GENE FUNCTION INFERENCE FROM EXPRESSION USING PROBABILISTIC TOPIC MODELS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BAHAR TERCAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
MEDICAL INFORMATICS

AUGUST 2016

Approval of the thesis:

GENE FUNCTION INFERENCE FROM EXPRESSION USING PROBABILISTIC TOPIC MODELS

submitted by **BAHAR TERCAN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Medical Informatics, Middle East Technical University** by,

Prof. Dr. Nazife Baykal
Director, **Graduate School of Informatics** ________________

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics, METU** ________________

Assist. Prof. Dr. Aybar Can Acar
Supervisor, **Health Informatics, METU** ________________

Examining Committee Members:

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics, METU ________________

Assist. Prof. Dr. Aybar Can Acar
Health Informatics, METU ________________

Assist. Prof. Dr. Rahime Belen Sağlam
Computer Engineering, Yıldırım Beyazıt University ________________

Assist. Prof. Dr. Murat Perit Çakır
Cognitive Science, METU ________________

Prof. Dr. Hasan Oğul
Computer Engineering, Başkent University ________________

Date: ________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: **Bahar Tercan**

Signature :

ABSTRACT

GENE FUNCTION INFERENCE FROM EXPRESSION USING PROBABILISTIC TOPIC MODELS

Tercan, Bahar
Ph.D., Department of Medical Informatics
Supervisor : Assist. Prof. Dr. Aybar Can Acar

August 2016, 81 pages

The main aim of this study is to develop a probabilistic biclustering approach which can help to elaborate on the question "Can we determine the biological context of a sample (tissue/condition etc.) using expression data and associate the contexts with annotation databases like Gene Ontology, KEGG and HUGE to discover annotations (like cell division, metabolic process, illness etc.) for these contexts?". We applied a nonparametric probabilistic topic model, Hierarchical Dirichlet Process (HDP), which was originally developed for text mining to extract unknown number of latent topics from documents, to gene expression data analysis. In this study, the analogy is the mRNA transcript to the word, the biological context to the topic and the sample to the document. This study builds on previous studies that have, to varying extents, been able to apply topic models to the problem of differential expression, and improves on the current state of the art by producing a comprehensive and integrative method to enhance HDP with prior information. The main areas of proposed improvement are the preprocessing of gene expression data for topic models and the introduction of informed priors to the HDP model. The results of experiments showed that prior improved HDP successfully reveals the hidden biclusters in gene expression data with higher robustness to changes in sparsity levels (number of samples) and prior strengths (η).

Keywords: Expression data analysis, Probabilistic topic models, Hierarchical Dirichlet Process, Prior smoothing, Biclustering

ÖZ

OLASILIKSAL TEMA MODELLERİ KULLANARAK GEN İFADESİNDEN İŞLEV ÇIKARIMI

Tercan, Bahar
Doktora, Tıp Bilişimi Programı
Tez Yöneticisi : Yrd. Doç. Dr. Aybar Can Acar

Ağustos 2016 , 81 sayfa

Bu çalışmanın temel amacı, "İfade verisi kullanarak bir örneğin (doku/durum vb.) biyolojik bağlamını belirleyebilir miyiz ve bu bağlamları Gene Ontology, KEGG, HUGE gibi yorumlama veritabanları ile ilişkilendirebilir miyiz?" sorusuna cevap bulmamıza yardımcı olabilecek olasılıksal bir ikili kümeleme yaklaşımı geliştirmektir. Başlangıçta dökümanlarda bulunan bilinmeyen sayıdaki gizli temaları çıkartmak için geliştirilen ve metin madenciliği metodu olan olasılıksal tema modeli Hiyerarşik Dirichlet Süreci (HDP)'ni gen ifadesi veri analizine uyguladık. Bu çalışmada analoji mRNA transkriptten kelimeye, biyolojik bağlamdan temaya, örnekten dökümanadır. Bu tez çalışması, tema modellerini farklılaşmış ifade problemine belirli bir ölçüde uygulamayı başarmış çalışmaların üzerine inşa edilmiştir ve tema modellerinin gen ifadesi analizinde kullanılması için HDP'yi öncül bilgi ile güçlendirerek kapsamlı ve bütüncül bir metot geliştirilmiştir. Önerilen iyileştirmenin temel alanları, gen ifade verisinin tema modelleri için ön işlemesinin yapılması ve Hiyerarşik Dirichlet Sürecine bilgilendirilmiş öncüllerin eklenmesidir. Sonuçlar, öncül iyileştirilmiş HDP'nin gen ekspresyon verisi içindeki gizli ikili kümeleri seyreklik seviyesi (örnek sayısı) ve öncül gücündeki (η) değişikliklerden etkilenmeden başarılı bir şekilde ortaya çıkardığını göstermiştir.

Anahtar Kelimeler: Ekspresyon veri analizi, Olasılıksal tema modelleri, Hiyerarşik Dirichlet süreci, Öncül düzgünleştirme, İkili kümeleme

ACKNOWLEDGMENTS

First and foremost, I would like to express my special appreciation and thanks to my advisor Dr. Aybar Can ACAR for his meticulous mentorship and endless support. His advices on both research and my career have been priceless.

I would also like to thank my committee members, Dr. Yeşim AYDIN SON, Dr. Murat Perit ÇAKIR, Dr. Hasan OĞUL and Dr. Rahime BELEN SAĞLAM for serving as my committee members and their insightful comments. I also thank Dr. Tuğba TAŞKAYA TEMİZEL for her contribution on the preparation of this thesis.

I thank to Dr. Resul KARA for his continuous support before and during my PhD study.

I also thank to my university, METU, for all the facilities it provided.

Special thanks go to my family members, who have always supported me in every aspect of my life. I appreciate my parents who have tolerated my absence during my PhD study and have wished and done best for me in my whole life. I would also like to thank all of my friends for their fellow feelings. At the end I would like to express appreciation to my beloved husband Mehmet TERCAN for tolerating my long work hours and stressful moments.

TABLE OF CONTENTS

CHAPTERS

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaiki Information Criterion
CRF	Chinese Restaurant Franchise
CRP	Chinese Restaurant Process
DNA	Deoxyribonucleic Acid
cDNA	Complementary DNA
DP	Dirichlet Process
GO	Gene Ontology
GSEA	Geneset Enrichment Analysis
HDP	Hierarchical Dirichlet Process
HuGENet	Human Genome of Epidemiology Network
KEGG	Kyoto Encyclopedia of Genes and Genomes
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MCMC	Markov Chain Monte Carlo
MEA	Modular Enrichment Analysis
miRNA	Mikro RNA
mRNA	Messenger RNA
pLSA	Probabilistic Latent Semantic Analysis
RMA	Robust Multiarray Analysis
RNA	Ribonucleic Acid
SEA	Singular Enrichment Analysis

CHAPTER 1

INTRODUCTION

Transcriptomics is the study area that aims to understand gene activity by measuring messenger Ribonucleic Acids (mRNAs) within a cell or organism [1]. By examining the transcriptome, total mRNA in a cell or organism, researchers can find out turned on/turned off genes at a given time in a cell and this examination provides information about gene's effect on a particular phenotype. It is possible to know which gene(s) to interfere with for curation of complex diseases such as cancer with this information. This merit of gene expression analysis is facilitated in drug discovery and this is just one of the usages of transcriptome data. Transcriptome data can also be used for class prediction, class discovery, pathway analysis, biomarker detection, development of prognostic tests, and disease-subclass determination, etc. [2].

The reason why transriptome is used for gene activity measurement is that gene regulation mostly occurs at transcription that is when Deoxyribonucleic Acid (DNA) is transfered into mRNA.

Although every cell in an organism has the same DNA, different cells have different gene expression profiles.This is differential gene expression and it is preceeded by gene regulation. Differential gene expression is responsible for the differences among tissue types. It causes brain cell to be different from liver cell and as a consequence, brain functions differently than liver.

Some of the differences between cancerous and normal cells can also be attributable to gene regulation. Besides many attributes differentiate the onco and normal cell like ability to metastasize, difference in appearance etc, we want to give an example of the cell death difference in onco and normal cells. The inactivation of tumor supressor genes like p53 will cause the cell not to undergo apoptosis (programmed cell death) and will continue to reproduce despite being too old or damaged, with increasing mutations causing malignant tumors [3].

The differential expression of a gene is assumed to give clues about existence of biological conditions in which the gene is known to take role in. Visa versa is also valid; once a gene is discovered and its function is not known clearly yet, the tissue or biological condition in which it is differentially expressed gives information about the gene's function.

Besides investigating single gene effects, we know that biological activity is usually carried out by a coordination of several genes and our aim is to find gene expression patterns seen in a cellular activity or biological function, we need to find the group of genes expressed in a correlated way and we can cluster genes according to their expression levels under different conditions (samples) for this purpose. Samples can also be clustered according to their expression profiles to find out similarities between conditions. Traditional clustering algorithms like k-means, SOM, etc. can perform clustering on all the feature set, that is sample profiles of all genes for sample clustering and gene expression profiles of all samples for gene clustering. This process can be visualized as in Figure 1.1.

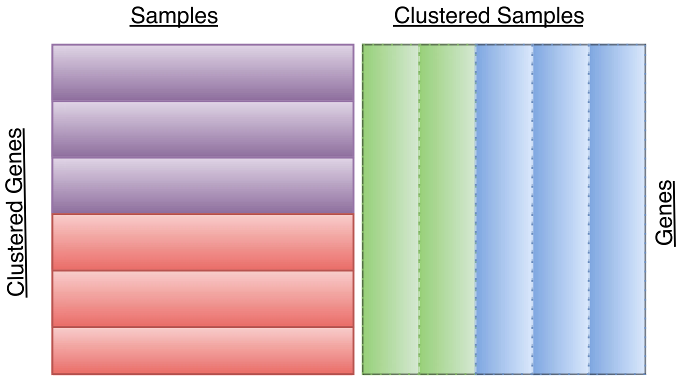


Figure 1.1: Traditional clustering of gene expression matrix, gene clustering on the left and sample clustering on the right side.

Traditional clustering approach has some shortcomings like restricting samples or genes to a single cluster, and causing noisy genes to join a cluster causing deteoriation of the clustering process. There is a study [4] which can get rid of the obligation to cluster a gene into a single cluster. By this study, genes can be clustered into multiple or none clusters. To us, apart from these shortcomings, the most important limitation of traditional clustering algorithms is their global modeling. Global modeling does not always suffice to model gene behavior. A group of genes may be differentially expressed in some conditions and they may be totally uncorrelated in other conditions due to being co-expressed in a cellular process which is active under only a subset of conditions.

There is an obvious need for local modeling. Local patterns (correlated expression of a subset of genes in a subset of conditions) in gene expresssion data can be found by clustering genes and samples simultaneously. This clustering approach is called biclustering [5, 6].

1.1 Motivation

There are different biclustering approaches. Traditional biclustering algorithms like Chen and Church's algorithm [5] and Spectral Biclustering [7] can extract biclusters in gene expression data but the extracted biclusters do not always model biological reality. The biclusters found by these algorithms are binary and exclusive that is a sample or a transcript can belong to one or no bicluster as can be seen on left side of Figure 1.2.

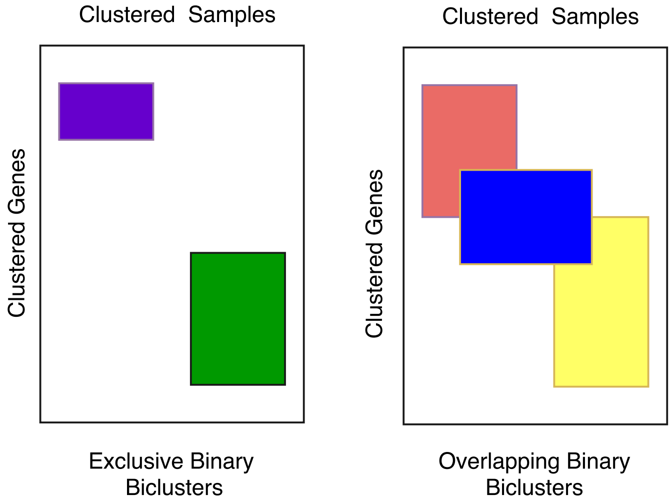


Figure 1.2: Traditional biclustering of gene expression matrix

The aforementioned biclustering approach underestimates the complexity in the nature of gene expression data. A sample is not limited to a single biological event, samples have a mixture of these events where each of these biological events can be seen as a differential expression of a group of genes. A very similar condition applies to the gene side of the biclustering process. Genes (especially regulatory genes) differentially express due to different roles in many different contexts for different reasons. Transcripts are "polysemic" or "context-sensitive" in other words. This "context sensitiviy" is mentioned in several publications. A few examples of these publications are on miRNAs. Gabriely et. al. [8], have stated that miRNA-10b acts differently in different cancer types, for example it fosters metastasis in breast cancer but apoptosis in glioblastoma. Blenkiron et al. [9] have reported that overlapping subsets of a group of miRNAs have context-specific roles in different types of breast cancer (luminal, basal). Zhou et al. [10] have used context-specific miRNA activity as feature space for SVM classifier and achieved much more accurate prognosis prediction on breast and brain cancer than using feature space on mRNA expression.

The context-sensitivity in gene expression data requires it to be handled with overlapping biclusters where a gene or a sample can belong to more than one bicluster. This approach can be seen on the right side of Figure 1.2. PLAID model [11], FLOC model [12] and ISA model [13] are some examples for methods that can find overlapping biclusters.

There is even a better solution, context-sensitivity can be represented by soft biclusters, letting genes and samples have membership to different biclusters with different degrees. In order to achieve soft biclustering, we propose a model where samples are mixtures of biclusters and biclusters are mixtures of gene expressions. Namely, a bicluster is a probability distribution over transcripts and a sample is a probability distribution over biclusters. This model is called *Bayesian Mixed-membership Model* [14]. Bayesian Mixed-membership Models are commonly used in text mining and their domain-specific name is "Probabilistic Topic Models" [15]. Probabilistic Latent Semantic Analysis (PLSA) [16], Latent Dirichlet Allocation (LDA) [17], and the Hierarchical Dirichlet Process (HDP) [18] are among the most commonly used topic models.

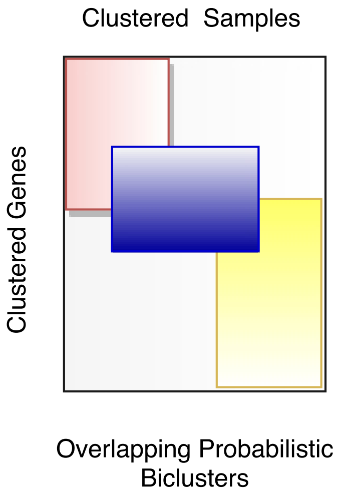


Figure 1.3: Our biclustering approach

In the text domain, a document is a mixture of words and a corpus is a mixture of documents. A document is treated as being "bag of words", this means the order of words is not important in a document and also the order of documents is not important in a corpus. This case is very similar to the problem at hand. The order of samples is not important in gene expression data and the order of transcripts in a sample is again of no importance. The "bag of words" assumption fits our problem even better than text domain since the order of words may be important in phrases but a sample is exactly a "bag of transcripts".

The topics in topic models can be considered as mixtures of semantically related words. Indeed, each topic is a probability distribution over word types in the entire corpus. We can sort the words based on their probability values in a topic in descending order and set a threshold on either number of words to include or a cumulative probability distribution of words (such as top 10 words or the top words that explain 50 percent of the topic), we can cut at the threshold and use the top words as the topic with or without their probability values. If we include the probability values, this will give us the weight of each word in a topic. In our case, a topic is a mixture of biologically related transcripts. Hence, the topics in mixed-membership models are analogues to soft biclusters that we are seeking. In our analogy, the topics can be functional modules of gene products, in other words biological contexts. Samples are mixtures of these biological contexts. We are going to call biological contexts as topics to be consistent with the literature throughout this thesis.

Different subsets of genes may over- or under- express in different topics and a certain gene may be significant in multiple contexts. All probabilistic topic models can handle this context sensitivity issue but parametric topic models like PLSI and LDA need many runs and model selection to find the number of topics that fits the data best. The number of topics has to be set beforehand and it is not easy to know the number of biological contexts active in a set of samples.

The HDP model is nonparametric and it infers both the topic distributions and number of topics from data. HDP assumes an infinite number of topics but concretizes a finite number of them. We worked on HDP not to be forced to estimate number of topics before running the algorithm.

The results of this study are expected to have beneficial impact on the study of the differential expression of highly context sensitive genes and gene products such as microRNAs, transcription factors and other genes with high centrality in cellular processes. As such genes typically regulate cell processes and have high impact in disorders like cancer; we hope that this study will indirectly benefit the research in these areas.

1.2 Problem Statement

The crux of our study is to find out the situation of a sample by biclustering gene expression data. We achieve this using the workflow which can be seen in Figure 1.4.

In the system, first gene expression data is converted into a format that is applicable to topic model which is originally a text mining method. Since we are not just using the method as is, we also prepare the prior information in the pre-processing step.

The simplified graphical representation of the Probabilistic Biclustering (Topic Model) Mechanism we propose can be seen in Figure 1.5, please see Figure 3.1(d) for the detailed representation.

In Figure 1.5, S is the sample, G is the gene and z is the topic assignment, M is the number of samples and N_j is the number of observations in sample j. Shaded variables are observed, unshaded variable is latent to be inferred during biclustering process. The number of topics is inferred by the topic model as well.

After running the probabilistic biclustering algorithm, we have two distributions as its outputs. First is sample-topic, $P(topic|sample)$, and the second is gene-topic, $P(gene|topic)$, distribution. Note that the topics are the pivot elements between samples and genes, they provide the connection between genes and samples. We are also interested in annotation-topic distribution, $P(annotation|topic)$, which can be inferred from annotation databases by using $P(annotation|gene)$ of most representative genes of each topic with geneset enrichment.

The outputs of the biclustering can be used in several ways. First usage utilizes all of the distributions mentioned above to label samples, we can annotate topics by gene set enrichment of the top genes of each topic. Thus, we have the biological meaning of topics. The topics which are dominant in each sample, top topics of each sample, can explain the sample's situation (cancer etc.), in turn.

On the other hand, genes' features are topics and samples' features are topics, again. Unlike using one side of the gene expression matrix as features of the other side, the features which are composed of topics are local modeled. That is, if samples and genes are classified or clustered according to this new feature set, it is possible to recover gene expression similarities due to their correlated expression in only subset of samples and sample similarities due to their gene expression patterns in a subset of genes and this alleviates the problem caused by global modeling approach of traditional clustering

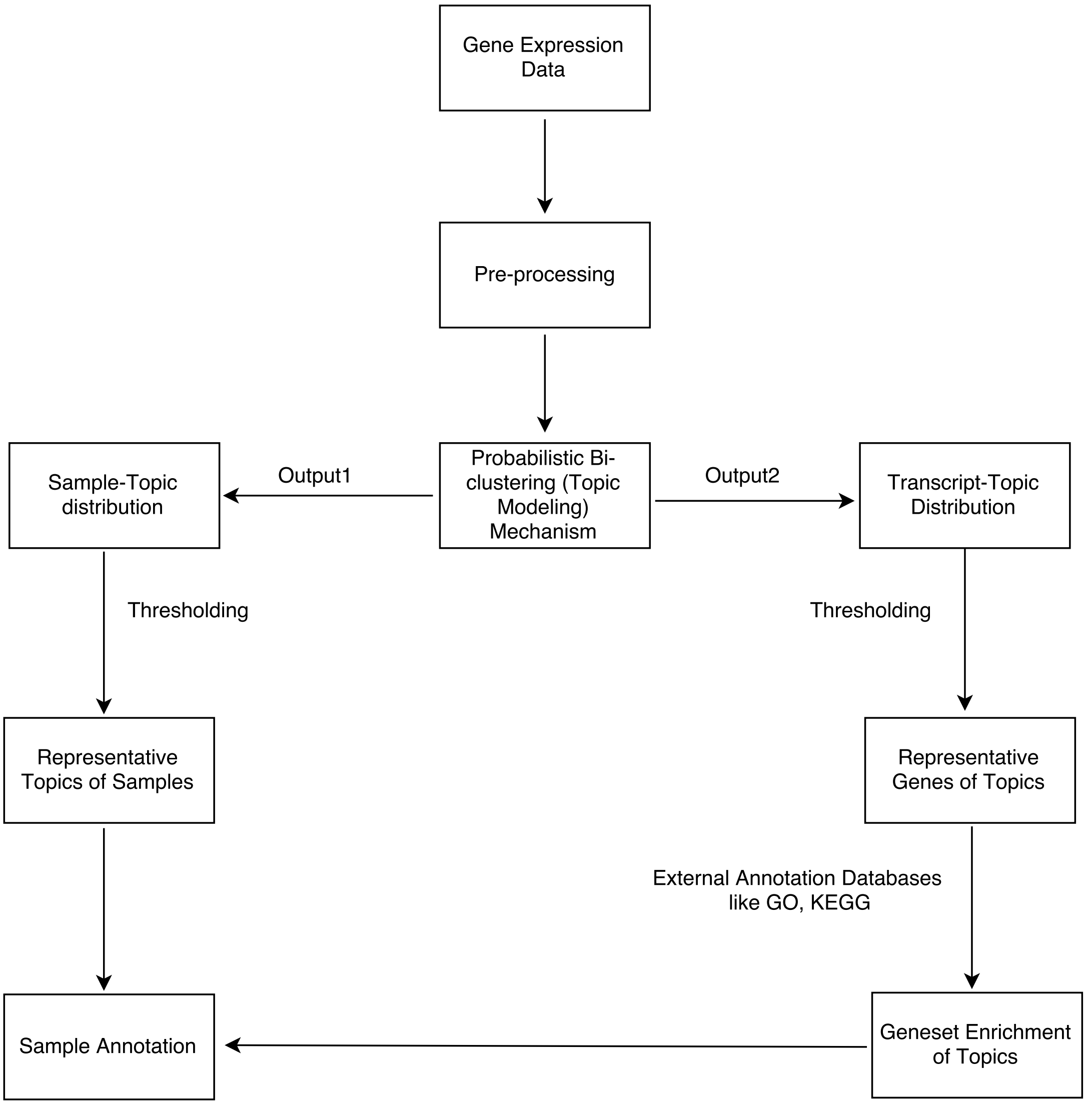

Gene Expression Data
Pre-processing
Probabilistic Bi-clustering (Topic Modeling) Mechanism
Output1
Output2
Sample-Topic distribution
Transcript-Topic Distribution
Thresholding
Thresholding
Representative Topics of Samples
Representative Genes of Topics
External Annotation Databases like GO, KEGG
Sample Annotation
Geneset Enrichment of Topics

Figure 1.4: Workflow of the system

algorithms. We do not suffer from the disadvantage of traditional clustering/classification algorithms, although we use them. Especially in sample clustering/classification, we alleviate the problem of curse of dimensionality because we will be using tens of topics instead of thousands of genes.

For another usage of topic models in bioinformatics, the top genes in a topic can be used for gene regulatory network construction or module detection. This can be achieved by combining the gene-topic distribution output of topic models with the information about transcription binding sites and tf-gene interaction, to our best knowledge there is no such a study.

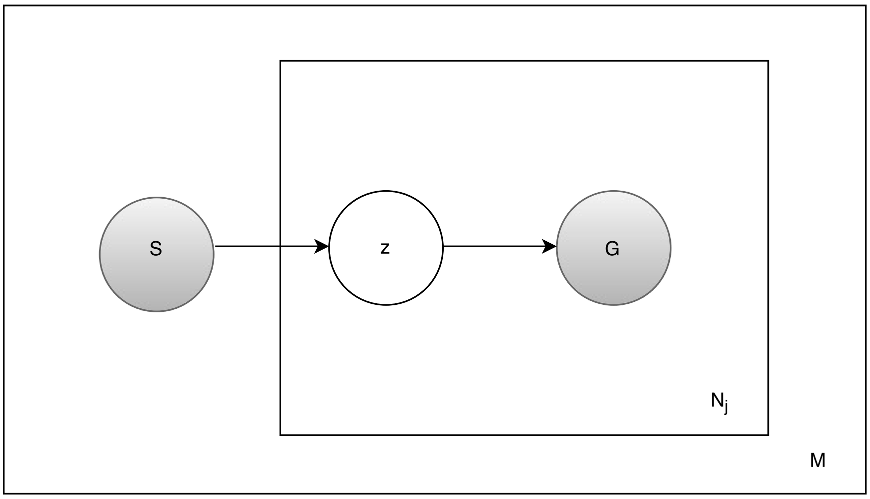


Figure 1.5: The Plate Model of Probabilistic Topic Model

1.3 Contribution

The main contribution proposed in this study is the use of nonparametric topic model, Hierarchical Dirichlet Process (HDP), in biclustering microarray data while taking account the prior information. The prior information we have incorporated to standard HDP is taken from either an external gene regulatory network or co-expression information calculated over the correlation of gene expression matrix.

HDP is a nonparametric Bayesian model. Bayesian models have prior belief about their parameter distributions and update this belief with observations. If the number of observations is enough to represent the tendency in data, posterior parameter distribution is sound. If the number of observations is small compared to number of parameters ($n << p$), the posterior won't be well-defined. Starting from an accurate prior belief enables the model to work on a smaller space configurations and this gives better results with less number of observations.

The previous applications of HDP in gene expression data analysis have used the method as is with respect to prior distribution. Standard HDP assumes flat prior distribution over transcript-topic distribution. So the previous applications do not take into account gene co-expression or co-regulation information in their prior transcript-topic distribution. These studies will be summarized in Section 2.8. In this study, we proved that prior informed (using this information) HDP can mitigate data sparsity problem and also the model becomes robust to hyperparameter changes. Incorporating informed prior into HDP enables modeling mixed-memberships on sparse data more successfully. Our model works without necessity to specify the number of topics in advance, this is an inherent attribute of HDP and left intact in our model.

In transcriptomics studies, number of genes is in the thousands and number of samples is generally in the tens. The standard approach used in order to evade curse of dimensionality problem, genes that are not differentially expressed or having high correlation to each other can be removed with gene selection methods, the rest of analysis can be done with the remaining distinguishing genes. The removal of genes that are not differentially expressed is handled through our preprocesing approach defined in Section 3.1.2.1.

1.4 Thesis Organization

The rest of this thesis is organized as follows: In Chapter 2, we gave background information about microarray and RNA-seq data which are valid input to our preprocessing for topic models. We mentioned different topic models: unigram, mixture of unigrams, PLSI, LDA; nonparametric Bayesian models, Dirichlet processes and HDP and also two metrics used in topic model evaluation, perplexity and topic coherence. We explained gene set enrichment analysis, hypergeometric distribution of genes and gene set enrichment analysis tools. In Section 2.8, we summarized previous work on the application of topic models to gene expression and the incorporation of priors to topic models.

In Chapter 3, we described our model "Externally Smoothed HDP" with comparison to previous mixed membership models, Probabilistic Latent Semantic Indexing (PLSI), Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP). We explained the preprocessing approach we adapted to be able to use topic models for gene expression data analysis. We proposed two methods for encoding the priors: "Co-expression smoothing" and "Network-based smoothing". "Co-expression smoothing" uses the correlation of data and "Network-based smoothing" uses external gene-gene interaction networks as prior information. We proved that both smoothing approaches enhance success of original HDP algorithm in gene expression data analysis.

In Chapter 4, we mentioned several experiments which were carried out using HDP and Smoothed HDP algorithms. First experiment was performed to establish the performance of HDP in finding pre-seeded biclusters. HDP algorithm successfully recovered the biclusters in two dimensional datasets of 2, 4 and 8 biclusters. Second experiment was performed to make sure prior smoothing works as we expect. We tested HDP and Smoothed HDP algorithms on a test platform which was originally created by Yee Whye Teh [19] to test HDP algorithm. The results showed that Smoothed HDP algorithm finds topic distributions over words and number of topics much more successfully than original HDP algorithm in every sparsity level (number of documents) and prior strength (η) value combination. In the third experiment, we performed quantitive evaluation of Smoothed HDP algorithm using the model comparison metric Akaike Information Criterion (AIC) which provides a trade off between model complexity and likelihood of model given data. According to results of AIC experiment, Smoothed HDP algorithm is of more quality compared to HDP algorithm under every sparsity level (number of documents) and prior strength (η) value combination. The fourth experiment was performed on a semi-synthetic dataset generated using Syntren [20]. In this experiment, we used a yeast transcriptional network, we first fully perturbed some of the hub genes individually to know the gene expression profile generated owing to each perturbation. Resultant gene expression profile of perturbation of each hub gene fully was regarded as a topic. We perturbed different hub genes at different levels to generate gene expression data of each sample. We then ran HDP, Co-expression Smoothing HDP and Network-based Smooothing HDP algorithms and tried to find the latent topics we had seeded in the experiment set. Both Co-expression Smoothing

HDP and Network-based Smoothing HDP revealed the latent topics more successfully than HDP under every sparsity level (number of samples in this experiment) and prior strength(η) combination. In the fifth and last experiment, we worked on a dataset from prostate cancer study by Dhanasekaran et. al. [21]. In this experiment, we evaluated comparative success of HDP and Co-expression smoothing HDP in two metrics. First is dependent on sample-topic distribution and sample labels. We compared each sample's label with that of the sample which is most similar to it according to its topic distribution. Second is based on the topic coherence metric given in Section 2.5, this metric measures how often the top words of each topic are seen in the same documents in the original dataset, its implicit assumption is that if a group of words are seen together in the same documents in the original corpus, it means that they are related, so they should be found in the same topic. In both evaluation approaches, Co-expression Smoothing HDP provided more successful and consistent results over different η values.

In Chapter 5, we discussed our approach, contribution and the results of experiments, commented on future directions and concluded the thesis.

CHAPTER 2

BACKGROUND

In this chapter, I give brief background information to enable readers to familiarize themselves with the materials covered in the rest of this thesis study. We begin with basic biology information in Section 2.1 and continue with the data types, microarray data and RNA-Seq data in Section 2.2. In Section 2.3, we provide introductory information about probabilistic topic models and list the probabilistic topic models in evolutionary order: unigram, mixture of unigrams, PLSA and LDA. In Section 2.4, we point out the differences between frequentist and Bayesian approaches and parametric and nonparametric Bayesian methods. We touch upon properties of the Dirichlet Processes, different metaphors for constructing the Dirichlet Processes and Hierarchical Dirichlet Process. In Section 2.5, we address the post evaluation issue of topic models and mention a topic coherence metric defined by Mimno et.al [22]. In Section 2.6, we give definition and formula of perplexity which is a measure for prediction power of probabilistic models. In Section 2.8, we refer to the previous studies on both mixed membership model usage in transcriptomics studies and informed priors in text domain.

2.1 Background Biology

The cell is the smallest unit in an organism and it contains all genetic information of the organism in it. The genetic information in the cell is stored in a nucleic acid type called deoxyribonucleic acid (DNA). The gene is a segment of the DNA and it contains necessary information to create functional structures, that are proteins. The other nucleic acid is the ribonucleic acid (RNA) and a special type of RNAs, messenger RNA (mRNA), maintains the flow of information in protein synthesis. First, the gene on DNA is transcribed into mRNA and mRNA is translated into building blocks of protein, that are amino acids. The flow of information can be summarized as follows [23]:

DNA → mRNA → Amino acid → Protein → Phenotype (cell) → Phenotype (sample)

Samples are the biological materials like tissues and phenotypes are the visible characteristics like cancer, non-cancer.

If the gene is transcribed into its mRNA, it means that it is expressed and the transcription level is its gene expression and can be measured by microaarray or RNA-seq technology. The importance of measuring the amount of mRNA is that it gives information about the amount of protein and proteins' functions determine the phenotype.

We call mature mRNAs (containing only exons) mRNA transcripts. Gene and transcript terms are used interchangeably throughout this thesis study.

2.2 Data Sources

The genome is the blueprint of all cellular processes and activities in a living thing. Although every cell contains a copy of whole genome, not all the genes are expressed equally in each cell every time. This is called differential gene expression [24]. We are interested in differential gene expression because a successful understanding of gene expression will lead us to understand cell function and pathology. Gene expression levels can be measured via mRNA amounts and mRNAs are captured with microarray chips in microarray technology. There are two main types of microarray chips: first type is spotted or cDNA microarrays and second type is oligonucleotide chips [25]. In spotted or cDNA microarrays, a probe is a complementary copy of original DNA and corresponds to one gene. Two classes of tissues (for example, healthy vs. cancer) are dyed with different colors and they compete to hybridize with probes. In oligonucleotide chips, a gene is represented by a probe set. In this technology, a sample is hybridized on one chip.

In microarray technologies, raw microarray data are scanned as images, fluorescence readings from these images are transformed into mRNA expression values [26]. The datasets used in our experiments come from oligonucleotide chips. One of the methods for normalization array images into mRNA values for data retrieved with oligonucleotide chips is Robust Multiarray Analysis (RMA) [27]. RMA is a technique that consists of background correction, normalization across arrays, probe level intensity calculation and probe set summarization. After this method is applied to raw data, the gene expression matrix where the mRNA expression values are stored is obtained. In a gene expression matrix, rows represent genes and columns represent samples (like tissues, experimental conditions), hence each cell represents the expression level of a particular gene in a particular sample [28]. Genes can be clustered according to their expression levels under different conditions (samples) especially for discovery of regulatory motifs and conditions (samples) can be clustered according to their expression profiles to find out condition similarities [5]. Analysis of local expression patterns in gene expression matrix is also essential because genes may co-express under a subset of conditions and be independent in other conditions; this simultenous clustering of genes and conditions is called biclustering of microarray expression data [6].

Before analyzing the gene expression matrix, it should be cleaned from genes exhibiting little variation across samples (like house-keeping genes). This is especially important in topic modeling because house keeping genes play a similar role to stop words in text mining. If this process is skipped, the prevelant genes across the experiment will dominate all topics and the differences among topics will be obscured.

An alternative to microarray technology is high-throughput sequencing of cDNA (RNA-Seq). RNA-Seq counts the number of discrete sequence reads while hybridization-based array methods (microarray) measure continuous probe intensities [29]. Raw RNA-Seq data is usually in FASTQ format. It contains an ID number for each read, the read sequence, and a quality score [30]. Low quality reads are removed and rest are mapped to a reference genome. After splice junction detection and gene/isoform expression quantification, differential expression analysis is performed [31].

RNA-Seq data has many advantages over microarray expression data. Both RNA-Seq and microarray data can detect differentially expressed genes but only RNA-Seq data can detect abundances of alternative isoforms. RNA-Seq data favors larger dynamic range and less background and technical variation [32]. Microarray data is usually used for comparing the same gene across multiple samples/conditions but not expression levels of different genes in a single sample because of cross-hybridization effects on probe intensities. RNA-Seq data makes it possible to do such analyses. RNA-Sequencing is possible for any organism while microarray platforms are only available for model organisms [33].

Besides these advantages, RNA-Seq data has its own limitations. The reads are not uniform along genome, more reads are mapped to longer genes and there is an artificial correlation between differential expression and gene length; this effects within sample analysis. Reads Per Kilobase of transcript per Million (RPKM) and Fragments Per Kilobase of transcript per Million (FPKM) are used methods for normalizing expression of genes with different length within a sample. Dependence to sequencing depths and library sizes effects comparison among samples and different normalization algorithms like Trimmed Mean of M-values (TMM) [34] is used to make the same genes in different samples comparable [32].

2.3 Probabilistic Topic Models

Probabilistic topic models automatically extract hidden topics from document sets, in other terms, corpora. They achieve this by considering a topic as a probability distribution over words and a document as a mixture of topics [35]. These models define a joint probability distribution on both latent and observed variables; conditional distributions of hidden variables are calculated given the observed ones [15].

Probabilistic topic models are both generative and discriminative. As a generative model, a document can be composed by sampling words from topics according to weight given to each topic; and as a discriminative model, they can be used for statistical inference of topics that have generated the observed words in corpora.

The probabilistic topic models from the most naive one, the unigram model, up to the Latent Dirichlet Allocation are explained below:

2.3.1 Unigram

In this model, the "bag of words" assumption is essential as it is in all other topic models. For every document $d_{1..M}$ in the corpus, its observations $w_{1..N}$ are sampled independently from a single multinomial [17]. The graphical representation of the unigram model can be seen in Figure 2.1.

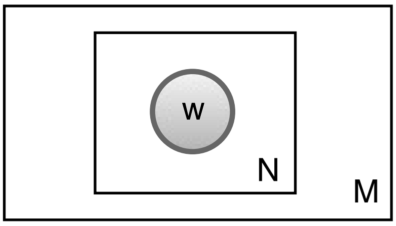


Figure 2.1: The Plate Diagram of the Unigram Model

The probability of a document can be calculated as follows:

$$P(w) = \prod_{n=1}^{N} P(w_n) \tag{2.1}$$

Unigram model can also be explained in a geometrical perspective. Like all topic models, unigram acts in the space of distributions over words. Each distribution is a point on the $(V - 1)$ simplex where V is the number of word types and this simplex is known as word simplex. The unigram model selects a single point on the word simplex and assumes that all the observations in the corpus originate from this distribution [17].

2.3.2 Mixture of Unigrams

This model is an extension of the unigram model [17] with a latent variable topic z. In this model; for each document, first a topic z is chosen and words $w_{1...N}$ are sampled from this topic. The graphical representation of the mixture of unigrams model can be seen in Fig. 2.2.

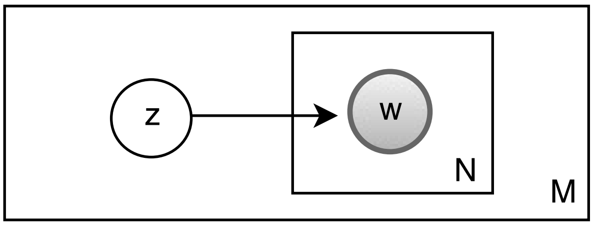


Figure 2.2: The Plate Diagram of the Mixture of Unigrams Model

The probability of a document can be calculated as follows:

$$P(w) = \sum_{\mathbf{z}} P(z) \prod_{n=1}^{N} P(w_n|z) \tag{2.2}$$

This model assumes that a document can be related to only one topic and it is unrealistic to assume that each document is relevant to a single topic.

From a geometrical point of view, the mixture of unigrams model selects one of the points on the word simplex for each document and observations of the document originate from this point, in other words, each observation of the document is drawn according to this distribution [17].

2.3.3 Probabilistic Latent Semantic Analysis (pLSA)

Latent Semantic Analysis (LSA) is a method used for data reduction in text mining [36, 37]. Let a term-document matrix have terms in rows and documents in columns. Each cell consists of the number of occurance of the row-indexed word in the column-indexed document, we can find term similarities by correlation of rows and document similarities by correlation of columns. The number of features of a document is the number of terms, and the number of features of a term is the number of documents. This is a very sparse matrix and LSA offers to reduce the number of features of both terms and samples by using the Singular Value Decomposition method. This method breaks the term-document matrix A into linearly independent components:

$$A = USV^T \tag{2.3}$$

where A is the term-document matrix, U is the matrix whose columns are orthonormal eigenvectors of matrix AA^T, S is the diagonal matrix having square roots of eigenvalues of U or T in descending order, V is the matrix whose columns are orthonormal eigenvectors of matrix A^TA.

In this equation, the first few columns of U and V carry information that accounts for the most variation in the data. We can select a number of eigenvalues beginning from the largest one and the corresponding eigenvectors in the columns of the matrices U and V. The number of features of words and documents become this value, hence the data reduction is realized.

The word and document are represented as rows of reduced form of U and V respectively and their similarities can be computed over this matrices instead of original term-document matrix.

This method is effective in data reduction and also noise filtering but it has limitations. The selection of the number of features is arbitrary and this method can not handle polysemy which is very common in text domain.

This limitations are overcome by pLSA model. pLSA has a statistical foundation, model selection is possible and this attribute can alliveate the problem of choosing the number of latent factors randomly. In pLSA model, each observation in a document is sampled from a topic, which can be viewed as a multinomial random variable. Each observation is generated from a single topic, and different observations in a document may be generated from different topics and pLSA can handle polysemy. A document can be viewed as a mixture of topics with a weight given to each topic [16]. The graphical representation of the pLSA model can be seen in Fig. 2.3.

In this model, document d and word w_n are conditionally independent given latent variable z.

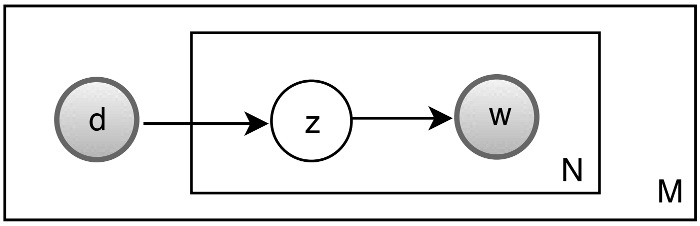


Figure 2.3: The Plate Diagram of pLSA where d represents the document, z represents the topic and w represents the word. N is the number of words and M is the number of documents.

The joint probability of document d and word w_n can be calculated as follows:

$$P(d, w_n) = P(d) \sum_{z_i \in \mathbf{Z}} P(z = z_i | d) P(w_n | z = z_i) \tag{2.4}$$

The pLSA model allows a document to represent a mixture of topics but it has two serious limitations which were later overcome by the LDA model. First limitation is that pLSA learns topic proportions $p(z|d)$ only for the documents in the training set and does not provide generalization to unseen documents. The second limitation is the number of parameters, $kV + kM$ (k is the number of topics, V vocabulary size, M is number of documents), grows linearly with the number of documents, which causes overfitting [17,38].

If we turn back to the geometric interpretation, we now have a new definition: topic simplex. A sub - simplex is built on the word simplex by selecting k points where k is the number of topics, the sub-simplex formed by these k points is called the topic simplex. For each document, pLSA finds a document specific distribution over topics and this means a point on the topic simplex. pLSA uses Expectation-Maximization approach to find the distributions that maximize the parameters $P(z)$, $P(w|z)$ and $P(z|d)$.

LDA instead relates a document with a k-parameter hidden variable and builds a Dirichlet distribution over it. This enables LDA to have distribution for both training and unobserved documents. LDA is a truly generative model. LDA uses $k + kV$ parameters, the number of parameters do not grow with the number of documents and LDA does not suffer from overfitting.

The distribution used over LDA's word-topic and topic-document distributions is the Dirichlet distribution and it can be defined as follows [39, 40]: Random variables $X_1, X_2, \ldots, X_r$ have a Dirichlet distribution if they have a density function with parameters $\alpha_1, \alpha_2, \ldots, \alpha_r$ and $N = \sum_{k=1}^{r} \alpha_k$:

$$p(x_1, x_2, \ldots, x_r) \sim Dir(\alpha_1, \alpha_2, \ldots, \alpha_r) = \frac{\Gamma(N)}{\prod_{k=1}^{r} \Gamma(\alpha_k)} \prod_{k=1}^{r} x_k^{\alpha_k - 1} \tag{2.5}$$

where $0 \leq x_k \leq 1$ and $\sum_{k=1}^{r} x_k = 1$

$$E(X_k) = \frac{\alpha_k}{N} \tag{2.6}$$

Let X be a random variable with event space $1, 2, \ldots, r$,

$$P(X = k) = E(X_k)$$

Dirichlet distribution is a probability distribution over probability distributions and a draw from a Dirichlet distribution is a probability distribution.

2.3.4 Latent Dirichlet Allocation (LDA)

LDA clusters words into topics and documents into mixtures of topics just like pLSA. This is, in fact, a three level hierarchical Bayesian model where each document is associated with a probability distribution over topics and each topic is a probability distribution over words. The probability of whole words in a topic sum up to one and the topic ratios of a document are, again, additive.

The plate representation of LDA can be seen in Figure 2.4. In this figure, α and β are hyper parameters on θ and ϕ . θ is the per-document topic distribution, ϕ is the per-topic word distribution and z is the topic assignment of each observation to be estimated.

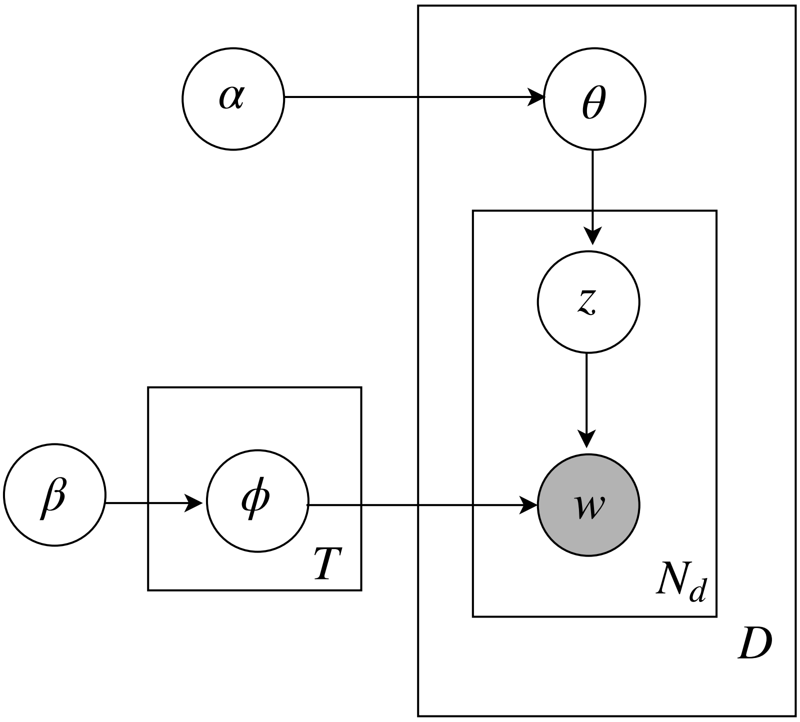


Figure 2.4: The Plate Diagram of LDA where D is number of documents, N_d is the number of tokens in a document, T is the number of topics

When using topic models as generative model, θ_d is a document level variable and sampled once per document, ϕ_k is a topic level parameter and sampled once per topic. $z_{d,n}$ and $w_{d,n}$ are observation level variables and sampled once for each observation in each document.

The generative process of LDA can be summarized as follows [41]:

1. For each topic, draw a distribution over words: $\phi_k \sim Dir(\beta)$
2. For each document, draw a distribution over topics, $\theta_d \sim Dir(\alpha)$
3. For each observation of each document,
   (a) Draw a topic assignment $z_{d,n} \sim Mult(\theta_d)$ where $z_{d,n} \in 1, ..., K$.
   (b) Draw a word $w_{d,n} \sim Mult(\phi z_{d,n})$ where $w_{d,n} \in 1, ..., V$.

The joint probability of θ, ϕ, w and z given α and β is calculated as follows:

$$p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) = p(\theta|\alpha) \prod_{j=1}^{T} p(\phi_j|\beta) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|\phi_{z_n}) \tag{2.7}$$

The exact posterior distribution of LDA is not tractable and there are variational, Laplacian and sampling based approximation methods for inference in LDA [17]. In this thesis study we used a sampling-based method "Gibbs sampling". By applying Gibbs Sampling to the LDA model, we are seeking the conditional probability $p(z_{(i,j)}|z_{\neg(i,j)}, w, \alpha, \beta)$ where $z_{(i,j)}$ is the topic assignment for the j^{th} word of the i^{th} document and $z_{\neg(i,j)}$ is the topic assignment of every observation in the corpus except the current observation-topic assignment pair. Therefore for each topic k,

$$p(z_{(i,j)} = k|z_{\neg(i,j)}, w, \alpha, \beta) \propto \frac{\left(n_{\neg(i,j)}^{(w_i)} + \beta\right)}{\left(n_{\neg(i,j)}^{(\cdot)} + V\beta\right)} \frac{\left(n_{\neg(i,j)}^{(d_i)} + \alpha\right)}{\left(n_{\neg(i,.)}^{(d_i)} + T\alpha\right)} \tag{2.8}$$

where $n_{\neg(i,j)}^{(w_i)}$ count for word type w_i assigned to topic k, $n_{\neg(i,j)}^{(\cdot)}$ is the total number of observations assigned to topic k, $n_{\neg(i,j)}^{(d_i)}$ is the number of observations assigned to topic k in document d_i, $n_{\neg(i,.)}^{(d_i)}$ is the document size, all not including the current observation w_i, V is the corpus size and T is the number of topics.

This formula is calculated for every observation of each document iteratively until it reaches a stable state. The other latent variables $\theta_d^{(k)}$, the topic-document distribution and $\phi_k^{(w)}$, the topic-word distribution are calculated from z as follows:

$$\hat{\theta}_d^{(k)} = \frac{n_k^{(d)} + \alpha}{n_{(\cdot)}^{d} + T\alpha} \tag{2.9}$$

$$\hat{\phi}_k^{(w)} = \frac{n_w^{(k)} + \beta}{n_{(\cdot)}^{k} + V\beta} \tag{2.10}$$

where $n_k^{(d)}$ is the number of observations assigned to topic k in document d, $n_{(\cdot)}^{d}$ is the number of observations in document d, $n_w^{(k)}$ is the number of observations of word type w under topic k and $n_{(\cdot)}^{k}$ is the total number of assignments to topic k. The most representative words can be extracted out of ϕ and the prominent topics of documents can be determined out of θ.

2.4 Non-Parametric Bayesian Methods

Two main approaches for solving statistical problems are frequentist and Bayesian. In frequentist analyses, parameters are fixed, in Bayesian ones, prior distributions are placed on parameters [42]. Parametric and nonparametric Bayesian models differ in model selection. If the data is to be explored with parametric Bayesian models, different models with different number of parameters are fit to the data; a model comparison metric that measures which one of the models fits the data best and a penalty score that is higher in complex models are calculated and a trade off between these two metrics is used to select the best model. However, complexity is adapted to the data with Bayesian nonparametric models and number of parameters is estimated by the model [43].

2.4.1 Dirichlet Process

One of the most commonly used prior distributions in nonparametric Bayesian models is the Dirichlet Process (DP). DP is a distribution over distributions and has two parameters. First is the base distribution as the prior belief and the second is the concentration parameter as its strength [44]. It is symbolized as $G \backsim DP(\alpha, H)$ where G is the Dirichlet process distributed with base distribution, H and concentration parameter, α. If the concentration parameter α is small, the samples of Dirichlet Process will cumulate around small number of units, if it is large, the distribution of samples will be similar to H.

Dirichlet process is indeed infinite dimensional generalization of Dirichlet distribution.

Some important attributes of a Dirichlet Process can be listed as follows [44, 45]:

- $E(G) = H$ that is the base distribution is the mean of the DP.
- Draws from a DP are discrete and probabilities are additive. So identical draws are possible.
- The posterior distribution of a Dirichlet distribution given observations $\theta_1, \ldots, \theta_n$ is

  $$G|\theta_1, \ldots, \theta_n \sim DP(\alpha + n, \frac{\alpha}{\alpha + n}H + \frac{n}{\alpha + n}\frac{\sum_i^n \delta_{\theta i}}{n}) \tag{2.11}$$

  where δ_{θ_i} is the unit mass function concentrated at θ_i. This is a weighted average of the base distribution H and the emprical distribution

  $$\frac{\sum_i^n \delta_{\theta i}}{n} \tag{2.12}$$

  The weights are α and n respectively. When the number of observations are large enough, $n \gg \alpha$, the posterior DP becomes more and more close to the underlying distribution of data.
- Posterior distribution given $\theta_1, \ldots, \theta_n$ is the predictive distribution of θ_{n+1}.

2.4.2 Construction of Dirichlet Process

There are some methaphors used to explain the construction of the Dirichlet Process. These are Blackwell - MacQueen Urn Schema, the Stick Breaking Construction and Chinese Restaurant Process.

2.4.2.1 Blackwell – MacQueen Urn Schema

This metaphor is established by Blackwell and MacQueen in 1973 [46]. At the begining, there is an empty urn G. A color is drawn from the base distribution H and a ball is painted with this color and placed into the urn. In the subsequent steps either this process is repeated or a ball is drawn from the urn and another ball is painted the same color as the just drawn ball and both are dropped into the urn. At the $n + 1st$ draw:

- Either, a new color is drawn with probability $\frac{\alpha}{\alpha+n}$ from the base distribution and a ball is painted with this color and dropped into the urn.
- Or, a ball is drawn from urn with probability $\frac{n}{\alpha+n}$ and a new ball with the same color as the just drawn is dropped along with the drawn ball. Drawing of a ball with a specific color is proportional to number of previous draws of balls wih its color.

If $\{\theta_i^n\}$ are successive draws from the urn:

$$\theta_{n+1}|\theta_1, .., \theta_n, \alpha, H = \sum_{k=1}^{K} \frac{m_k}{n+\alpha} \delta_{\theta_k^*} + \frac{\alpha}{n+\alpha} H \tag{2.13}$$

where m_k is the number of previous draws of the ball colored k from the urn G and $\delta_{\theta_k^*}$ is the unit mass function concentrating at $\delta_{\theta k}$, K is number of different colors in the urn.

2.4.2.2 Stick Breaking

This definition was established by Sethuraman in 1994 [47]. Suppose that we have a stick of unit length, we break it at a random proportion β_1 and assign π_1 to the just broken piece's length. Repeat this process to get π_2, π_3, ... on the remaining stick recursively [48].

$$\beta_k \sim Beta(1, \alpha) \quad \pi_k = \beta_k \textstyle\prod_{j=1}^{k-1}(1 - \beta_j) \tag{2.14}$$

An infinite sequence of weights $\pi = \{\pi_k\}_{k=1}^{\infty}$ is referred to be distributed according to $GEM(\alpha)$, where GEM stands for Griffiths-Engen-McCloskey.

The random discrete probability G is said to be a Dirichlet Process symbolized as $G \sim DP(\alpha, H)$ if

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \quad where \quad \theta_k^* \sim H \tag{2.15}$$

2.4.2.3 Chinese Restaurant Process

In Chinese restaurant process analogy, a Chinese restaurant has countably infinite number of tables, countably infinite number of customers can sit at a table. A customer can sit at an occupied table or at the next unoccupied table. When customer X_{n+1} comes into the resturant,

- Either, they sit at an already occupied table k with probability $\frac{m_k}{n+\alpha}$ where m_k is the number of customers at table k.
- Or, they sit at the next unoccupied table with probability $\frac{\alpha}{n+\alpha}$

$$\theta_{n+1}|\theta_1, .., \theta_n, \alpha, H = \sum_{k=1}^{K} \frac{m_k}{n+\alpha} \delta_{\theta_k^*} + \frac{\alpha}{n+\alpha} H \tag{2.16}$$

where m_k is the number of customer sitting at table k, K is number of allocated tables. The tables can be thought as the colors in the Black$-$ MacQueen Urn schema namely, G; and the unallocated tables compose the base distribution H.

2.4.3 Hierarchical Dirichlet Process

Hierarchical Dirichlet Process (HDP) is a model that is built on the recursive construction of Dirichlet Processes and it handles cases where data consist of groups (i.e.documents) and each data point in a group (i.e observation) belongs to a latent cluster (i.e. topic) and the latent clusters are shared across groups [49]. HDP-LDA can be considered as nonparametric counterpart of LDA and while the number of topics is given to LDA, the HDP-LDA model assumes the number of topics is infinite and that can be inferred from data.

The graphical model of HDP-LDA can be seen in Figure 2.5. In this model H is the prior distribution over topics. θ_{ji} is a parameter specifying the topic associated with x_{ji}, the i^{th} observation of j^{th} document. G_0 is the set of topics and G_j samples a subset of topics to use in document j from its base distribution G_0. γ and α are concentration parameters for G_0 and G_j respectively. These concentration parameters govern variability.

In HDP construction the Dirichlet process $G \sim DP(\alpha, G_0)$ is drawn from another Dirichlet distribution $G_0 \sim DP(\alpha_0, H)$ which forces G to replace its atoms on discrete places determined by G_0 [50] because support of each draw from the G distribution has to be a subset of the support of its base distribution G_0. This enables sharing atoms of G_0 distribution across G_j distributions.

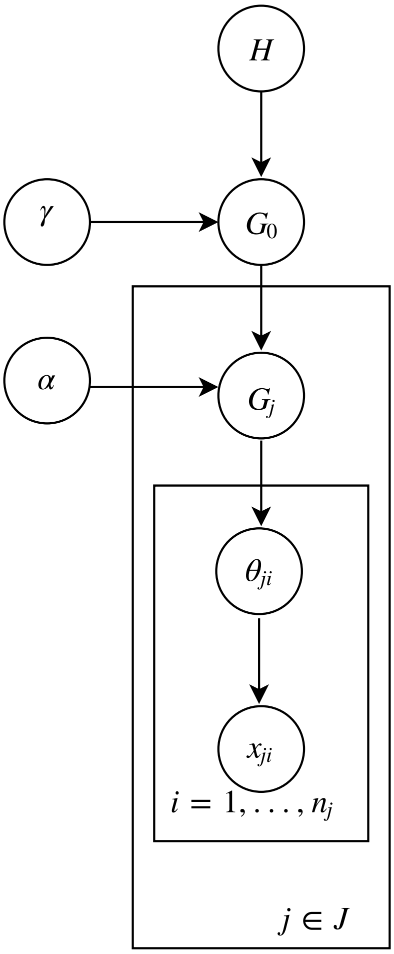


Figure 2.5: HDP-LDA model for topic modeling

The HDP model can be summarized as follows:

$$G_0 \sim DP(\alpha_0, H) \tag{2.17}$$
$$G_j|G_0 \sim DP(\alpha, G_0) \tag{2.18}$$
$$\theta_{ji}|G_j \sim G_j \tag{2.19}$$
$$x_{ji}|\theta_{ji} \sim F(\theta_{ji}) \tag{2.20}$$

where $F(\theta_{ji})$ is the distribution of x_{ji} given θ_{ji}.

Chinese Restaurant Franchise Sampling is a metaphor used for inference in Hierarchical Dirichlet Process. In this analogy, there is a two level hierarchy of Chinese Restaurant Processes. It uses a seperate Chinese Restaurant Process in each group (i.e. document). These are document level CRPs. Since latent variables are shared across groups, a corpus level CRP is defined as the upper level and the dish (i.e. topic) of the tables in the customer level CRPs are sampled from this layer.

Probabilities for the lower (customer) level CRP calculated using Gibbs sampling are shown in the following equations: [50].

The probability of the last customer sitting in a previously selected table.

$$p(t_{ji} = t) \sim \frac{n_{jt.}^{\neg ji}}{n_{j..}^{\neg ji} + \alpha} f_{k_{jt}}(\{x_{ji}\}) \tag{2.21}$$

The probability of the last customer to open a new table but with a previously sampled topic.

$$p(t_{ji} = t^{new} and \quad k_{jt^{new}} = k) \sim \frac{\alpha}{n_{j..}^{\neg ji} + \alpha} \frac{m_{.k}^{\neg ji}}{m_{..}^{\neg ji} + \gamma} f_k(\{x_{ji}\}) \tag{2.22}$$

The probability of the last customer to open a new table and sample a new topic.

$$p(t_{ji} = t^{new} and \quad k_{jt^{new}} = k^{new}) \sim \frac{\alpha}{n_{j..}^{\neg ji} + \alpha} \frac{\gamma}{m_{..}^{\neg ji} + \gamma} f_k{}^{new}(\{x_{ji}\}) \tag{2.23}$$

where t_{ji} is the table at which customer i in restaurant j sits, $n_{jt.}$ is the number of customers sitting at table t in restaurant j, $n_{j..}$ is the number of customers in restaurant j, $m_{.k}$ is the number of tables serving dish k, $m_{..}$ is the total number of tables, $\neg ji$ means that customer i in restaurant j is removed from CRF.

Probabilities calculated for the upper (menu) level CRP by using Gibbs sampling is as follows: [50].

$$p(k_{jt} = k) \sim \frac{m_{.k}^{\neg jt}}{m_{..}^{\neg jt} + \gamma} f_k(\{x_{ji} : t_{ji} = t\}) \tag{2.24}$$

$$p(k_{jt} = k^{new}) \sim \frac{\gamma}{m_{..}^{\neg jt} + \gamma} f_k{}^{new}(\{x_{ji} : t_{ji} = t\}) \tag{2.25}$$

2.5 Topic Coherence

After data is clustered into topics, quality of topics should be evaluated in order to get rid of incoherent topics. Since words with highest probability in each topic are representative for the topic, in a high quality topic, it is expected that each most representative word's conditional probability given the other most representative word should be high. A topic coherence metric in a pairwise fashion is defined as follows [22]:

$$C(t, V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{2.26}$$

where $D(v)$ is document frequency of word type v that is the number of documents that word type v seen at least once, $D(v_m, v_l)$ is the co-document frequency of word types v_m and v_l i.e the number of documents containing both v_m and v_l. $V^{(t)} = (V_1^{(t)}, \ldots, V_M^{(t)})$ is a list of most probable M words in topic t.

2.6 Perplexity

Perplexity is a measure of prediction power of a probabilitistic model on test data and monotonically decreases with likelihood. Lower perplexity values mean better generalization. In probabilistic topic models, for a test set of M documents, perplexity is:

$$perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M} logp(w_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{2.27}$$

where w_d represents the words in document d, N_d is the number of words in document d.

2.7 Gene Set Enrichment Analysis

After genes are clustered together, we consult gene databases to find out the biological interpretation of the relevant genes. Three of the databases are the Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Human Genome Epidemiology Network (HuGENet).

Gene Ontology(GO) [51] has a directed acyclic graph structure where the specificity increases from root to leaves. It has three main divisions: Cellular Compartment, Molecular Function and Biological Process. Cellular Compartment is the place where a gene product is active in a cell. Ribosome and nuclear membrane are example terms of cellular compartment. Molecular Function is the biological activiy of a gene product. Enzyme and transporter are example terms of molecular function. Biological Process is the biological objective which a gene/gene product takes role in. Translation and cAMP biosynthesis are example terms for biological process.

The KEGG database relates gene information to pathways and groups genes according to the biological pathways that they take role in. It has three databases: GENES for gene catalogues, PATHWAY for functions in terms of interacting module network and LIGAND for cellular chemical compounds, enzyme molecules and enzymatic reactions [52].

HuGENet maintains a database for published epidemiologic studies of human genes extracted from PubMed. Each article is indexed with MESH terms (by MESH hierarchical structure) and gene information from the NCBI Gene database [53].

2.7.1 Hypergeometric Distribution of Genes

If we can answer if a specific GO term or KEGG pathway is enriched in the gene list, the resulting terms can be biologically meaningful in describing the set of differentially expressed genes or the gene clusters found. In that sense, an overrepresentation test of genes can be achieved by using the hypergeometric distribution.

Let N be the total number of genes in the universe of the experiment, M be the number of genes annotated with a specific GO term or KEGG pathway, n genes are differentially expressed or form a cluster, k of these n genes are annotated by the specific GO term or KEGG pathway.

The probability for each k is:

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \tag{2.28}$$

The probability of having at most k genes:

$$P(X \leq k) = \sum_{y=0}^{k} \frac{\binom{M}{y}\binom{N-M}{n-y}}{\binom{N}{n}} \tag{2.29}$$

In the GO database, less specific nodes contain more specific nodes, so the nodes at root or in upper levels can achieve significant p values although these terms are not very informative. We need to set a limit for the number of genes annotated with the terms, and exclude the terms exceeding the limit [54].

2.7.2 Enrichment Analysis Tools

The co-expression of genes should be biologically analysed, interpreted and the results should be visualized. This process is called *geneset enrichment analysis* and as of 2009, there were more than 60 tools developed for this purpose [55].

These tools can be studied under three categories:

1. Singular enrichment analysis (SEA)

   SEA counts the enrichment of each term for the geneset and it compares the number of differentially expressed genes to the term and compares the result with random assignments, calculating the p-value based on statistical tests (e.g. Fisher exact, Chi-square, Hyper-geometric etc.). The most well-known examples of SEA tools are GoMiner, Onto-Express, DAVID, EASE, GOEAST and GFinder.

2. Geneset enrichment analysis (GSEA)

   GSEA works on all the genes from the experiment without using any threshold. It takes experimental results of all genes and the extent of differential expression is important, unlike SEA methods which only consider differential expression as a binary state. The Kolmogorov-Smirnov test, t-test and Z-score are commonly used in GSEA tools like Fatiscan, T-profiler, and GOdist.

3. Modular enrichment analysis (MEA)

   MEA combines SEA results with network information. This captures term-term co-occurrences which distinguish between biological conditions. Statistical methods used in MEA tools are Kappa statistics, the Czekanowski-Dice distance and Pearson Correlation. Examples of MEA tools are Ontologizer, topGO, ADGO and GENECODIS.

A new gene set enrichment analysis tool enrichr [56] approaches gene set enrichment analysis differently. While most of the gene set enrichment analysis tools rely on only Gene Ontology, this tool uses 35 different gene libraries which can be divided into six categories: transcription, pathways, ontologies, diseases/drugs, cell types and miscellanous.

2.8 Related Work

The usage of parametric topic models in gene expression data analysis has been reported in several publications.

Bicego et. al. [57] have applied PLSA and LDA for biclustering microarray data. They used the analogy between word-document and gene-sample pairs, their preprocessing approach does not have a biological foundation. They have applied these topic models on cancer classification (classification of leukemia and colon cancer) and tested their system with 10 fold cross validation. The average error rate is 9.24% for PLSA, 18.45% for LDA and 14.08% for PCA and error rates in LDA varies much depending on the number of genes. The results of this straightforward application of topic models show that LDA is not so successful when it is not optimized.

Chen et. al. [58] have applied LDA to analyze genome level composition of DNA sequences in order to understand whether genes with similar functional roles contain the similar latent topics. They applied LDA on N-mer sequence data of 635 genomes acquired from NCBI database with symmetric priors. In this method, N-mers are taken as words with N-letters and genome as the document.

Chen et. al. [59] have used LDA in metagenomics area. They have found top ranked taxa of each latent topic by biclustering NMERs and the top ranked latent topics of samples. They have discovered microbial groups in each sample.

Bicego et.al. [60] have used PLSA to extract biclusters in microarray expression data and highly correlated sample and gene groups. They determined overrepresented GO terms in the biclusters using GOstat tool and evaluated the study with both real data and synthetic benchmark.

Flaherty et. al. [61] have developed LLDA (Labeled Latent Dirichlet Allocation) model in order to get rid of the limitation of traditional algorithms to classify a gene in a single class when classifying pleiotropic genes. This study aimed to make drug-target predictions and its area was chemogenomics.

Chheng [62] have used topic modeling to extract gene relations from medical literature (PUBMED). The idea behind this study was that genes with similar research topics would be functionally related.

Caldas et. al. [63] have developed a system in order to find related experiments given a particular experiment. This application makes it possible to use a dataset itself as a query to search for similar experimental data. This study enables to overcome problems stemmed from e meta-data usage. In this study, high biological coherence was achieved and related experiments were captured. The average precision was 82%

while random base was 40%.

Bicego et. al. [64] have used probabilistic topic modeling in generative embedding phase of renal cancer cell classification on tissue microarray images. They took visual features as words and nuclei as documents.

Liu et. a. [65] have used Correspondence Latent Dirichlet Allocation (Corr-LDA) for the purpose of identifying FMRM (functional miRNA regulatory modules).Their algorithm allows to bicluster heterogenous data of miRNA and mRNA both with and without binding information but they reported only result of the implementation without binding information in this paper.

Perina et. al. [66] have handled usage of generative-discriminative approaches in microarray data classification task. They derived features of samples as being The Fisher Score, TOP kernel scores, Log Likelihood Ratio, Score Space, Free Energy Score Space and Posterior Divergence Spaces on PLSA model. They performed Support Vector Machine classification with linear kernel on Colon Cancer, Ovariance Cancer and DLBCL datasets by defining similarity of two samples as inner product of their scores.

Rogers et. al. [67] have developed a model called Latent Process Decomposition (LPD). In this study, the LDA model has been modified to be able to express the continuous nature of gene expression data better. The word- topic distributions are Gaussians instead of multinomials in original LDA.

Perina et. al. [68] have proposed a method called BaLDA (Biologically - aware latent dirichlet allocation) for the classification of microarray expression, it is a modification of LPD where the dependencies among genes are integrated to the system with a clustering module. This study takes external information into account but uses external information as constraints not as priors. If the prior information is not true, the model can not overcome this problem. In our approach, if data contradicts with prior information, prior information is set aside.

Pino et. al. [69] have used LDA to predict gene annotations. They used gene-annotation matrix A, as their corpus where gene i has an annotation to annotation term j $A[i, j] = 1$ else $A[i, j] = 0$. The interesting point in this study, they used asymmetric prior on topic-annotation distribution. Their approach is different from ours, their aim is not to incorporate external gene co-operation information. In our study, we use a matrix whose individual row has a gene's co-expression/co-regulation information, we use a single row for each topic. In this study, they use the same asymmetric vector for all topics. They were inspired by the term frequency- inverse document frequency (tf-idf) concept of text mining domain. They use inverse gene frequency and the prior favors the biological annotation terms associated with fewer genes and they contend that this improvement contributes to generate more specific topics.

Nonparametric Bayesian methods have also been used in microarray and RNA-seq data studies. Vavoulis et. al. [70] developed a software package called DGEclust for clustering and differential expression analysis of RNA-Seq data. In this software, they have implemented HDP algorithm but they have modified HDP in a way that allows to draw expression profiles of genes from Negative Binomial Distribution.

Gerber et. al. [71] have developed a software called GeneProgram. They modified HDP in a way to collect tissues into tissue groups. Their model collects tissues into groups and genes into overlapping topics on time series data.

Wang&Wang [72] have applied HDP in order to segment regulatory network and clustering gene expression data of yeast cell cycle.

Caldas&Kaski [73] have used Nested Chinese Restaurant Processes for hierarchical biclustering of miRNA expression data.

These nonparametric Bayesian studies either use HDP or a very similar variant, but they use flat priors unlike our study.

In the document clustering domain, Wallach et. al. [74] claim that asymmetric Dirichlet priors over the document-topic distribution can improve LDA's performance while asymmetric priors over the word-topic distribution can not. Although words in documents and transcripts (genes) in biological samples have many common features; in at least one aspect they differ: causal relationship among words is not as strong as that among genes. We are claiming and will prove that asymmetric priors over gene-topic distribution have an impact on performance.

Furthermore, Chen et. al. [75] have objected Wallach et al.'s claim in the text domain as well. There are a number of studies (e.g. [75–84]) that improve different parametric and nonparametric topic models with priors on the word-topic distribution in the text domain.

CHAPTER 3

MATERIALS AND METHODS

3.1 Method

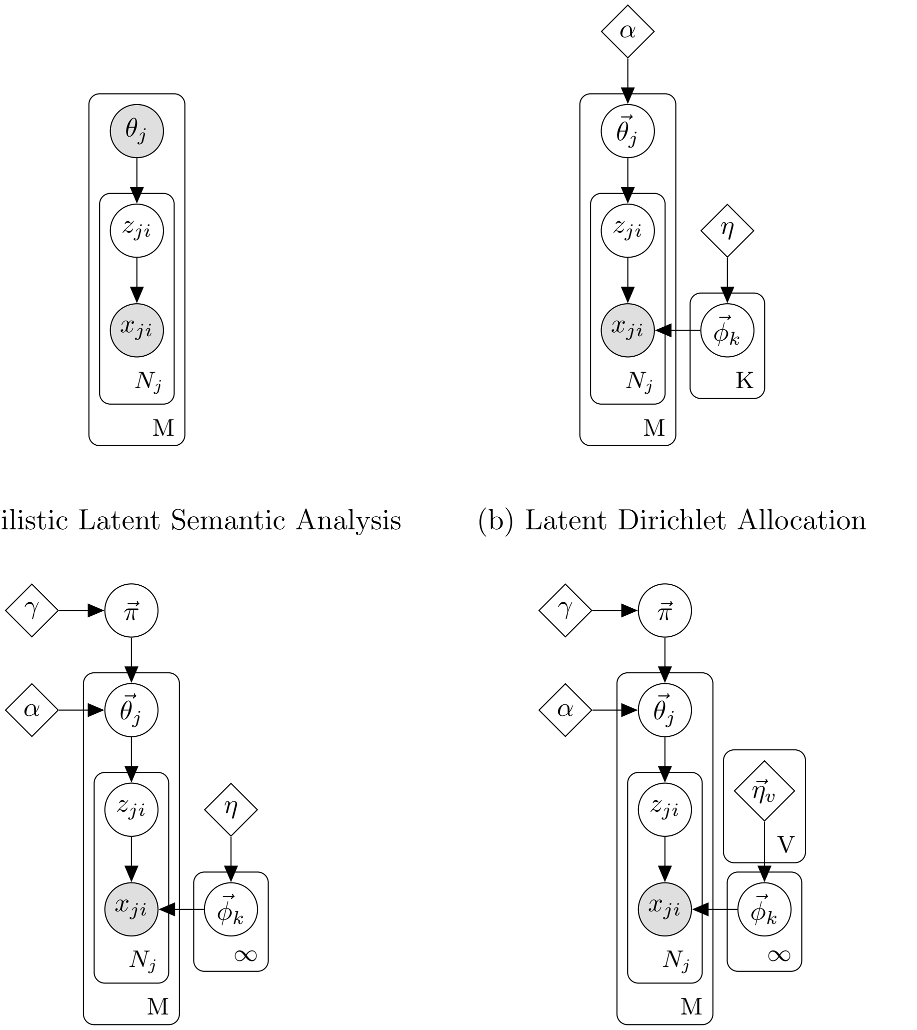
The model we have proposed in this thesis study is supposed to have the following features:

1. It should be able to find biclusters, in other words, the clusters found by the model must cover both samples and transcripts simultaneously. A bicluster should be a mixture of transcripts and a sample should be a mixture of biclusters.
2. The clusters should be overlapping and should have fuzzy memberships. By this way, context specific behavior of genes can be modeled.
3. It should not need any knowledge or guess about the number of biclusters.
4. It should be capable of using external information about gene co-regulation and co-expression (like gene regulatory networks). This enables the model to work with sparser data.

3.1.1 Mixed-Membership Models

Mixed-membership models are models that can find latent factors in grouped data. Grouped data is data which can be represented in a frequency table. In the text mining domain, the groups are the documents, the frequencies are the number of occurances of different word types in each document. Topics are the latent factors. In our analogy, we represent expression levels as the observations; samples as the groups and the biological contexts as the factors. We give the graphical representation of our model in Figure 3.1 with the other topic models to illustrate the position of our model in the evolution of topic models.

Although mixed membership models have been explained in detail in Section **??**, we find it useful to review them in order to make our model more understandable. First mixed membership model is the Probabilistic Latent Semantic Analysis (PLSA) [16] and its plate notation can be seen in Figure 3.1(a). In PLSA, a latent mixture of topics ($z_{j.}$) which is in fact a point in the topic simplex, is assigned for each sample, the observations (i.e. gene expression levels) x_{ji} s are drawn from this distribution.



(a)Probabilistic Latent Semantic Analysis

(b) Latent Dirichlet Allocation

(c) Hierarchical Dirichlet Process

(d) Externally Smoothed Hierarchical Dirichlet Process(this study)

Figure 3.1: **Mixed-membership models.** Observed variables are shaded, latent variables are unshaded, and deterministic entities are framed by diamonds. In each plate diagram M denotes the number of samples (assays), N_j the number of transcripts (observations) assayed in sample j, K the number of topics, and V the number of transcript types (i.e. the vocabulary). $\vec{\theta_j}$ is the topic distribution of j^{th} sample (or the index of the j^{th} sample in the case of PLSA), and x_{ji} is the observed transcript type of the i^{th} transcript observed in the j^{th} sample. z_{ji} is the topic membership of each observation in each sample. $\vec{\phi}_k$ is the distribution of the k^{th} topic over transcript types.

Expectation-Maximization approach is used for finding this mixture of topics as well as the correct word distribution, which is a point in the word simplex, for each topic. The main shortcoming of PLSA is that it is not a generative model and can not assess an unseen sample.

The second model whose plate notation can be seen in Figure 3.1(b) is Latent Dirichlet Allocation (LDA) and it overcomes the aferomentioned problem. In LDA model, samples are modeled as Dirichlet distributions ($\vec{\theta_j}$ is the Dirichlet distribution for the j^{th} sample) over the topic simplex; similarly, topics are Dirichlet distributions ($\vec{\phi_k}$ is the Dirichlet distribution for the k^{th} topic) over the word simplex. The hyperparameter α is the concentration parameter for the sample distribution. It defines the smoothness of topic distribution in a sample; if it is too small, the distribution is peaky and a few topics have high probability while the others have very small probability. If it is too large the topic distribution becomes uniform. Similarly, η is the concentration parameter for the topic distribution and it defines the smoothness of word distribution over a topic. If it is too small, few words tend to have high probability and if it is too large, every word has nearly same probability in a topic. In generative modeling, LDA's mechanism is as follows:

$$\begin{array}{rcl} \vec{\theta_j} & \sim & Dir(\alpha) \\ \boldsymbol{\phi}_k & \sim & Dir(\eta) \\ z_{ji} & \sim & Multinomial(\vec{\theta_j}) \text{ for each obs. } i \text{ in sample } j \\ x_{ji} & \sim & Multinomial(\vec{\phi}_{1\ldots k} \mid z_{ji}) \end{array} \tag{3.1}$$

Both PLSA and LDA need the number of topics to be set before running the algorithms. It is hard to guess a plausible number of topics because one may have no idea about the number of active cellular contexts in a sample set. Model selection is the approach to get over this handicap and it is performed by running LDA and PLSA with different number of topics. Cross validation using an external index like perplexity can be used to decide on the best number of topic among these runs.

Hierarchical Dirichlet Process as a nonparametric Bayesian Model achieves model selection on a single run instead of trying to optimize the number of topics across runs, it finds the number of topics which is likely to result in the simplest accurate model. HDP is the nonparametric counterpart of LDA and the Dirichlet distributions in LDA are replaced by Dirichlet Processes in HDP. The graphical representation of HDP can be seen in Figure 3.1(c).

In HDP model, the global (experiment level) topic distribution is composed by draws from a Dirichlet process $\vec{\pi}$ whose concentration parameter is γ. Topic distribution of sample j is given by $\vec{\theta_j}$. $\vec{\theta_j}$ is a Dirichlet distribution sampled from the Dirichlet distribution $\vec{\pi}$ with a concentration parameter α. Each observation of a sample (let's say i^{th} observation of j^{th} sample) is assumed to be generated by first sampling a topic k ($k = z_{ji}$) according to the distribution $\vec{\pi}$ and then sampling an observation according to the parameter distribution of topic k, ϕ_k.

This recursive construction of Dirichlet Processes ($\vec{\pi}$ and $\vec{\theta_j}$) provides shrinkage among samples that is all the samples are mixtures of the global topics but the distribution of topics over each sample is different. This can be shown as follows:

$$\begin{aligned}\vec{\pi} &\sim GEM(\gamma)\\ \vec{\theta}_j &\sim DP(\alpha\vec{\pi})\end{aligned} \tag{3.2}$$

The metaphor used for construction of HDP in our implementation is the Stick Breaking Construction [47] and this representation can be summarized as follows:

Let's start with the base measure $\vec{\pi}$.

$$\vec{\pi} = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k^*} \tag{3.3}$$

GEM(γ)(Griffiths–Engen–McCloskey [85]) refers to the joint distribution on the infinite sequence (β_1, β_2, β_3...), where $\beta_k \sim Beta(1,\gamma)\left(1-\sum_{l=1}^{k-1}\beta_l\right)$. This process is the recursive partitioning of a unit probability mass into infinite number of parameters and k^{th} partition's probability is β_k. The partitioning of the unit probability mass is distributed according to the Dirichlet process $GEM(\gamma)$ and $\vec{\pi}$ is the resulting distribution. More information on Stick Breaking Construction can be found in Section 2.4.2.2.

The topic distribution over j^{th} sample ($\vec{\theta}_j$) is generated by a Dirichlet Process, but this time the base measure is $\vec{\pi}$ and the concentration parameter is α. This results in a hierarchy of Dirichlet processes. In other words, topic distribution over each sample ($\vec{\theta}_j$) is drawn from the global distribution $\vec{\pi}$ which is the distribution over all topics. By this way, support of each $\vec{\theta}_j$ is in the support of $\vec{\pi}$. The topics in the samples are inherited from the global topic distribution and no other topic than the topics in the global topic distribution can be sampled in each sample and since $\vec{\pi}$ is a discrete distribution, the topics are shared across samples. Topics are indexed by $k = 1 \ldots \infty$.

$$\vec{\theta}_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k^*} \tag{3.4}$$

There are technically an infinite number of topics and $\vec{\phi}_k$ is the distribution of observation (transcript) types over the k^{th} topic. These topics are the ones mentioned in the definitions of $\vec{\pi}$ and $\vec{\theta}_j$.

$$\begin{aligned}z_{ji} &\sim Multinomial(\vec{\theta}_j)\\ \vec{\phi}_k &\sim DP(\eta)\\ x_{ji} &\sim F(\vec{\phi}_{1\ldots\infty}|z_{ji})\end{aligned} \tag{3.5}$$

x_{ji} is the i^{th} observation of j^{th} sample and it is drawn from a categorical distribution. It is the "type" of the observed transcript which can be a gene or probe name depending on the experimental observation type.

The Dirichlet Processes included in HDP are modeled by an algorithmic construction. The metaphors used for this construction are Stick Breaking and the Chinese Restaurant Process [86]. The Stick Breaking metaphor is explained above and the Chine Restaurant Process is explained in Section 2.4.2.3. For HDP construction, two hierarchically constructed CRPs can be used for $\vec{\pi}$ and $\vec{\theta}_j$ and this hierarchical structure is called Chinese Restaurant Franchise (CRF) [18]. The posterior inference is done with Gibbs sampling in this study. More detailed information about CRF and posterior inference can be found in Section 2.4.3 and also [50] and [87].

The Gibbs update step for assignment of observations to topics can be formulated as follows:

$$p(z_{ji} = k \mid \cdot) \propto \left(n_{jk}^{\neg i} + \alpha \pi_k\right) \cdot \frac{n_{k,v=x_{ji}}^{\neg i} + \eta}{n_k^{\neg i} + V\eta} \tag{3.6}$$

where $n_{j,k}^{\neg i}$ is the total number of observations assigned to topic k in j^{th} sample; $n_{k,v=x_{ji}}^{\neg i}$ is the total number of observations in the experiment with the same type as the observation in question, which are also assigned to topic k; and $n_k^{\neg i}$ is the total number of observations which are assigned to topic k in the experiment. The observation in question, x_{ji} is excluded from these counts.

3.1.2 Proposed Model

The proposed model offers novelties in two aspects, first is how gene expression data is converted into sample-transcript counts (mentioned in Section 3.1.2.1) and the second is how the prior information is handled and incorporated into the model (mentioned in Section 3.1.2.2).

3.1.2.1 Preprocessing

First of all, we needed to find a way to represent gene expression levels of transcripts in samples. Gene expression data takes continuous values and it is natural to be inclined to use a continuous distribution for its representation. For example, Rogers et.al. [67] used Gaussian distributions. In microarray image readings, the gene expression data seems to be continuous but in fact the number of transcript in a cell is discrete. In an assay of a cell, the reads are multinomial variables. Therefore the distribution we have used to model gene expression data is multinomial distribution (more specifically categorical distribution).

The multinomial distribution's support set is the set of positive numbers and this causes difficulty in centering. When we center transcript quantities in samples against a reference value, over-expressed values become positive and under-expressed values become negative. The magnitude of these values represent the extent of differential expression.

In transcriptomics experiments, we have $V^{'}$ different observations per sample (one for each transcript, exon, probe etc.). We first center these observations against a reference. We now have positive and negative values for each of $V^{'}$ observations. Since we are using multinomial distribution, we can only use positive integers. To be able to use negative values in multinomial distribution, the individual observation variable is split into two different multinomial variables. Thus, the transcript type vocabulary size is doubled, we now have a transcript type vocabulary of size $V = 2V^{'}$. Over-expressed observations are represented in the first of the pairs, that is in the first half of the V size vocabulary; under-expressed observations are represented in the second of the pairs. In other words, over-expressed observations only take 0 values in the second of the pairs and vice versa applies to the under-expressed observations. These pairs need to be expressed in integer values and these integer values give the number of times that corresponds to differential expression level of a given transcript. There are different ways for preprocessing gene expresssion data in our approach. One possible way is Multiples of Median (MoM) measure. Another way may be rounded fold change. The preprocessing for Multiples of Reference where the reference can be median over all samples or a control assay is:

$$\{e^{'}_{jv}, e^{'}_{j(V'+v)}\} = \begin{cases} \{\|\frac{e_{jv}-\tilde{e}_v}{\tilde{e}_v}\|, 0\}, & \text{if } e_{jv} - \tilde{e}_v \geq 0 \\ \\ \{0, \|\frac{|e_{jv}-\tilde{e}_v|}{\tilde{e}_v}\|\}, & \text{otherwise} \end{cases} \tag{3.7}$$

where e_{jv} is the expression level of v^{th} transcript type of j^{th} sample, $\tilde{e}_v$ is the reference level for the v^{th} transcript type.

Our approach is not the only option for modeling gene expression data as a multinomial distribution. For instance, Gerber et. al. [71], have used a multinomial variable for magnitude and a Bernoulli indicator for the direction (over-expression or under-expression) of differential expression for each transcript type. This usage does not decrease the number of parameters because a transcript type is represented with two variables like it is in our model. This representation doesn't model the biological contexts where some of the genes' behaviors are volatile among samples (i.e. cases where the same context is defined by an "oscillation" of one or more genes).

3.1.2.2 Prior Improvement

The improvements we have made on standard HDP (Figure 3.1(c)) algorithm is enhancing transcription distribution over topics. In the standard HDP model a flat distribution is used, thus the prior information is only parametrized by η. This parameter is the degree of belief in prior information.

We named our model Smoothed HDP(Figure 3.1(d)), in our model, each topic prior is drawn from a set of $\vec{\eta}_v$ multinomials. Each of $\vec{\eta}_v$ multinomials is a distribution based on perturbation of a transcript type. When a new topic is drawn, our model is informed about which transcript types are likely to be co-expressed whereas standard HDP assumes uniform prior that is all transcripts are equally likely in a topic distribution. In both models, prior information may be overridden by a sufficient number of observations. If the observations are sparse, starting from an accurate prior belief enables the model to work on a smaller space configurations and this gives better results with less number of observations. The generative process of standard HDP (Equation 3.5) is extended as follows:

$$\begin{aligned} \vec{\eta}^k &\sim Multinomial(\{\vec{\eta}_1,\ \ldots,\ \vec{\eta}_V\}) \\ \vec{\phi}_k &\sim DP(\vec{\eta}^k) \\ x_{ji} &\sim F(\vec{\phi}_{1\ldots\infty}|z_{ji}) \end{aligned} \tag{3.8}$$

$\vec{\eta}^k$ is sampled from V distributions of the same type multinomially based on the centrality of transcript types because the likelihood of a transcript type to originate a new topic is proportional to the prior density of it. The distribution of topic k in question will have a prior distribution that is based on perturbation of the transcript type which is the first observation assigned to it.

A $\vec{\eta}^k$ is drawn for each topic when it is originated. The subsequent draws of $\vec{\eta}^k$ among $\vec{\eta}_{1\ldots v}$ and $\vec{\phi}_k$ is similar to a draw from Imprecise Dirichlet Process [88] but different in that the number of distributions is finite.

In the implementation of smoothed HDP, a new topic is materialized by assignment of first observation (of the transcript type whose perturbation is related to η_k) to it. The η_k vector composes the pseudo counts of the transcript types in this topic. The Gibbs update of standard HDP given in Equation 3.6 is modified as follows:

$$p(z_{ji} = k \mid \cdot) \propto \left(n_{jk}^{\neg i} + \alpha\pi_k\right) \cdot \frac{n_{k,v=x_{ji}}^{\neg i} + \eta_{v=x_{ji}}^k}{n_k^{\neg i} + \sum_{v=1}^{V} \eta_v^k} \tag{3.9}$$

where η_v^k is the element of the vector $\vec{\eta}^k$ which corresponds to transcript type v.

3.1.3 Creating The Prior

In our approach, the $\{\vec{\eta}_1,\ \ldots,\ \vec{\eta}_V\}$ set is represented in a $V \times V$ matrix H, which has individual $\vec{\eta}_i$ s in its rows. We have proposed two kernels for preparing H matrix. First is co-expression smoothing which is based on correlation of gene expression data and the second is network-based smoothing which is based on transcriptional regulatory network.

3.1.3.1 Co-expression Smoothing

This approach is an example of empirical Bayes estimation [89] where the priors (hyperparameters) are estimated from data and the estimation of priors from data in co-expression smoothing approach can be explained as follows, let $\boldsymbol{E}$ be the original gene expression matrix, each row representing sample profile of a gene, each column representing gene expression profile of a sample. The correlation matrix (e.g. Pearson correlation) of this $\boldsymbol{E}$ matrix is $\boldsymbol{\rho}$ where $\rho_{i,j} = cor(E_{i\cdot}, E_{j\cdot})$. The $V \times V$ matrix $\boldsymbol{H}$ can then be created by the following concatenation of transformed copies of $\boldsymbol{\rho}$ as also described in [90]:

$$\boldsymbol{H} = \left[\begin{array}{c|c} 0.5 + 0.5\boldsymbol{\rho} & 0.5 - 0.5\boldsymbol{\rho} \\ \hline 0.5 - 0.5\boldsymbol{\rho} & 0.5 + 0.5\boldsymbol{\rho} \end{array}\right]^{\beta} \tag{3.10}$$

This H matrix is normalized to have total of $V\eta$ value in each of its rows and β is set to 8 in our experiments. The over-expression counts of transcripts are represented in the first V' and the under-expression counts are represented in the indices offset by V' because of the arrangement in Equation (3.7). For example, if the 1^{st} row of the H matrix represents the counts for transcripts of A (over-expressed A), the $V' + 1^{st}$ of H matrix represents the counts for $\neg A$ (under-expressed A). So the likelihood of over and under-expression of the same transcript type are inversely correlated.

Individual rows of H matrix is assured to have a sum of $V\eta$. Owing to this, prior information can be treated as it is in the standard HDP. The individual rows of H matrix becomes the η_v^k's in the Equation (3.9), the $\sum_{v=1}^{V} \eta_v^k$ expression in the denominator of this equation can be replaced with $V\eta$.

The correlation (ρ) can be computed from the dataset to be biclustered or another dataset including more samples from a large database like Array-Express [91]or Gene Expression Omnibus [92] to be able to be more precise in correlation calculation.

3.1.3.2 Network-based Smoothing

The second way of generating H matrix is to use an external transcriptional regulatory network information. They are available for most of the model organisms and also can be built by gathering information from literature or experimental tools like Chip-Seq.

How to convert a transcriptional regulatory network into H matrix can be explained as follows: A network is represented as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of transcript types and $\mathcal{E}$ is the links between them. so $|\mathcal{V}| = \mathcal{V}'$. Each element of $\mathcal{E}$ is a triplet of (i, j, w) and this triplet means that there is an edge from transcript type i to transcript type j with weight w. The weights are in the range [-1,1], positive w values mean activation and negative w values mean inhibition of transcript type j by transcript type i.

First, we build the undirected adjacency matrix $\boldsymbol{A}$ whose elements are:

$$a_{ij} = \begin{cases} w & \text{if } (i,j,w) \in \mathcal{E} \\ w & \text{if } (j,i,w) \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases}$$

This matrix A, holds the relations between transcript types and their immediate neighbor transcript types (radius 1). If one is interested in relations in more than 1 neighborhood, this can be found by taking appropriate power of A (e.g. A^2 for radius 2),

The adjacency matrix becomes:

$$\boldsymbol{A} \times \ldots \times \boldsymbol{A} = \boldsymbol{A}^r = \begin{bmatrix} a_{11}^{(r)} & u_{12}^{(r)} & \ldots & a_{1V'}^{(r)} \\ a_{12}^{(r)} & u_{22}^{(r)} & \ldots & a_{2V'}^{(r)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{V'1}^{(r)} & u_{V'2}^{(r)} & \ldots & a_{V'V'}^{(r)} \end{bmatrix} \tag{3.11}$$

where $a_{ij}^{(r)}$'s are elements of the matrix. The matrix A^r is normalized to get the ρ matrix.

$$\rho_{ij} = \frac{a_{ij}^{(r)}}{\sqrt{a_{ii}^{(r)} a_{jj}^{(r)}}} \tag{3.12}$$

The rest of network-based smoothing approach is similar to co-expression smoothing. The ρ value is substituted in the Equation 3.10, hence H matrix is calculated.

Both Co-expression smoothing and Network-based smoothing approaches make the HDP model more accurate and more robust to changes in hyperparameter settings. In the next chapter, we will report the results of experiments that show this.

CHAPTER 4

EXPERIMENTS

We have performed various experiments on synthetic and biological datasets using HDP, Co-expression smoothing HDP and Network-based smoothing HDP.

The first experiment which is explained in detail in Section 4.1 was performed for ensuring that HDP solves the biclustering problem. This proof of concept was achieved by using datasets consisting of pre-seeded biclusters. In this experiment, we tested if HDP could find the positions of biclusters correctly in two-dimensional datasets.

The second experiment which is explained in detail in Section 4.2 is the starting point of the main contribution of this thesis study. In this experiment, we proved that informed priors enhance success of HDP algorithm in finding overlapping biclusters.

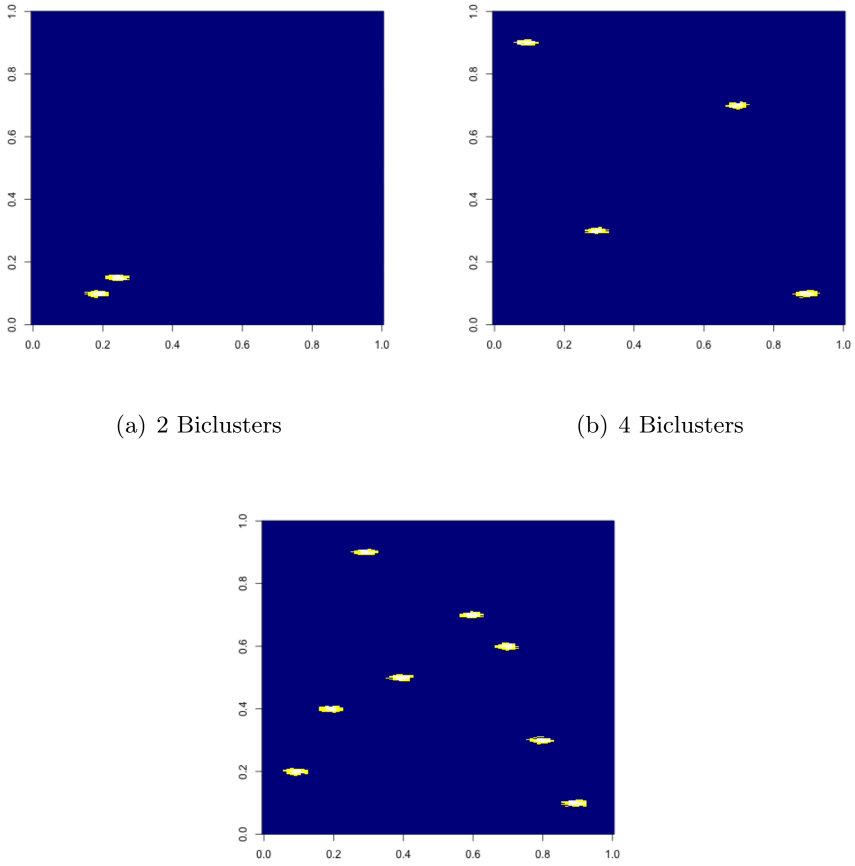
The third experiment which is explained in detail in Section 4.3 was carried out for quantitive evaluation of HDP and Smoothed HDP and it reports the Akaike Information Criterion (AIC) values for datasets with different sparsity levels (number of samples) and prior strengths (η values).

The fourth experiment which is explained in detail in Section 4.4 was accomplished using a semi-synthetic dataset created using Syntren [20]. In this experiment different genes were perturbed and each perturbance simulated a unique biological context. Smoothed HDP algorithms outperformed standard HDP algorithm in finding these biological contexts with robustness to changes in prior strength and in sparsity levels.

The fifth experiment which is explained in detail in Section 4.5 was performed on a dataset from a prostate cancer study by Dhanasekaran et. al. [21]. In this experiment, we evaluated comparative success of HDP and Co-expression smoothing HDP in two metrics. First is dependent on sample-topic distribution and sample labels. We compared each sample's label with that of the sample which is most similar to it according to its topic distribution. Second is based on the topic coherence metric given in Section 2.5. In both evaluation approaches, Co-expression smoothing HDP provided more successful and consistent results over different prior strengths (η values).

4.1 HDP Trials

We wanted to establish the performance of HDP in finding biclusters. We seeded different numbers of biclusters into two dimensional datasets, as well as noise, in a way that every observation type is seen in every sample at least once. The datasets of 2, 4 and 8 biclusters can be seen in Figure 4.1.



(a) 2 Biclusters

(b) 4 Biclusters

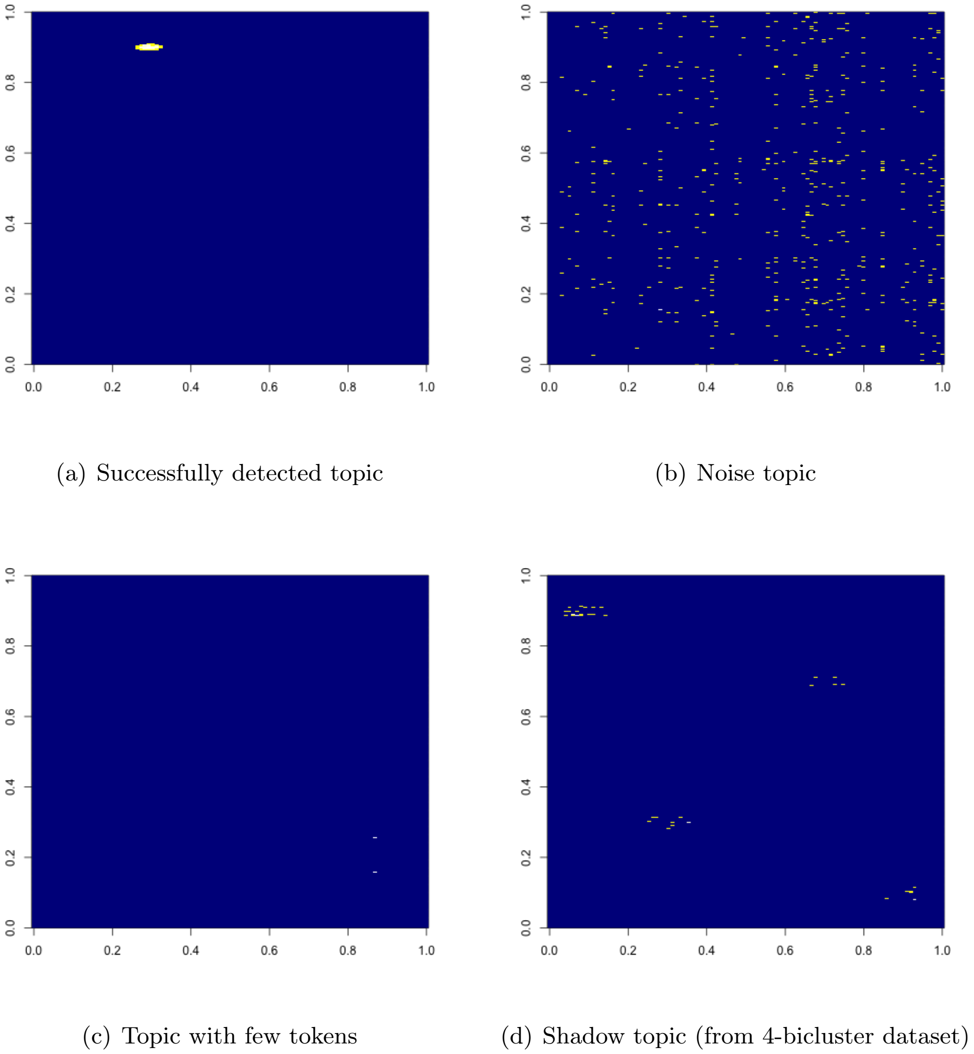
(c) 8 Biclusters

Figure 4.1: Seeded Biclusters

We ran HDP algorithm on these datasets and retrieved the biclusters from sample - gene - topic triples and visualized them. All the biclusters we seeded were recovered after HDP runs. There were also some topics with few observations, some noise topics and some shadow topics consisting tokens from some or all of the seeded biclusters. One example for each type of topics can be seen in 4.2.

The topics retrieved after running HDP with the datasets of pre-seeded 2, 4 and 8 biclusters with default hyperparameters ($\eta = 0.5$, $\gamma = 1$ and $\alpha = 1$) can be seen in Figure 4.3.

Topics with few tokens can be filtered just by removing topics having number of ob-



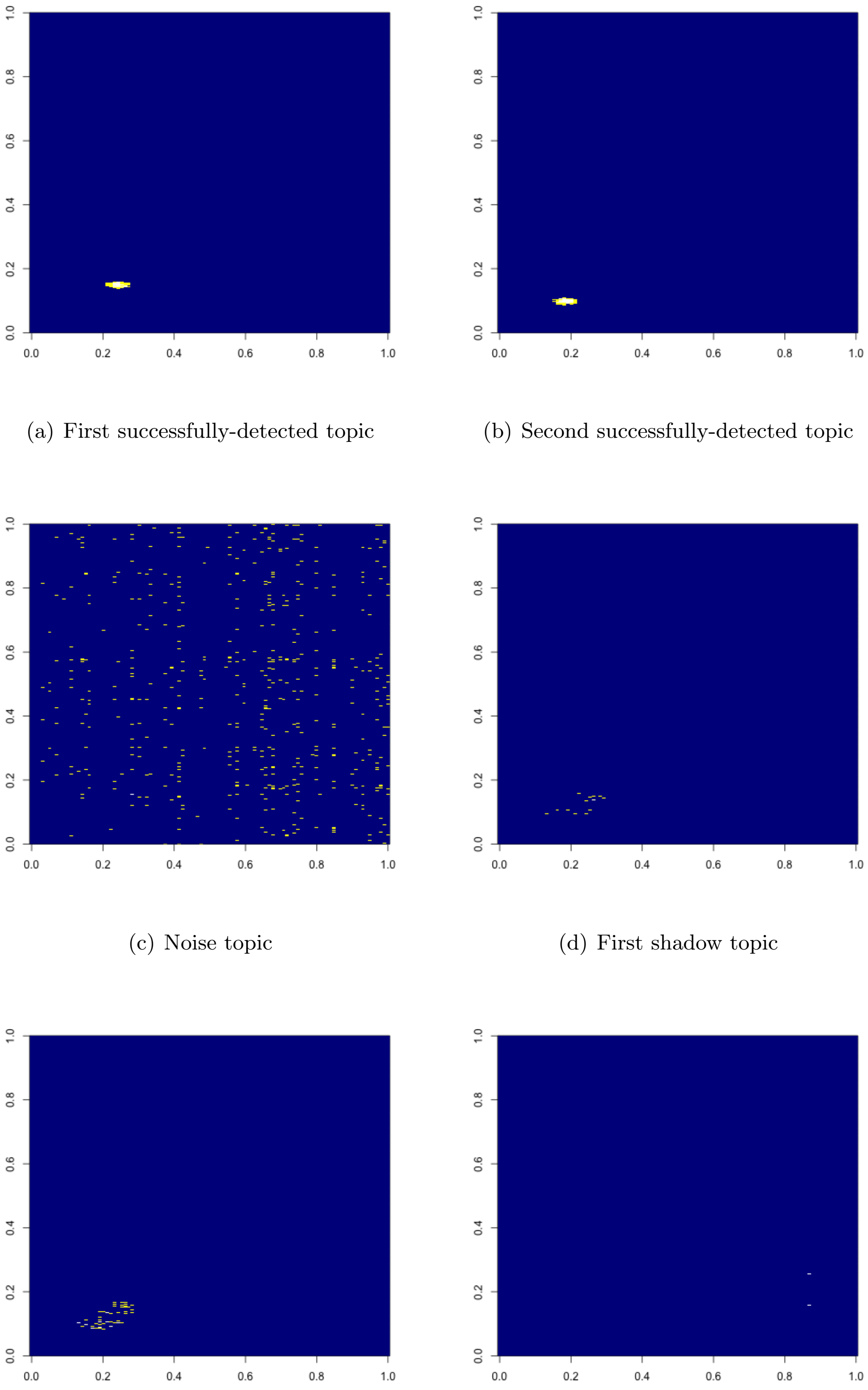
(a) Successfully detected topic

(b) Noise topic

(c) Topic with few tokens

(d) Shadow topic (from 4-bicluster dataset)

Figure 4.2: Types of Retrieved Topics



(a) First successfully-detected topic

(b) Second successfully-detected topic

(c) Noise topic

(d) First shadow topic

(e) Second shadow topic

(f) Topic with few tokens

Figure 4.3: Topics Retrieved After 2-Bicluster Dataset Run

servations under a specific threshold. Shadow topics can be filtered by means of the topic coherence metric defined in Section 2.5.

4.2 HDP - Prior Improvement Trials

We used HDP test cases created by Yee Whye Teh [19] to test our prior improvement approach. The test platform works as follows: it generates a dataset with the number of documents and the number of word types in the vocabulary entered by the user. The resultant clustering should satisfy bars problem. That is, let us say we have 9 different types of words in our dataset, the number of topics will be 6 -that is the total number of columns and rows- and each topic must be one of the rectangles which can be seen in Figure 4.4. For example, words encoded with 1,2 and 3 form a topic, 2, 5, 8 form another topic etc.

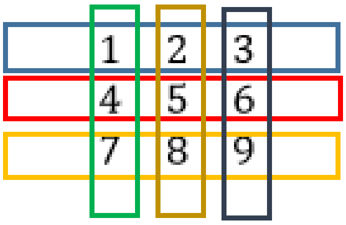


Figure 4.4: Distribution of words among topics

We implemented our preliminary Smoothed HDP approach by modifying the HDP program [93] which has been written by Wang and Blei in C++ programming language. We ran the Smoothed HDP program with uniform priors and got the same results with the original HDP program using the same random seed meaning that we did not change any behavior of the program other than we had planned to.

In the prior set-up, the total of η_v^k values (pseudo counts) in the topic k, $\sum_{v=1}^{V} \eta_v^k$, is distributed heavily among the words that are likely to be in the same topic with the word which has originated topic k. Let's say word type 1 originates a new topic, word types 1, 2, 3, 4 and 7 take greater η_v^k values than the rest of the vocabulary. Note that the prior information does not directly feed topics to the model, it just gives information about co-existence of words.

We used terms "word" and "document" to refer the original testbench but from this paragraph on, we will be using terms in bioinformatics. Namely sample instead of document, transcript instead of word. The test was organized as 30, 40, 50, 60 and 70, 80 samples (sparsity levels), each sample having a Dirichlet topic distribution with hyper-parameter 0.1. The total number of topics in the corpus was 20. The vocabulary size in the corpus (experiment set) was 100. The experiments were repeated for the η values (prior strengths) of 0.05, 0.25, 0.5, 0.75 and 1. The tests were run 100 times for each combination of algorithm, sparsity level and prior strength resulting in 6000 runs.

Transcript distribution of each original topic was compared to that of each topic retrieved in a run. The cosine similarity between the original topic and the most similar topic was regarded as the measure of topic accuracy.

Since there were 20 topics in the experiment, a properly working algorithm should have found around 20 topics.

The topic accuracy values and the number of topics with respect to different sparsity levels can be seen in Figure 4.5(a) and Figure 4.5(b) respectively. Likewise the topic accuracy values and the number of topics with respect to different prior strengths can be seen in Figure 4.6(a) and Figure 4.6(b) respectively.

The results showed that Smoothed HDP algorithm outperforms HDP in every combination of sparsity level and prior strength providing more consistent results under varying prior strengths and sparsity levels. It also finds the correct number of topics more accurately than standard HDP.

4.3 Akaike Information Criterion Test for HDP and Smoothing HDP Algorithm

We used the same experimental set up with the previous test to analyze our approach quantitively. Akaike Information Criterion (AIC) [94]was used for model selection between original HDP and Smoothed HDP. The AIC criterion given each dataset is calculated with Equation 4.1.

$$AIC = 2k - 2L \tag{4.1}$$

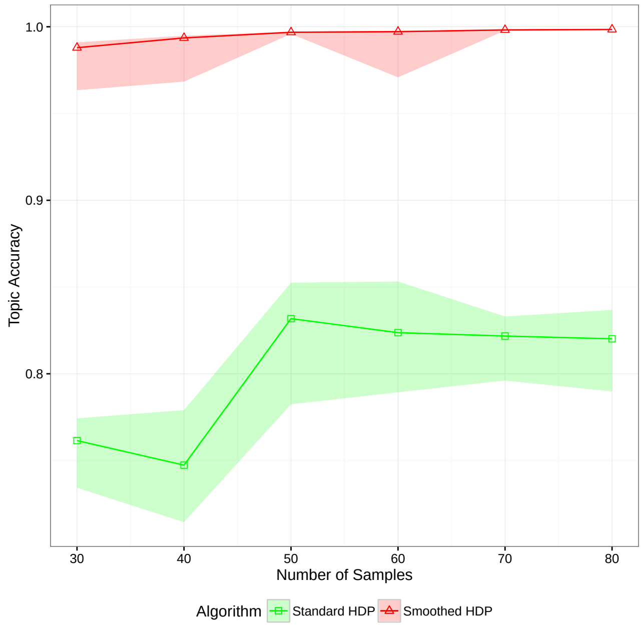
where k is the number of parameters and L is the log-likelihood of the model given data. AIC can give us an idea on relative quality of a model among the other models applied on the same dataset and it has no absolute meaning. It measures relative information loss when the model is applied on dataset, and it provides a trade-off between fit on data and model complexity, it has higher values for models with worse fit and more parameters. Lower AIC values mean better clustering performance. AIC value of HDP was higher than that of Smoothed HDP algorithms for all sparsity level and prior strength combinations. The AIC values with respect to sparsity levels and prior strengths can be seen in Figure 4.7(a) and Figure 4.7(b) respectively.

4.4 Experiments with Semi-synthetic Dataset

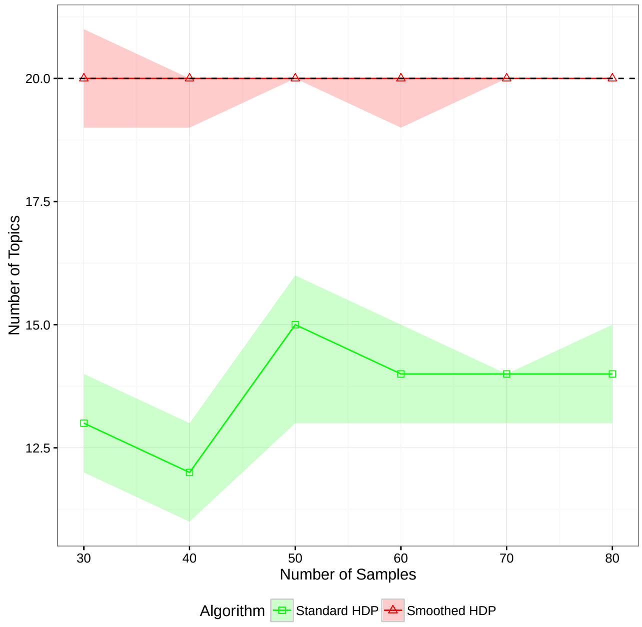
In this experiment, we tested our prior improvement approach with a semi-synthetic dataset whose preparation details, implementation and evaluation are explained in Section 4.4.1, Section 4.4.2 and Section 4.4.3 respectively. We used R and C++ programming languages, C++ code was modified from HDP code by Wang and Blei [93].

4.4.1 Data Preparation

We generated gene expression data using Syntren [20] which is able to simulate gene expression data in accordance with network interactions given and perturbation of se-

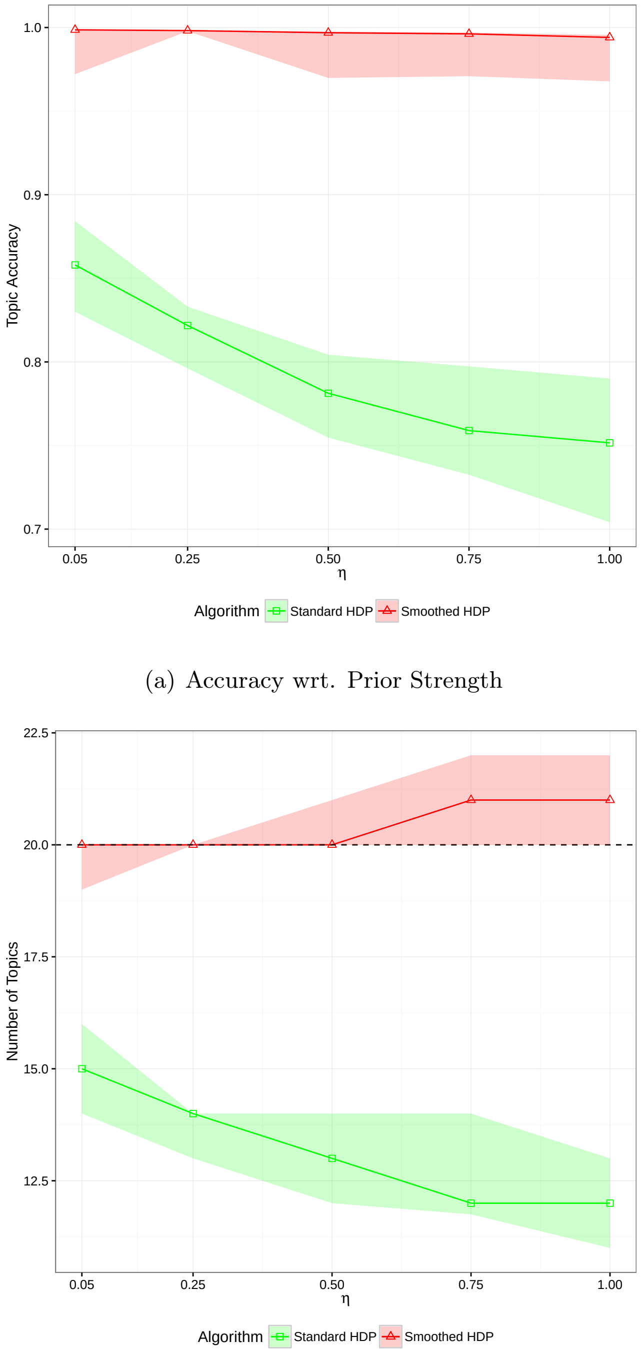


(a) Accuracy wrt. Sparsity



(b) Number of Topics wrt. Sparsity

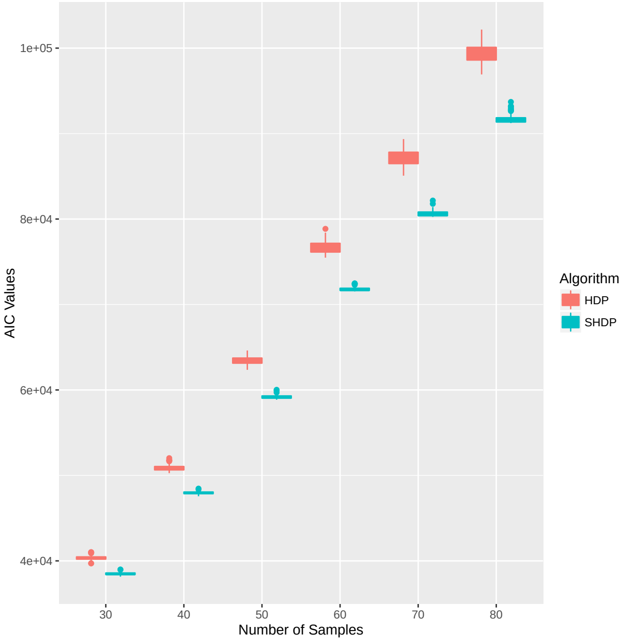
Figure 4.5: Prior Improvement Experiment: Accuracy and the Number of Topics with respect to Sparsity ($\eta = 0.25$). The data points shows medians over all runs, the bands show the interquartile ranges, the horizontal dashed lines represents the true number of topics (20)



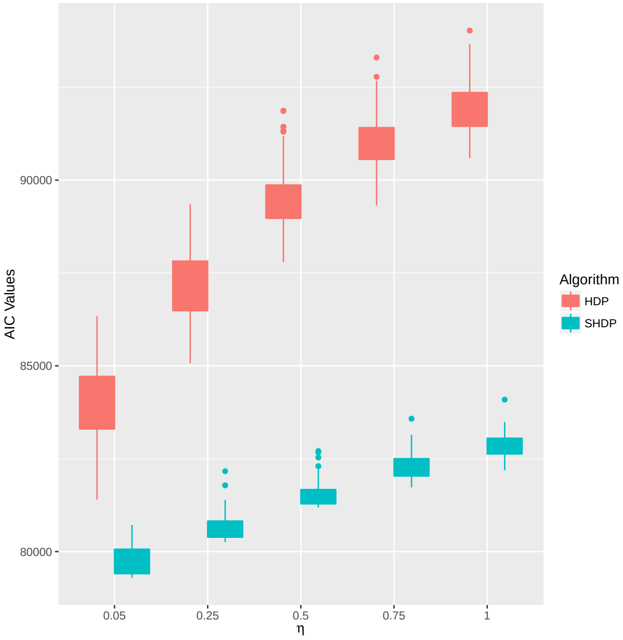
(a) Accuracy wrt. Prior Strength

(b) Number of Topics wrt. Prior Strength

Figure 4.6: Prior Improvement Experiment: Accuracy and the Number of Topics with respect to Prior Strength (The number of samples=70)



(a) AIC wrt. Sparsity



(b) AIC wrt. Prior Strength

Figure 4.7: AIC values with respect to Sparsity ($\eta = 0.25$) and Prior Strength(Number of samples $= 70$)

lected regulators. We used a yeast transcriptional network available in GeneNetWeaver [95] for both simulation of gene expression data and inference of prior information. This network was composed of 4440 genes and 12872 edges. 6342 of these edges were exhibitory and 6330 edges were inhibitory.

Since Syntren allows perturbation of genes with no indegree, we continued to work with the genes, YAL051W, YBL054W, YBR240C, YDL048C, YDL056W, YDR081C, YDR213W, YDR253C, YDR266C, YDR421W, YER108C, YER169W, YFL044C, YHR206W, YJL056C, YJL206C, YKL032C, YKR064W, YLR014C, YLR098C, YML113W, YMR021C, YMR042W, YOL067C, YOL089C, YOR113W, YOR363C, YOR380W, YPL038W, YPL089C, YPR199C. We generated gene expression data by fully activating each of these genes at different experiments while the rest are deactivated. In another experiment, we deactivated all regulators and used the gene expression data of this experiment as reference. Examining the difference between each activation experiment result and reference gave us the number of genes effected by the activation of each gene. The number of genes effected after perturbation of the regulators were 2224, 2, 40, 8, 2288, 3, 14, 46, 1, 32, 1, 19, 10, 62, 10, 10, 10, 1, 15, 14, 1, 8, 3, 1, 6, 1, 13, 1, 1, 7, 1 respectively. We selected the regulators which effect the number of genes between 13 and 62. YBR240C, YDR213W, YDR253C, YDR421W, YER169W, YHR206W, YLR014C, YLR098C, YOR363C were the selected regulators. We perturbed these 9 regulators to generate gene expression data with Syntren [20]. The gene expression induced by perturbation of individual regulator was considered as being a topic. By this way, we obtained true gene expression profile of each topic. The gene expression profiles of samples in the experiment were generated based on perturbation of genes with a Dirichlet distribution whose hyperparameter was 0.05.

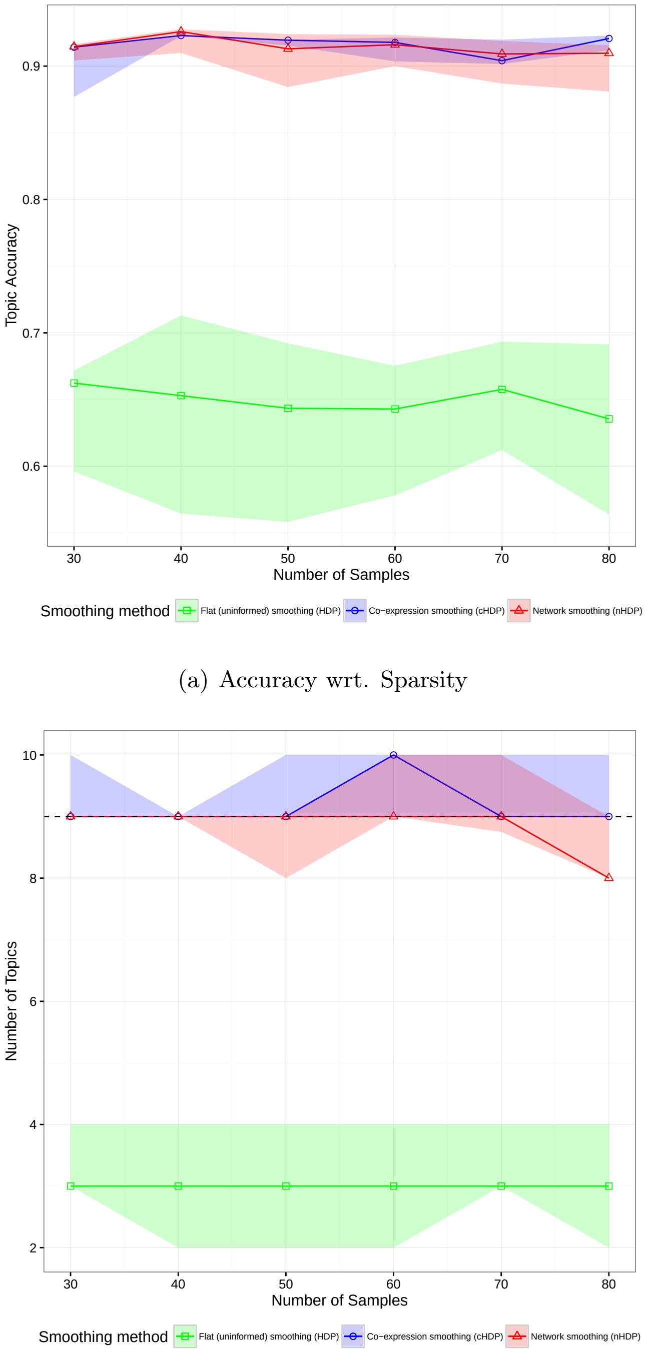
4.4.2 Implementation

The test was organized as 30, 40, 50, 60 and 70, 80 samples (sparsity levels). The experiments were repeated for the η values (prior strengths) of 0.05, 0.25, 0.5, 0.75 and 1. The tests were run 100 times for each combination of algorithm, sparsity level and prior strength resulting in 9000 runs.

4.4.3 Results

Gene expression profile of each original topic was compared to that of each topic retrieved in a run. The cosine similarity between the original topic and the most similar topic was regarded as the measure of topic accuracy.

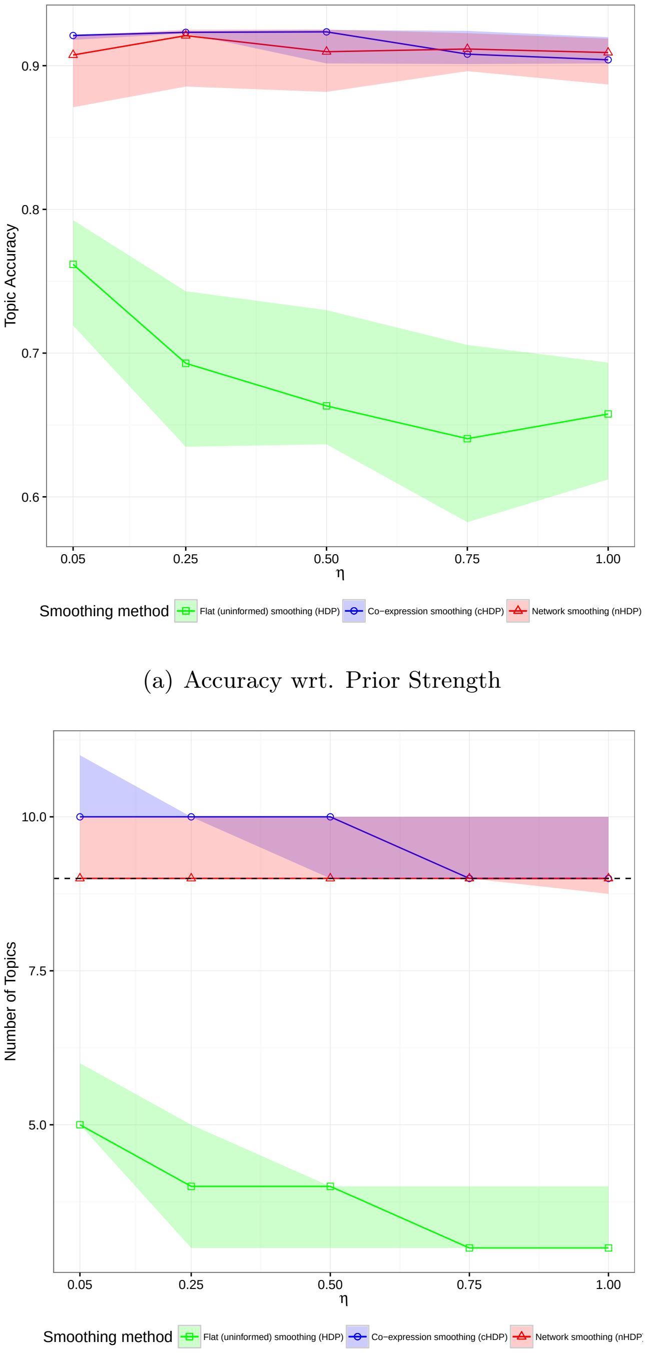
The topic accuracy values and the number of topics with respect to different sparsity levels can be seen in Figure 4.8(a) and Figure 4.8(b) respectively. Likewise the topic accuracy values and the number of topics with respect to different prior strengths can be seen in Figure 4.9(a) and Figure 4.9(b) respectively.



(a) Accuracy wrt. Sparsity

(b) Number of Topics wrt. Sparsity

Figure 4.8: Semi-synthetic Dataset Experiment: Accuracy and the Number of Topics with respect to Sparsity ($\eta = 1$). The data points shows medians over all runs, the bands show the interquartile ranges, the horizontal dashed lines represents the true number of topics (9)



(a) Accuracy wrt. Prior Strength

(b) Number of Topics wrt. Prior Strength

Figure 4.9: Semi-synthetic Dataset Experiment: Accuracy and the Number of Topics with respect to Prior Strength (The number of samples=70)

4.5 Evaluation on Prostate Cancer Dataset

4.5.1 Dataset

The dataset used for evaluation was taken from a prostate cancer study by Dhanasekaran et. al. [21]. In this dataset, there are 53 samples and 9984 genes. The distribution of the samples for classes is as follows: 14 samples for benign prostatic hyperplasia (BPH), 3 samples for normal adjacent prostate (NAP), 1 sample for normal adjacent tumor (NAT), 14 samples for localized prostate cancer (PCA), 1 sample for prostatitis (PRO) and 20 samples for metastatic tumors (MET). For evaluation purpose, the samples were considered in 3 macro-classes: non-cancer (BPH, NAP, PRO), cancer (NAT,PCA) and metastatic (MET) as it had been in [67] and [68].

4.5.2 Implementation and Results

In this study we used 500 genes which have the highest variance across samples. We applied the multiples-of-reference approach which is defined in Equation 3.7 to in order to convert gene expression data into how many times a given transcript type is seen in a sample. We prepared the H matrix as defined in Equation 3.10.

We repeated our experiment 100 times for each η value (0.1, 0.5, 2.5, 12.5) resulting in 400 runs. We performed evaluation by two different metrics. First is dependent on labels of samples. The topics that explain 0.9 probability of each sample were extracted and each sample's distance to remaining samples was calculated as Euclidean distance between their values on these topics. The most similar sample was found in 1 neighborhood and compared in terms of macro-class label (non-cancer, cancer, metastatic). The comparative success rate of HDP and Co-expression Smoothed HDP with respect to different η values can be seen in Figure 4.10.

The second evaluation criterion is the topic coherence metric which was defined by Mimno et. al. [22] and given in Section 2.5. According to this metric, it is expected that the likelihood of each representative gene given the other representative genes should be high in a high-quality topic. We extracted the genes that explain 30 percent of each topic and used as representative genes. The topic coherence values of HDP and Co-expression Smoothed HDP with respect to different η values can be seen in Figure 4.11.

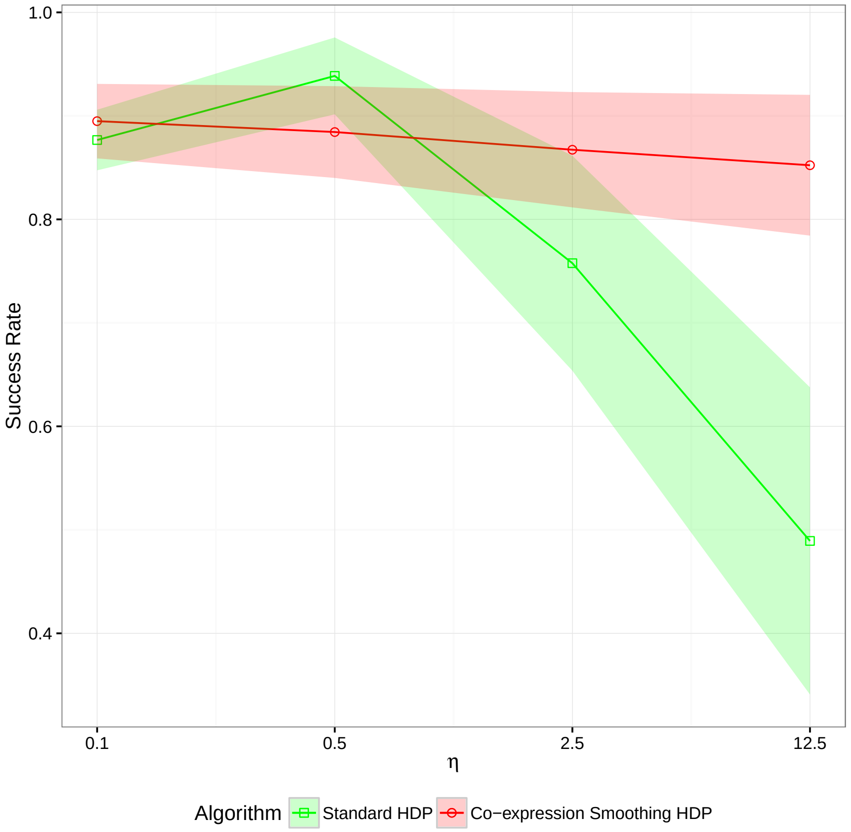


Figure 4.10: Accuracy vs. Prior Strength(η) The data points shows means over all runs, the bands show mean± sd

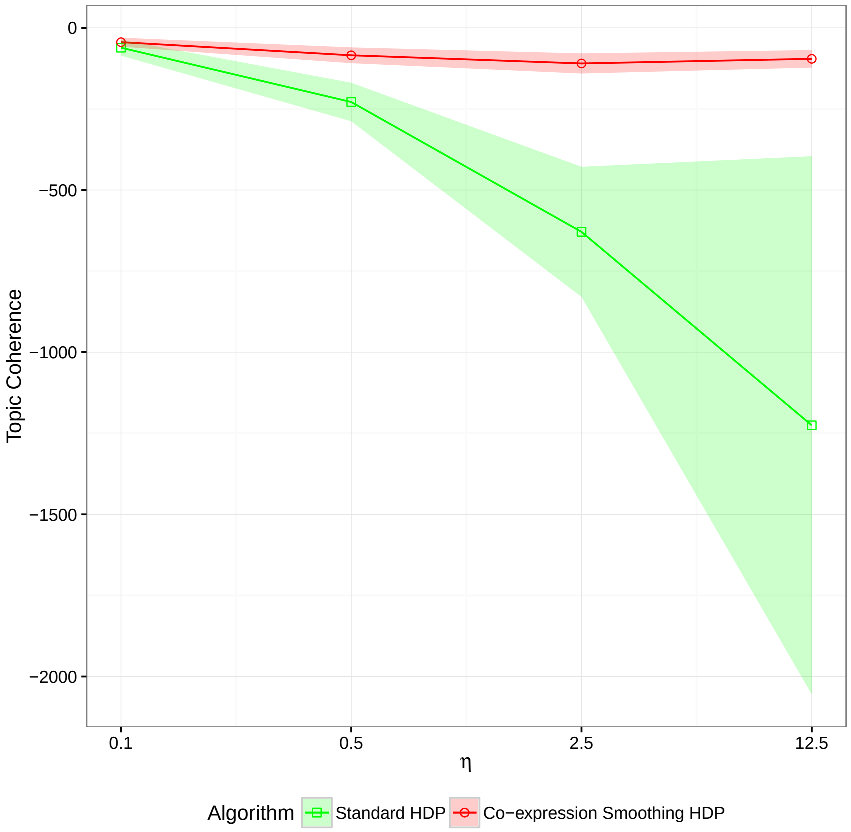


Figure 4.11: Topic Coherence vs. Prior Strength(η) The data points shows means over all runs, the bands show mean± sd

CHAPTER 5

DISCUSSION AND CONCLUSION

In this dissertation, we applied the nonparametric topic model, HDP, to gene expression data biclustering problem. We enhanced the original HDP model by incorporating gene-gene co-expression and co-regulation information as prior improvement. We defined two different prior improvement approaches, first is encoding an external gene regulatory network into co-regulation information and the second is encoding the correlation of gene expression matrix into co-expression information. We proved, through experiments, that HDP informed with either of the two prior improvement approaches, is more successful at biclustering gene expression data and robust to changes in hyper-parameter on transcript-topic distribution (η) compared to standard HDP.

The reason why we used a topic model, which is originally a text mining method, is that text mining and gene expression data analysis have a lot in common. First, in text mining there are stop words, like *the*, *an*, *is*, etc. which have almost the same frequency in every document and do not distinguish documents, we have house keeping (maintenance) genes whose expression level have low variation across different cells [96, 97], like $ERCC3$, $NR3C1$, $HPS1$ [97]. We filter housekeeping genes before starting gene expression analysis; the text miner, likewise, removes the stop words before text analysis.

The topics, co-occurance of semantically related words, are conceptually very similar to biclusters in bioinformatics domain. Recall the reason which forced the bioinformaticians to move on from clusters to biclusters, it was the need for capturing local patterns in gene expression data. The same applies to text domain, too. If a group of words together compose a semantically meaningful co-occurance, we do not expect to observe this togetherness in every document, we see the words together in only relevant documents, this is text domain's local pattern.

The gene expression matrix is similar to document-term matrix. Like the gene expression matrix represents mRNA extents of each gene of a sample in a row, document-term matrix represents the frequency of a term in a document in each of its rows. So the analogy is from gene expression data matrix to document-term matrix.

There is a similarity between genes and words. The words, the topics' constituents, may have polysemic usages, for example the word "fine" means well enough, on the other hand it also means punisment in terms of money for an offense. Namely, the word "fine" has different meaning in different contexts. Genes in different context contribute to different activities in a cell, their context sensitivity is like polysemy in text domain.

A topic is a probability distribution over genes in our analogy. This provides membership values of genes to biclusters. We can make GSEA as well as SEA since we can use the ranked list of genes in a bicluster with respect to their membership to the bicluster, traditional biclustering algorithms allow only SEA. GSEA's main advantage over SEA is avoiding the consequences of selecting an arbitrary threshold to decide "interesting genes" for gene set annotation [98]. The differences between these enrichment analysis methods were mentioned in 2.7.2.

Another similarity is between samples and documents. A document can be about more than one topic; for example, it may be about the impact of doing exercise on school success. So, we expect to see words of both sports and education topics in this document. Likewise, there may be more than one biclusters active in a sample. The weights of biclusters are also informative than the binary variable whether the bicluster is active in a sample. Let us give an example, if we work on a time series data, we can regard time points as conditions and we can observe the transition between different biological activities during the passage over the sequential time points. At least, having probabilistic features provides more information about sample similarities than binary features when we use bicluster distribution in a sample as the feature set.

Our study has three stages, first is preprocessing, that is converting gene expression data into transcript counts in samples to be able use topic model. Second is incorporating priors into topic model. Third is using the output of topic model at bioinformatics domain's disposal. At different steps, it has some commonalities and discrepancies with other studies. We want to explain the reasons why we have not followed similar path with the following studies. Namely, we will focus on discrepancies.

At the preprocessing we differ from the study by Rogers et. al. [67], they did not modify gene expression data to be able to use as input to the topic model, they instead modified the parametric topic model, LDA, to make it compatible with gene expression data and they named their model LPD. They used Gaussians for topic-gene distribution to model continuous nature of gene expression data. To us, it is more plausible to model gene expression data as multinomials because number of transcripts in a cell is discrete, despite gene expression seems to be continuous in microarray images. The study by Gerber et. al. [71] discretized gene expression data and used the nonparametric topic model, HDP, as is in this aspect. While they used multinomial variable for magnitude to represent gene expression level in a sample like us, their study differs in under- and over- expression definition. They used a Bernoulli indicator for the direction of differential expression for each transcript type, we instead centered the gene expression values against a reference, and used positive and negative values for observations in samples. Only positive integers can be used in multinomial distribution and to be able to use negative values, we splitted the individual observation variable into two different multinomial variables. Thus, the transcript type vocabulary size was doubled, first half represented over- and second half represented under-expression. So, both approaches use two variables for a transcript type. Our approach handles one issue that was ignored in their study. In the biological contexts where some of the genes' behaviors are volatile among samples (i.e. cases where the same context is defined by an "oscillation" of one or more genes), both under-expression and over-expression of the same transcript type can have high probabilities in a topic because they are independent variables.

Let us mention our main contribution; we incorporated prior gene co-regulation and co-expression information into HDP. There is another method (BaLDA) [68] which incorporated the gene dependencies into the LPD model [67] with a clustering module. This study takes external information into account but uses external information as constraints not as priors. If the prior information is not true, the model can not overcome this problem. In our approach, if data contradicts with prior information, prior information is set aside.

There is a recent study [69] which also used asymmetric priors in bioinformatics domain. Their approach is different from ours. In our study, we used a matrix whose individual row has a gene's co-expression/co-regulation information, we used a single row for each topic. In this study, they use the same asymmetric vector for all topics and this vector does not carry co-expression or co-regulation information. They used inverse gene frequency and the prior favors the biological annotation terms associated with fewer genes and they stipulate that this improvement contributes to generate more specific topics.

5.1 Discussion on Results of Experiments

In the first experiment (Section 4.1), we established HDP's performance in finding pre-seeded biclusters. In the result of biclustering two dimensional datasets of 2, 4 and 8 biclusters with HDP, we observed that all biclusters were recovered, there were additional topics (biclusters), some had few observations, some had the background image of all biclusters, and some were noise biclusters. To us, this is acceptable because need for removing junk topics have been mentioned and metrics have been defined for finding these topics in several papers [22,99–103]. One of these metrics [22] was mentioned in Section 2.5. All of these metrics are for text mining domain and some of them like [22] are applicable to our study, too. We have a stronger incoherence metric in our domain, incoherent topics are unlikely to annotate GO terms or KEGG pathways with significant p-values, so we can identify and remove them.

The second experiment was performed to prove that informed priors enhance the success of HDP algorithm in finding overlapping biclusters. We used the testbench created by Yee Whye Teh [19] to test the performance of standard HDP. We compared standard HDP and Smoothed HDP under different spasity levels (number of samples) and prior strengths (η values). Smoothed HDP outperformed standard HDP in each sparsity level and prior strength combination in both finding the correct number of topics and the transcript distribution over topics. We can see that HDP underestimated the correct number of topics. The reason for this, it merged different topics in the same topic because each transcript type has membership to two topics and the model could not differentiate the topics which have commonalities but also differences. In gene expression analysis, this shortcoming may cause to make gene set enrichment analysis results too general. Smoothed HDP algorithms will prevent us from overlooking topics which have commonalities but also differences from other topics. We are thus more likely to hit more specific (closer to the leaf terms) GO terms if we use Smoothed HDP algorithms in geneset enrichment analysis. The Smoothed HDP found a few additional topics with fewer tokens than the rest of the topics in high η values (0.75 and 1) and they can easily be removed by thresholding which was also the case in HDP runs but unrecognized because it already underestimated the number of topics. Note that the interquantile band of Smoothed HDP algorithm is tighter than HDP's in all experiments, this means the variability of results across runs is lower, thus the results are more reliable.

The third experiment was performed whether a model selection criterion, AIC, would favor SHDP rather than HDP. AIC value which provides a trade-off between fit on data and model complexity was calculated for each run. Lower AIC values mean better clustering performance and AIC value of HDP was higher than that of SHDP algorithm for all sparsity level and prior strength combinations. This experiment proved with internal indices, that SHDP is preferable over HDP. Note that the AIC values over SHDP runs have lower interquantile range than HDP runs and this is consistent with the other experiments' results.

The fourth experiment was performed on a semi-synthetic dataset generated by Syntren [20]. This experiment is the backbone of our study because it is a controlled gene perturbation experiment that is we already know the ground truth. In addition, the gene expression profiles of topics and samples were generated by an independent platform [20] which generates gene expression data similar to biological experimental data. We tested both of our prior smoothing approaches (Co-expression Smoothing, Network-based Smoothing) in this experiment. The results showed that prior smoothed HDP outperformed standard HDP by far. The results of this experiment were consistent with the results of the first experiment and this experiment set-up also have overlapping topics. The cosine similarity between topics based on their gene expression profiles can be seen in Figure 5.1.

We can see that Smoothed HDP is more successful at recovering the preseeded topics and also finding the correct number of topics. Smoothed HDP is robust to changes in prior strength levels. In contrast, HDP is effected heavily by the change in prior strength. Let us explain the reason for it. The prior information in standard HDP is flat that is probability of each transcript type in a topic is the same and the prior strength is the belief in this flat distribution. So, if the prior strength is high, the topics will start with a strong belief that every gene has the same chance to be in a topic. This causes underfitting and the model can not capture the relations between genes. If the prior strength is very small, a few genes will dominate the topic and the sampling process will start to open new topics for each small group of observations. This will cause high number of topics and execution time.

A topic is a Dirichlet compund multinomial distribution over transcript types that is the prior is Dirichlet distribution and the likelihood is multinomial distribution. In our approach, the prior distribution is asymmetric Dirichlet that is each topic starts with a belief that a specific group of genes is active in a topic. The initial probability of a transcript type in a topic is proportional the η_v^k which is the transcript type's element of the vector $\vec{\eta}^k$ (prior distribution of topic k) and first draw will be accordingly. The subsequent draws of the transcript v in sampling are proportional to number of times the transcript type v has been drawn in topic k plus η_v^k; so, the topic will have tendency to form around the genes with higher probabilities in the prior distribution. This is

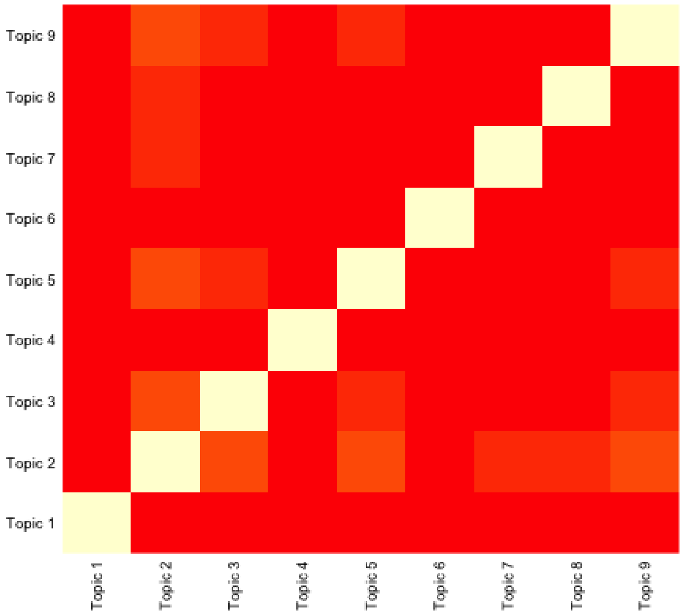


Figure 5.1: Original Topic Cosine Similarity

the key for success of Smoothed HDP algorithms. The reason why smoothed HDP algorithm is robust to changes in η values that the prior is consistent with likelihood, that is the degree of belief in the correct distribution does not effect the results. If the prior information was incorrect, the increase in η values would negatively effect the success.

The posterior distribution is the weighted average of prior and empiricial distributions that is the empricial distribution is smoothed by the prior distribution to compute the posterior distribution. As the evidence gets more, the prior distribution is dominated by the empirical distribution. There are two extreme cases. First is if there is no evidence, posterior distribution will the same with the prior distribution. Second is if there is infinite evidence, the posterior will be the same with the empirical distribution. In standard HDP, posterior distribution tends to be similar to uniform prior distribution in lack of evidence, but in smoothed HDP, as long as the prior information is correct,the prior reflects the underlying distribution of data, thus smoothed HDP is more succesful at finding the inherent structure in data even if there is not sufficent data to represent it.

We did not set a specific number of iterations for HDP and Smoothed HDP runs, we continued training for more 300 iterations after the best likelihood achieved by the models in each run. So, the low success rate of HDP is not attributable to convergence.

The time (till each algorithm reached its best likelihood) analyses of HDP, Co-expression Smoothing HDP and Network-based Smoothing HDP with respect to sparsity levels and prior strengths can be seen in Figure 5.2 and 5.3 respectively. The computer which the experiments were run on has a processor of 2,8 GHz Intel Core i7 and RAM of 16 GB 1600 MHz DDR3.

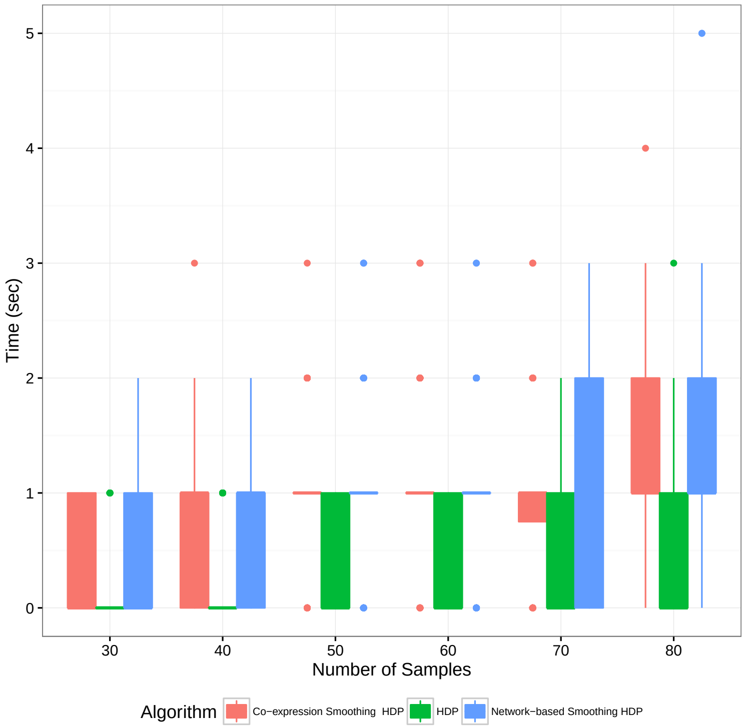


Figure 5.2: Time wrt. Sparsity Levels ($\eta = 1$)

The fifth experiment was performed on a real biological dataset, it was a prostate cancer dataset collected by Dhanasekaran et. al. [21]. In this experiment, we evaluated Co-expression Smoothing approach with two different criteria. First is based on their labels and the second is the topic coherence metric. In the first evaluation, HDP and Co-expression Smoothing HDP were used to reconstruct samples' feature set in order to reduce dimensionality. The topic proportions of each sample became its features. The topics that explain 0.9 probability of each sample were extracted and each sample's distance to remaining samples was calculated as Euclidean distance between their values on these topics. The most similar sample was found in 1 neighborhood and compared in terms of class label. In the second evaluation, the coherence metric which was defined by Mimno et. al. [22] and given in Section 2.5 was used. The genes that explain 30 percent of each topic were used as the representative genes and substituted in Equation 2.26. In both evaluation approaches, Co-expression smoothing HDP were robust to changes in hyperparameter η. In the first evaluation, Co-expression Smoothing HDP outperformed HDP in every η values but $\eta = 0.5$. Indeed, the label evaluation is less reliable than the rest of evaluation approaches we have used in this dissertion, because the aim of the unsupervised topic models is to maximize the posterior probability of the corpus. If the prevalent formation in the corpus is not relevant to labels, the reduced feature set will not be succesful at predicting the labels. Blei & McAuliffe [104] developed a topic model, supervised latent Dirichlet allocation (sLDA), to model documents and responses/labels together to overcome this problem.

The second evaluation approach was dependent on an internal topic coherence index.

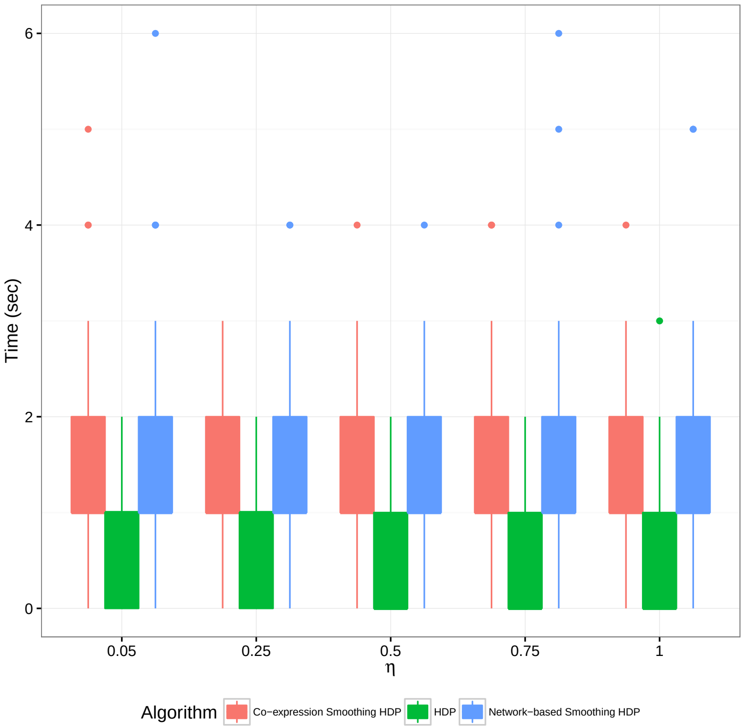


Figure 5.3: Time wrt. Prior Strength (Number of samples=80)

In this evaluation approach the representative genes of each topic were extracted and their co-occurance frequency in the samples were calculated and used as the comparison metric. This evaluation provided to measure quality of topics without any external labeling which may have been misleading. The results showed that Co-expression Smoothing HDP outperformed standard HDP in finding coherent topics with robustness to changes in prior strength η.

5.2 Future Work

The prior improvement we have proposed in this dissertation enables to use a single gene regulatory network in HDP prior and up to now, genome-wide gene regulatory network is used, tissue specific networks can be incorporated into prior information and each sample can use prior information derived from its own tissue type's network. GNAT [105] can be a source for tissue specific gene regulatory networks.

The model can further be improved by incorparating sample similarities either on the hyperparameter on sample-topic distribution(α), to our best knowledge there is no such a study, or grouping of samples into sample groups based on sample topic distribution during iterations, this approach was handled in Gerber et. al.'s study [71].

This study can be converted into a web application, which allows uploading gene expression data and gene regulatory network and gives sample-topic, gene-topic distribution in turn, for the users interested in probabilistic biclustering of gene expression data.

We have used HDP for discriminative purposes but in generative sense, new samples can be generated from the fitted model and we can simulate gene expression data similar to the gene expression data used in the input of the system and this can be used testing the robustness of algorithms developed for module detection or network extraction from gene expression data like WGCNA [106] and CoRegNet [107].

The proposed model can be used to find the most correct gene regulatory network from the existing ones by evaluating the output of Network-based Smoothing HDP. This evaluation can be done with either the topic coherence metric we have mentioned in Section 2.5 and applied to evaluation of prostate cancer example in Section 4.5 or the p-values of the geneset enrichment of the topics retrieved after the Network-based Smoothing HDP runs with different gene regulatory networks.

The top genes in a topic can be used for gene regulatory network construction or module detection. This can be achieved by combining the gene-topic distribution output of topic model (HDP or Smoothed HDP) with the information from transcription binding site databases like JASPAR [108] and tf-target gene databases like TRANSFAC [109], to our best knowledge there is no such a study.

The integrated analysis of different types of omics (genomics, proteomics, metabolomics etc.) data from a group of samples can be achieved using topic models by defining more than one word-topic distribution. Topic models specialized for this purpose can also be developed by modifying the original topic models according to the distribution of the data types of interest or relations among them.

The gist of our study is to prove that prior improvement works in exploring gene expression data with probabilistic topic models, we offered two differerent prior improvement approaches. Different prior improvement approaches can also be developed, for example the R package GOSim [110] defines gene similarities based on their GO term annotations. Their calculation methods can be kernels to prepare the H matrix defined in Section 3.1.3.

Researchers prefer to use HDP because it does not require the number of topics to be set beforehand but it may produce uneven size distribution over topics because of rich gets richer property of Dirichlet process, that is topics having high number of observations tend to attract new observations than topics with low number of observations, meaning that each topic may not have similar number of observations, there is a study by Wallach et.al. [111] to fix this issue but it is still an open problem.

There are topic model variants developed for specialized purposes, like the Author Topic Model [112] which was developed for exploring the relationships between authors, documents, topics, and words and Collaborative Topic-Model [113] which was developed as a recommender system matching readers, and scientific articles, both systems can be empowered with our prior improvement approach using WordNet [114] which is a network of semantic relations and similarities among words and can together form a helpful academic search engine.

REFERENCES

[1] Richard P Horgan and Louise C Kenny. 'Omic'technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician and Gynaecologist*, 13(3):189–195, 2011.

[2] Atul Butte. The use and analysis of microarray data. *Nature reviews drug discovery*, 1(12):951–960, 2002.

[3] Robert A. Weinberg. Tumor suppressor genes. *Science*, 254(5035):1138–1146, 1991.

[4] Eran Segal, Alexis Battle, and Daphne Koller. Decomposing gene expression into cellular processes. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 8, pages 89–100, 2002.

[5] Yizong Cheng and George M Church. Biclustering of expression data. In *The Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103. AAAI Press, 2000.

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

[7] Yuval Kluger, Ronen Basri, Joseph T Chang, and Mark Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, 13(4):703–716, 2003.

[8] Galina Gabriely, Nadiya M Teplyuk, and Anna M Krichevsky. Context effect: microRNA-10b in cancer cell proliferation, spread and death. *Autophagy*, 7(11):1384–1386, 2011.

[9] Cherie Blenkiron, Leonard D Goldstein, Natalie P Thorne, Inmaculada Spiteri, Suet-Feung Chin, Mark J Dunning, Nuno L Barbosa-Morais, Andrew E Teschendorff, Andrew R Green, Ian O Ellis, Simon Tavaré, Carlos Caldas, and Eric A Miska. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol*, 8(10):R214, 2007.

[10] Xionghui Zhou, Juan Liu, Changning Liu, Simon Rayner, Fengji Liang, Jingfang Ju, Yinghui Li, Shanguang Chen, and Jianghui Xiong. Context-specific mirna regulation network predicts cancer prognosis. In *Systems Biology (ISB), 2011 IEEE International Conference on*, pages 225–243. IEEE, 2011.

[11] Laura Lazzeroni and Art Owen. Plaid models for gene expression data. *Statistica Sinica*, 12(1):61–86, 2002.

[12] Jiong Yang, Haixun Wang, Wei Wang, and Philip Yu. Enhanced biclustering on expression data. In *Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on*, pages 321–327. IEEE, 2003.

[13] Sven Bergmann, Jan Ihmels, and Naama Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E*, 67(3):031902, 2003.

[14] Edoardo M Airoldi, David Blei, Elena A Erosheva, and Stephen E Fienberg. *Handbook of Mixed Membership Models and Their Applications.* CRC Press, 2014.

[15] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[16] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[17] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[18] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

[19] Yee Whye Teh. Nonparametric bayesian mixture models - release 1. `http://www.stats.ox.ac.uk/~teh/software.html`, 2004.

[20] Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(1):43, 2006.

[21] Saravana M Dhanasekaran, Terrence R Barrette, Debashis Ghosh, Rajal Shah, Sooryanarayana Varambally, Kotoku Kurachi, Kenneth J Pienta, Mark A Rubin, and Arul M Chinnaiyan. Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412(6849):822–826, 2001.

[22] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[23] Bani K Mallick, David Gold, and Veera Baladandayuthapani. *Bayesian analysis of gene expression data*, volume 131. John Wiley & Sons, 2009.

[24] Faramarz Valafar. Pattern recognition techniques in microarray data analysis. *Annals of the New York Academy of Sciences*, 980(1):41–64, 2002.

[25] Alex Sánchez and MC de Villa. A tutorial review of microarray data analysis. *Bioinformatics Tutorial, Universitat de Barcelona*, 2008.

[26] M Kathleen Kerr and Gary A Churchill. Statistical design and the analysis of gene expression microarray data. *Genetical Research*, 77(02):123–128, 2001.

[27] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4):e15–e15, 2003.

[28] Alvis Brazma and Jaak Vilo. Gene expression data analysis. *FEBS Letters*, 480(1):17–24, 2000.

[29] Robert Hitzemann, Daniel Bottomly, Priscila Darakjian, Nicole Walter, Ovidiu Iancu, Robert Searles, Beth Wilmot, and Shannon McWeeney. Genes, behavior and next-generation rna sequencing. *Genes, Brain and Behavior*, 12(1):1–12, 2013.

[30] Yongjun Chu and David R Corey. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics*, 22(4):271–274, 2012.

[31] Geng Chen, Charles Wang, and TieLiu Shi. Overview of available methods for diverse RNA-seq data analyses. *Science China Life Sciences*, 54(12):1121–1128, 2011.

[32] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):91, 2013.

[33] Matthew D Young, Davis J McCarthy, Matthew J Wakefield, Gordon K Smyth, Alicia Oshlack, and Mark D Robinson. Differential expression for RNA sequencing (RNA-seq) data: mapping, summarization, statistical analysis, and experimental design. In *Bioinformatics for High Throughput Sequencing*, pages 169–190. Springer, 2012.

[34] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, 11(3):R25, 2010.

[35] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440, 2007.

[36] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391, 1990.

[37] Susan T Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230, 2004.

[38] Wei Wang, Payam Mamaani Barnaghi, and Andrzej Bargiela. Probabilistic topic models for learning terminological ontologies. *Knowledge and Data Engineering, IEEE Transactions on*, 22(7):1028–1040, 2010.

[39] Richard E Neapolitan. *Learning Bayesian networks*, volume 38. Prentice Hall Upper Saddle River, 2004.

[40] Thomas Minka. Estimating a Dirichlet distribution, 2000.

[41] David M Blei and John D Lafferty. Topic models. *Text Mining: Classification, Clustering, and Applications*, 10(71):34, 2009.

[42] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.

[43] Samuel J Gershman and David M Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.

[44] Dilan Görür. *Nonparametric Bayesian discrete latent variable models for unsupervised learning.* PhD thesis, Berlin Institute of Technology, 2007.

[45] Yee Whye Teh. Dirichlet process. In *Encyclopedia of Machine Learning*, pages 280–287. Springer, 2010.

[46] David Blackwell and James B MacQueen. Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, pages 353–355, 1973.

[47] Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 1994.

[48] Emin Orhan. *Dirichlet Processes.* PhD thesis, Rochester University, 2012.

[49] Dongwoo Kim and Alice Oh. Accounting for data dependencies within a hierarchical dirichlet process mixture model. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 873–878. ACM, 2011.

[50] Yee Whye Teh and Michael I Jordan. Hierarchical bayesian nonparametric models with applications. *Bayesian Nonparametrics*, pages 158–207, 2010.

[51] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[52] Minoru Kanehisa. The kegg database. *In Silico Simulation of Biological Processes*, 247:91–103, 2002.

[53] Wei Yu, Marta Gwinn, Melinda Clyne, Ajay Yesupriya, and Muin J Khoury. A navigator for human genome epidemiology. *Nature Genetics*, 40(2):124–125, 2008.

[54] Thomas Engeitner. Enrichment analysis of go terms, kegg pathways and mirnas by using the hypergeometrical distribution. 2005.

[55] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009.

[56] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela V Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128, 2013.

[57] Manuele Bicego, Pietro Lovato, Barbara Oliboni, and Alessandro Perina. Expression microarray classification using topic models. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1516–1520. ACM, 2010.

[58] Xin Chen, Xiaohua Hu, Xiajiong Shen, and Gail Rosen. Probabilistic topic modeling for genomic data interpretation. In *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 149–152. IEEE, 2010.

[59] Xin Chen, Xiaohua Hu, Tze Yee Lim, Xiajiong Shen, EK Park, and Gail L Rosen. Exploiting the functional and taxonomic structure of genomic data by probabilistic topic modeling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(4):980–991, 2012.

[60] Manuele Bicego, Pietro Lovato, Alberto Ferrarini, and Massimo Delledonne. Biclustering of expression microarray data with topic models. In *2010 International Conference Pattern Recognition (ICPR)*, pages 2728–2731. IEEE, 2010.

[61] Patrick Flaherty, Guri Giaever, Jochen Kumm, Michael I Jordan, and Adam P Arkin. A latent variable model for chemogenomic profiling. *Bioinformatics*, 21(15):3286–3293, 2005.

[62] Tommy Chheng. Analyzing genes with topic modeling. 2007.

[63] José Caldas, Nils Gehlenborg, Ali Faisal, Alvis Brazma, and Samuel Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *BMC Bioinformatics*, 10(Suppl 13):P1, 2009.

[64] Manuele Bicego, Aydın Ulaş, Peter Schüffler, Umberto Castellani, Vittorio Murino, André Martins, Pedro Aguiar, and Mario Figueiredo. Renal cancer cell classification using generative embeddings and information theoretic kernels. In *Pattern Recognition in Bioinformatics*, pages 75–86. Springer, 2011.

[65] Bing Liu, Lin Liu, Anna Tsykin, Gregory J Goodall, Jeffrey E Green, Min Zhu, Chang Hee Kim, and Jiuyong Li. Identifying functional miRNA–mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics*, 26(24):3105–3111, 2010.

[66] Alessandro Perina, Pietro Lovato, Marco Cristani, and Manuele Bicego. A comparison on score spaces for expression microarray data classification. In *Pattern Recognition in Bioinformatics*, pages 202–213. Springer, 2011.

[67] Simon Rogers, Mark Girolami, Colin Campbell, and Rainer Breitling. The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(2):143–156, 2005.

[68] Alessandro Perina, Pietro Lovato, Vittorio Murino, and Manuele Bicego. Biologically-aware latent Dirichlet allocation (BaLDA) for the classification of expression microarray. In *Pattern Recognition in Bioinformatics*, pages 230–241. Springer, 2010.

[69] Pietro Pinoli, Davide Chicco, and Marco Masseroli. Latent dirichlet allocation based on gibbs sampling for gene function prediction. In *Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on*, pages 1–8. IEEE, 2014.

[70] Dimitrios V Vavoulis, Margherita Francescatto, Peter Heutink, and Julian Gough. Dgeclust: differential expression analysis of clustered count data. *Genome Biology*, 16(1):1, 2015.

[71] Georg K Gerber, Robin D Dowell, Tommi S Jaakkola, and David K Gifford. Automated discovery of functional generality of human gene expression programs. *PLoS Computational Biology*, 3(8):e148, 2007.

[72] Liming Wang and Xiaodong Wang. Hierarchical dirichlet process model for gene expression clustering. *EURASIP Journal on Bioinformatics and Systems Biology*, 2013(1):1–14, 2013.

[73] José Caldas and Samuel Kaski. Hierarchical generative biclustering for microRNA expression analysis. In *Research in Computational Molecular Biology*, pages 65–79. Springer, 2010.

[74] Hanna M Wallach, David M Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems*, pages 1973–1981, 2009.

[75] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Leveraging multi-domain prior knowledge in topic models. In *IJCAI*, 2013.

[76] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[77] James Petterson, Wray Buntine, Shravan M Narayanamurthy, Tibério S Caetano, and Alex J Smola. Word features for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 1921–1929, 2010.

[78] David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1171, 2011.

[79] Changyou Chen, Wray Buntine, Nan Ding, Lexing Xie, and Lan Du. Differential topic models. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):230–242, 2015.

[80] Wray L Buntine and Swapnil Mishra. Experiments with non-parametric topic models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 881–890. ACM, 2014.

[81] David Newman, Edwin V Bonilla, and Wray Buntine. Improving topic coherence with regularized topic models. In *Advances in Neural Information Processing Systems*, pages 496–504, 2011.

[82] David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32. ACM, 2009.

[83] Michael J Paul and Mark Dredze. Sprite: Generalizing topic models with structured priors. *Transactions of the Association for Computational Linguistics*, 3:43–57, 2015.

[84] Issei Sato and Hiroshi Nakagawa. Topic models with power-law using Pitman-Yor process. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 673–682. ACM, 2010.

[85] Jim Pitman. Combinatorial stochastic processes. Technical Report 621, Dept. of Statistics, UC Berkeley, 2002.

[86] David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

[87] Erik B Sudderth. *Graphical models for visual object recognition and tracking.* PhD thesis, Massachusetts Institute of Technology, 2006.

[88] Alessio Benavoli, Francesca Mangili, Fabrizio Ruggeri, and Marco Zaffalon. Imprecise Dirichlet process with application to the hypothesis test on the probability that x≤y. *Journal of Statistical Theory and Practice*, 9(3):658–684, 2015.

[89] Richard G Krutchkoff. Empirical Bayes estimation. *The American Statistician*, 26(5):14–16, 1972.

[90] Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, 13(1):328, 2012.

[91] Nikolay Kolesnikov, Emma Hastings, Maria Keays, Olga Melnichuk, Y Amy Tang, Eleanor Williams, Miroslaw Dylag, Natalja Kurbatova, Marco Brandizi, Tony Burdett, Karyn Megy, Ekaterina Pilicheva, Gabriella Rustici, Andrew Tikhonov, Helen Parkinson, Robert Petryszak, Ugis Sarkans, and Alvis Brazma. Arrayexpress update—simplifying data submissions. *Nucleic Acids Research*, page gku1057, 2014.

[92] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[93] Chong Wang and David Blei. Hierarchical Dirichlet process. `https://github.com/Blei-Lab/hdp`, 2010.

[94] Hirotugu Akaike. Akaike's information criterion. In *International Encyclopedia of Statistical Science*, pages 25–25. Springer, 2011.

[95] Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.

[96] Atul J Butte, Victor J Dzau, and Susan B Glueck. Further defining housekeeping, or "maintenance," genes focus on "a compendium of gene expression in normal human tissues". *Physiological genomics*, 7(2):95–96, 2001.

[97] Eli Eisenberg and Erez Y Levanon. Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574, 2013.

[98] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P. Mesirova. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[99] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

[100] Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. Topic significance ranking of LDA generative models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 67–82. Springer, 2009.

[101] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539, 2014.

[102] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pages 288–296, 2009.

[103] Nikolaos Aletras and Mark Stevenson. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 13–22, 2013.

[104] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 121–128, 2008.

[105] Emma Pierson, Daphne Koller, Alexis Battle, Sara Mostafavi, and GTEx Consortium. Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comput Biol*, 11(5):e1004220, 2015.

[106] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):1, 2008.

[107] Rémy Nicolle, François Radvanyi, and Mohamed Elati. CoRegNet: reconstruction and integrated analysis of co-regulatory networks. *Bioinformatics*, page btv305, 2015.

[108] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(suppl 1):D91–D94, 2004.

[109] Vea Matys, Ellen Fricke, R Geffers, Ellen Gößling, Martin Haubrock, R Hehl, Klaus Hornischer, Dagmar Karas, Alexander E Kel, and Olga V Kel-Margoulis. TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378, 2003.

[110] Holger Fröhlich, Nora Speer, Annemarie Poustka, and Tim Beißbarth. Gosim–an r-package for computation of information theoretic go similarities between terms and gene products. *BMC Bioinformatics*, 8(1):1, 2007.

[111] Hanna M Wallach, Shane Jensen, Lee H Dicker, and Katherine A Heller. An alternative prior process for nonparametric Bayesian clustering. In *AISTATS*, volume 9, pages 892–899, 2010.

[112] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence*, pages 487–494. AUAI Press, 2004.

[113] Chong Wang and David M Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 448–456. ACM, 2011.

[114] George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[115] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2012.

APPENDIX

Posterior Sampling By Direct Sampling

The Monte Carlo Markov Chain (MCMC) sampling mechanism used for posterior sampling in this study is Direct Assignment [115] and its mechanism is explained on a toy example.

Detailed Explanation of Direct Assignment Sampling Mechanism for SHDP on a Toy Example

We explain one iteration of posterior sampling on a toy example. Our example is from the bars problem which has been defined in Section 4.2. We are going to use a vocabulary of four word types, $V = (1, 2, 3, 4)$.

1	2	1^{st} topic
3	4	2^{nd} topic
3^{rd} topic	4^{th} topic	

Recall that each row and each column represented a topic.

Word distribution over 1^{st} topic is $[0.5, 0.5, 0, 0]$,

2^{nd} topic is $[0, 0, 0.5, 0.5]$,

3^{rd} topic is $[0.5, 0, 0.5, 0]$,

4^{th} topic is $[0, 0.5, 0, 0.5]$.

We generate 2 documents:

Doc1 is a mixture of 1^{st} and 4^{th} topics.

Doc1=[1,2, 2,4]

Doc2 is a mixture of 2^{nd} and 3^{rd} topics.

Doc2=[3 4 1 3]

We calculate the H matrix:

$$H = \begin{bmatrix} 0.5V\eta & 0.25V\eta & 0.25V\eta & 0 \\ 0.25V\eta & 0.5V\eta & 0 & 0.25V\eta \\ 0.25V\eta & 0 & 0.5V\eta & 0.25V\eta \\ 0 & 0.25V\eta & 0.25V\eta & 0.5V\eta \end{bmatrix}$$

$V = 4$ and $\eta = 0.5$ and we use a small constant (10^-6) instead of 0.

$$H = \begin{bmatrix} 1 & 0.5 & 0.5 & 10^-6 \\ 0.5 & 1 & 10^-6 & 0.5 \\ 0.5 & 10^-6 & 1 & 0.5 \\ 10^-6 & 0.5 & 0.5 & 1 \end{bmatrix}$$

Our hyperparameter for G_0 Dirichlet distribution is $\gamma = 1$ and G_j Dirichlet distributions is $\alpha = 1$.

One Iteration: Assume that we are running an iteration in Gibbs sampling, at this point there are 2 active topic, 1 table for each topic in each document($m_{1,1} = m_{1,2} = m_{2,1} = m_{2,2} = 1$, so $m_{.1} = 2$, $m_{.2} = 2$ where $m_{1,2}$ is the number of tables assigned to topic 2 in document 1 and $m_{.2}$ is the total number of tables assigned to topic 2). Doc1's first two tokens (1, 2) have been assigned to Topic1 whose prior is related to word type 1 and last two tokens (2,4) have been assigned to Topic2 whose prior is related to word type 2. Doc2's first two tokens have been assigned to Topic1 and last two tokens have been assigned to Topic2.

Sampling $\boldsymbol{\pi}$:

$$(\pi_1, \ldots, \pi_k, \pi_{new})|. \sim Dir(m_{.1}, \ldots, m_{.K}, \gamma) \qquad (.1)$$

1. **Calculate the global topic distribution, $\boldsymbol{\pi}$:**

$(\pi_1, \pi_2, \pi_{new}) = Dir(m_{.1}, m_{.2}, \gamma) = Dir(2, 2, 1)$

$\pi_1 = 0.6788572, \pi_2 = 0.2680592, \pi_{new} = 0.05308357$

2. **Calculate $\boldsymbol{z_{ji}}$ for each token of each document:**

   Sampling $\boldsymbol{z}$:

$$p(z_{ji} = k|.) \propto \begin{cases} (n_{jk}^{-ji} + \alpha\pi_k) f_k^{-x_{ji}}, & \text{if k previously used} \\ \\ \alpha\pi_{new} f_{k^{new}}^{-x_{ji}}(x_{ji}), & \text{if } k = k_{new} \end{cases} \tag{.2}$$

where n_{jk}^{-ji} is the number of tokens assigned to topic k in document j excluding the token in question.

Doc1's first token is 1:

$z_{1,1} = 1|. \propto n_{jk}^{\neg i} + \alpha\pi_k \cdot \frac{n_{k,v=x_{ji}}^{\neg i} + \eta_{v=x_{ji}}^k}{n_k^{\neg i} + \sum_{v=1}^V \eta_v^k} = (1 + 1 * 0.678857). \frac{(0+1)}{(3+2)} = 0.5357714$

$z_{1,1} = 2|. \propto (2 + 1 * 0.2680592). \frac{(1+0.5)}{(4+2)} = 0.5670148$

$z_{1,1} = k_{new}|. \propto \frac{\alpha\pi_{new}}{V} = \frac{1*0.053083574}{4} = 0.01327089$

Let's say this token is assigned to Topic2.

Doc1's second token is 2:

$z_{1,2} = 1|. \propto (0 + 1 * 0.678857). \frac{(0+0.5)}{(2+2)} = 0.08485713$

$z_{1,2} = 2|. \propto (3 + 1 * 0.2680592). \frac{(1+1)}{(5+2)} = 0.9337312$

$z_{1,2} = k_{new}|. \propto \frac{1*0.053083574}{4} = 0.01327089$

It is assigned to Topic2.

Doc1's third token is 2:

$z_{1,3} = 1|. \propto (0 + 1 * 0.678857). \frac{(0+0.5)}{(2+2)} = 0.08485713$

$z_{1,3} = 2|. \propto (3 + 1 * 0.2680592). \frac{(1+1)}{(5+2)} = 0.9337312$

$z_{1,3} = k_{new}|. \propto \frac{1*0.053083574}{4} = 0.01327089$

It is assigned to Topic2.

Doc1's fourth token is 4:

$z_{1,4} = 1|. \propto (0 + 1 * 0.678857). \frac{(1+10^{-6})}{(2+2)} = 0.1697144$

$z_{1,4} = 2|. \propto (3 + 1 * 0.2680592).\frac{(0+0.5)}{(4+2)} = 0.2723383$

$z_{1,4} = k_{new}|. \propto \frac{1*0.053083574}{4} = 0.01327089$

It is assigned to Topic1.

Doc2's first token is 3:

$z_{2,1} = 1|. \propto n_{jk}^{\neg i} + \alpha\pi_k \cdot \frac{n_{k,v=x_{ji}}^{\neg i} + \eta_{v=x_{ji}}^k}{n_k^{\neg i} + \sum_{v=1}^V \eta_v^k} = (1 + 1 * 0.678857).\frac{(0+0.5)}{(2+2)} = 0.2098571$

$z_{2,1} = 2|. \propto (2 + 1 * 0.2680592).\frac{(1+10^{-6})}{(5+2)} = 0.03829421$

$z_{2,1} = k_{new}|. \propto \frac{1*0.053083574}{4} = 0.01327089$

Let's say this token is assigned to Topic1.

Doc2's second token is 4:

$z_{2,2} = 1|. \propto (1 + 1 * 0.678857).\frac{(1+10^{-6})}{(2+2)} = 0.4197147$

$z_{2,2} = 2|. \propto (2 + 1 * 0.2680592).\frac{(0+0.5)}{(5+2)} = 0.1620042$

$z_{2,2} = k_{new}|. \propto \frac{1*0.053083574}{4} = 0.01327089$

Let's say, it has started a new topic (3^{rd}) and this topic's prior distribution is related to the fourth row of H matrix since word type 4 started the new topic.

Now we need to calculate π_3, $\pi_3 = Beta(1, \gamma) * \pi_{new} = 0.9054117 * 0.053083574 = 0.04806249$

π_{new} is updated, it is $0.053083574 - 0.04806249 = 0.005021084$

Doc2's third token is 1:

$z_{2,3} = 1|. \propto (1 + 1 * 0.678857).\frac{(0+1)}{(2+2)} = 0.1697143$

$z_{2,3} = 2|. \propto (1 + 1 * 0.2680592).\frac{(1+0.5)}{(4+2)} = 0.3170148$

$z_{2,3} = 3|. \propto (1 + 1 * 0.04806249).\frac{(0+10^{-6})}{(1+2)} = 3.493542 * 10^{-7}$

$z_{1,3} = k_{new}|. \propto \frac{1*0.005021084}{4} = 0.001255271$

It is assigned to Topic2.

Doc2's fourth token is 3:

$z_{2,4} = 1|. \propto (1 + 1 * 0.678857).\frac{(1+0.5)}{(2+2)} = 0.6295714$

$z_{2,4} = 2|. \propto (1 + 1 * 0.2680592).\frac{(0+10^{-6})}{(4+2)} = 2.113432 * 10^{-7}$

$z_{2,4} = 3|. \propto (1 + 1 * 0.04806249).\frac{(0+0.5)}{(1+2)} = 0.1746771$

$z_{2,4} = k_{new}|. \propto \frac{1*0.053083574}{4} = 0.01327089$

It is assigned to Topic1.

3. **Sample table counts of each topic for each document:**

   Sampling $\boldsymbol{m}$:

$$p(m_{jk} = m|.) = \frac{\Gamma(\alpha\beta_k)}{\Gamma(\alpha\beta_k + n_{jk})} s(n_{jk}, m)(\alpha\beta_k)^m \tag{.3}$$

where

$$s(n,m) = \begin{cases} 0, & \text{if } n = 0, \quad m = 0 \\ 1, & \text{if } n = 1, \quad m = 1 \\ 0, & \text{if } n > 0, \quad m = 0 \\ 0, & \text{if } m > n \\ 0, & \text{if } m > n \\ s(n-1, m-1) + (n-1)s(n-1, m) & \text{otherwise} \end{cases} \tag{.4}$$

For the first document:

For Topic1:

There is a single token assigned to Topic1 so $m_{1,1} = 1$, $m_{.1} = 2$

For Topic 2:

$P(m_{1,2} = 1|.) = \frac{\Gamma(1*0.2680592)}{\Gamma(1*0.2680592+3)} * 2 * (1 * 0.2680592)^1 = 0.6954022$

$P(m_{1,2} = 2|.) = \frac{\Gamma(1*0.2680592)}{\Gamma(1*0.2680592+3)} * 3 * (1 * 0.2680592)^2 = 0.2796134$

$P(m_{1,2} = 3|.) = \frac{\Gamma(1*0.2680592)}{\Gamma(1*0.2680592+3)} * 1 * (1 * 0.2680592)^3 = 0.02498432$

$m_{1,2} = 1$, $m_{.2} = 2$

For Topic 3:

Since there is no token assigned to Topic3 in Doc1 $m_{1,3} = 0$

For the second document:

For Topic1:

$P(m_{2,1} = 1|.) = \frac{\Gamma(1*0.678857)}{\Gamma(1*0.678857+2)} * 1 * (1 * 0.678857)^1 = 0.5956433$

$P(m_{2,1} = 2|.) = \frac{\Gamma(1*0.678857)}{\Gamma(1*0.678857+2)} * 1 * (1 * 0.678857)^2 = 0.4043567$

$m_{2,1} = 2$, $m_{.1} = 3$

For Topic2:

Since there is single token assigned to Topic2 in Doc2 $m_{2,2} = 1$ $m_{.2} = 2$

For Topic3:

Since there is single token assigned to Topic3 in Doc2 $m_{2,3} = 1$ $m_{.3} = 1$

Hence, one iteration is completed; the new iteration is going to start with π sampling as $(\pi_1, \pi_2, \pi_3, \pi_{new}) = Dir(m_{.1}, m_{.2}, m_{.3}, \gamma) = Dir(3, 2, 1, 1)$

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Çelik Tercan, Bahar

Nationality: Turkish (TC)

Date and Place of Birth: 19.04.1984, Mardin

Marital Status: Married

E-mail: baharc@metu.edu.tr, tercanbahar@gmail.com

EDUCATION

Degree	**Institution**	**Year of Graduation**
MS	Gazi University, Electronics and Computer Education	2009
BS	Selçuk University, Computer Systems Education	2006
BS	Anadolu University, Public Administration	2009
BS	Karabük University, Computer Engineering	2014
High School	Süleyman Demirel Anatolian High School	2002

PROFESSIONAL EXPERIENCE

Year	**Place**	**Enrollment**
2010 -. . .	METU, Department of Health Informatics	Research Assistant
2007-2010	Duzce University, Computer Systems Education	Research Assistant
2006-2007	Ministry of National Education	Technical Teacher

PUBLICATIONS

SCI Publications

Bahar Çelik, Nihal Fatma Güler and İnan Güler, "Design and Realization of a Microcontroller Based E-Test Strip Application Device", Instrumentation Science and Technology 37.6 (2009): 676-682.

Bahar Tercan, Aybar C. Acar, "Externally Smoothed Mixed Membership Models for Gene Expression Analysis", (sent to a journal)

International Conference Publications

Bahar Tercan, Aybar C. Acar, "Context-Dependent Gene Regulatory Mechanism Identification with Probabilistic Topic Models", HIBIT 2013, Ankara, TURKEY (Poster)

Bahar Tercan, Aybar C. Acar, "An LDA - WGCNA Hybrid Method for Better Biclustering of Expression Data", HIBIT 2015, Muğla, TURKEY (Poster)

National Conference Publications

Bahar Tercan, Cengiz Çelik, "Türkiye'de Sağlık Bilişimi Alanındaki Gelişmeler", 2. Ulusal Yönetim Bilişim Sistemleri Kongresi, Erzurum, TÜRKİYE, 8-10 Ekim 2015, cilt.1, no.978-975-442-738-7, ss.1453-1458

Projects

Project: "Enhancement of Student Cooperation in Computer Education" Institution: Ministry of National Education (ICT Seagulls) (Executer: Bahar Çelik) (2006-2007)

Project: "Definition of MicroRNA- Transcription Factor Regulating Networks from Gene Expression Data Using Probabilistic Topic Models", Institution: Middle East Technical University (Executer: Dr. Aybar C. Acar) (2013-2015)

Project: "Correct diet to inpatients", Institution: Ministry of Science, Industry and

Technology (Executer: Bahar Tercan) (2015-2016)