matrices generated from TIBP can serve as valid membership variables for infinite bi-clustering.

TIBP is a generic BNP prior related to the marked Beta process (Zhou et al. 2012). It would be applicable for many other problems, not limited to the infinite bi-clustering. For example, (Ross et al. 2014) employed a similar infinite bi-clustering prior for a mixture of Gaussian processes to learn clusters of disease trajectories. In this paper, we employ TIBP for totally different problem; unsupervised infinite bi-clustering of plaid models.

Inference without truncations

We usually use the variational inference (VI) or the MCMC inferences to estimate unknown variables in probabilistic models. The aforementioned work by (Ross et al. 2014) employed the VI, which utilizes a finite truncation of TIBP with the fixed K^+ model in its variational approximation, where K^+ is the assumed "maximum" number of sub-matrices. To do this, VI achieves an analytical approximation of the posterior distributions.

Instead, we developed a MCMC inference combining Gibbs samplers and Metropolis-Hastings (MH) samplers (c.f. (Meeds et al. 2007)), which approximate the true posteriors by sampling. Roughly speaking, the Gibbs samplers infer individual parameters (θ, ϕ) and hidden variables (z) while the MH samplers test drastic changes in the number of sub-matrices such as split-merge moves and a proposal of new sub-matrices. It is noteworthy that our MCMC inference does not require finite truncations of TIBP unlike VI. Thus it yields a posterior inference of the infinite bi-clustering that potentially address all possible numbers of sub-matrices, not truncated at K^+. In addition, we can infer the hyperparameters to boost bi-clustering performances (c.f. (Hoff 2005; Meeds et al. 2007)). Please consult the supplemental material for inference details.

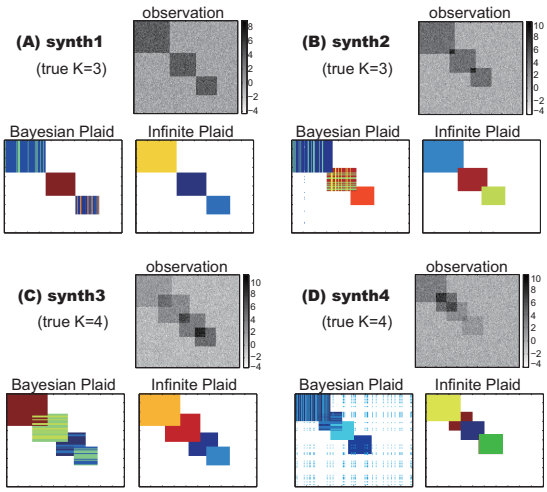

Figure 2: Observations and typical bi-clustering results on synthetic datasets. A colored rectangle indicates a sub-matrix k ((i, j) $s.t.\, z_{1,i,k} = z_{2,j,k} = 1$). Sub-matrix overlapping is allowed: for example, in Panel (B), three sub-matrices overlap each other on their corners. For better presentations we sort the rows and column so as to sub-matrices are seen as rectangles. In experiments row and column indices are randomly permuted.

Experiments

Procedure

We prepared four synthetic small datasets (Fig. 2). All data sets are $N_1(=100) \times N_2(=200)$ real-valued matrices, but differ in the number of sub-matrices and the proportion of sub-matrix overlap. For real-world datasets, we prepared the following datasets. The **Enron** E-mail dataset is a collection of E-mail transactions in the Enron Corporation (Klimt and Yang 2004). We computed the number of monthly transactions of E-mails sent/received between $N_1 = N_2 = 151$ employees in 2001. We used transactions of Aug., Oct., Nov., and Dec. when transactions were active. We also collected a larger **Lastfm** dataset, which is a co-occurrence count set of artist names and tag words. The matrix consists of $N_1 = 10,425$ artists and $N_2 = 4,212$ tag words. All real-world datasets are sparse: the densities of non-zero entries are, at most, 4% for Enron, and 0.3% for Lastfm. For dataset details, please consult the supplemental materials.

Our interest is that how good the Infinite Plaid models can solve bi-clustering without knowing the true number of sub-matrices, K^{true}. Thus we set the same initial hyper-parameter values for the baseline Bayesian Plaid models and the proposed Infinite Plaid models. For the choices of K, the Bayesian Plaid models conduct inferences with the fixed number of sub-matrices K. The Infinite Plaid models are initialized with K^{init} sub-matrices, then adaptively infer an appropriate number of sub-matrices through inferences. All other hyperparameters of two models are inferred via MCMC inferences.

Table 1: Average NMI values on synthetic data with different K and K^{init}. B.P. indicates the baseline Bayesian Plaid models, and Inf. P. indicates the proposed Infinite Plaid models. Bold faces indicate statistical significance.

	$K, K^{\text{init}} = K^{\text{true}}$		$K, K^{\text{init}} = 5$		$K, K^{\text{init}} = 10$	
	B. P.	Inf. P.	B. P.	Inf. P.	B. P.	Inf. P.
Synth 1	0.848	**0.970**	0.791	**0.860**	0.700	**0.791**
Synth 2	0.782	**0.976**	0.754	**0.924**	0.659	**0.823**
Synth 3	0.938	0.973	0.777	**0.972**	0.660	**0.971**
Synth 4	0.975	0.981	0.754	**0.963**	0.647	**0.940**

Table 2: Two additional criteria scores on synthetic and real data. B.P. indicates the baseline Bayesian Plaid models, and Inf. P. indicates the proposed Infinite Plaid models. Bold faces indicate statistical significance.

	Distinctiveness		F-measure	
	B. P.	Inf. P.	B. P.	Inf. P.
Synth 1	2.8	**3.5**	0.94	0.94
Synth 2	2.6	**3.7**	0.94	0.94
Synth 3	0.4	**1.8**	0.79	**0.84**
Synth 4	0.1	**1.4**	0.69	**0.72**
Enron Aug.	1.1	1.2	0.32	**0.42**
Enron Oct.	0.98	0.96	0.47	**0.54**
Enron Nov.	1.1	1.0	0.38	**0.53**
Enron Dec.	0.93	1.1	0.35	**0.48**
Lastfm	**1.3**	0.89	0.11	**0.18**

For quantitative evaluations, we employed the **Normalized Mutual Information (NMI)** for overlapping clustering (Lancichinetti, Fortunato, and Kertesz 2009). NMIs take the maximum value of 1.0 if and only if the two clustering are completely the same, including the number of clusters. NMIs yield a precise evaluation of bi-clustering but require ground truth sub-matrix assignments, which are unavailable in general. Thus we also consider two generic criteria that work without ground truth information. First, we want sub-matrices that have remarkably different θ_k compared to the background ϕ. For that purpose, we compute a **distinctiveness**: $\min_k |\theta_k - \phi|$. Next, we expect the sub-matrices include many informative entries but less non-informative entries. Fortunately, it is relatively easy to define (non-)informative entries without the ground truth for sparse datasets: the dominant "zero" entries might be non-informative while non-zero entries are possibly informative. Let us denote E as the set of all entries extracted by K sub-matrices and I as the set of all non-zero entries. Then we compute the **F-measure** by: $\frac{2\text{Recall}\cdot\text{Precision}}{\text{Recall}+\text{Precision}}$ where Recall $= \frac{\#\cap(E,I)}{\#I}$ and Precison $= \frac{\#\cap(E,I)}{\#E}$. Larger values of these criteria imply distinct and clean sub-matrices.

Synthetic Data Experiments

Table 1 presents averages of NMIs on synthetic datasets with several K and K^{init}. We see that the proposed Infinite Plaid models achieved significantly better NMIs against the baseline Bayesian Plaid models. Also the proposed model obtained NMI $\approxeq 1.0$ i.e. the perfect bi-clustering in many cases, regardless of K^{init} values. This makes contrast with Bayesian

Plaid models suffer sharp falls in NMIs as K increases. This is what we expect: we designed the Infinite Plaid model so that it is capable of inferring a number of sub-matrices, while the baseline model and many existing bi-clustering methods cannot perform such inference.

We observe that the first two synthetic data (synth1, synth2) are more difficult than the later two data (synth3, synth4). This is reasonable since all the sub-matrices of the first two data were designed to have the same value θ_k while the later two data were designed with different θ_k values (θ_ks are illustrated in Fig. 2 as the color depth of rectangles). We also observe that Bayesian Plaid model does not necessarily obtain the perfect recovery even if $K = K^{\text{true}}$. This may be explained by the fact that the MCMC inferences on BNP exhaustive bi-clustering models are easily trapped at local optimum in practice (Albers et al. 2013). We expect the performance will be improved with more MCMC iterations.

The upper half of Table 2 presents the additional two criteria ($K, K^{\text{init}} = 10$). The Infinite Plaid models are significantly better than the baseline for many cases in both of the distinctiveness and the F-measure. Combining the NMI results, we can safely say that the proposed model is superior to the fixed-K bi-clustering.

Fig. 2 shows examples of bi-clustering results ($K, K^{\text{init}} = 10$). The Bayesian Plaid models often extract noisy background factors, or divide sub-matrices into multiple blocks unnecessarily. Those are exactly the outcomes we were afraid of, induced by the sub-optimal K. In contrast, Infinite Plaid models achieved perfect bi-clusters, as expected from the NMI scores.

Real-world Datasets Experiments

For the real-world datasets, we rely on the distinctiveness and F-measure as there are no ground truth sub-matrix assignments. Evaluations are presented in the lower half of Table 2 (because we do not know K^{true}, we presented the best scores among several choices of K, K^{init}, in terms of the distinctiveness). We confirmed that the Infinite Plaid models always outperform the baseline in terms of the F-measure. This effectively demonstrates the validity of our infinite bi-clustering model for sparse real-world datasets.

Enron dataset Next, we qualitatively examine the results of the Infinite Plaid models on Enron datasets. Please refer to the supplemental material for the results of Enron Aug. and Enron Nov. datasets.

Fig. 3 presents an example of sub-matrices from Enron Oct. data. The $k = 2$nd sub-matrix (orange colored) highlights the event in which the COO sent many mails to Enron employees. This pattern is also found by existing exhaustive bi-clustering works (Ishiguro et al. 2010; Ishiguro, Ueda, and Sawada 2012) but our model reveals that there are a few additional employees who behaved like the COO. We also find two VIP sub-matrices ($k = 8, 10$). We may distinguish these two by the presence of legal experts and the founder. The $k = 8$th sub-matrix (light-blue colored) includes these people, but the $k = 10$th sub-matrix (dark-blue colored) does not. A few people join both sub-matrices, thus two sub-matrices may be exchanging e-mails about a

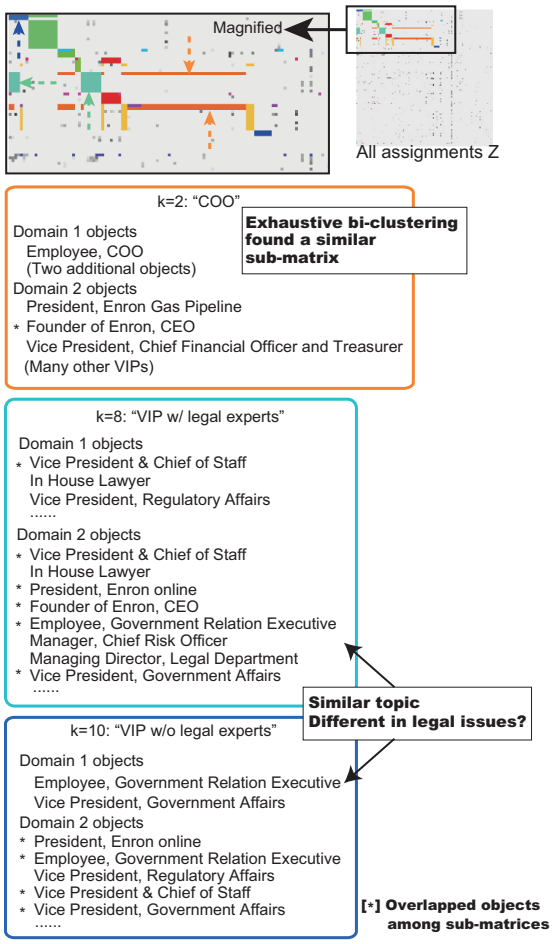

Figure 3: Results on Enron Oct. data.

similar topic, but from different viewpoints concerning legal issues. Interestingly, members in these two sub-matrices are grouped in one sub-matrix in August data (presented in the supplement).

Fig. 4 presents an example of sub-matrices from Enron Dec., the final days of the bankruptcy. As the Enron became bankrupt, human resources mattered a lot. The $k = 3$rd sub-matrix (green colored) may capture this factor. The main sender of the sub-matrix is the Chief of Staff. The receivers were CEOs, Presidents and directors of several Enron group companies. In this month, we found another interesting pair of the sub-matrices: the $k = 1$st and the $k = 6$th sub-matrices (red and sky-blue colored). In the $k = 1$st sub-matrix, the President of Enron Online sent many mails to VIPs. We don't know the content of these e-mails, but they may be of significant interest as they were sent by the president of one of the most valuable Enron group companies. Interestingly, in the 6th sub-matrix the president is the sole receiver of e-mails and the number of senders (the 1st domain membership) is smaller than that of receivers (the 2nd domain membership) of the $k = 1$st sub-matrix. This may imply that the $k = 6$th sub-matrix captured the responses to the e-mails

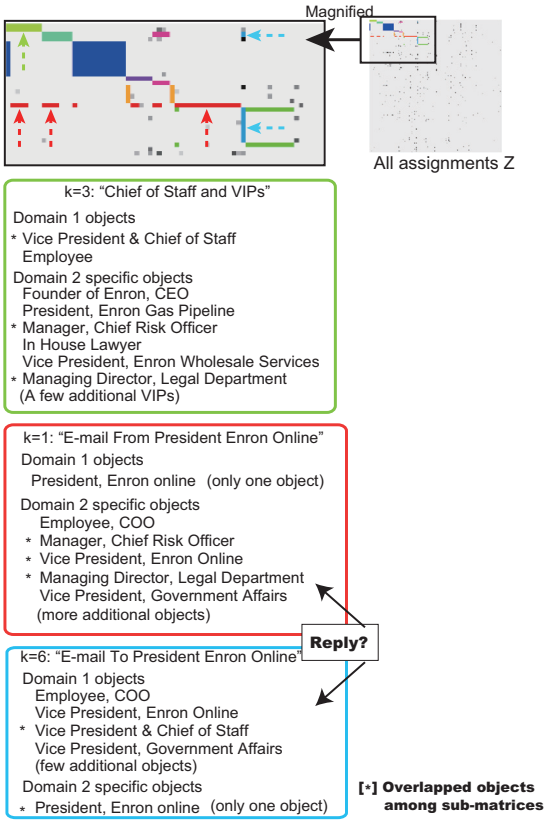

Figure 4: Results on Enron Dec. data.

in the $k = 1$st sub-matrix.

Lastfm dataset Fig. 5 shows an example of the results of applying the Infinite Plaid models to the Lastfm dataset. The $k = 3$rd sub-matrix consists of black metal bands (1st domain objects). This sub-matrix selects only one tag (2nd domain object): "black metal", which might be appropriate for these bands. The $k = 8$th sub-matrix is a sub-matrix of relatively pop music. This sub-matrix has tags such as "alternative rock", "pop punk", and "grunge". Indeed, there are some alternative rock bands such as "Coldplay", popular punk bands such as "Green Day", while "Nirvana" is one of the best grunge bands. We also found a sub-matrix of older rock artists at $k = 4$. Included classical artists are: "Beatles", "John Lennon", "Deep Purple" (Hard Rock), "Judas Priest" (Metal), and others.

Finally we present slight strange but somewhat reasonable assignments. The $k = 2$nd sub-matrix consists of various top female artists. Chosen tags are "electronic", "pop", "dance", "female vocalist", and "rnb" (R&B). Among the chosen artists, however, there are two male artists; "Michael Jackson" and "Justin Timberlake". These choices have somewhat reasonable aspects: both are very famous pop and dance music grandmasters, and both are characterized by their high-tone voices like some female singers.

These deep analyses of sub-matrices would cost much

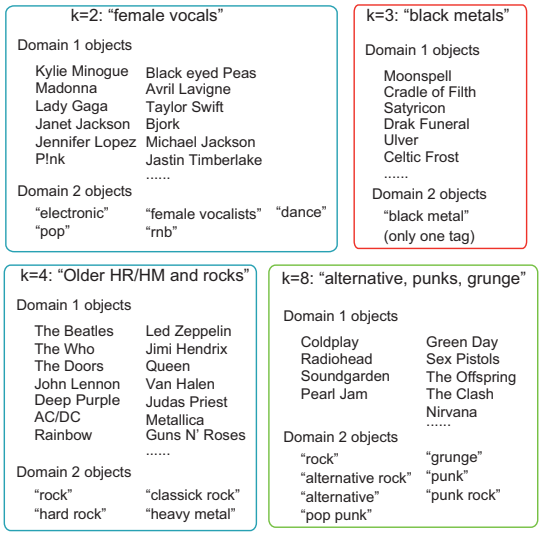

Figure 5: Results on Lastfm data.

with existing exhaustive bi-clustering methods which require manual checking of all partition blocks, or with existing NEO bi-clustering methods which require model selection. The Infinite Plaid models reduced the cost of knowledge discovery by extracting remarkably different sub-matrices automatically.

Conclusion and Future Works

We presented an infinite bi-clustering model that solves the NEO bi-clustering without knowing the number of sub-matrices. We proposed the Infinite Plaid models, which extend the well-known Plaid models to represent infinite sub-matrices by a BNP prior. Our MCMC inference allows the model to infer an appropriate number of sub-matrices to describe the given matrix data. Our inference does not require the finite truncation as it is required in the previous method (Ross et al. 2014); thus the inference algorithm correctly addresses all possible patterns of the infinite bi-clusters. We experimentally confirmed that the proposed model quantitatively outperforms the baseline method fixing the number of sub-matrices a priori. Qualitative evaluations showed that the proposed model allows us to conduct more in-depth analysis of bi-clustering in real-world datasets.

Plaid models have been intensively employed in gene expression data analysis, and also we are interested in the purchase log bi-clustering and social data analyses as discussed in the introduction. In addition, infinite bi-clustering can be used for multimedia data such as images and audios. For example, it is well known that the nonnegative matrix factorization can extract meaningful factors from image matrices (Cichocki et al. 2009), such as eyes, mouths, and ears from human face images. The infinite bi-clustering may improve in finding meaningful parts by just focusing on distinctive sub-images. Some audio signal processing researchers are interested in audio signals with the time-

frequency domain representations that are real- or complex-valued matrix data. We can apply the infinite bi-clustering to those matrices to capture time-depending acoustic harmonic patterns of music instruments, which are of interest of music information analysis((Sawada et al. 2013; Kameoka et al. 2009)).

Finally we list a few open questions. First, explicitly capturing relationships between sub-matrices would be of interest for further deep bi-clustering analysis. Second, it is important to verify the limitation of the Infinite Plaid models against the number of hidden sub-matrices, the noise tolerance, and overlaps. Finally, computational scalability matters in today's big data environment. We are considering stochastic inferences (e.g (Hernandez-Lobato, Houlsby, and Ghahramani 2014; Wang and Blei 2012)) for larger matrices.

A supplemental material and information for a MATLAB demo program package can be found at: http://www.kecl.ntt.co.jp/as/members/ishiguro/index.html

The part of research results have been achieved by "Research and Development on Fundamental and Utilization Technologies for Social Big Data", the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

Albers, K. J.; Moth, A. L. A.; Mørup, M.; and Schmidt, M. N. . 2013. Large Scale Inference in the Infinite Relational Model: Gibbs Sampling is not Enough. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*.

Ben-Dor, A.; Chor, B.; Karp, R.; and Yakhini, Z. 2003. Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of computational biology* 10(3-4):373–384.

Caldas, J., and Kaski, S. 2008. Bayesian Biclustering with the Plaid Model. In *Proceedings of the IEEE International Workshop on Machine Leaning for Signal Proceesing (MLSP)*.

Cheng, Y., and Church, G. M. 2000. Biclustering of Expression Data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 93–103.

Cichocki, A.; Zdunek, R.; Phan, A. H.; and Amari, S.-i. 2009. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley.

Eren, K.; Deveci, M.; Küçüktunç, O.; and Çatalyürek, Ü. V. 2013. A ccomparative analysis of biclustering algorithms for gene expression dat. *Briefings in Bioinformatics* 14(3):279–292.

Erosheva, E.; Fienberg, S.; and Lafferty, J. 2004. Mixed-membership Models of Scientific Publications. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 101(Suppl 1):5220–5227.

Fu, Q., and Banerjee, A. 2009. Bayesian Overlapping Subspace Clustering. *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM)* 1:776–781.

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

Gu, J., and Liu, J. S. 2008. Bayesian biclustering of gene expression data. *BMC genomics* 9 Suppl 1:S4.

Hernandez-Lobato, J. M.; Houlsby, N.; and Ghahramani, Z. 2014. Stochastic Inference for Scalable Probabilistic Modeling of Binary Matrices. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32.

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.

Kameoka, H.; Ono, N.; Kashino, K.; and Sagayama, S. 2009. Complex NMF: A new sparse representation for acoustic signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, 2353–2356.

Kemp, C.; Tenenbaum, J. B.; Griffiths, T. L.; Yamada, T.; and Ueda, N. 2006. Learning Systems of Concepts with an Infinite Relational Model. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*.

Kim, M., and Leskovec, J. 2013. Nonparametric Multigroup Membership Model for Dynamic Networks. In *Advances in Neural Information Processing Systems 26 (Proceedings of NIPS)*, 1–9.

Klimt, B., and Yang, Y. 2004. The Enron Corpus : A New Dataset for Email Classification Research. In *Proceedings of the European Conference on Machine Learning (ECML)*.

Lancichinetti, A.; Fortunato, S.; and Kertesz, J. 2009. Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics* 11(3).

Lazzeroni, L., and Owen, A. 2002. Plaid Models for Gene Expression Data. *Statistica Sinica* 12:61–86.

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.

Miller, K. T.; Griffiths, T. L.; and Jordan, M. I. 2009. Nonparametric Latent Feature Models for Link Prediction. In Bengio, Y.; Schuurmans, D.; Lafferty, J.; Williams, C. K. I.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 22 (Proceedings of NIPS)*.

Nakano, M.; Ishiguro, K.; Kimura, A.; Yamada, T.; and Ueda, N. 2014. Rectangular Tiling Process. In *Proceedings*

of the 31st International Conference on Machine Learning (ICML), volume 32.

Oghabian, A.; Kilpinen, S.; Hautaniemi, S.; and Czeizler, E. 2014. Biclustering methods: biological relevance and application in gene expression analysis. *PloS one* 9(3):e90801.

Palla, K.; Knowles, D. A.; and Ghahramani, Z. 2012. An Infinite Latent Attribute Model for Network Data. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*.

Ross, J. C.; Castaldi, P. J.; Cho, M. H.; and Dy, J. G. 2014. Dual Beta Process Priors for Latent Cluster Discovery in Chronic Obstructive Pulmonary Disease. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 155–162.

Roy, D. M., and Teh, Y. W. 2009. The Mondrian Process. In *Advances in Neural Information Processing Systems 21 (Proceedings of NIPS)*.

Sawada, H.; Kameoka, H.; Araki, S.; and Ueda, N. 2013. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech and Language Processing* 21(5):971–982.

Shabalin, A. A.; Weigman, V. J.; Perou, C. M.; and Nobel, A. B. 2009. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics* 3(3):985–1012.

Wang, C., and Blei, D. M. 2012. Truncation-free Stochastic Variational Inference for Bayesian Nonparametric Models. In *Advances in Neural Information Processing Systems 25 (Proceedings of NIPS)*.

Whang, J. J.; Rai, P.; and Dhillon, I. S. 2013. Stochastic blockmodel with cluster overlap, relevance selection, and similarity-based smoothing. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 817–826.

Zhou, M.; Hannah, L. A.; Dunson, D. B.; and Carin, L. 2012. Beta-Negative Binomial Process and Poisson Factor Analysis. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*.