Yin Yang Mixture Model

Daniel Rugeles

Supervisor: Asst. Professor Gao Cong
Dr. Shonali Krishnaswamy

School of Computer Engineering

Nanyang Technological University

Extended Report

March 2016

Abstract

The extraction of a short description of the relationship between two variables from a given matrix is useful for understanding the dependencies between the variables. Although, many bicluster algorithms have been proposed to solve this problem, they provide little information about the importance of each bicluster, as well as the importance of each variable in a bicluster. To address these problems, we propose a novel mixture model for bivariate data. We test the ability of our model to extract biclusters using a synthetic data and we demonstrate the applicability of our model using text data as well as gene expression data.In addition, we prove that our model generalizes the Latent Dirichlet Allocation model.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Background

In Data Mining, clustering corresponds to the task of dividing the data into groups of similar objects. Representing the data with clusters causes information loss but it achieves simplification and it allows discovering the structure of the data. Clustering is applied to any dataset that can be represented by a matrix where the rows correspond to the set of variables that we want to cluster and the columns correspond to the set of attributes in the dataset. Every cell in the matrix corresponds to the value of one of the attributes assigned to one of the variables. Clustering algorithms group the variables by measuring the similarities between the entire set of attributes.

In some datasets, only a subset of the attributes influences the similarity between some of the variables. Therefore, it may be convenient to study which attributes influence which variables. Attempting to cluster these datasets with the entire set of attributes may lead to diffuse results [33]. Under this motivation, biclustering algorithms have been developed to cluster simultaneously the attributes and the variables into biclusters. A bicluster represents a subset of attributes that is highly related with a subset of the variables. The most common datasets where biclustering has been applied includes mobility data, microarray data and

Dataset	Clustering	Motivation	Biclustering
Check-in	Groups the users based on how frequent theY visit the entire set of places.	A subset of the users might commonly be found in a subset of the places. E.g. A user who is a student might visit places inside a school.	We obtain groups of users that visit a sets of places.
MicroArray	Group the genes based on their expression level across all the biological samples.	Even though all samples may be selected from the same part of the body, the cell's physiological state among other conditions might lead to some samples being unable to identify the typical expression of the cells extracted from such body part.	We obtain groups of genes that identify the expression pattern of tumor cells and we also obtain the subset of the samples that support this evidence.
Text	Group the authors based on the similarities between the words that they use to write their papers.	Authors use different words depending on the target audience of their work.	We obtain groups of authors who write about similar 'topics'. In this case, a 'topic' represents a group of words.

Table 1.1 Comparison between Clustering and Biclustering in three different scientific fields.

text data. In table 1.1, we present a comparison between clustering, biclustering and the motivation to use biclustering.

Biclustering algorithms aim to extract a set of biclusters, which are short descriptions of the interdependencies between two variables [33]. Given a matrix as the input, a bicluster is represented by a subset of the rows and a subset of columns that exhibit some relationship. Many types of relationships have been studied in the literature. For example, a submatrix whose values are constant is known as the constant bicluster and a submatrix whose rows are linearly dependent is known as the perfect additive bicluster [56]. All types of biclusters that have been studied in the literature will be presented in section 2.3.2.

In count data and datasets where the magnitude of a value indicates the importance of the relationship between two variables, it is indispensable to account for the magnitude of the values between variables. This is because a high magnitude represents a strong evidence of the relationship between a pair of variables. For instance, in mobility data the users tend to visit their preferred places. In text data, the authors tend to use the words of their preference. In Microarray data, a gene might respond to pathogens or other organisms when its value has a high magnitude. Therefore, it is required that we extract biclusters that maximize the sum of its values. We refer to such type of biclusters as *heavy biclusters*.

Alternatively, we may represent the input dataset using a bipartite graph. In the graph representation, the rows represent one set of vertices, the columns represent the other set of vertices and the edges represent weights extracted from the corresponding entries in the matrix. In the bipartite graph representation, the *heavy biclusters* correspond to extracting quasi-bicliques. This is because a high count represents a high weight in an edge, and so this edge will have higher probability to be assigned to one of the quasi-bicliques.

For illustration, we present a mobility dataset in table 1.2. The rows correspond to the set of users, and the columns correspond to the set of places. The values in each cell of the table correspond to the number of times that a user visited a place. The expected biclusters corresponds to the sub-matrices that include the higher values across the entire matrix. We visualize them in the table using blue and red color respectively. We observe that users U_1, U_2 and U_3 tend to visit frequently places P_1, P_2 and P_3. Therefore, they are assigned to one bicluster. Similarly users U_3, U_4 and U_5 tend to visit places P_4, P_5 and P_6 with a high frequency and they are assigned to another bicluster. We observe how user U_3 visits all the places in the dataset and thus it is expected that he may be assigned to both biclusters. On the contrary, there is not enough visits recorded for user U_6, and thus he should not be assigned to any bicluster. A similar case occurs for Place P_7, however a robust model might assign partial membership for P_7 to the red bicluster. Furthermore, it is desirable to extract additional information about these relationships. For instance, the importance of each user and place in each bicluster and the importance of each bicluster. In the example, we observe

	P_1	P_2	P_3	P_4	P_5	P_6	P_7
U_1	20	22	23	0	0	1	0
U_2	17	20	18	0	0	2	0
U_3	21	19	19	10	13	7	0
U_4	0	0	0	12	9	8	15
U_5	0	1	0	11	11	7	0
U_6	0	0	2	0	3	0	0

Table 1.2 Sample Mobility dataset for 6 users and 7 places. The expected biclusters are presented in blue and red.

that P_6 is not as typical as the other places in the red bicluster. Also, the blue bicluster is more important than the red bicluster because it has a stronger dependence as indicated by a higher number of visits.

In summary, it is important to discover these inter-dependent relationships and the structures of the biclusters should be flexible, i.e., it is important that the model supports cases where a row and column may be assigned to (1) only one bicluster (2) many biclusters, and (3) no bicluster. Furthermore, it is desirable to discover the importance of each row/column in biclusters, and the importance of the discovered biclusters. It is also beneficial that the algorithm does not assume the specification of the total number of biclusters a priori.

Some existing biclustering algorithms are able to extract this type of bicluster. However, these techniques typically do not support a flexible assignment of columns and rows to the biclusters and they assume the number of biclusters is known in advance. For example, Tanay et al takes a binary matrix as the input [64], Dhillon enforces that every row and column must be assigned to one bicluster [14], Wang et al assumes that every cell in the matrix must be part of one bicluster, and Mallela et al assumes a hard clustering [16]. More discussions can be found in Section 2. We conclude that current methods impose constraints that limit the discovery of the underlying relationships in the datasets.

1.2 Motivation

With the objective of relaxing the constraints imposed by related algorithms, we propose a novel algorithm for mining heavy biclusters from bivariate categorical data using a mixture model. We construct the model by observing that a mixture model may cluster the rows without taking into account the information of the columns or it may cluster the columns without taking into account the information of the rows. Nonetheless, we cannot do both at the same time despite they take the same input data. Mutual reinforcement of both approaches has shown improvement over a single approach in non-probabilistic clustering in the mobility domain [1] and the text domain [14]. Our approach uses one mixture model to cluster the rows and a second mixture model to cluster the columns. In addition, we model the interdependence between the topics with the objective of modeling this mutual reinforcement. The interdependence between topics is interpreted as a bicluster because it extracts the information of a group of rows related with a group of the columns. On the example presented, this information corresponds to: “Which group of users visit which groups of places."

Our formulation can be used to extract biclusters without constraining their structure and without specifying the number of biclusters in advance. In addition, the probability distributions characterizing the topics and its interrelation capture additional information not revealed by other algorithms. That is, the importance of the variables in each of the extracted biclusters and the importance of each bicluster.

To understand how we obtain these advantages, we have to start by analyzing the advantages of probabilistic clustering over non-probabilistic clustering in univariate data. Probabilistic models clusters univariate data using a mixture model. A mixture model uses a latent variable to model the data as a mixture of probability distributions. In the model, every probability distribution corresponds to a ’cluster’. Using this formulation, we obtain the following advantages:

First, a distribution provides the information about how relevant are each of the objects in the cluster. Second, mixture models provide an explanation for how the data was generated. In this case, the explanation is that the data is generated by sampling from a mixture of probability distributions. As a third advantage, mixture models can be used as components in more complex models. A good example is the Latent Dirichlet Allocation model (LDA) [8]. LDA uses a categorical mixture model to represent every document. In addition, LDA constraints the latent variables to be shared across all the documents in the corpus. Since the LDA uses a mixture model in its structure, it inherits the other advantages of mixture models. First, the latent variables (also known as topics) cluster the set of words and each 'cluster' includes the relevance of each of the words in it. Second, LDA uses the Mixture model to explain how the words are written in a document.

In biclustering, we can also inherit the advantages of mixture models:

Advantage 1. Since biclustering clusters the rows and the columns of a matrix. we propose to use a latent variable to cluster the rows into row-topics and another latent variable to cluster the columns into column-topics. From this modeling, we can obtain information about which rows are relevant in each row-topic and which columns are relevant in each column-topic.

Advantage 2. If we use every latent variable independently, we would obtain two mixture models clustering independently each of the dimensions. However, our goal is to extract biclusters that represent a strong relationship between the rows and the columns. To achieve this, we mix the two latent variables together. In other words, the mixture will encode which row-topics are relate to which column-topics. Because the mixture is also a probability distribution, we obtain the information about which biclusters exhibit a stronger relation between the topics.

Advantage 3. Being the model a Generative model, we obtain description of how the data is generated. Therefore, we can use the model to solve additional problems such the extraction of the joint and the conditional distributions between the rows and the columns.

Advantage 4. Due to the soft-clustering nature of probabilistic approaches, our model does not assume that every row and every column must belong to exactly one bicluster

We evaluate these advantages in section 4.1

1.3 Aim

The aim of our research is to design, and implement topic models that can be used to solve the following questions about the field of human mobility:

- Which POIs are the most popular given a group of users?
- Which are the most common groups of users that visit a group of similar places?
- Which POI would a user be likely to visit next?
- When do users prefer visiting a specific type of places?
- What is relationship between urban planning areas of Singapore in terms of mobility?

The model that we present in this report is able to solve the first two questions mentioned above, we will target on solving the other questions in our Future work as explained in section **??**

1.4 Objectives

The steps that we will take to achieve our aims are:

1. To model, derive and implement a topic model for the generation of users and places.
2. To design and implement a probabilistic model for the automatic labeling of location data obtained from cellular towers.
3. To model, derive and implement a topic model for simulating human mobility.
4. Formulate and implement an architecture for the inference of a family of topic models using distributed computing.

In this report we explained with details how we accomplish the first objective. We explain how do we plan to accomplish objective 2 in Section **??**, Objective 3 in Section **??** and Objective 4 in Section **??**

1.5 Challenges

Building a model that captures the interrelation between column-topics and row-topics has two main challenges. First, we must model the mutual dependency between topics representing distributions over different variables. Existing work has explored modeling the correlation between the same kind of topics [42, 6]. The proposed approaches use the parameters of the covariance matrix of the Multivariate Log-Normal distribution to store the covariance between each pair of topics and the variance of each topic. However, this approach can not be used to model different kinds of topics. In this report, we propose a bivariate extension of the categorical distribution to be able to identify the relationship between pairs of topics.

As a second challenge, we must solve the intractability of the inference process for our model. Because our model uses the definition of a new probability distribution, we must derive analytically the posterior probability of observing the topics given the conditions described by our model. To solve this challenge we have found a mathematical relationship that links our defined distribution and the categorical distribution. Using this relationship we can add a prior distribution to our model and we can also approximate the posterior probability of observing the given topics using the collapsed gibbs sampling formulation.

1.6 Organization

The rest of this report is organized as follows. In chapter 2, we present the related work in the fields of Biclustering as well as Topic Models. In Chapter 3, we show the formal definitions of our model and we include the inference updating equations, which are derived based on Collapsed Gibbs Sampling. In addition, we demonstrate how LDA corresponds to a particular case of our model and we provide an explanation for why our model extracts heavy biclusters. Chapter 4 presents the applications of of our model including: a) Biclustering on Gene Expression datasets, b) Biclustering on synthetic data, c) Author classification, and d) The approximation of the joint distribution between two variables. In addition, we present experimental evaluation for each of the applications. Finally, Chapter 6 concludes our current work and discusses future research directions and plans.

Chapter 2

Literature Review

In this chapter, we review several existing proposals related to modeling human mobility using topic models. We divide these proposals into two categories. First, we introduce the proposals on how to model topic models and how these models have been applied in the Human mobility field. Then, we review the research work on how to solve the intractability of the inference of topic models. Understanding the existing inference algorithms is important because the inference process is the main bottleneck in topic modeling development.

By reviewing the existing work, we find that the existing topic models have improved Latent Dirichlet Allocation by either relaxing its assumptions or extending the model to include additional variables. Though some of this models capture the dependency between topics of the same kind (intra-topic dependency). None of these models have considered capturing the topic dependencies between different kind of topics (inter-topic dependency).

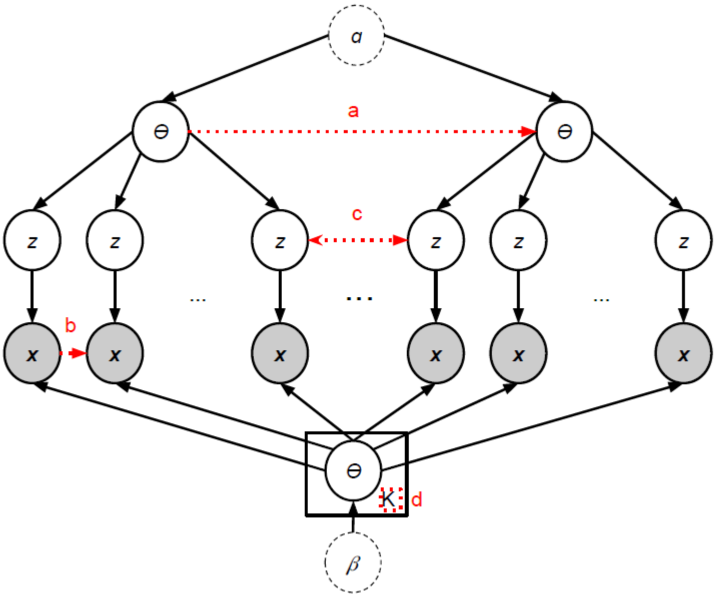


Fig. 2.1 Expanded graphical representation of the LDA model. The red arrows indicate the main independence assumptions about the LDA model. a) The order of Documents is irrelevant b) The order of words is irrelevant c) Topic dependency is only given by a Dirichlet prior distribution parameterized by α d)The number of topics must be given a priori.

2.1 Topic Models and its applications in Human Mobility Field

Topic models are commonly used to organize documents by using the set of words written in them [3, 63], some researchers studying human mobility behavior adopted some of these models with the objective of extracting information about the mobility of users within places. Topic models started with the introduction on the Latent Dirichlet Allocation model (LDA) [8], this model has been adapted to the human mobility field by some authors with the intention of extracting place-topics z by modeling a user Θ as a document and the set of places $\mathbf{x}$ visited by the user as the set of words in the document [23, 45, 70]. The topics are parameterized by a multinomial distribution φ with a Dirichlet distribution parameterized by β, Each document is represented by a multinomial distribution Θ with a different Dirichlet distribution parameterized by α. In figure 2.1, we show an extended model of LDA pointing at the main assumptions made by the model. Reducing these assumptions has shown a positive impact in the application of topic models in several fields, we will now review how researchers have relaxed these assumptions and how these improved models are being used to model human mobility.

2.1.1 N-gram Topic Models

LDA is a bag of words model, this assumption implies that the ordering of the words inside a document is irrelevant [4]. Some researchers have addressed this problem by conditioning the generation of a word on the previous n-words. Hanna Wallach proposed a bigram topic model, where topics would be extracted as pairs of words [67]. This idea was extended by the Distant N-Gram Topic Model [72], where the authors focused in extending the bigram models by considering the dependencies between a word and the *n-th* previous words. The N-gram topic model tends to create an exponential number of parameters as N-increases. One possible solution is to use a Hidden Markov Model to model the dependency between words. This idea was developed by Griffiths et al, their composite model switches between LDA and a standard HMM so that the dependencies between words are governed by the HMM and the content of the words is governed by the LDA [31]. An alternative modeling relaxing the bag of words assumption is the Hidden Topic Markov Model (HTMM) [32]. HTMM considers the relationship between the topic assigned to the current word and the topics assigned to several previous words instead of considering the relationship between the current word and the previous word. In the field of human mobility, N-gram topic models have made possible the extraction of place patterns by considering the relationship between the current place and the *n-th* previous places [22, 21].

2.1.2 Bayesian non-parametric topic models

LDA assumes that the number of topics must be known a priori. This assumption has been relaxed with the introduction of the hierarchical Dirichlet process (HDP) that allows the extraction of the number of topics of an LDA model without a explicit specification [5]. HDP operates by assuming that there exists a finite number of groups sharing the same set of mixture components with different mixing proportion. To describe the mixing proportions of each group, one probability measure is assigned to each of the groups using a Dirichlet Process. The base distribution shared as a parameter among the Dirichlet Processes cannot control the trade-off between the desired property of sharing clusters among groups and allowing all possible mixing combinations of clusters in a given group. As solution, Teh et al considered using a prior over this base distribution. Because of the benefits of using conjugate priors, the authors propose using another Dirichlet Process as prior distribution, hence the name of the Hierarchical Dirichlet Process [65]. The number of topics is modeled as infinite and the relevance of each topic is represented by the sampling from Dirichlet process. One sample of the Dirichlet process will yield a discrete distribution whose support has finite dimensionality, since every dimension represents a topic, the extracted number of topics will be finite as well. The Dirichlet process has several construction definitions being the stick-breaking construction the simplest known [38].

Several human mobility models have been based on the [49, 50] Some human mobility models have also used Dirichlet Processes as to mine mobility patterns. J.Joseph et al, proposed a bayesian non parametric approach to model mobility patterns justified by the fact that Dirichlet process extract overlapping routes where Markov models will normally get confused [39], however they did not use topics in their work [50].

2.1.3 Correlated Topic Models

Another assumption of LDA is the near independence among the topics, the Dirichlet hyper prior used in LDA as prior knowledge about the topic distributions can always be replaced by one independent Gamma distribution per topic. The Correlated topic model [6] has been designed to relax this assumption by changing the Dirichlet hyper-prior for a Multivariate Log-normal hyper-prior, this distribution models the correlation between the topics by using its parametrization given by a covariance matrix and a mean vector. Every parameter of the covariance matrix is used to encode the correlation between each pair of topics, while the mean vector represents the importance of the topics. The correlated topic model has been applied to the modeling of human mobility behavior by modeling a document as a geographical region and the set of words as the set of places, the topics found correspond to geographically related place topics [77].

As an alternative for modeling the correlation between topics, Li et al proposed the Pachinko Allocation Topic Model (PAM) which uses a directed acyclic graph to represent the topics with each non-leaf node and the words with the leaf nodes [42]. Hence, The relationship between all nodes represents the distributions of topics over topics and distributions of topics over words. PAM is a generalization of LDA since a graph where all nodes are connected to only one node will correspond to the standard LDA. The correlation between topics is given by the hierarchical relations existent between topics. PAM has an advantage over the correlated topic model given by the possibility of a sparse parametrization of the topic intra-correlation. In the Correlated topic models the correlation between all topics must be parameterized.

2.1.4 Temporal Dynamics of the Topic Models

A final assumption of LDA is the exchangeability of the documents. i.e. The ordering of documents is irrelevant when doing the inference of the topics and the topics will not exhibit changes in time. A set of topic models represent the temporal dynamics of the data by modeling the latent topics as changing throughout time.

Blei and Lafferty presented the Dynamic Topic Model (DTM) that uses a Kalman filter to model the temporal alignment among topics across collections of documents [7]. As consequence, DTM uses the Markov assumption that the future state of the topics only depends on the inference values of the present state of the topics. Several models have proposed different ways to improve DTM. In [73], the authors proposed the Dynamic Mixture Model to force temporal dependence between documents rather than the temporal dependence between collections of documents, this allows a finer modeling of the evolution of the per document topic distributions. Similar approaches are the Sequential LDA (SeqLDA) [17] and the Evolutionary HDP (EvoHDP) [79]. Both models propose changing the Dirichlet prior of DMM. SeqLDA uses a Pitman-Yor process to improve of the space and computational complexity of the inference process, while EvoHDP uses a Hierarchical Dirichlet Process to automatically extract the number of topics including emergence and disappearance.

Also, In Topics over time model, the authors relaxes the Markov assumption of DTM by treating time as an observed continuous variable such as the topics [71]. This helps avoiding a Markov model's risk of inappropriately dividing a topic in two when there is a brief gap in its appearance. However, with all topics are drawn from one distribution, TOT and DMM are not appropriate for identifying topics that change in time.

The Dynamic Topic Model remains as the only topic model that models topics and its changes of topics over time. However DTM, also named as the Discrete Time Dynamic Topic Model (dDTM) splits the data into groups based on temporal discretization and thus, it requires to choose a discretization value. Because of computational limitations, the discretization value is chosen based on the trade off between accuracy and computational complexity, nonetheless choosing the right discretization value should be a decision based on the data. To solve this issue, Wong and Blei proposed the Continuous Time Dynamic Topic

Model (cDTM) which corresponds to the infinitesimal discretization of the dDTM. cDTM uses a Brownian motion to model the evolution of the topics in time [69].

2.1.5 Extending LDA with additional variables

Aside from solving the assumptions of LDA, other researchers have studied extending an LDA model by including additional variables in order to be able to use topic models for particular supervised learning applications such place recommendation or next place prediction. For geographic-textual data, researchers have studied different variables as user, geographical and functional region, time and text [76, 2, 52, 78]. For check-in data researchers have focused on users, time, item preference, geographical and functional regions [43]. All the mentioned models represent a user as a document, hence we recognize these models as individual-based approaches.

Other extensions of the LDA model have allowed the exploration of non-personalized models. Non-personalized models are important, as they provide essential information in recommendation tasks [9]. The author topic model (ATM) is an extension of LDA where each word in a document is assumed to be written by a specific user, and a document can be written by several authors [59]. This model has been used in the human mobility behavior field by modeling the document as time of the day and the authors as the users. As result, ATM model can extract place-topics of interest for a set of users in a particular time of the day [20]. The author topic model has also been used for aggregating locations instead of users in the place-recommendation task [44].

2.2 Improving the Inference of Topic Models

Modeling topic models requires more than understanding the relationships between the problem variables and relaxing assumptions to create more realistic models. The computational complexity of LDA-based topic models such all the ones mentioned in section 2.2.2 has been proved to be NP-hard [62]. Additionally, different inference algorithms have different advantages and there is no solution on how to generalize the inference process in topic modeling. As consequence, deep understanding of inference methods is necessary in order to design and apply new topic models to large data-sets. In this section, We review the approximation algorithms that have been proposed in the literature to solve the intractability problem of topic models. We classify these algorithms in two categories: The model-based inference methods and the family-based inference methods.

2.2.1 Model-based Inference methods

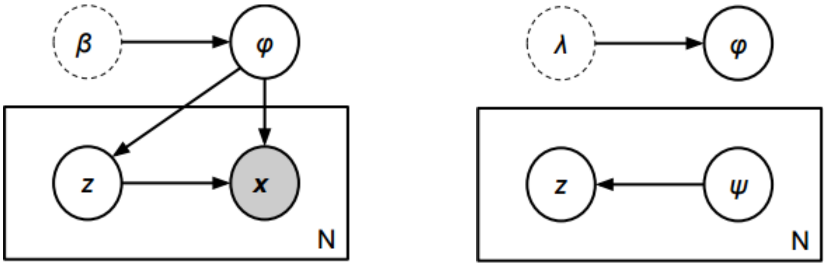
Model-based methods correspond to those techniques that must be tailored for each particular topic model. To solve the intractability of topic models, Hoffman et al proposed using an Expectation-Maximization algorithm (EM) [36] for an LDA model with no priors also known as probabilistic Latent Semantic Analysis. The algorithm aims to find the corresponding values of the latent variables and the given parameters of the model based on the conditional dependencies implied by the model. The same approach works for LDA using Dirichlet priors [53]. A pragmatic improvement over EM has been achieved by using the variational EM framework [8] in which a new simplified tractable model is used to find the parameters of the intractable model. The simplified model makes the assumption that the latent variables and the parameters are mutually independent. Rather than solving the simplified model, the variational bayesian inference (VB) method solves the simplified model whose Kullback-Leibler Divergence to the original model is minimized. This problem is equivalent to the problem of maximizing the free variational energy which is a known framework for solving the inference of probabilistic models. The equivalence between the problems can be shown using Jensen's Inequality [8] or by manipulation of the equations [24].

In [30], Griffiths and Steyvers proposed an MCMC approach based on collapsing or integrating out the parameters of the model which is analytically feasible since the prior distributions are conjugate with the distributions over the variables of the model, the model was labeled Collapsed Gibbs Sampling (CGS). The advantage of CGS over VB is the existence of theoretical bounds over the convergence of the inference process, however CGS requires conjugate priors in order to collapse the parameters of the model. Teh et al proposed Collapsed Variational Bayesian Inference [**?**]. CVB is a variational bayesian algorithm that uses the Collapsed Gibbs Sampling approach of collapsing the parameters by integrating them out from the joint probability distribution of the model. So instead of approximating the model with a simplified version of the model, CVB proposes collapsing the parameters of the model and then solving the collapsed version of the model using variational inference. As result CVB is able relax the independence assumptions between the parameters used in VB, thought the latent variables are still assumed to be mutually independent.

2.2.2 Family-based Inference methods

More recent algorithms aim to develop inference techniques for a family of probabilistic models. Blei et al proposed Stochastic Variational Inference [35], an algorithm to approximate the posterior distribution using subsets of the data rather than the entire dataset. Their method works for a family of models with i) Conjugate priors ii) The generative process follows a conditional chain. e.g. In LDA, Θ is condition by α, Φ is condition by β and w is conditioned by Φ_z iii) the distributions assigned to the random variables must belong to the exponential family. The works by algorithm can be applied to any model a complete conditional is the set of conditional distributions of a single hidden variable in terms of all other hidden and non-hidden variables in the model.

The algorithm works by recognizing that all models that follow conditions i,ii, and iii can be expressed in terms of a simple topic model as displayed in Figure 2.2a. The simple topic



(a) A simple topic model with a local variable z and a global variable φ

(b) Approximate variational model of the simple topic model

Fig. 2.2 A simple topic model and an approximate topic model assuming independences between all parameters in the model

model p is composed of a local latent variable α linked to every instance of the data and global latent variables φ governing all instances of the local latent variables. Being this a Bayesian formulation, β corresponds to the given hyperparameters of φ. Solving the simple model will be enough to solve a family of topic models. The authors of this work have decided to solve the inference of the simple topic model using Variational inference. The approximate model q is given in figure 2.2b. Ideally we want the model $\hat{q}$ that is more similar to the original model given by p, that is $\hat{q} = argmin_q \; KL(q(z,\varphi)||p(z,\varphi|x))$. Using Jensen's inequality, the authors show how this problem is equivalent to maximizing the evidence lower bound (ELBO). i.e. Maximizing the right hand side of equation 2.2

$$\log p(x) \geq E_q[\log p(x,z,\varphi)] - E_q[\log q(z,\varphi)] = L(q) \tag{2.1}$$

Because this equation depends on p, solving the inference problem will depend on p having conjugate priors and using distributions in the exponential family (conditions i and iii). This model can be applied to models such Bayesian mixture models [28], hidden Markov models [29, 26], Kalman filters [40, 25], probabilistic factor analysis/matrix factorization models [55, 13] and some Bayesian nonparametric mixture models [19, 66].

More recently, Black Box Variational Inference [58] was introduced as a generalization of Stochastic Variational Inference by removing conditions i) and iii). The relaxation is possible since the objective function $L(q)$ is solved using Stochastic Optimization and Monte Carlo estimation. Stochastic optimization is used to find the maximum value of $L(q)$ using noisy estimates of $\nabla_{\varphi} L$ in a framework provided by the Robbins-Monro. The framework guarantees convergence if the following iterative process is followed:

$$q(\varphi)_{t+1} \leftarrow q(\varphi)_t + \rho h_t(q) \tag{2.2}$$

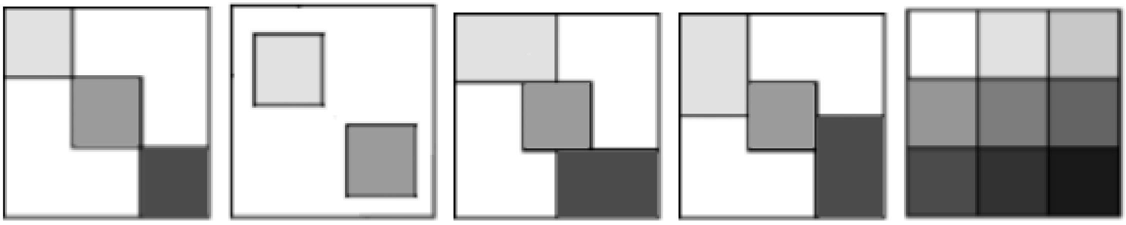
where ρ corresponds to the learning rate and $h_t(q)$ is a distribution whose expectation approximates $\nabla_{\varphi} L$. The authors use Monte Carlo estimation to approximate $\nabla_{\varphi} L$ as follows $h_t(q) = \frac{1}{S}\sum_{s=1}^{S} \nabla_{\varphi} \log q(z_s|\varphi)(\log p(x,z_s) - \log q(z_s|\varphi))$ where z_s corresponds to a sample from the variational distribution as given by $z_s \sim q(z|\varphi)$

2.3 Biclustering

Biclustering algorithms have been used to cluster simultaneously two variables [33]. While it is understandable that clustering one variable is beneficial for organizing or summarizing information, it is not intuitive why do we need to extract Biclusters. To see why extracting biclusters can be useful, let us consider a simple example:

Tom is planning a party for his new four-room flat. Each room has a separate sound system, so he wants to play a movie in each room. As a host, Bob wants everyone to enjoy the movie. Therefore, he needs to distribute movies and guests to each room in order to ensure that each guest watch their favorite movie. Tom has invited forty guests, and he owns twenty movies. He sends out a survey to each guest asking to rate from 1 to 5 each movie. After receiving the guest's responses, Tom collects the data into a 40×20 matrix A, where $A_{i,j}$ stores the ranking of $j-th$ movie given by the $i-th$ guest.

A Biclustering algorithm would take as an input the matrix A and it would output a list of biclusters where each bicluster is defined by a subset of the guests and a subset of the movies.



(a) Exclusive row and column (b) Disjoint Biclusters (c) Exclusive row biclusters (d) Exclusive column biclusters (e) Checkerboard structure

Fig. 2.3 Bicluster Structures. Figure modified from [47]

After biclustering, the rows and columns of the data matrix may be reordered to show the assignment of guests and movies to the rooms in the flat.

Biclustering can be applied to any pair of variables whose type of relationship is described as *many-to-many*. For example, users and movies, documents and words, or users and webpages. We can categorize the families of algorithms by structure and by type [47].

2.3.1 Biclustering Structures

The Biclustering structure represents the shape of the biclusters that can be extracted from a matrix. The different bicluster structures are shown in figure 2.3. Their description is given below. :

Exclusive row and column Every user and every place is assigned to exactly one bicluster.

Disjoint Biclusters Every user and every place is assigned to at most one bicluster.

Exclusive row biclusters Every user is assigned to one bicluster, but every place is assigned to at least one bicluster.

Exclusive columns biclusters Every user is assigned to at least one bicluster, but every place is assigned to one bicluster.

Checkerboard structure Every user, place pair is assigned to exactly one bicluster.

2.3.2 Biclustering Types

The biclustering type refers to the relationship in the matrix that forces the rows and the columns to form a bicluster. There are two main types of Bicluster: Biclusters with constant values and biclusters with coherent values. The first type correspond to those biclusters whose elements have the same or a similar value. There exists several subtypes. For example, biclusters with same values on the rows or biclusters with same values on its columns.

The second type corresponds to those biclusters whose elements have some coherence. For example, either the rows, or the columns are dependent according to some mathematical condition or the values in the bicluster follow a known distribution. In particular, when the coherence is defined to extract biclusters whose rows and columns have a high-frequency of occurring in the dataset, the biclusters obtained corresponds to the bicliques of the dataset.

2.3.3 Related biclustering Algorithms

In this section, we briefly review some of the biclustering algorithms that best performed on tested data with varying conditions, including varying noise and varying numbers of biclusters [18, 47].

Cheng and Church

Cheng and Church proposed the first biclustering algorithm. Their algorithm is famous for having the theoretical guarantee of finding the bicluster where its rows and columns are shifted versions of each other. i.e. The bicluster is represented by a matrix with rank equals to one [12]. Their algorithm is the most common baseline among biclustering algorithms [47]. The algorithm operates by assuming that all the data corresponds to a bicluster, and then iteratively it removes the row or the column with lowest similarity score to the rest of the bicluster. At each iteration, the algorithm evaluates the candidate bicluster using a function score and when the bicluster reaches its smallest size (last iteration); it retrieves the bicluster with highest score. If a second bicluster needs to be found, the data from the first algorithm is replaced by noisy data and the algorithm is restarted. In Short, Cheng and Church biclustering algorithm extracts a bicluster at a time.

xMOTIF

xMOTIF [54] aims to find subset of rows that occur simultaneously across the columns under a set of conditions. e.g. a linear order across the columns. xMOTIF retrieves The largest bicluster that contains the maximum number of conserved rows. Similarly to Cheng and Church algorithm, xMOTIF extracts one bicluster at a time.

QUBIC

While the methods explained above identify patterns under parameterized conditions such a linear scaling among the rows or columns. Quantitative Biclustering (QUBIC) attempts to find similar patterns based on positive and negative correlation, so that the algorithm is less sensitive to outliers in the data [46]. The algorithm uses a combinatorial technique that has demonstrated an improvement over other combinatorial techniques such SAMBA [64], ISA [37], Bimax [57]. The discovery of biclusters follows the same procedure as [12] and [54], that is, biclusters are discovered one at a time.

Spectral Bipartite Graph

This method proposes a novel approach to solve the biclustering problem by transforming the matrix model into a bipartite graph where every row is now part of one set of elements, the columns are another other set of elements and the values in the matrix corresponds to the weight of the relationship between the elements in both sets. The cliques in the graph will correspond to biclusters. However, because in most cases there are not cliques in the

Algorithm	Type	Structure	Discovery	Approach
Block Clustering	Constant	a	One at a time	Divide and Conquer
Bimax [57]	Coherent Values	b	One at a time	Greedy
FLOC [48]	Coherent Values	a,b	Simultaneous	Greedy
PLAID [41]	Coherent Values	a,b	One at a time	Probabilistic
CTWC [11]	Coherent Values	b	One at a time	Combinatorial
ITWC [27]	Coherent Values	b	One at a time	Combinatorial
Spectral [15]	Coherent Values	a	Simultaneous	Optimization
QUBIC [46]	Coherent Values	a,b,c	One at a time	Optimization
xMotif [54]	Coherent Values	a,b,c	Simultaneous	Greedy
SAMBA [64]	Coherent Evolution	a,b,c	One at a time	Combinatorial
Bayesian [61]	Coherent Values	e	Simultaneous	Probabilistic
Cheng [12]	Constant	a,b,c	One at a time	Iterative
Mondrian [60]	Coherent Values	e	Simultaneous	Probabilistic

graph, Dhillon proposes an optimization function taking into account the weight between the vertices and the sizes of the clique to extract groups of vertices that better fit a clique.

Chapter 3

Yin Yang Mixture Model

We start this chapter by introducing the type of biclusters that we aim to extract and how we can use a probabilistic model to extract them. Then, we explain how we construct our model using two complementary mixture models and a novel probability distribution. Afterwards, we explain the generative process of the resulting model and the inference equation that describes how we estimate the parameters of the model given a training dataset. In addition, we use this inference equation to proof that our model corresponds to a generalization of LDA. Finally, we explain how we can use the model to extract biclusters and why the model is able to extract the information corresponding to *Which groups of users visit which groups of places*.

3.1 Problem definition

Madeira et al give the following definition of biclustering [47]. That is: *The goal of biclustering methods is to identify subgroups of rows and subgroups of columns in a matrix, by performing simultaneous clustering of both dimensions, instead of clustering these two dimensions separately.*

Given a matrix $|U| \times |P|$, we may use a categorical mixture model to cluster the rows U by using the information in all the columns P. The mixture model would represent the rows as random mixture of latent topics Z_u, where the $i-th$ topic is characterized by a distribution over the rows, denoted by $\Phi_{u(i)}$. Based on the same matrix, we may use a categorical mixture model to cluster the columns P by using the information in all the rows U. This mixture model would represent the columns as random mixture of latent topics Z_p, where the $j-th$ topic is characterized by a distribution over the columns, denoted by $\Phi_{p(j)}$. We refer to this mixture model as the Dual Mixture model since we use the same input data to obtain a complementary result. We label the topics Z_u obtained by the mixture model as *row-topics* and the topics Z_p obtained by the Dual Mixture model as *column-topics*.

In order to perform simultaneous clustering of both dimensions, we need to combine the mixture model with its dual . In other words, we must model a mutual dependency between the row-topics Z_u and the column-topics Z_p. We may use additional parameters $\Theta_{i,j}$ to capture the relationship between the *i-th* row-topic and the *j-th* column-topic.

For example, using the sample dataset presented in table 1.2, we aim to find the groups of users as user-topics, the groups of places as place-topics and the relationship between each user topic and each place topic. Those relationships with higher probability will correspond to the extracted biclusters. The groups of users will be obtained from Z_u. Assuming two groups of users, we would expect $Z_u = 1$ and $Z_u = 2$ to be characterized respectively by the following user-topics distributions:

$$\Phi_{u(1)} = [0.36, 0.32, 0.34, 0, 0, 0.01]$$

$$\Phi_{u(2)} = [0, 0, 0.34, 0.33, 0.33, 0]$$

Assuming two groups of places, we would expect $Z_p = 1$ and $Z_p = 2$ to be characterized respectively by the following place-topics distributions:

$$\Phi_{p(1)} = [0.33, 0.33, 0.34, 0, 0, 0, 0]$$
$$\Phi_{p(2)} = [0, 0, 0, 0.32, 0.32, 0.26, 0.1]$$

Because Θ must encode the relationship between every user-cluster and every place-cluster, we expect to find the following values:

$$\Theta = \begin{bmatrix} 0.60 & 0.05 \\ 0.05 & 0.30 \end{bmatrix}$$

If we consider a relationship to be a bicluster when $P(Z_u, Z_p|\Theta) > 0.25$, then we would extract two biclusters. The first bicluster corresponding to the group of users extracted from the first user-topic $Z_u = 1$ and the group of places extracted from the first place-topic $Z_p = 1$. The second bicluster corresponding to the groups of users extracted from the second user-topic $Z_u = 2$ and the group of places extracted by the second place topic $Z_p = 2$. We may use a similar criteria to determine which users are extracted from the user-topic and which places are extracted from the place-topic. For instance, in the first bicluster we may select a user to be part of the bicluster if $P(u|\Phi_{u(1)}) > 0.25$. In this case U_1, U_2 and U_3 will be part of the bicluster. For the places, we may select a place to be part of the bicluster if $P(p|\Phi_{p(1)}) > 0.25$. In this case P_1, P_2 and P_3 are part of the bicluster.

In conclusion, we aim to use row-topics Z_u characterized by the parameters Φ_u to obtain groups of rows and the information about which rows are more important in each bicluster. Similarly, we aim to use column-topics Z_p characterized by the parameters Φ_p to obtain groups of columns and the information about which columns are more important in each bicluster. Finally, we expect to have some parameters Θ from where we can obtain the infor-

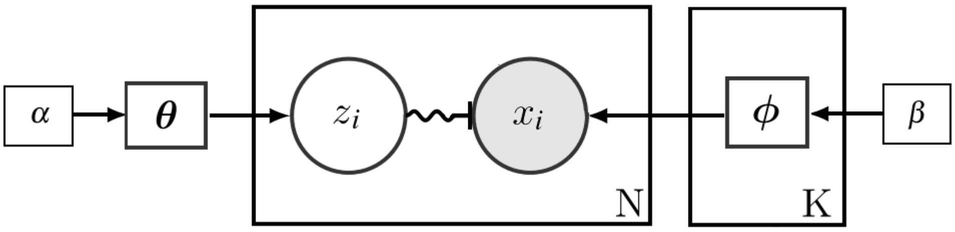


Fig. 3.1 Graphical Representation of the Categorical Mixture Model.

mation about the relationships between the row-topics and the column-topics. Successively, we may extract those relationships with higher probability as biclusters.

3.2 Categorical Mixture Model

The Categorical Mixture Model (CMM) constitutes a building block for the Latent Dirichlet Allocation(LDA). In particular, the CMM provides an architecture for extracting topics represented by distributions over a random variable. Using a similar intuition, We will use the CMM as a building block component for our model.

The CMM models the observations of one random variable by combining a finite number of probability distributions. In the case of the Categorical Mixture Model (CMM), the objective is to represent how N observations x_i are generated by mixing K categorical distributions parameterized by ϕ where the mixing proportions are given by another categorical distribution with parameters θ. The topics in this model correspond to the information extracted by the distributions parameterized by ϕ. The term *topic* originated when this model was applied to a set of words because in that scenario, the parameters ϕ represent distributions over the vocabulary of words. We depict this information in figure 3.4a.

When we apply the CMM to other random variables, we obtain other kinds of topics. For example, in the Mobility field, when we use the CMM to obtain a mixture of distributions over the set of places, the parameters ϕ characterize *place-topics*. Similarly, when we use the

CMM to obtain a mixture of distributions over the set of users, the parameters ϕ characterize *user-topics*

3.3 Yin Yang Mixture Model Model

Specifications of our model:

1. Group the columns using a mixture model with latent variable Z_u.

2. Group the rows using a mixture model with latent variable Z_p.

3. Model the mutual dependency between Z_u and Z_p.

3.3.1 Using mixture models to implement specifications 1 and 2

We construct the model by implementing the three specifications. For the implementation of specifications 1 and 2, we use the Categorical Mixture model (CMM) [51]. The CMM clusters a discrete variable using a latent variable. Hence, we use one CMM to obtain column-topics and another CMM to obtain row-topics. A graphical representation of the model implementing the first two specifications is presented in figure 3.2. From the figure, we observe Θ which describes the importance of each of the topics and Φ_u, Φ_p which describes the relationship of each pair of topics $< Z_u, Z_p >$. V represents the total number of observations.

3.3.2 Introducing a novel distribution to implement specification 3

So far, we have obtained a model that clusters independently the columns and the rows. To model the mutual dependency between the cluster descriptors Z_u and Z_p, we assume that the two kinds of topics are independent only if we know the parameters that describe their dependency. That is, we assume a common cause dependency between the latent variables

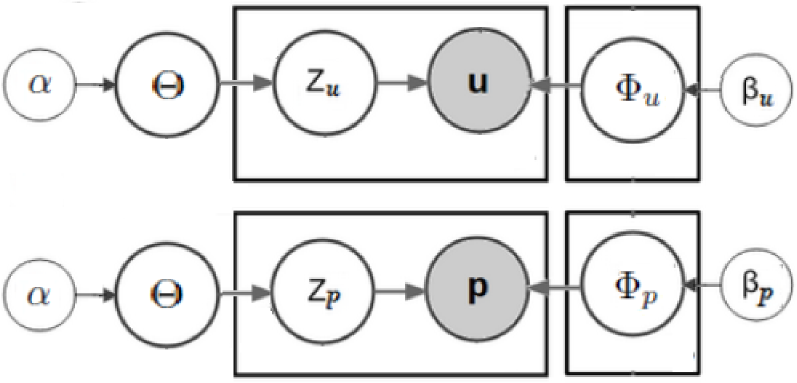


Fig. 3.2 Implementation of specifications 1 and 2.

[34]. The mutual dependency between two types of topics has not been studied by Topic Models. Hence, We address this necessity by defining a probability distribution.

Because we require a probability distribution to describe the simultaneous generation of two variables, we will define a Bivariate Probability distribution. Additionally, it would be ideal to use a categorical distribution because categorical distributions are commonly used to capture the mixing components of the topics. In our case, with a categorical distribution, we can use the parameter $\Theta_{i,j}$ to capture the relationship from the $i-th$ row-topic to the $j-th$ column-topic. Using these requirements, we introduce the *Bivariate Categorical Distribution* in Definition 1

In order to define a probability distribution, we need to specify the support, parameters and probability mass function of the distribution. The support of a probability distribution represents the set of outcomes that may appear when sampling from the distribution; the parameters corresponds to the set of variables that describe the probability mass function and the probability mass function maps every element in the support to a probability value. Using these components, we present the following description of the *Bivariate Categorical Distribution*:

Support The distribution has a two dimensional support since we require simultaneous sampling of two random variables.

Parameters The distribution uses $K_u \cdot K_p$ parameters. Each parameter represents the relationship between one place-topic and one user-topic.

Probability Mass Function $f_\Theta(\vec{x}) = \sum_i^{K_u} \sum_j^{Kp} \mathbb{1}[x_0 = i \wedge x_1 = j]\Theta_{i,j}$ This function represents the probability of observing every user-topic, place-topic pair ($\vec{x}$) using a parameter.

Now we present the formal definition of the Bivariate Categorical Distribution:

Definition 1. A Bivariate Categorical Distribution represented by $Categorical^2(\Theta, K_u, K_p)$, is a discrete probability distribution parameterized by $\Theta \in \mathbb{R}_{[0,1]}^{K_u \cdot K_p}$ and $K_u, K_p \in \mathbb{N}_{>0}$. The support of the distribution is $X \in \mathbb{N}_{[1:K_u]} \times \mathbb{N}_{[1:K_p]}$. The probability mass function is given by $f_\Theta(\vec{x}) = \sum_i^{K_u} \sum_j^{Kp} \mathbb{1}[x_0 = i \wedge x_1 = j]\Theta_{i,j}$

Proof. We prove that the Bivariate Categorical Distribution is a valid distribution by showing how we can represent the Bivariate Categorical Distribution in terms of the Categorical distribution:

Given a random variable $x \sim Categorical^2(\Theta, K_u, K_p)$. We need to use a mapping g so that $y = g(x) \sim Categorical(\Theta)$. This mapping is given by $g : \mathbb{N}_{[1:k]} \to \mathbb{N}_{[1:K_u]} \times \mathbb{N}_{[1:K_p]}$ where $g(x) = (x \div K_p, x \bmod K_p)$. Because g is a bijection, we are guaranteed that we can always represent a Bivariate Categorical Distribution in terms of the Categorical Distribution and viceversa. ■

3.3.3 Graphical representation of the model

By linking the implementation of the three specifications, we obtain the Yin Yang Mixture Model (YMM). YMM assumes knowing the number of row-topics K_u and the number of column-topics K_p in advance. In figure 3.3, we present the graphical model that encodes the dependencies between the random variables used in YMM. We also present the formal definition of all the variables used by our model in table 3.1.

Variable	Dimension	Description	Distribution
v	1	Index to a visit record	NA
V	1	Number of columns	NA
K_p	1	Number of column-topics	NA
K_u	1	Number of row-topics	NA
U	1	Number of rows	NA
P	1	Number of columns	NA
Φ_p	$P \times K_p$	column per column-topic distribution	$Dirichlet(\beta_p)$
β_p	$P \times 1$	Hyper-parameter of Φ_p	NA
p_v	1	Observed column in the v-th visit record	$Categorical(\Phi_p, K_p)$
Φ_u	$U \times K_u$	row per row-topic distribution	$Dirichlet(\beta_u)$
β_u	$U \times 1$	Hyper-parameter of Φ_u	NA
u_v	1	Observed row in the v-th visit record	$Categorical(\Phi_u, K_u)$
θ	$K_u \times K_p$	Joint distribution of row-topic and column-topics	$Dirichlet(\alpha)$
α	$K_u \times K_p$	Hyper-parameter of θ	NA
$z_{u(v)}, z_{p(v)}$	1,1	row-topic and column-topic assigned to the v-th visit record	$Categorical^2(\Theta, K_p, K_u)$

Table 3.1 Variable definitions for the Yin Yang Mixture Model.

From the structure of the model, we observe that we can obtain the joint probability of observing two categorical variables as given by $P(u,p) = \sum_{Z_u=1}^{Ku} \sum_{Z_p=1}^{Kp} P(u|\Phi_{Z_u})P(p|\Phi_{Z_p})P(Z_u, Z_p|\Theta)$

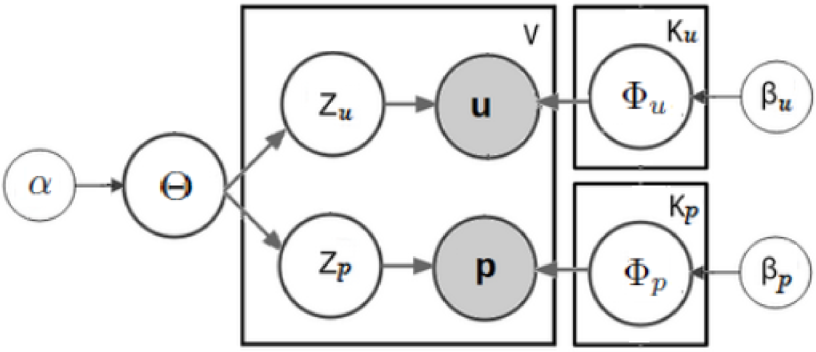


Fig. 3.3 Yin Yang Mixture Model graphical model

3.3.4 Generative Process

We use the generative process to explain how to sample pairs of values corresponding to observations of two random variables. First, we use a Dirichlet distribution parameterized by α to generate the low dimensionality representation of the joint distribution of the two variables. So, we obtain $K_u \times K_p$ parameters representing the partition of the data into $K_u \times K_p$ sections. Second, we use the bivariate categorical distribution to sample one section of the partition which will point to one row-topic and one column-topic. Third, we use the row-topic to select a row following the same process as the categorical mixture model, and similarly we use the column-topic to sample one column. In algorithm 1, we present the algorithmic description of the generative process.

Algorithm 1 YMM Generative Process

```
Draw a distribution over topics $\Theta \sim Dirichlet(\alpha)$
for i in 1 $\cdots$ $K_u$ do
    Draw a distribution over the rows $\Phi_i \sim Dirichlet(\beta_u)$
end for
for j in 1 $\cdots$ $K_p$ do
    Draw a distribution over the columns $\Phi_j \sim Dirichlet(\beta_p)$
end for
for v in 1 $\cdots$ $|V|$ do
    Draw simultaneously a row-topic and a column-topic $z_{p(v)}, z_{u(v)} \sim Categorical^2(\Theta)$
    Draw a row $u \sim Categorical(\Phi_{z_{u(v)}})$
    Draw a column $p \sim Categorical(\Phi_{z_{p(v)}})$
end for
```

3.3.5 Inference Process

We use Collapsed Gibbs Sampling to obtain the posterior distribution that we use to update the row-topic $z_{u(v)}$ and the column-topic $z_{p(v)}$ assigned to the *v-th* record. For each record, we must sample from this distribution to update its topic assignments; the posterior distribution must be recomputed after each assignment. Updating all the records corresponds to one iteration. After several iterations, the posterior distribution will converge, and we have found the row-topic and the column-topic assignment for every record. The posterior distribution is presented in Equation 3.1. All details of the derivation can be found at the appendix.

$$P\left(z_{u(v)}, z_{p(v)} | z_{u(-v)}, z_{p(-v)}, u, p, \alpha, \beta_u, \beta_p\right) \propto$$

$$\frac{\mathscr{Z}\,\mathscr{P}(v)\,\mathscr{U}(v)}{\sum_{j=1}^{|P|}\mathscr{P}(j)\sum_{i=1}^{|U|}\mathscr{U}(i)} \tag{3.1}$$

$$\begin{aligned}
\mathscr{Z} &= \sum_{m=1}^{|U|}\sum_{n=1}^{|P|} c\left(z_{p(v)}, z_{u(v)}, m, n\right)^{(-v)} + \alpha_{z_{u(v)}, z_{p(v)}} \\
\mathscr{P}(i) &= \sum_{y=1}^{Kp}\sum_{n=1}^{|P|} c\left(z_{p(v)}, y, p_{(i)}, n\right)^{(-v)} + \beta_{p(i)} \\
\mathscr{U}(i) &= \sum_{x=1}^{Ku}\sum_{m=1}^{|U|} c\left(x, z_{u(v)}, m, u_{(i)}\right)^{(-v)} + \beta_{u(i)} \\
c(x,y,m,n) &= \sum_{j=1}^{|V|} \mathbb{1}(z_{u(j)} = x, z_{p(j)} = y, \\
&\qquad u_{(j)} = m, p_{(j)} = n)
\end{aligned}$$

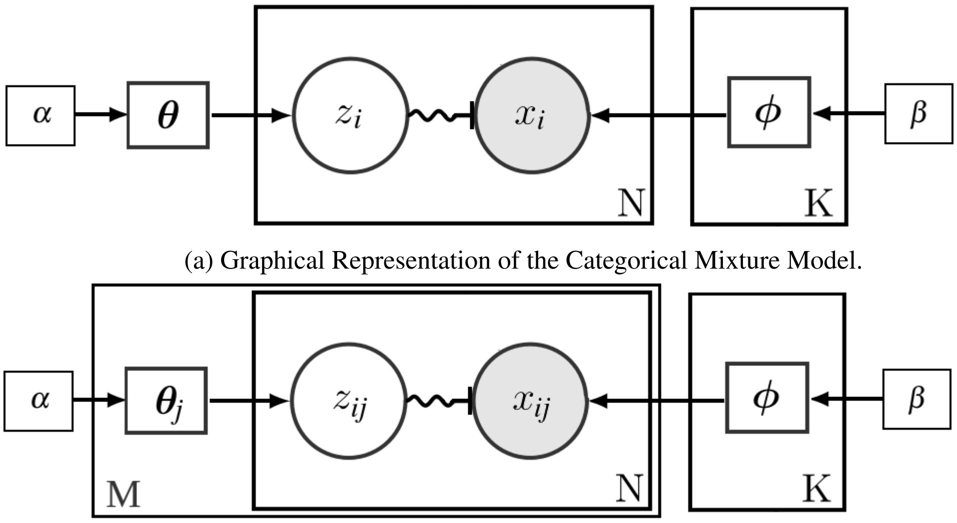
The inference equation has an intuitive explanation. First, the ratio $\frac{\mathscr{U}(v)}{\sum_{i=1}^{|U|}\mathscr{U}(i)}$ expresses the probability of the row $u_{(v)}$ under row-topic $z_{u(v)}$. Second, the ratio $\frac{\mathscr{P}(v)}{\sum_{j=1}^{|P|}\mathscr{P}(j)}$ expresses the probability of the column $p_{(v)}$ under column-topic $z_{p(v)}$. Finally a counting over the current relationship between row-topics and column-topics $\mathscr{Z}$ captures the mutual dependence between the topics.

3.4 Relationship with Latent Dirichlet Allocation

In this section, we introduce the Latent Dirichlet Allocation (LDA) as an extension of the Categorical Mixture Model (CMM). Then, we present the inference equation of LDA and we use it to proof that LDA is a particular case of the Yin Yang Mixture Model.

3.4.1 Latent Dirichlet Allocation

While the Categorical Mixture Model (CMM) is able to extract topics by looking at the frequencies of the words over the entire collection of documents, the Latent Dirichlet



(a) Graphical Representation of the Categorical Mixture Model.

(b) Graphical Representation of the Latent Dirichlet Allocation Model

Fig. 3.4 Information extracted by an LDA (left) and a Dual LDA (right). The width of the arrow represent the preference of a topic, the ovals represents the topics.

Allocation (LDA) model extends the CMM by dividing the words into M different documents. Therefore, instead of parameterizing the entire set of words with one categorical distribution Θ; LDA parameterizes the documents – indexed with j – using a categorical distribution with parameters Θ_j. We depict this information in figure 3.4b.

In the mobility field we can use an LDA to split the set of places into $|U|$ different users. As consequence, every user is defined as the places that he visit. In this setting, LDA's topics represent distributions over the set of places and therefore we obtain a explicit grouping of the set of places as given by the topics. In addition LDA gives a direct representation of the distribution of places visited by any user u:

$$P(p|u) = P(p|\Theta_u) = \sum_{Z_p=1}^{Kp} P(p|\Phi_{Z_p})P(Z_p|\Theta_u) \tag{3.2}$$

However, LDA hinders the representation of the distribution of users visiting any of the places p ($P(u|p)$). In addition, the model does not extract explicit information about grouping the set of users.

The above-mentioned limitations of an LDA are solved using a Dual LDA. The Dual LDA is also an LDA model that interchanges the role of the users and places. In the Dual LDA, instead of splitting the set of places, we split the set of users into $|P|$ different places. As consequence, every place is defined as the users that visited the place. In this case, the topics represent distributions over the set of users and the distribution of users visiting any of the places p ($P(u|\Theta_p)$) is directly represented by the model.

In summary, we can express the relationship between users and places into two approaches.

i) LDA or user-based approach:

- A user is defined by the places that he visited.
- We obtain topics represented by distributions over the set of places.
- We can represent the distribution of the places for a given user.

ii) Dual LDA or location-based approach:

- A place is defined by the users who visit the place.
- We obtain topics represented by distributions over the set of users.
- We can represent the distribution of the users for a given place.

For the comparison, we obtain the following information if we model the two variables using YMM:

- We obtain topics represented by distributions over the set of places.
- We obtain topics represented by distributions over the set of users.
- We can represent the joint distribution of the users and places. Hence, we can compute the distribution of the users for a given place and the distribution of the places for a given user.

3.4.2 Yin Yang Mixture Model generalizes Latent Dirichlet Allocation Model

We start by presenting the inference equation used to update the topics in the LDA model. Then, we compare with the inference equation obtained by the YMM model. In conclusion, we find that YMM generalizes LDA.

Inference of the Latent Dirichlet Allocation

The topic assignment to the $v-th$ visit record $\{u(v), p(v)\}$ is given by sampling from the following derived posterior probability [8]:

$$P\left(z_{p(v)}|z_{p(-v)}, u, p, \alpha_u, \beta_p\right) \propto \frac{\bar{\mathscr{Z}}\,\mathscr{P}(v)}{\sum_{j=1}^{|P|}\mathscr{P}(j)}$$

where

$$\bar{\mathscr{Z}} = \sum_{n=1}^{|P|} c\left(z_{p(v)}, u_{(v)}, n\right)^{(-v)} + \alpha_{z_{p(v)}}$$

$$\mathscr{P}(i) = \sum_{m=1}^{|U|} c\left(z_{p(v)}, m, p_{(i)}\right)^{(-v)} + \beta_{p(i)}$$

$$c(x,m,n) = \sum_{j=1}^{|V|} \mathbb{1}(z_{u(j)} = x, u_{(j)} = m, p_{(j)} = n)$$

Since the function c corresponds to sums of indicator functions, it is easy to interpret the two factors responsible for the topic assignment:

i) $\bar{\mathscr{Z}}$: Counting of topics assigned to user u_v.

ii) $\frac{\mathscr{P}(v)}{\sum_{j=1}^{|P|}\mathscr{P}(j)}$: Probability of observing the place p_v given a topic.

Comparison with the inference equation of YMM

We show that LDA is a particularization of YMM by proving that LDA's inference equation can be reduced to the inference equation of YMM when we assign every user to a single and unique user-topic. i.e., users and user-topics are assigned in a *one-to-one* mapping.

Proof. Because we are given the user-topic assignments for every user, the inference of YMM only depends on $z_p(v)$. So, recalling equation 3.1, we begin by analyzing the term $\frac{\mathscr{U}(v)}{\sum_{i=1}^{|U|} \mathscr{U}(i)}$. Because $\mathscr{U}(v)$ does not depend on $z_p(v)$, we can replace $\mathscr{U}(v)$ with a constant value. Then, we analyze the term $\mathscr{Z}$. Since $z_u = u \ \forall u \in 1 : U$, we can replace z_u with u and as result, we end up with the LDA inference equation. That is:

$$
\begin{aligned}
&P\left(z_{p(v)} | z_u, z_{p(-v)}, u, p, \alpha, \beta_u, \beta_p\right) = \\
&P\left(z_{p(v)} | z_{p(-v)}, u, p, \alpha_u, \beta_p\right) \propto \frac{\bar{\mathscr{Z}}\, \mathscr{P}(v)}{\sum_{j=1}^{|P|} \mathscr{P}(j)} \text{ where} \\
&\bar{\mathscr{Z}} = \sum_{m=1}^{|U|} \sum_{n=1}^{|P|} c\left(z_{p(v)}, u(v), m, n\right)^{(-v)} + \alpha_{u(v), z_{p(v)}}
\end{aligned}
$$

and α_u corresponds to a K_p dimensional vector conditioned on the user u. In LDA, the hyper-parameter of the Dirichlet prior on the per-user topic distributions $\bar{\alpha}$ is independent of the given user to avoid over-fitting. We can break the dependence of α and the given user by setting α to the same value for all users.

■

Finally, we note that if we start off by assigning a *one-to-one* mapping between the places and the place-topics, YMM's inference equation would transform into the inference equation of the Dual LDA.

3.4.3 Bicluster Extraction using YMM

We obtain the biclusters from the estimated values of the parameters Θ, Φ_u, Φ_p. First, we obtain the groups of rows from the parameter Φ_u. The parameter Φ_u represents K_u distributions over the set of rows. Therefore, we must set a threshold γ_u to select which rows are more relevant in each distribution. Second, we obtain the groups of columns from the parameter Φ_p. The parameter Φ_p represent K_p distributions over the set of rows. Therefore, we must set a threshold γ_p to select which rows are more relevant in each distribution.

Finally, we extract which groups of rows have a stronger relationship with which groups of columns. In other words, we extract the biclusters. This information is revealed by the parameter Θ. The reason is that Θ encodes the relationship between every distribution in Φ_u and every distribution in Φ_p. We set a threshold γ to select the more relevant relationships in the dataset.

All γ_u, γ_p, and γ correspond to percentiles. More formally:

1) We extract the subset of rows for each row-topic from $U^m = \{u \in U | P(u|\phi_{u,(m)}) > \gamma_u\} \forall m \in 1 : K_u$.

2) We extract the subset of columns for each column-topic from $P^m = \{p \in P | P(p|\phi_{p,(n)}) > \gamma_p\} \forall n \in 1 : K_p$.

3) Once we partition the rows into groups of rows and the columns into groups of columns, we may extract the top K strongest relationships between the groups of rows and groups of columns. We can do that by selecting the top K parameters $\Theta_{i,j}$ with the largest value. More preferably, we may use γ so that we do not need to specify the number of biclusters a priori. Using γ, we can extract the list of the most relevant biclusters from $\mathscr{B} = \{U^m, P^n | P(z_u = m, z_p = n | \Theta, \alpha) > \gamma \, \forall m \in 1 : K_u \, \forall n \in 1 : K_p\}$.

These percentiles control the trade-off between the quantity and the quality of the extracted biclusters. If we set $\gamma = 0$, then we obtain the maximum number of biclusters, which corresponds to $K_u \times K_p$. However, not all the biclusters represent strong relationships. If we set $\gamma > 0.5$, then we might extract at most one bicluster. Empirically, we find that setting $\gamma = \frac{1}{K_u \times K_p} \sum_{i=1}^{K_u} \sum_{j=1}^{K_p} \Theta_{i,j}$ works well in most situations.

Using a probabilistic approach we avoid constraining the model to a particular structure. This is because every bicluster is defined with a mixture over the set of rows and a mixture over the set of columns, the biclusters might overlap, or simply have sparse distributions. Additionally, we avoid clustering the data into a pre-specified number of biclusters.

3.5 Why does YMM works?

We show that YMM is able to extract quasi-bicliques by analyzing which records influence the topic assignment of which other records. We discover that the records belonging to the same quasi-biclique have the highest mutual influence and so they are more likely to be assigned to the same user-topic and the same place-topic.

3.5.1 Topic influence between a pair of records

We start by rewriting the inference equation (Eq. B.22) in terms of three factors:

$$P(z^u, z^p | u, p, \alpha, \beta) = P(z^u, z^p | \Theta) P(u | \Phi^u) P(p | \Phi^p)$$

The topic influence from the record $< u, p >$ to the record $< \hat{u}, \hat{p} >$ is given by the following equation:

$$P(z^u, z^p | \hat{u}, \hat{p}, \alpha, \beta) = P(z^u, z^p | u, p, \alpha, \beta) \frac{P(p | \Phi^p)}{P(\hat{p} | \Phi^p)} \frac{P(u | \Phi^u)}{P(\hat{u} | \Phi^u)} \tag{3.3}$$

We know that the influence is maximal if $\frac{P(p | \Phi^p)}{P(\hat{p} | \Phi^p)} \frac{P(u | \Phi^u)}{P(\hat{u} | \Phi^u)} = 1$ because the topic assignment of $< u, p >$ will follow the same equation as the topic assignment for the record $< \hat{u}, \hat{p} >$.

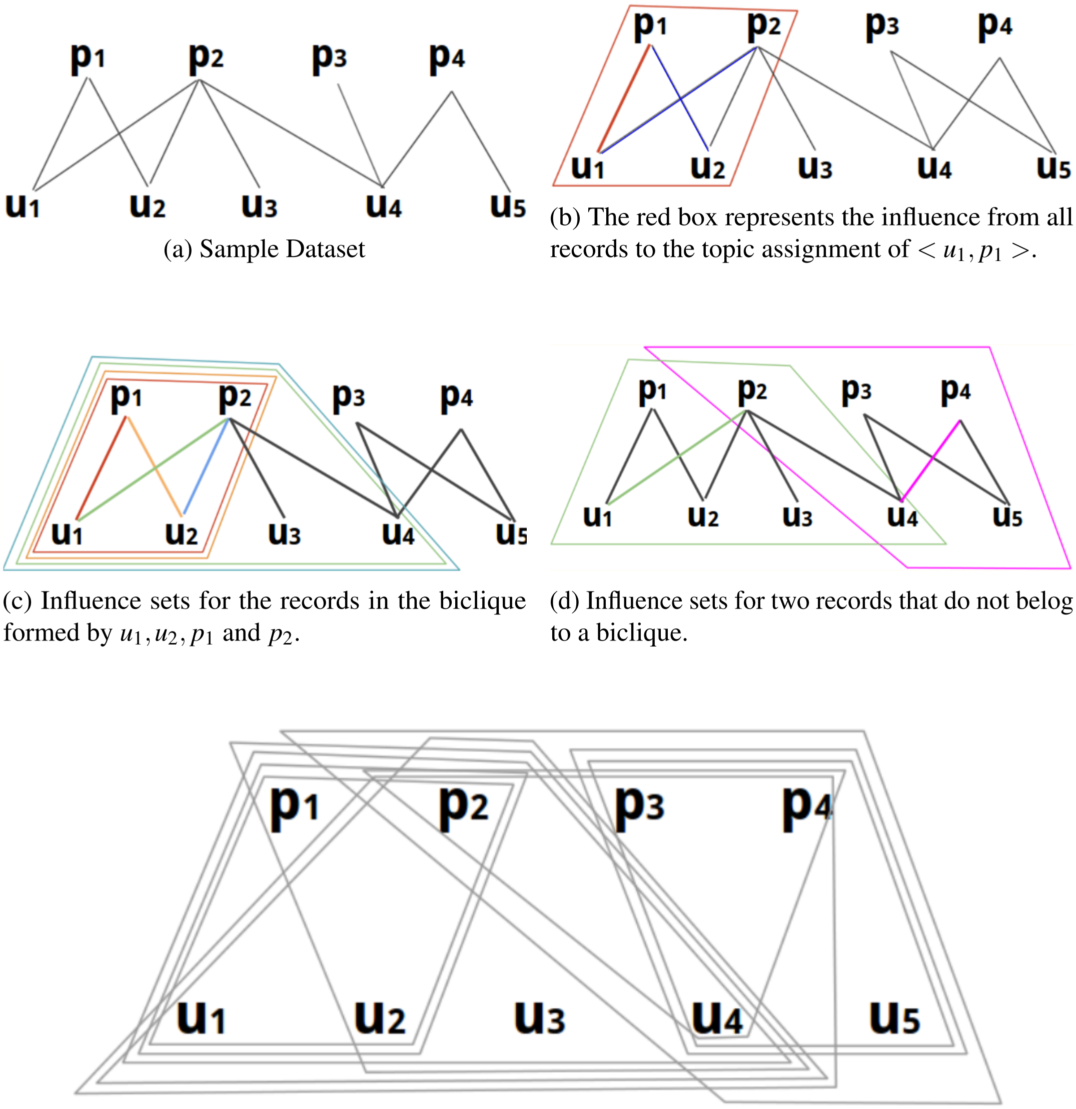
If the two records belong to the same quasi-biclique, then it must follow that either i) The user in both records is the same ($u = \hat{u}$) ii) The place in both records is the same ($p = \hat{p}$) iii) Trivially, the user and the place is the same in both records.

In these three cases, we can simplify equation 3.3 because either $\dfrac{P(u|\Phi^u)}{P(\hat{u}|\Phi^u)} = 1$ or $\dfrac{P(p|\Phi^p)}{P(\hat{p}|\Phi^p)} = 1$. It is only in the case when two records are not in the same quasi-biclique that equation 3.3 will not be simplified. If equation 3.3 cannot be simplified, both records will not have a strong influence in the topic assignment of each other. In conclusion, those records that share the same user or the same place have the highest influence in the topic assignment of each other.

3.5.2 Records in the same quasi-biclique have the strongest mutual influence

We have showed that a pair of records in the same quasi-biclique will have higher influence when compared with a pair of records that do not belong to the same quasi-biclique. Therefore, all records in a quasi-biclique will mutually reinforce the topic assignment of each other and so the user-topic, place-topic assignment will tend to be the same for all these records. For a better understanding of this fact, we provide a graphical example in figure 3.5.

First, we represent the visit records in a graph so that each vertex represents the visit of a user to a place (Figure 3.5a). Second, we visualize with a colored set those records that influence the topic assignment of the record $< u_1, p_1 >$. In this case, we note that the records $< u_2, p_1 >$ and $< u_1, p_2 >$ will have higher influence on $< u_1, p_1 >$ because those records have either the same place or the same user (Figure 3.5b). Third, we visualize the influence of those records in the quasi-biclique composed by $u_1, u_2; p_1, p_2$. We observe how all the records in the quasi-biclique mutually influence each other (Figure 3.5c). Fourth, we display an example of two records who do not belong to the same quasi-biclique. The records $< u_1, p_2 >$ and $< u_4, p_4 >$ influence the topic assignment of the record $< p_2, u_4 >$ But they will not have a mutual infuence (Figure 3.5d). Fifth, we display all the influence sets generated by the records in the dataset. We observe that from the aggregation of all the influence sets, the records including $u_1, u_2; p_1, p_2$ are likely to be assigned to the same user-topic, place-topic pair. We conclude the same for the records including $u_4, u_5; p_3, p_4$ (Figure 3.5e).



(a) Sample Dataset

(b) The red box represents the influence from all records to the topic assignment of $< u_1, p_1 >$.

(c) Influence sets for the records in the biclique formed by u_1, u_2, p_1 and p_2.

(d) Influence sets for two records that do not belog to a biclique.

(e)

Fig. 3.5 Demonstration of how YMM extracts bicliques.

Chapter 4

Applications and Experimental Evaluation

In section 1, we have specified the advantages of the YMM with respect to other biclustering algorithms. In this section we present the applications and experiments that benefit from these advantages. Table 4.1 list the applications that benefit from each of the advantages.

4.1 Datasets

Foursquare. The first type of data contains check-in data. Every check-in record can be represented as a <user_id, place_id, category> tuple. The records include 227428 check-ins

Advantage	Section						
	4.3	4.5	4.6	4.7	4.8	??	??
Identify the most relevant rows in each bicluster	✓			✓	✓	✓	
Identify the most relevant columns in each bicluster	✓			✓	✓	✓	
Infer the relevance of the biclusters	✓					✓	
Extract Bicluster with different structures	✓	✓					
Infer the joint and conditional distributions between the variables			✓		✓		✓

Table 4.1 Applications that benefit from the advantages of YMM.

from 1082 users into 38333 venues collected from Apr 2012 to Feb 2013 in New York [75], 573703 check-ins recorded from 2293 users into 61858 venues collected from Apr 2012 to Feb 2013 in Tokyo [75], and 342850 check-ins from 2341 users into 44173 venues collected from Aug 2010 to Jul 2011 in Singapore.

Singtel Cell Location. The second data set contains cell-tower location records which can be represented as a <user_id, cell_tower lat, cell_tower lon> tuple. We used 6880128 records from 20830 Singapore's residents and visitors into 4328 cell-towers located across Singapore collected during 2013. This records were provided by Singtel .

NIPS. The third dataset is the text data obtained from papers accepted by NIPS from 1988 until 2003 [**?**]. The data set contains 40552970 records generated by 2865 authors who wrote 2483 documents using a vocabulary of 14036 words.

MicroArray. The last type of data contains eight different Micro Array gene expression datasets from the Gene Expression Omnibus (GEO). GEO is a public repository designed to be used by many analysis applications. The dataset adheres to academic and industrial standards. Each dataset can be represented by a matrix, where the rows represent the genes and the columns represent samples of DNA strands. Every cell in the matrix corresponds to the regulation of the expression level of a particular gene in a particular sample. The regulation might be positive indicating that the gene is expressed in normal cells. The description and break down of how many genes and samples were used in each of the datasets is presented in table 4.2.

[0] http://info.singtel.com/

Dataset	Genes	Samples	Description
GDS181	12559	84	Human and mouse
GDS589	8799	122	Rat peripheral and bran regions
GDS1027	15866	154	Rat Lung sulphur mustard exposure model
GDS1319	22548	123	C blastomere mutant embryos
GDS1406	12422	87	Mouse brain regions
GDS1490	12422	150	Mouse neural and body tissue
GDS3715	12559	110	Human skeletal muscles
GDS3716	22215	42	Breats epithelia: cancer patients

Table 4.2 GEO Datasets.

4.2 Biclustering using MicroArray Data

4.2.1 Application

YMM may be used to mine information from DNA Microarray Data. In this experiment, we followed the evaluation procedure proposed by Eren et al [18]. That is, the extraction of biclusters and the execution of Gene Ontology enrichment analysis on the extracted biclusters. In their work, they analyze this procedure for 12 algorithms. Therefore we compare results with the algorithms used in their study, namely Cheng and Church, PLAID, OPSM, ISA, Spectral, xMOTIFs, bimax, bayesian clustering, COALESCE, CPB, QUBIC and FABIA.

4.2.2 Settings

For YMM, we set the number of both row-topics (gene-topics) and column-topics (sample-topics) to be 15. Also, we set γ,γ_u,γ_p as explained in Section 3.4.3. The data must be pre-processed since Microarray data contains negative values and our model is designed for count data. We propose analyzing the positive values and the negative values separately. First, we replaced the negative values with zero so that we extract those genes and samples where there is a high expression of normal cells relative to tumor cells. Then, we replaced the positive values with zero so that we extract those genes and samples where there is a high expression of tumor cells relative to normal cells. A bicluster is considered to be enriched if

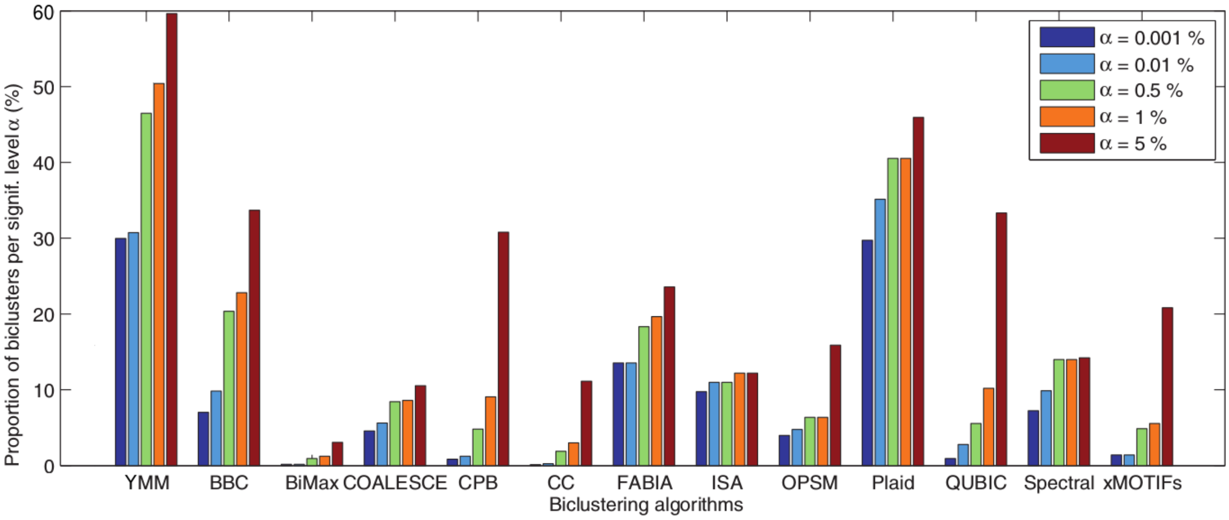


Fig. 4.1 Proportion of the enriched biclusters for YMM. The results are compared with the results of the study made by Eren et al on five different significance level α [18].

at least one term from the Biological Process Gene Ontology was enriched at the $P = 0.05$ level after Benjamini and Hochberg multiple test correction [18].

4.2.3 Results and Analysis

In figure 4.1, we show the proportions of the enriched biclusters aggregated for the eight datasets. The results show that the heavy bicluster is a bicluster type worth studying in the Gene Expression data. The number of biclusters extracted by YMM is relatively low, although, the proportion of biclusters extracted is relatively high. This is particularly true when the α level is selected to be greater than 0.1%. Table 4.3 shows the terms associated for the biclusters with the lowest p-value for the GDS581 dataset. YMM revealed new terms not discovered by other bicluster algorithms. This is because YMM targets the extraction of heavy biclusters with a flexible structure, while most algorithms target other types of biclusters and bicluster structures. The number of Genes associated with each bicluster is higher than other algorithms while maintaining low p-values; This indicates that YMM is a complementary model capable of extracting biclusters that may not be easy to extract with other algorithms.

Rows,cols	Terms	P-value
502,20	Hydrogen ion transmembrane transport	2.52e-17
566,4	ATP metabolic process	1.17e-15
674,18	Cellular amide metabolic process	2.72e-15
685,20	Hydrogen ion transmembrane transport	4.21e-14
643,10	Peptide metabolic process	4.17e-11

Table 4.3 Five most enriched terms on GDS589.

4.3 Modeling Visit Records

4.3.1 Application

YMM can generate new simultaneous samples for a pair of categorical variables. To generate new data using YMM we need to run the inference process to extract the topics and then we can use the generative process described in section 3.3.4 to sample instances of new records.

4.3.2 Experimental Settings

To evaluate YMM's ability to model data, we can compute the perplexity on a held-out test set. The perplexity on a test set is a commonly used metric for the predictability of the given test set. The following equation is used to find the perplexity for a set of records $\mathscr{W}$:

$$\text{Perplexity } (\mathscr{W}) = \exp^{-\frac{\sum_{v \in \mathscr{W}} \log P(\mathscr{W}_v, |\Phi, \beta)}{|\mathscr{W}|}}$$

We compute the perplexity for evaluating the predictability of words, authors, users and places using the NIPS and Foursquare datasets. We compare the perplexity of YMM with an LDA model, its dual formulation and the perplexity value at the initialization. Since YMM generates both word and authors simultaneously, we require to extract independently the perplexity of predicting users given by $P(u|\Phi_u, \alpha)$ and the perplexity of predicting places as given by $P(p|\Phi_p, \alpha)$. In the appendix we show how to compute these probabilities when using YMM.

For LDA, we set the number of places-topics and user-topics to fifty, we select $\alpha_i = 0.5\ \forall i \in 1 : K$ and $\beta_j = 0.1\ \forall j \in W$. Where W corresponds to the vocabulary set. In YMM models, We set $\alpha_i = 0.05\ \forall i \in K_u \times K_p$, all other parameters are set identically. For foursquare's data, we split the dataset by time, so that the 40% more recent records are held-out for testing. For the NIPS dataset, we select 20% of the documents at random.

We repeat the experiment with different initialization methods. We use uniform sampling as a naive initialization method, then we use K-means and finally we explore the initialization of YMM using the extracted results of LDA. The latter method will help us check if there is an improvement in the topic extraction given by the mutual reinforcement of location-based models and individual-based models.

4.3.3 Results and Analysis

We present the results in table 4.4. We observe that YMM converges to the best results. When we use the results of an LDA and its dual LDA to initialize YMM, YMM yields overall acceptable results and substantial faster convergence. The improvement is relatively larger for the evaluation of place topics when compared to the user topic, this is a consequence of the larger sparsity in the set of places. This effect is also observed on the results for the NIPS dataset. YMM generalizes the way that LDA models the data. As consequence, YMM obtains lower perplexity values.

Initialization	Model	Fsq New York		Fsq Tokyo		Fsq S'pore		NIPS	
		User	Place	User	Place	User	Place	Document	Word
Uniform	K-means	1559	992	4341	1922	2674	11810	3302	2290
K-means	LDA	1411	820	3994	1532	2457	11746	2583	1529
K-means	YMM	1411	788	4002	1412	2395	11731	2514	1452
K-means LDA	YMM	1411	783	4005	1349	2381	11730	2543	1496

Table 4.4 Average perplexity computed over twenty trials for four different data sets.

4.4 Reviewer Recommendation.

4.4.1 Application

We use text data to evaluate the ability of YMM of modeling the joint distribution of two categorical random variables under sparsity. As suggested by Rosen-Zvi et al, Automated Reviewer Recommendations is an application that requires extracting author similarity as well as word similarity under sparse conditions [59].

Since we do not have labels for the appropriate reviewers of a paper, we propose to do author classification. The objective is to identify the authors of a paper given the words written in it. Our assumption is that if we can correctly identify the authors of a document, then we should be able to identify similar authors who may have written the document and hence they could serve as reviewers for the document.

We use YMM to model the rows as authors, the columns as words and the values in each cell as the number of times an author wrote a word. With this configuration, we capture the similarity between authors with author-topics z_a, the similarity between words with word-topics z_w and the relationships between these topics.

More formally, given a document $\mathscr{D}$ represented by a set of words w, we can use YMM to extract the probability of an author a to write a word w from equation 4.1

$$P(a,w) = \sum_{z_a} \sum_{z_w} P(a|\Phi_{z_w}, \beta_w) P(a|\Phi_{z_a}, \beta_a) P(z_a, z_w|\Theta) \tag{4.1}$$

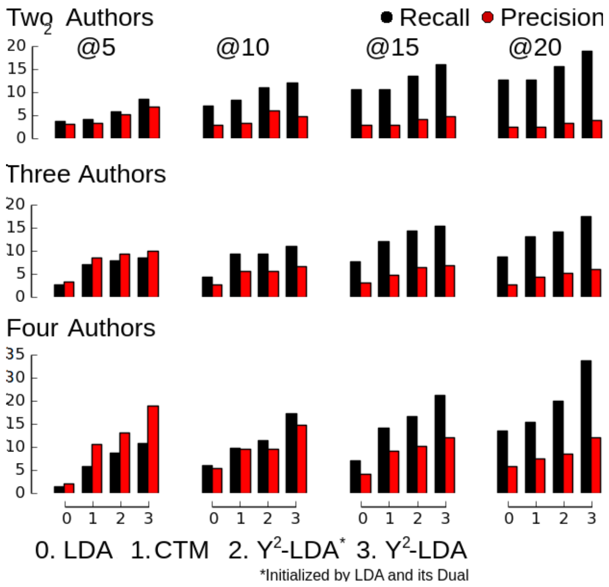


Fig. 4.2 Reviewer recommendation

Then, we use $\prod_{w\in\mathscr{D}} P(a|w)$ to find the probability of an author to have written the document $\mathscr{D}$. For comparative purposes, we also extract $P(a|w)$ using an LDA by representing the document level using words and the word-level using the set of all authors in the corpus. i.e. we use an LDA to represent a word as the set of authors who have written this word. In this model, the probability of an author a to write a word w is calculated from equation 4.2.

$$P(a|w) = \sum_{z_a} P(a|\Phi_{z_a})P(z_a|\Theta_w) \tag{4.2}$$

We use a similar formulation to estimate $P(a|w)$ using the Correlated Topic Model (CTM) [6].

4.4.2 Experimental Settings

For this experiment, we split the papers in the NIPS corpus into a training set and a testing set, after deleting all papers in the test set with one author and those papers where the authors are not in the training set, we obtain as the test set 525 papers with two authors, 305 with three authors and 111 with four authors. We evaluate the precision and recall at n by selecting

Trial	Uniform			Kmeans		
	LDA	YMM	LDA+YMM	LDA	YMM	LDA+YMM
1	0.032	0.032	$\underline{0.016}$	0.104	0.024	$\underline{0.008}$
2	0.032	0.272*	$\underline{0.016}$	$\underline{0.296}$*	0.304*	$\underline{0.296}$*
3	$\underline{0.008}$	0.272*	$\underline{0.008}$	0.208	0.336	$\underline{0.000}$
4	$\underline{0.016}$	0.016	$\underline{0.016}$	0.448	0.416	$\underline{0.208}$
5	0.040	0.024	$\underline{0.016}$	0.224	0.112	$\underline{0.008}$
6	0.312*	0.288*	$\underline{0.264}$*	0.288*	$\underline{0.144}$	0.272*
7	0.176	0.016	$\underline{0.008}$	0.112	0.128	$\underline{0.048}$
8	$\underline{0.000}$	0.152	$\underline{0.000}$	$\underline{0.016}$	$\underline{0.016}$	$\underline{0.016}$
9	0.304*	$\underline{0.272}$*	0.304	0.024	0.016	$\underline{0.008}$
10	0.184	$\underline{0.000}$	0.016	0.344*	0.304*	$\underline{0.160}$
Average	0.1104	0.1344	0.0664	0.2064	0.1800	0.1024

Table 4.5 Average misclassification error for ten random initializations. The initialization method is specified in the headers of the table. The asterisk represents those cases where the models completely confused two topics.

the top n authors from $P(a|\mathscr{D})$. In all the models we set the number of topics equal to fifty. In the case of YMM both author-topics and word-topics are set to fifty as well.

4.4.3 Results and Analysis

We present the results in figure 4.2. We observe and improvement of YMM over CTM and an improvement of CTM over LDA. This is expected as both CTM and YMM generalize LDA. Overall we can see how papers with larger number of authors have a better performance when trying to identify the authors of the paper. As expected, the precision decreases as we increase the list of recommended authors and in contrast, the recall increases. Finally, we note that initializing YMM with a uniform distribution leads to better performance that initializing it with the result of an LDA and its dual LDA, though the convergence time is about an order of magnitude slower.

4.5 User and Place Classification

4.5.1 Application

YMM can be used to classify data using the latent topics as the categories that we use to represent the data. In a hard classification scenario, we can assign class z to user u_i using $\hat{z_u} = \arg\max_{z_u} P(u = u_i|\Phi_u, z_u)$. In a soft-classification mode, one could extract $P(z_u|u = u_i) = \frac{P(u=u_i|\Phi_{u,z_u})P(\Phi_{u,z_u})}{P(u=u_i)}$ where $P(u = u_i)$ can be found by counting from the given data and $P(\Phi_{u,z_u}) = \frac{\prod_{i=1}^{K}\Gamma(\beta_i)}{\Gamma(\sum_{i=1}^{K})\beta_i}\prod_{i=1}^{K}\Phi_{u,i}{}^{\beta_i - 1}$ as specified by the Dirichlet prior with parameter β

4.5.2 Experimental Settings

In this experiment, we test YMM's ability to extract individual user-topics and individual place-topics that may be used in a classification scenario. We selectively filter Singtel cell-tower data so that we can obtain clear topics of users and places. We select all the visit records of five groups of 250 users whose home and work correspond to the same cell tower location. We assume as ground truth that there exists five user-topics, each topic representing a group of 250 users. For every user u, we assign the user-topic $\hat{z_u}$ using $\hat{z_u} = \arg\max_{z_u} P(u = u_i|\Phi_u, z_u)$. For the evaluation, we compare the average misclassification error of YMM against an LDA. We initialize the models uniformly, using K-Means, and using an LDA and its dual LDA.

4.5.3 Results and Analysis

The results of the experiment are presented in table 4.5. As in experiment 4.3, we observe consistent improvement of YMM model when initialized using K-means and the independent LDA models. However, in those cases where K-means initialization is not so convenient such in trials 7 and 8, we cannot expect YMM to solve the confusion between the extracted

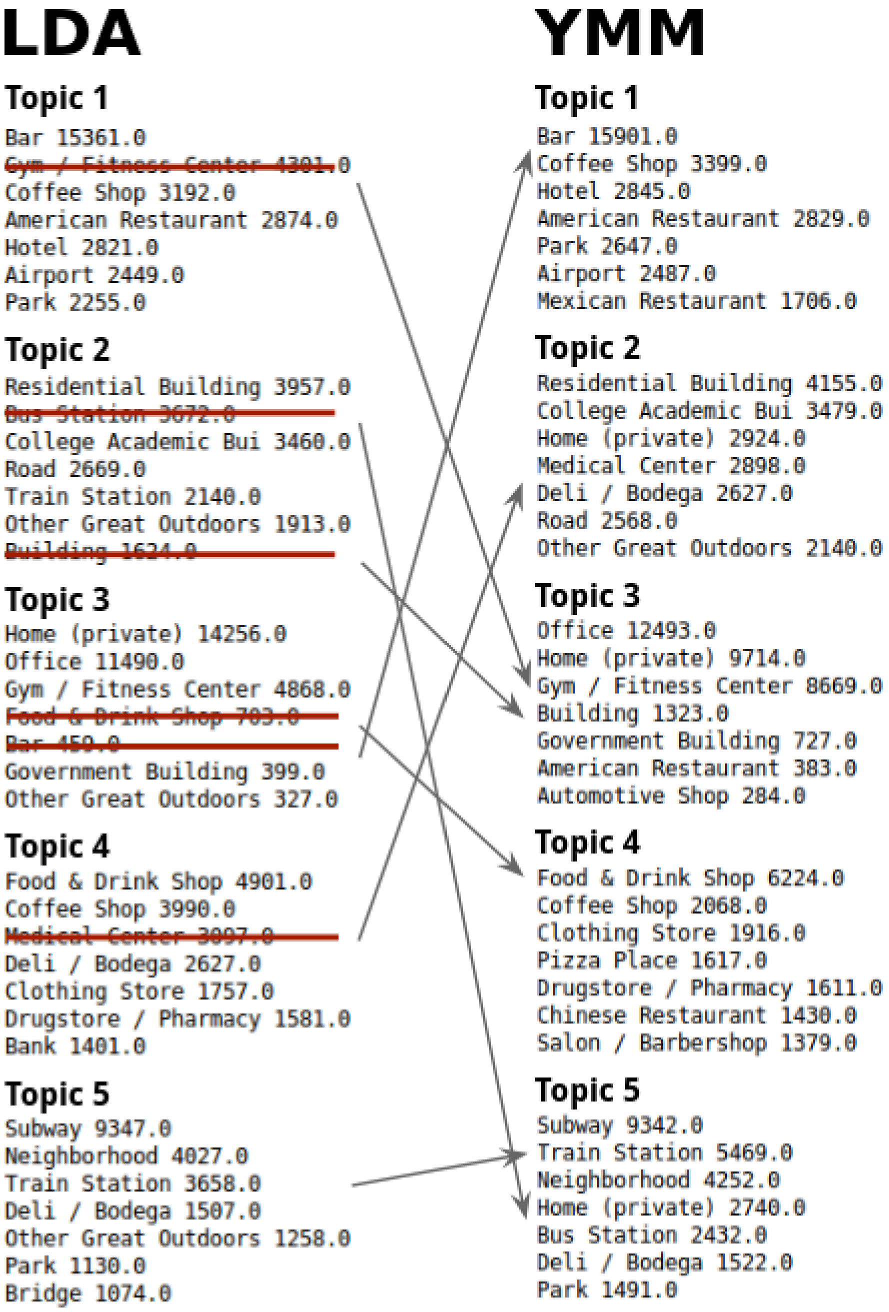


Fig. 4.3 Comparison for place-topic extraction between LDA and YMM. YMM was initialized with the results of the LDA. The numbers represent the number of samples allocated to a topic

topics of the independent LDA models. We also observe that LDA is in this case a better way to initialize YMM model. We attribute the improvement of YMM over LDA to the fact that YMM models an iterative inference of both user-based models and location-based model. In contrast, the LDA is purely location-based model.

Additionally, we used the New York Foursquare's dataset to compare the extracted place-topics of LDA and YMM. In figure 4.3 We point out all differences that we consider to be an improvement. Typically, YMM is able to solve some confusion between the places represented by two topics. For example, after convergence of the LDA model, *Gym* appeared in topic 1 and 3. However, when we used YMM to improve the convergence of LDA, *Gym* was moved to topic 3 where it is better represented. Our model also seems to improve the ranking of the places inside a topic. For example, in topic 5, YMM allocated more instances to *Train Station* in a topic that clearly corresponds to places related to transportation.

4.6 Biclustering

4.6.1 Application

In this experiment, we evaluate the flexibility of our model to extract *heavy biclusters* assuming different structures in a synthetic dataset. We study three biclustering structures: i) Every column and every row must be assigned to exactly one bicluster, ii) Every column must be assigned to one bicluster, but the rows may be assigned to more than one bicluster, iii) Every column/row may or may not be assigned to any bicluster. We use normal distributions to generate the data. The cells corresponding to a bicluster naturally have a higher mean when compared with those cells that do not belong to any bicluster. The variance is used to control amount of noise in the data. The specification of each dataset is shown in Figure 4.5.

4.6.2 Experimental Settings

For the experiment, we assigned the hyper priors α with a value of 5, β_u and β_p with a value of 0.01. We set γ, γ_u, γ_p as explained in Section 3.4.3. We initialize our model uniformly and we also tried to initialize our model using two LDA models. One LDA considering the columns as the documents, and the rows as words to initialize Z_u, and one LDA considering the rows as the documents and the columns as words to initialize Z_p.

For the evaluation metric, we use recovery $S(\mathscr{E},\mathscr{B})$ and relevance $S(\mathscr{B},\mathscr{E})$ as an established metric used to validate the biclustering extraction [18, 74]. $\mathscr{E}$ as the list of expected biclusters, $\mathscr{B}$ is the list of extracted biclusters and the function S is defined by

$$S(\mathscr{B}_1,\mathscr{B}_2) = \frac{1}{|\mathscr{B}_1|}\sum_{b_1\in\mathscr{B}_1}\max_{b_2\in\mathscr{B}_2} s(b_1,b_2)$$

Here $s(b_1,b_2)$ is the Jaccard similarity applied for the submatrix elements defined by each bicluster. Intuitively, these metrics are an analogy to Recall and Precision.

We compared with the spectral bipartite graph [14] and SAMBA [64], since these algorithms extract similar types of bicluster as we do. We also compared with the most commonly used baseline in the biclustering task [33] and with the PLAID model [41] because of its good performance on Gene Expression data. Finally, we compared with probabilistic approach called xMOTIF [54].

4.6.3 Results and Analysis

We repeated the experiment ten times and we report the average results in Figure 4.5. In the first dataset, we observe an improvement of the spectral bipartite graph over other models, that is because Spectral bipartite graph assumes that all columns and all rows must be assigned to exactly one bicluster and so the algorithm is fitted for the assumptions of the first dataset. Nonetheless the performance of YMM is also competitive in this dataset. But,

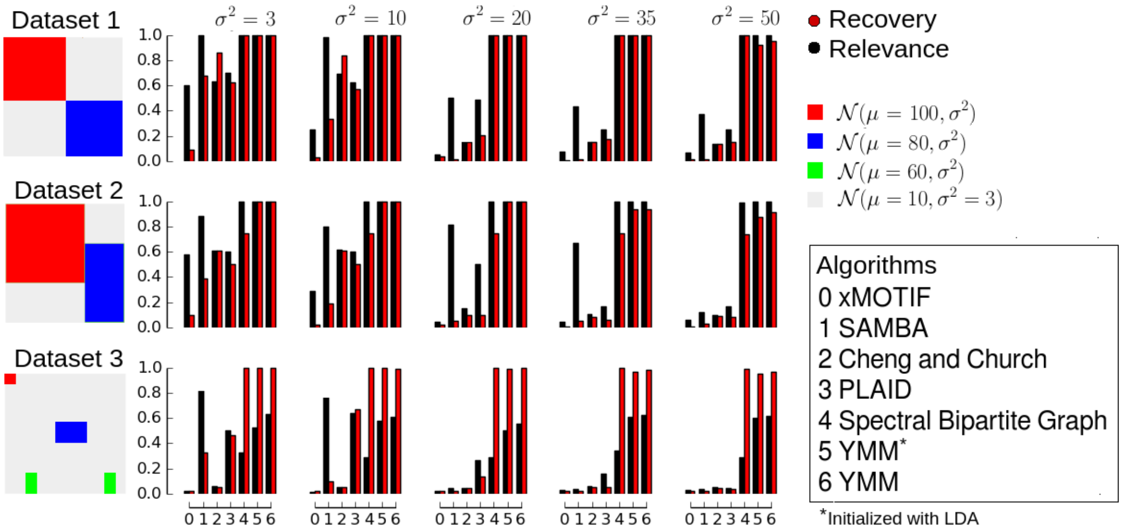


Fig. 4.4 Recovery and Relevance results for bicluster classification. The datasets are visualized with a matrix whose color represent the number of visits of users (rows) to places (columns).

when we relax this assumption in datasets 2 and 3, the results of the spectral bipartite graph degrade compared to our model. SAMBA is able to recognize the biclusters, but since it does not use the values in the matrix, it is not robust to noise and it does not often discover all the rows and columns involved in the bicluster. Overall YMM yields better results even when initialized using LDA.

The spectral bipartite graph [15] algorithm performs much better than other baseline models because the spectral bipartite graph algorithm models the relation between users and places with a bipartite graph weighted proportional the number of visits. Therefore, this algorithm is designed for extracting *heavy biclusters*.

In addition, we visualize one sample of the biclusters extracted by YMM to compare the difference between initializing YMM using two LDA models and YMM using two k-means algorithms. The visualization is shown in figure 4.4, every color represent a different bicluster, on the right we can see the input dataset where the number of visits from user $i-th$ to place $j-th$ is given by a value between 0 and 200.

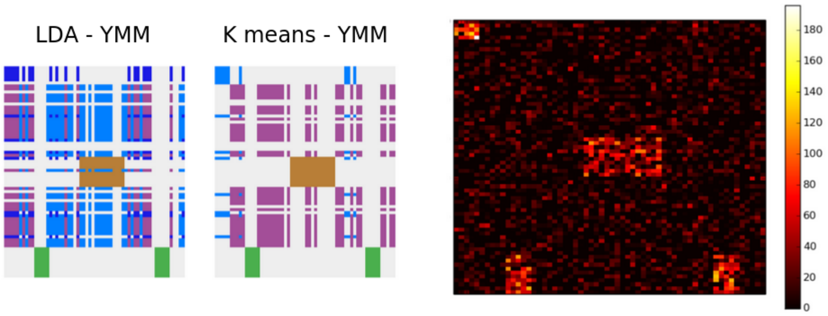


Fig. 4.5 Different initialization methods for extracting biclusters with YMM

As result, we observe that initializing YMM with K means is a better way to find more relevant biclusters for two reasons. First, the number of extracted biclusters is more precise since K means-YMM adds more weight to the topic-correlation that represents the ground truth biclusters. Second, K means-YMM extracts biclusters that better fit the shape of the ground truth biclusters. This is because K means-YMM extracts topics that better represent the subsets of rows and subset of columns that we expect as ground truth.

4.7 Estimating Conditional Distributions

4.7.1 Application

When the relationship between categorical variables is not sparse, we may estimate their conditional distribution using maximum likelihood estimation or maximum a posteriori estimation. However, estimation methods suffer from over parametrization [10]. Topic Models can be used to model the conditional distribution between two categorical random variables with a sparse relationship. In the mobility field, we can use LDA extracted place-topics to cluster the places $\mathscr{P}$) and we assigned the users ($\mathscr{U}$) with membership probabilities to the place-clusters. In other words, we describe $\hat{P}(\mathscr{P}|\mathscr{U})$ using the topics as a mean to solve the sparsity existing in the relationship between users and places. An optimal estimation of the conditional distribution would be as close as possible to the truth distribution $P(\mathscr{P}|\mathscr{U})$.

N^o Topics	$P(\mathscr{P}\vert\mathscr{U})$			$P(\mathscr{U}\vert\mathscr{P})$		
	YMM	LDA	Improvement	YMM	LDA	Improvement
3	.261 (.003)	.262 (.005)	44.7%	.644 (.006)	.647 (.005)	58.2%
6	.227 (.002)	.245 (.006)	76.3%	.592 (.006)	.621 (.005)	69.9%
9	.214 (.003)	.242 (.005)	78.1%	.565 (.007)	.623 (.005)	89.4%
12	.204 (.004)	.559 (.008)	99.7%	.550 (.007)	1.00 (.004)	99.8%
15	.193 (.004)	.528 (.011)	99.9%	.523 (.006)	1.01 (.005)	99.9%

Table 4.6 On the left side of the table, we show the average Battacharyya distance between the approximated distributions $\hat{P}_{LDA}(\mathscr{P}|\mathscr{U})$, $\hat{P}_{YMM}(\mathscr{P}|\mathscr{U})$ and the estimated distribution $P(\mathscr{P}|\mathscr{U})$. We assume the maximum likelihood estimation of $P(\mathscr{P}|\mathscr{U})$ as ground truth. On the right side, we display the Battacharyya distance between the approximated distributions $\hat{P}_{LDA}(\mathscr{U}|\mathscr{P})$, $\hat{P}_{YMM}(\mathscr{U}|\mathscr{P})$ and the estimated distribution $P(\mathscr{U}|\mathscr{P})$. We assume the maximum likelihood estimation of $P(\mathscr{U}|\mathscr{P})$ as ground truth.

4.7.2 Experimental Settings

In this experiment, we use Singtel Cell Location dataset and we assume the ground truth to be the maximum likelihood estimation of $P(\mathscr{P}|\mathscr{U})$. To keep this assumption in place, we select the most active users and the most visited places. We choose the users whose visit count is above 0.05% of the total number of visit counts and we use the same procedure to choose the places. For the evaluation, we extract the estimated value of $P_{(}\mathscr{P}|\mathscr{U})$ using the LDA model as $\hat{P_{lda}}(\mathscr{U}|\mathscr{P}) = \sum_{z_u}^{Ku} P(z_u|\Theta,\alpha)P(\mathscr{U}|\Phi_{z_u}\beta_u)$. Similarly we use the dual LDA formulation to estimate $P(\mathscr{U}|\mathscr{P})$. For the YMM model we approximate the conditional distribution by first extracting the join distribution $P_YMM\hat{(}\mathscr{U},\mathscr{P}) = \sum_{z_u}^{Ku}\sum_{z_p}^{Kp} P(z_u,z_p|\Theta,\alpha)P(\mathscr{U}|\Phi_{z_u}\beta_u)P(\mathscr{P}|\Phi_{z_p}\beta_p)$ and then we extract the conditional distributions by normalizing the joint distribution. We use the Bachattaryya distance to measure the distance between distributions, the Bachattaryya distance is a metric commonly used to measure the distance for discrete distributions [**?**]. In the case of multinomial distributions the Bachattaryya distance between p and q corresponds to $Battacharyya(p,q) = -ln\left(\sum_x \sqrt{p(x),q(x)}\right)$. We report results based on averaging the results of running the experiment 50 times.

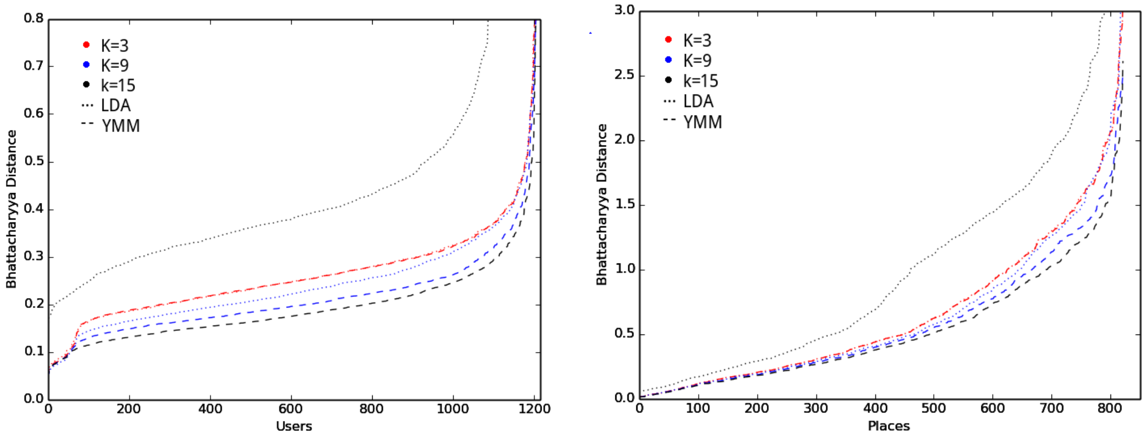


Fig. 4.6 Results of table 4.6 for every user, we sort the users and places by their average Bhattacharyya distance to facilitate the visualization of the results.

4.7.3 Results and Analysis

In table 4.6, we observe the average Bhattacharyya distance between the approximated distributions $\hat{P}_{LDA}(\mathscr{P}|\mathscr{U})$, $\hat{P}_{YMM}(\mathscr{P}|\mathscr{U})$ and the ground truth distribution $P(\mathscr{P}|\mathscr{U})$. We also show the percentage of the users whose average distance to the ground truth distribution is improved as well as the percentage of places improved by the YMM model.

YMM improves the place-distribution for around 75% of all the users and 70% of the user-distributions per place. These results hold independently of the number of topics. However, when we use a small number of topics, the performance of YMM and LDA tend to converge. Naturally, as we increase the number of topics, we obtain smaller distances to the truth distributions. When we surpass 10 topics, LDA decreases its performance dramatically. This is because when we use LDA to extract larger number of topics, several topics end up being repeated. This is a phenomena observed in [6]. The cause is that the relationship between LDA's topics are independent. The Dirichlet distribution that parameterizes the relation between the topics can be always replaced by independent gamma distributions. In our model, the bivariate categorical distribution parameterizes the relationship between different kinds of topics and it also holds implicitly the relationship between the same kind of topic.

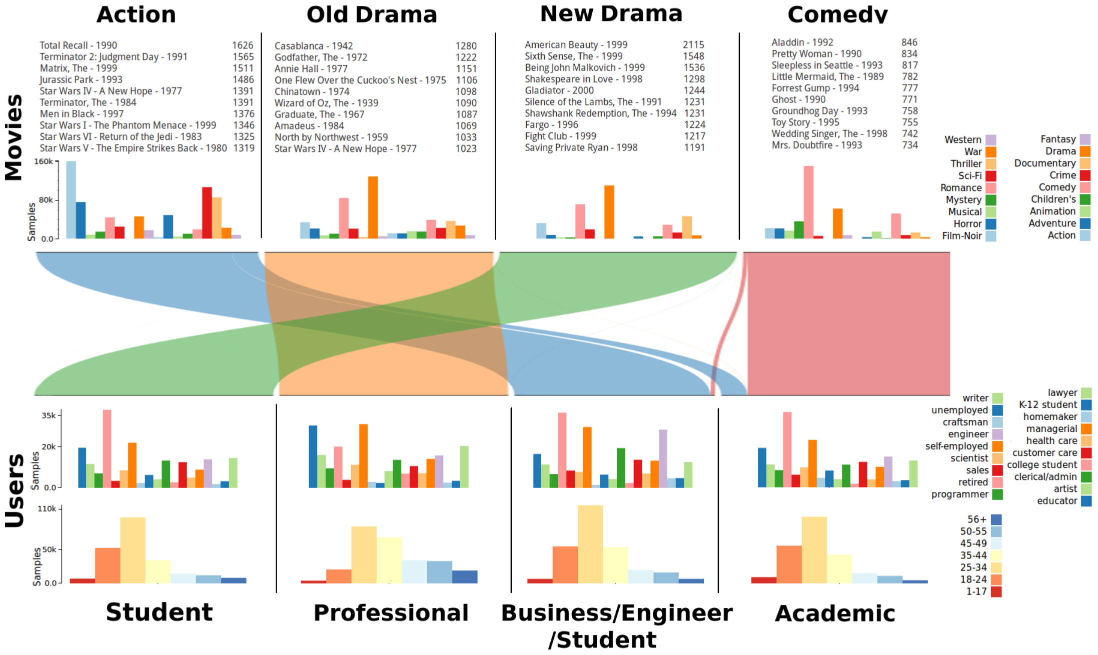
Figure 4.6 displays the distance of the place-distribution to the ground truth for every individual and the distance of the user-distribution to the ground truth for every place. For our model, the figure shows how increasing the number of topics reflects an homogeneous improvement of performance for all the users and all the places. For the LDA model, the figure shows an homogeneous increase of the performance when we increase the number of topics from 3 to 9. However, when we increase the number of topics from 9 to 15, the performances decreases.

4.8 Topic Visualization

Finally, we created a visualization to show the extracted topics and their correlation. We applied this visualization on two domains. First, we used a Movie rating dataset using one million ratings from 6000 users on 4000 movies [1]. We selected this dataset since we also have available user demographic information that can be used to explain the user topics. We applied YMM to the user-movie domain in order to extract *Which groups of users watch which types of movies*. The results are displayed in figure 4.7 We use the parallel sets visualization to display the correlation of topics [2]. In this experiment, We obtained a one to one relation between the user-topics and the movie-topics. The label of the topics is based on our judgments. Our main findings indicate that users related to academia have higher preference towards drama and comedy movies, while users related to the fields of business and engineering have a higher interest in action movies. As expected, Professionals and older users naturally prefer to watch older movies.

Secondly, we tried this visualization on the NIPS dataset, we filtered those words that don't exhibit a particular meaning e.g. data, figure, image. The results are displayed in

[1] http://grouplens.org/datasets/movielens/



Action	
Total Recall - 1990	1626
Terminator 2: Judgment Day - 1991	1565
Matrix, The - 1999	1511
Jurassic Park - 1993	1486
Star Wars IV - A New Hope - 1977	1391
Terminator, The - 1984	1391
Men in Black - 1997	1376
Star Wars I - The Phantom Menace - 1999	1346
Star Wars VI - Return of the Jedi - 1983	1325
Star Wars V - The Empire Strikes Back - 1980	1319

Old Drama	
Casablanca - 1942	1280
Godfather, The - 1972	1222
Annie Hall - 1977	1151
One Flew Over the Cuckoo's Nest - 1975	1106
Chinatown - 1974	1098
Wizard of Oz, The - 1939	1090
Graduate, The - 1967	1087
Amadeus - 1984	1069
North by Northwest - 1959	1033
Star Wars IV - A New Hope - 1977	1023

New Drama	
American Beauty - 1999	2115
Sixth Sense, The - 1999	1548
Being John Malkovich - 1999	1536
Shakespeare in Love - 1998	1298
Gladiator - 2000	1244
Silence of the Lambs, The - 1991	1231
Shawshank Redemption, The - 1994	1231
Fargo - 1996	1224
Fight Club - 1999	1217
Saving Private Ryan - 1998	1191

Comedy	
Aladdin - 1992	846
Pretty Woman - 1990	834
Sleepless in Seattle - 1993	817
Little Mermaid, The - 1989	782
Forrest Gump - 1994	777
Ghost - 1990	771
Groundhog Day - 1993	758
Toy Story - 1995	755
Wedding Singer, The - 1998	742
Mrs. Doubtfire - 1993	734

Fig. 4.7 On top, we show four movie-topics represented by a) Most representative movies and its year of release. b) Histogram of the movie genres weighted by the number of instances that a movie was allocated to a topic. On the middle we represent the correlation between topics using a parallel sets visualization. On the bottom, we show the user-topics represented by a) Histogram of the user's profession b) Histogram of the user's age. Both histograms are weighted by the number of instances that a user was allocated to a topic.

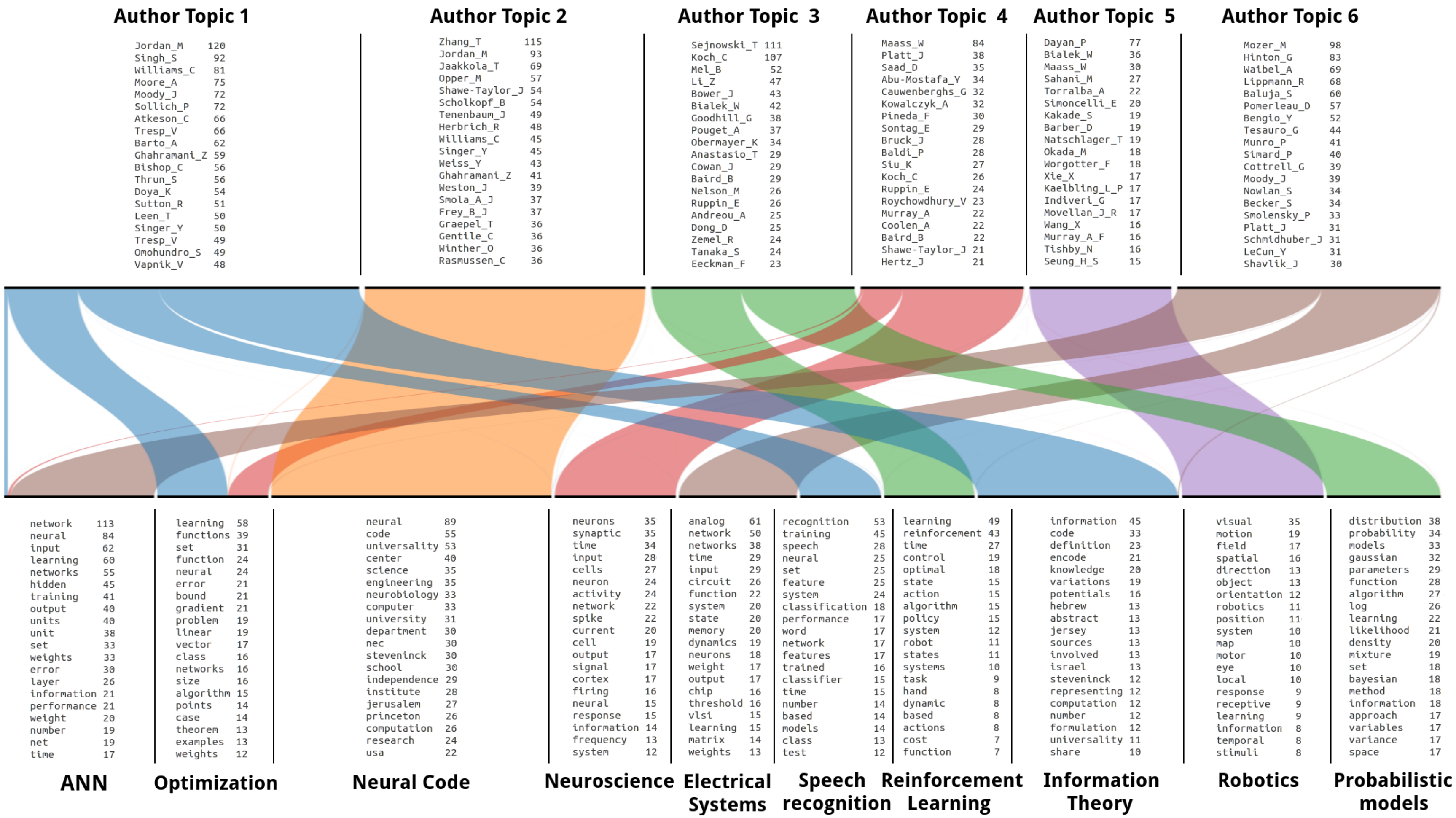


Fig. 4.8 Author-topics, word-topics and its interrelation. The width of the box surrounding each topic represents the total amount of samples allocated to the topic. For each topic, we list the top elements and the number of times that each element is allocated to the topic in thousands.

figure 4.8. In this experiment, we extracted six author topics and ten words topics to be able to compare with NIPS technical areas [3]. As result, we observe high correlation between the extracted word topics and NIPS technical areas as the fields of Hardware Technologies, Learning theory, Neuroscience and Speech Recognition emerge as topics. Different from the expected NIPS technical areas, we find a topic categorization that falls into solving approaches rather than technical areas. For instance, probability, optimization and artificial neural networks emerge each as topics, while in NIPS they are all considered to be part of the Algorithms and Architectures technical area.

[3] https://nips.cc/Conferences/2008/CallForPapers

[3] https://www.jasondavies.com/parallel-sets/

References

[1] Bao, J., Zheng, Y., and Mokbel, M. F. (2012). Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '12, pages 199–208, New York, NY, USA. ACM.

[2] Bauer, S., Noulas, A., Seaghdha, D. O., Clark, S., and Mascolo, C. (2012). Talking places: Modelling and analysing linguistic content in foursquare. In *SocialCom/PASSAT*, pages 348–357. IEEE.

[3] Blei, D. M. (2012a). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

[4] Blei, D. M. (2012b). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

[5] Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30.

[6] Blei, D. M. and Lafferty, J. D. (2006a). Correlated topic models. In *In Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. MIT Press.

[7] Blei, D. M. and Lafferty, J. D. (2006b). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA. ACM.

[8] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

[9] Bobadilla, J., Ortega, F., Hernando, A., and GutiéRrez, A. (2013). Recommender systems survey. *Know.-Based Syst.*, 46:109–132.

[10] Brown, M. B. and Fuchs, C. (1983). On maximum likelihood estimation in sparse contingency tables. *Computational Statistics and Data Analysis*, 1:3 – 15.

[11] Chandra, B., Shanker, S., and Mishra, S. (2006). A new approach: Interrelated two-way clustering of gene expression data. *Statistical Methodology*. Bioinformatics.

[12] Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103. AAAI Press.

[13] Collins, M., Dasgupta, S., and Schapire, R. E. (2002). A generalization of principal components analysis to the exponential family. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 617–624. MIT Press.

[14] Dhillon, I. S. (2001a). Co-clustering documents and words using bipartite spectral graph partitioning. KDD.

[15] Dhillon, I. S. (2001b). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 269–274, New York, NY, USA. ACM.

[16] Dhillon, I. S., Mallela, S., and Modha, D. S. (2003). Information-theoretic co-clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 89–98, New York, NY, USA. ACM.

[17] Du, L., Buntine, W., and Jin, H. (2010). Sequential latent dirichlet allocation: Discover underlying topic structures within a document. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 148–157.

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

[19] Escobar, M. D. and West, M. (1994). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

[20] Farrahi, K. and Gatica-Perez, D. (2011). Discovering routines from large-scale human locations using probabilistic topic models. *ACM Trans. Intell. Syst. Technol.*, 2(1):3:1–3:27.

[21] Farrahi, K. and Gatica-Perez, D. (2012). Extracting mobile behavioral patterns with the distant n-gram topic model. In *Wearable Computers (ISWC), 2012 16th International Symposium on*, pages 1–8.

[22] Farrahi, K. and Gatica-Perez, D. (2014). A probabilistic approach to mining mobile phone data sequences. *Personal Ubiquitous Comput.*, 18(1):223–238.

[23] Ferrari, L., Rosi, A., Mamei, M., and Zambonelli, F. (2011). Extracting urban patterns from location-based social networks. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, pages 9–16, New York, NY, USA. ACM.

[24] Fox, C. and Roberts, S. (2012). A tutorial on variational Bayesian inference. *Artificial Intelligence Review*, 38(2):85–95.

[25] Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2011). Bayesian nonparametric inference of switching dynamic linear models. *Signal Processing, IEEE Transactions on*, 59(4):1569–1585.

[26] Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2008). An hdp-hmm for systems with state persistence. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 312–319, New York, NY, USA. ACM.

[27] Getz, G., Levine, E., and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences*, (22):12079–12084.

[28] Ghahramani, Z. and Beal, M. J. (2000). Variational inference for bayesian mixtures of factor analysers. In *In Advances in Neural Information Processing Systems 12*, pages 449–455. MIT Press.

[29] Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden markov models. *Mach. Learn.*, 29(2-3):245–273.

[30] Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235.

[31] Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. (2005). Integrating topics and syntax. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press.

[32] Gruber, A., Weiss, Y., and Rosen-Zvi, M. (2007). Hidden topic markov models. In *AISTATS*, volume 2 of *JMLR Proceedings*, pages 163–170. JMLR.org.

[33] Hartigan, J. A. (1972). Direct clustering of a data matrix. 67(337):123–129.

[34] Hitchcock, C. (2012). Probabilistic causation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition.

[35] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347.

[36] Hofmann, T. (1999). Probabilistic latent semantic analysis. In *In Proc. of Uncertainty in Artificial Intelligence, UAI'99*, pages 289–296.

[37] Ihmels, J., Bergmann, S., and Barkai, N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20:2004.

[38] Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173.

[39] Joseph, J., Doshi-Velez, F., Huang, A. S., and Roy, N. (2011). A bayesian nonparametric approach to modeling motion patterns. *Auton. Robots*, 31(4):383–400.

[40] Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME – Journal of Basic Engineering*, (82 (Series D)):35–45.

[41] Lazzeroni, L. and Owen, A. (2000). Plaid models for gene expression data. *Statistica Sinica*, 12:61–86.

[42] Li, W. and McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 577–584, New York, NY, USA. ACM.

[43] Liu, B., Fu, Y., Yao, Z., and Xiong, H. (2013). Learning geographical preferences for point-of-interest recommendation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1043–1051, New York, NY, USA. ACM.

[44] Liu, B. and Xiong, H. (2013). Point-of-interest recommendation in location based social networks with topic and location awareness. In *Proceedings of the 13th SIAM International Conference on Data Mining, May 2-4, 2013. Austin, Texas, USA.*, pages 396–404.

[45] Long, X., Jin, L., and Joshi, J. (2012). Exploring trajectory-driven local geographic topics in foursquare. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 927–934, New York, NY, USA. ACM.

[46] Ma, Qin M, T. H. T. P. A. Y. X. (2009). Qubic: a qualitative biclustering algorithm for analyses of gene expression data. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*.

[47] Madeira, S. and Oliveira, A. (2004). Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 1(1):24–45.

[48] Mahfouz, M. and Ismail, M. (2012). Soft flexible overlapping biclustering utilizing hybrid search strategies. In Hassanien, A., Salem, A.-B., Ramadan, R., and Kim, T.-h., editors, *Advanced Machine Learning Technologies and Applications*, volume 322 of *Communications in Computer and Information Science*, pages 315–326. Springer Berlin Heidelberg.

[49] McInerney, J., Zheng, J., Rogers, A., and Jennings, N. R. (2013a). Modelling heterogeneous location habits in human populations for location prediction under data sparsity. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 469–478, New York, NY, USA. ACM.

[50] McInerney, J., Zheng, J., Rogers, A., and Jennings, N. R. (2013b). Modelling heterogeneous location habits in human populations for location prediction under data sparsity. In *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 469–478.

[51] McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics, New York.

[52] Mei, Q., Liu, C., Su, H., and Zhai, C. (2006). A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 533–542, New York, NY, USA. ACM.

[53] Minka, T. and Lafferty, J. (2002). Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pages 352–359, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[54] Murali, T. M. and Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data. In *Pac. Symp. Biocomput*, pages 77–88.

[55] Paisley, J. and Carin, L. (2009). Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 777–784, New York, NY, USA. ACM.

[56] Pontes, Beatriz, G., Raúl, and Aguilar-Ruiz, J. S. (2015). Biclustering on expression data: A review. *Journal of Biomedical Informatics*.

[57] Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129.

[58] Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pages 814–822.

[59] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States. AUAI Press.

[60] Roy, D. M. and Teh, Y. W. (2009). The mondrian process. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1377–1384.

[61] Shan, H. and Banerjee, A. (2008). Bayesian co-clustering. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, pages 530–539.

[62] Sontag, D. and Roy, D. (2011). Complexity of inference in latent dirichlet allocation. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 1008–1016. Curran Associates, Inc.

[63] Steyvers, M. and Griffiths, T. (2007). *Probabilistic Topic Models*. Lawrence Erlbaum Associates.

[64] Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. In *ISMB*, pages 136–144.

[65] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2004). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101.

[66] Teh, Y. W., Newman, D., and Welling, M. (2007). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 1353–1360. MIT Press.

[67] Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 977–984, New York, NY, USA. ACM.

[68] Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, New York, NY, USA. ACM.

[69] Wang, C., Blei, D. M., and Heckerman, D. (2008). Continuous time dynamic topic models. In McAllester, D. A. and Myllymäki, P., editors, *UAI*, pages 579–586. AUAI Press.

[70] Wang, C., Wang, J., Xie, X., and Ma, W.-Y. (2007a). Mining geographic knowledge using location aware topic model. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, GIR '07, pages 65–70, New York, NY, USA. ACM.

[71] Wang, X. and McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 424–433, New York, NY, USA. ACM.

[72] Wang, X., McCallum, A., and Wei, X. (2007b). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 697–702, Washington, DC, USA. IEEE Computer Society.

[73] Wei, X., Sun, J., and Wang, X. (2007). Dynamic mixture models for multiple time series. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 459, pages 2909–2914.

[74] Xiong, H., Karypis, G., Thuraisingham, B. M., Cook, D. J., and Wu, X., editors (2013). *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*. IEEE Computer Society.

[75] Yang, D., Zhang, D., Zheng, V., and Yu, Z. (2015). Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, 45(1):129–142.

[76] Yin, Z., Cao, L., Han, J., Zhai, C., and Huang, T. (2011). Geographical topic discovery and comparison. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 247–256, New York, NY, USA. ACM.

[77] Yuan, J., Zheng, Y., and Xie, X. (2012). Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 186–194, New York, NY, USA. ACM.

[78] Yuan, Q., Cong, G., Ma, Z., Sun, A., and Thalmann, N. M. (2013). Who, where, when and what: Discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 605–613, New York, NY, USA. ACM.

[79] Zhang, J., Song, Y., Zhang, C., and Liu, S. (2010). Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1079–1088, New York, NY, USA. ACM.

Appendix A

Perplexity of an Individual Topic when Using Y^2-LDA

To compare GLDA's perplexity with an LDA model, we need to extract the user-topics $P(u|\Phi_u,\alpha)$ and the place-topics $P(p|\Phi_p,\alpha)$ separately. Because these distributions are intractable, we use Chib-Estimation method [68] to obtain an approximate result. The estimation method requires that we derive analytically $P(u,z_u|\Phi_u,\alpha)$, and $P(p,z_p|\Phi_p,\alpha)$. We can get the former distribution since $P(u,z_u|\Phi_u,\alpha) = P(u|\Phi_u,z_u)P(z_u|\alpha)$. We can get $P(u|\Phi_u,z_u)$ from the inference process. With respect to $P(z_u|\alpha)$, we can start by writing out $P(z_u,z_p|\alpha) = \int_{\Theta} P(z_u,z_p|\theta)P(\Theta|\alpha)$ which can be re-written as $P(z_u,z_p|\alpha) = \int_{\Theta_i}\int_{\Theta_j} P(z_u|\Theta_j)P(\Theta_j|\alpha)P(z_p|\Theta_i)P(\Theta_i|\alpha)$. Finally, we to obtain $P(z_u|\alpha) = \int_{\Theta_j} P(z_u|\Theta_j)P(\Theta_j|\alpha) \propto \frac{\prod_{i=1}^{K_p}\left(\Gamma\left(c\left(i,z_u,*,*\right)^{-v}+\alpha_{z_u\times K_p+i}\right)\right)}{\Gamma\left(\sum_{i=1}^{K_u}\left(c\left(i,k_u,*,*\right)^{-v}+\alpha_{z_u\times K_p+i}\right)\right)}$

Appendix B

Inference Process

Another challenge of the YMM is the inference of the posterior distribution $P(z_u, z_p|u, p)$. This distribution is intractable to compute due to the fact that the topics are dependent on each other. Furthermore, applying commonly used inference techniques such Variational Inference or Collapsed Gibbs Sampling is not viable since the generative process that we describe requires a simultaneous sampling. This type of sampling has not been studied in topic modeling.

Equation B.1 correspond to the posterior distribution that we use to update the user-topic $z_{u(v)}$ and place-topic $z_{p(v)}$ assigned to the *v-th* visit record. For each visit record, we must sample from this distribution to update its topic assignments, the posterior distribution must be recomputed after each assignment. Updating all the records corresponds to one iteration. After several iterations, the posterior distribution will converge, and we have found the topic assignment for each record. All details of the posterior derivation can be found in section B.1.

$$P\left(z_{u(v)}, z_{p(v)}|z_{u(-v)}, z_{p(-v)}, u, p, \alpha, \beta_u, \beta_p\right) \propto \frac{\mathscr{Z}\,\mathscr{P}(v)\,\mathscr{U}(v)}{\sum_{j=1}^{|P|}\mathscr{P}(j)\sum_{i=1}^{|U|}\mathscr{U}(i)} \tag{B.1}$$

$$\mathscr{X} = \sum_{m=1}^{|U|}\sum_{n=1}^{|P|} c\left(z_{p(v)}, z_{u(v)}, m, n\right)^{(-v)} + \alpha_{z_{u(v)}, z_{p(v)}}$$

$$\mathscr{P}(i) = \sum_{y=1}^{Kp}\sum_{n=1}^{|P|} c\left(z_{p(v)}, y, p_{(i)}, n\right)^{(-v)} + \beta_{p(i)}$$

$$\mathscr{U}(i) = \sum_{x=1}^{Ku}\sum_{m=1}^{|U|} c\left(x, z_{u(v)}, m, u_{(i)}\right)^{(-v)} + \beta_{u(i)}$$

$$c(x,y,m,n) = \sum_{j=1}^{|V|} I(z_{u(j)} = x, z_{p(j)} = y, u_{(j)} = m, p_{(j)} = n)$$

The inference equation depends on three types of counting. First, a normalized counting of the user assignment to the user-topics $\mathscr{U}$ capturing user co-occurrence. Second, a normalized counting of the place assignment to the place-topics $\mathscr{P}$ capturing place co-occurrence. And finally a counting over the current relationship between user-topics and place-topics $\mathscr{X}$.

B.1 Derivation of the Inference Equation using Collapsed Gibbs Sampling

In this section we derive the posterior distribution used to assign the topics. The derivation includes two main sections. First, we apply collapsed gibbs sampling to derive analytically the posterior distribution given the conditions of the model. Second, we use the fact that every record is independent and identically distributed so that we can find the posterior derivation for each particular record in the dataset.

B.1.1 Posterior Derivation given the conditions of the model

First we write down the target distribution that we want to derive:

$$P\left(z_{u,v}, z_{p,v} | z_{u,-v}, z_{p,-v}, u, p, \alpha, \beta_u, \beta_p\right) \tag{B.2}$$

The idea is to find the updating topics z_p and z_u for a given visit record v, given the training data. The definition of the variables used in this derivation can be found in figure 3.3.

Now, we start by formulating the joint distribution over all the variables in the model.

$$P(u,p,z_u,z_p,\Phi_u,\Phi_p,\Theta|\alpha,\beta_u,\beta_p) = \tag{B.3}$$

To obtain the target distribution using collapsed gibbs sampling, we need to integrate out the unknown parameters Φ_u, Φ_p and Θ.

$$\int\limits_{\Theta}\int\limits_{\Phi_u}\int\limits_{\Phi_p} P(u,p,z_u,z_p,\Phi_u,\Phi_p,\Theta|\alpha,\beta_u,\beta_p)\,\mathrm{d}\Phi_p\mathrm{d}\Phi_u\mathrm{d}\Theta = \tag{B.4}$$

Next, we apply YMM's independence assumptions.

$$\int\limits_{\Theta}\int\limits_{\Phi_u}\int\limits_{\Phi_p} P(\Theta|\alpha)\,P(z_p|\Theta)\,P(z_u|\Theta)\,P(p|\Phi_p)\,P(\Phi_p|\beta_p)\,P(u|\Phi_u)\,P(\Phi_u|\beta_u)\,\mathrm{d}\Phi_p\mathrm{d}\Phi_u\mathrm{d}\Theta \tag{B.5}$$

We split the integrals into three independent cases:

Case 1 (Θ):

$$\int\limits_{\Theta} P(\Theta|\alpha)\,P(z_p|\Theta)\,P(z_u|\Theta)\,\mathrm{d}\Theta = \tag{B.6}$$

We use the template to describe the V number of visit records:

$$\begin{aligned}\int\limits_{\Theta} P(\Theta|\alpha)\prod_{v=1}^{V}\left(P(z_{p,v}|\Theta)P(z_{u,v}|\Theta)\right)\mathrm{d}\Theta = \\ \int\limits_{\Theta} P(\Theta|\alpha)\prod_{v=1}^{V}\left(P(z_{p,v},z_{u,v}|\Theta)\right)\mathrm{d}\Theta =\end{aligned} \tag{B.7}$$

Θ is a $K_u x K_p$ dimensional variable. As demonstrated in section **??**, Θ_k can be transformed into $\Theta_{i,j}$ using a bijection function g where $i \in \mathbb{N}_{1:K_u}, j \in \mathbb{N}_{1:K_p}$.

$$\int\limits_{\Theta_k} P(\Theta_k|\alpha)\prod_{v=1}^{V}\left(P\left(z_{p,v},z_{u,v}|\Theta_{i,j}\right)\right)\mathrm{d}\Theta_k = \tag{B.8}$$

We replace the probability distribution with the definition of the Dirichlet distribution with equal number of parameters α as the dimensionality of Θ.

$$\iint\limits_{\Theta_{i,j}} \frac{\Gamma\left(\sum_{k=1}^{K_u x K_p}(\alpha_k)\right)}{\prod_{k=1}^{K_u x K_p}(\Gamma(\alpha_k))}\prod_{k=1}^{K_u x K_p}\left(\Theta_k^{\alpha_k - 1}\right)\prod_{v=1}^{V}\left(\Theta_{i,j}\right)\mathrm{d}\Theta_{i,j} = \tag{B.9}$$

For simplicity, We define the counting function $c(a,b,c,d)$ using an SQL query:

```
Select count(*)
From VisitRecords
Where usertopic=a
      and placetopic=b
      and user=c
      and place=d
```

If *a,b,c* or *d* are equal to *, then the correspondent filter is removed from the *Where* clause.

$$\int\limits_{\Theta_{i,j}} \frac{\Gamma\left(\sum_{k_u,k_p=1}^{K_u x K_p}\left(\alpha_{ku,kp}\right)\right)}{\prod_{k_u,k_p=1}^{K_u x K_p}\left(\Gamma\left(\alpha_{ku,kp}\right)\right)} \prod_{k_u,k_p=1}^{K_u x K_p}\left(\Theta_{k_u,k_p}^{\alpha_{ku,kp}-1}\right) \prod_{k_u,k_p=1}^{K_u x K_p}\left(\Theta_{k_u,k_p}^{c\left(k_p,k_u,*,*\right)}\right) \mathrm{d}\Theta_{k_p} \propto \tag{B.10}$$

Now, we arrange the variables so that we can describe a Dirichlet distribution which will integrate to one. This method has been used to derive the updating topics LDA [1]. After the integration we find the following:

$$\frac{\prod_{k_u,k_p=1}^{K_u x K_p}\left(\Gamma\left(c\left(k_p,k_u,*,*\right)+\alpha_{ku,kp}\right)\right)}{\Gamma\left(\sum_{k_u,k_p=1}^{K_u x K_p}\left(c\left(k_p,k_u,*,*\right)+\alpha_{ku,kp}\right)\right)} \tag{B.11}$$

Case 2 (Φ_u):

Following a similar formulation that the one used for case 1.

$$\int\limits_{\Phi_u} P\left(u|\Phi_u\right) P\left(\Phi_u|\beta_u\right) \mathrm{d}\Phi_u \tag{B.12}$$

$$\prod_{k_u=1}^{K_u}\left(\int\limits_{\Phi_u} P\left(u|\Phi_u\right) P\left(\Phi_u|\beta_u\right) \mathrm{d}\Phi_u\right) \tag{B.13}$$

We obtain:

$$\prod_{k_u=1}^{K_u}\left(\frac{\prod_{v=1}^{V}\left(\Gamma\left(c\left(*,k_u,*,u_v\right)+\beta_{u,v}\right)\right)}{\Gamma\left(\sum_{v=1}^{V}\left(c\left(*,k_u,*,u_v\right)+\beta_{u,v}\right)\right)}\right) \tag{B.14}$$

Case 3 (Φ_p):

[1] http://lingpipe-blog.com/2010/07/13/collapsed-gibbs-sampling-for-lda-bayesian-naive-bayes/

Following the same formulation that the one used for case 2.

$$\int\limits_{\Phi_p} P(p|\Phi_p)\,P(\Phi_p|\beta_p)\,\mathrm{d}\Phi_p = \tag{B.15}$$

$$\prod_{k_p=1}^{K_p}\left(\int\limits_{\Phi_p} P(p|\Phi_p)\,P(\Phi_p|\beta_p)\,\mathrm{d}\Phi_p\right) = \tag{B.16}$$

We obtain:

$$\prod_{k_p=1}^{K_p}\left(\frac{\prod_{v=1}^{V}\left(\Gamma\left(c\left(k_p,*,p_v,*\right)+\beta_{p,v}\right)\right)}{\Gamma\left(\sum_{v=1}^{V}\left(c\left(k_p,*,p_v,*\right)+\beta_{p,v}\right)\right)}\right) \tag{B.17}$$

In Summary:

$$P\left(z_u,z_p,u,p|\alpha,\beta_u,\beta_p\right)\propto$$

$$\begin{aligned}\frac{\prod_{k_u,k_p=1}^{K_uxK_p}\left(\Gamma\left(c\left(k_p,k_u,*,*\right)+\alpha_{ku,kp}\right)\right)}{\Gamma\left(\sum_{k_u,k_p=1}^{K_uxK_p}\left(c\left(k_p,k_u,*,*\right)+\alpha_{ku,kp}\right)\right)}&\prod_{k_u=1}^{K_u}\left(\frac{\prod_{v=1}^{V}\left(\Gamma\left(c\left(*,k_u,*,u_v\right)+\beta_{u,v}\right)\right)}{\Gamma\left(\sum_{v=1}^{V}\left(c\left(*,k_u,*,u_v\right)+\beta_{u,v}\right)\right)}\right)\\&\prod_{k_p=1}^{K_p}\left(\frac{\prod_{v=1}^{V}\left(\Gamma\left(c\left(k_p,*,p_v,*\right)+\beta_{p,v}\right)\right)}{\Gamma\left(\sum_{v=1}^{V}\left(c\left(k_p,*,p_v,*\right)+\beta_{p,v}\right)\right)}\right)\end{aligned} \tag{B.18}$$

B.1.2 Posterior Derivation for each record

Since every visit record is considered to be independent, we can update every single record's user-topic and place-topic instead, we refine our objective probability function to the following:

$$P\left(Z_u^{(v)}, Z_p^{(v)} | Z_u^{(-v)}, Z_p^{(-v)}, u_v, p_v, \alpha, \beta_u, \beta_p\right) \tag{B.19}$$

Where $Z_u^{(v)}, Z_p^{(v)}$ represents the user and place topic for the *v-th* record and $Z_u^{(-v)}, Z_p^{(-v)}$ represent the user and place topic for all other records. Once again, we split equation B.18 into three cases.

Case 1 (Θ): Given $u_v, p_v, Z_u^{(-v)}, Z_p^{(-v)}$:

$$\frac{\prod_{k_u,k_p=1}^{K_u x K_p}\left(\Gamma\left(c\left(k_p,k_u,*,*\right)^{-v}+\alpha_{ku,kp}\right)\right)}{\Gamma\left(\sum_{k_u,k_p=1}^{K_u x K_p}\left(c\left(k_p,k_u,*,*\right)^{-v}+\alpha_{ku,kp}\right)\right)} \tag{B.20}$$

We use the property of the gamma distribution $\Gamma(x+1) = \Gamma(x)\dot{x}$ in order to extract the components that correspond to our *v-th* record of interest.

$$\frac{\prod_{k_u,k_p \neq z_{p,v},z_{u,v}}^{K_u x K_p}\left(\Gamma\left(c\left(k_p,k_u,*,*\right)^{-v}+\alpha_{ku,kp}\right)\Gamma\left(c\left(z_{p,v},z_{u,v},*,*\right)^{-v}+\alpha_{zu_v,zp_v}\right)\right)}{\Gamma\left(1+\sum_{k_u,k_p=1}^{K_u x K_p}\left(c\left(k_p,k_u,*,*\right)^{-v}+\alpha_{ku,kp}\right)\right)}$$
$$\frac{c\left(z_{p,v},z_{u,v},*,*\right)^{-v}+\alpha_{zu_v,zp_v}}{\Gamma\left(1+\sum_{k_u,k_p=1}^{K_u x K_p}\left(c\left(k_p,k_u,*,*\right)^{-v}+\alpha_{ku,kp}\right)\right)} \propto$$

Finally, after simplifying all terms that are not proportional to the topics $Z_u^{(v)}, Z_p^{(v)}$

$$c\left(z_{p,v},z_{u,v},*,*\right)^{-v}+\alpha_{zu_v,zp_v} \tag{B.21}$$

We use the same formulation for the other cases and we obtain the updating equation for $Z_u^{(v)}, Z_p^{(v)}$

$$P\left(Z_u^{(v)}, Z_p^{(v)} | Z_u^{(-v)}, Z_p^{(-v)}, u_v, p_v, \alpha, \beta_u, \beta_p\right) \propto$$

$$\frac{\left(c\left(z_{p(v)}, z_{u(v)}, *, *\right)^{(-v)} + \alpha_{z_{u(v)}, z_{p(v)}}\right)\left(c\left(z_{p(v)}, *, p_{(v)}, *\right)^{(-v)} + \beta_{p(v)}\right)\left(c\left(*, z_{u(v)}, *, u_{(v)}\right)^{(-v)} + \beta_{u(v)}\right)}{\sum_{j=1}^{|P|}\left(c\left(z_{p(v)}, *, p_{(j)}, *\right)^{(-v)} + \beta_{p(j)}\right)\sum_{i=1}^{|U|}\left(c\left(*, z_{u(v)}, *, u_{(i)}\right)^{(-v)} + \beta_{u(i)}\right)} \tag{B.22}$$