Bayesian Nonparametric Learning
for Complicated Text Mining

Junyu Xuan

A thesis submitted for the Degree of
Doctor of Philosophy

Faculty of Engineering and Information Technology
University of Technology Sydney
October, 2016
CERTIFICATE OF AUTHORSHIP/ORGINALITY

This thesis is the result of a research candidature conducted jointly with another University as part of a collaborative Doctoral degree. I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree except as fully acknowledged within the text.

I also certify that this thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate
ACKNOWLEDGEMENTS

I would like to express my earnest thanks to my principal supervisor, Professor Jie Lu, and my co-supervisor, Professor Guangquan Zhang. Their comprehensive guidance has covered all aspects of my PhD study, including research methodology, research topic selection, experiments, academic writing skills and thesis writing, even sentence structure and formulas. Their critical comments and suggestions have strengthened my study significantly. Their strict academic attitude and respectful personalities have benefited my PhD study and will be a great treasure throughout my life. Without their excellent supervision and continuous encouragement, this research could not have been finished on time. Thanks to you both for your kind help! I would like to thank Dr. Richard Yi Da Xu for his advice on this study. Without his help it would not have been possible for me to plunge into this area and conduct this research.

I am grateful to all members of the Decision Systems and e-Service Intelligent (DeSI) Lab in the Centre for Quantum Computation and Intelligent Systems (QCIS) for their careful participation in my presentation and valuable comments on my research. I would like to thank Ms. Sue Felix and Jem Moore for helping me to proofread my publications and this thesis.

I wish to express my appreciation for the financial support I received for my study. Special thanks go to the China Scholarship Council (CSC) and the University of Technology Sydney (UTS).

Last but not least, I would like also to thank my family members. Thanks to
my mother, father and sisters for their conscious encouragement and generous support.
ABSTRACT

Text mining has gained the ever-increasing attention of researchers in recent years because text is one of the most natural and easy ways to express human knowledge and opinions, and is therefore believed to have a variety of application scenarios and a potentially high commercial value. It is commonly accepted that Bayesian models with finite-dimensional probability distributions as building blocks, also known as parametric topic models, are effective tools for text mining. However, one problem in existing parametric topic models is that the hidden topic number needs to be fixed in advance. Determining an appropriate number is very difficult, and sometimes unrealistic, for many real-world applications and may lead to over-fitting or under-fitting issues. Bayesian nonparametric learning is a key approach for learning the number of mixtures in a mixture model (also called the model selection problem), and has emerged as an elegant way to handle a flexible number of topics. The core idea of Bayesian nonparametric models is to use stochastic processes as building blocks, instead of traditional fixed-dimensional probability distributions. Even though Bayesian nonparametric learning has gained considerable research attention and undergone rapid development, its ability to conduct complicated text mining tasks, such as: document-word co-clustering, document network learning, multi-label document learning, and so on, is still weak. Therefore, there is still a gap between the Bayesian nonparametric learning theory and complicated real-world text mining tasks.
To fill this gap, this research aims to develop a set of Bayesian nonparametric models to accomplish four selected complex text mining tasks. First, three Bayesian nonparametric sparse nonnegative matrix factorization models, based on two innovative dependent Indian buffet processes, are proposed for document-word co-clustering tasks. Second, a Dirichlet mixture probability measure strategy is proposed to link the topics from different layers, and is used to build a Bayesian nonparametric deep topic model for topic hierarchy learning. Third, the thesis develops a Bayesian nonparametric relational topic model for document network learning tasks by a subsampling Markov random field. Lastly, the thesis develops Bayesian nonparametric cooperative hierarchical structure models for multi-label document learning task based on two stochastic process operations: inheritance and cooperation. The findings of this research not only contribute to the development of Bayesian nonparametric learning theory, but also provide a set of effective tools for complicated text mining applications.
# TABLE OF CONTENT

CERTIFICATE OF AUTHORSHIP/ORGINALITY ................................................................. iii

ACKNOWLEDGEMENTS ......................................................................................... v

ABSTRACT ........................................................................................................... vii

TABLE OF CONTENT ......................................................................................... ix

LIST OF FIGURES .............................................................................................. xiii

LIST OF TABLES .................................................................................................. xix

Chapter 1 Introduction .................................................................................... 1
  1.1 Background ............................................................................................... 1
  1.2 Research questions and objectives ............................................................ 5
  1.3 Research contributions ............................................................................. 9
  1.4 Research significance ............................................................................... 10
  1.5 Thesis structure ....................................................................................... 11
  1.6 Publications related to this thesis ............................................................. 14

Chapter 2 Literature Review .......................................................................... 17
  2.1 Definitions of Bayesian nonparametric learning ....................................... 17
  2.2 Basic ingredients: stochastic processes ..................................................... 19
  2.3 Manipulations of stochastic processes ....................................................... 23
    2.3.1 Marginalization .................................................................................. 23
    2.3.2 Layering ........................................................................................... 25
    2.3.3 Superposition ................................................................................... 27
    2.3.4 Subsampling ..................................................................................... 28
    2.3.5 Point-transition ............................................................................... 30
    2.3.6 Nesting .............................................................................................. 30
  2.4 Supervised learning ................................................................................. 31
  2.5 Tree structure learning .......................................................................... 32
  2.6 Nonparametric extensions ...................................................................... 33
# TABLE OF CONTENT

2.7 Posterior inference .................................. 36  
  2.7.1 Slice sampling .................................. 36  
  2.7.2 Variational inference ........................... 37  
  2.7.3 Scalability .................................... 39  
  2.7.4 Others ........................................ 40  
2.8 Application scenarios ................................. 40  
  2.8.1 Text mining .................................. 41  
  2.8.2 Natural language processing ..................... 42  
  2.8.3 Computer vision ............................... 42  
  2.8.4 Biology .................................. 43  
  2.8.5 Music analysis ................................ 44  
2.9 Summary ....................................... 45  

Chapter 3 Bayesian Nonparametric Sparse Nonnegative Matrix Factorization for Document-word Co-clustering 47  
3.1 Introduction ...................................... 47  
3.2 Preliminary knowledge .............................. 49  
  3.2.1 Sparse nonnegative matrix factorization ........ 49  
  3.2.2 Indian buffet process .......................... 51  
3.3 Doubly sparse nonparametric NMF .................. 51  
  3.3.1 Doubly sparse nonparametric NMF framework ...... 52  
  3.3.2 Implementation by GP-based dIBP .............. 54  
  3.3.3 Implementation by bivariate Beta distribution-based dIBP 56  
  3.3.4 Implementation by copula-based dIBP .......... 59  
  3.3.5 Models discussion ............................ 60  
3.4 Model inference .................................... 62  
  3.4.1 Inference for the GP-based dIBP .............. 62  
  3.4.2 Update stick weights .......................... 64  
  3.4.3 Update binary matrices ....................... 67  
  3.4.4 Update loading matrices ....................... 68  
  3.4.5 Update model parameter ....................... 68  
3.5 Experiments ...................................... 70  
  3.5.1 Evaluation on sparsity and nonparametric properties ... 70  
  3.5.2 Evaluation on correlation flexibility .......... 74  
  3.5.3 Real-world task: document-word co-clustering .... 76  
3.6 Summary ....................................... 83  

Chapter 4 Bayesian Nonparametric Deep Topic Model for the Topic Hierarchy Learning 85  
4.1 Introduction ..................................... 85  
4.2 Problem statement ................................ 87
TABLE OF CONTENT

4.3 Bayesian nonparametric deep topic model ................................. 88
4.4 Model inference ........................................................................ 98
4.5 Experimental evaluation ............................................................. 106
  4.5.1 Synthetic data ................................................................... 108
  4.5.2 Real-world data .................................................................. 110
4.6 Summary ................................................................................. 115

Chapter 5 Bayesian Nonparametric Relational Topic Model for Document Network Learning 119
  5.1 Introduction ............................................................................ 119
  5.2 Preliminary knowledge ............................................................. 122
    5.2.1 Relational topic model ..................................................... 122
    5.2.2 Markov random field ....................................................... 123
  5.3 Bayesian nonparametric relational topic model ......................... 124
  5.4 Model inference ................................................................. 132
    5.4.1 Gibbs sampling .............................................................. 133
    5.4.2 Slice sampling .............................................................. 136
  5.5 Experiments ........................................................................... 139
    5.5.1 Evaluation metrics .......................................................... 140
    5.5.2 Experiments on synthetic data ......................................... 141
    5.5.3 Experiments on real-world data ....................................... 147
  5.6 Summary ................................................................................. 153

Chapter 6 Bayesian Nonparametric Cooperative Hierarchical Structure Models for Multi-label Document Learning 155
  6.1 Introduction ............................................................................ 155
  6.2 Cooperative hierarchical structure ............................................ 157
  6.3 Cooperative hierarchical Dirichlet processes ......................... 161
    6.3.1 Stick-breaking representation .......................................... 166
    6.3.2 International restaurant process representation ............... 173
  6.3.3 Model analysis ................................................................. 180
  6.4 Mixed Gamma-negative binomial processes ............................. 184
    6.4.1 Model description .......................................................... 184
    6.4.2 Model inference ............................................................ 189
    6.4.3 Model analysis ............................................................. 196
  6.5 Experimental evaluation .......................................................... 201
    6.5.1 Author-topic model task .................................................. 201
    6.5.2 Clinical free text labeling task ........................................ 205
  6.6 Summary ................................................................................. 208

Chapter 7 Conclusion and Further Study ............................................. 209
## TABLE OF CONTENT

7.1 Conclusions .............................................. 209  
7.2 Further study .............................................. 212

**REFERENCES**  
215

**ABBREVIATIONS**  
243
LIST OF FIGURES

1.1 Thesis structure ..................................................... 12

2.1 The relations between stochastic processes. .................. 19

3.1 Graphical representations for (a) the original Indian Buffet Process, (b) GP-based dependent Indian Buffet Processes and (c) Bivariate Beta distribution-based or Copula-based dependent Indian Buffet Processes. ........................................ 53

3.2 FGM copula density surfaces with different values of \( \rho = \{-1, 0, 1\} \) and the marginal distributions are both Beta distributions: \( \text{Beta}(1, 1) \) (means that \( \alpha_1 = \alpha_2 = 1 \) in Eq. (3.15)). ......................... 59

3.3 Illustration of the meaning of learned correlation. The red/solid curve denotes a focused distribution on words from a large negative correlation; the blue/dashed curve denotes a non-focused distribution on words from a large positive correlation. ........ 61

3.4 Comparison of the sparsity on synthetic dataset between traditional NMF (NMF), traditional sparse NMF (sNMF), single IBP-based sparse NMF (sIBP-NMF), and doubly IBP-based sparse NMF (dIBP-NMF). Note that dIBP-NMF here is based on bivariate beta distribution. .................................................. 73

3.5 Comparison of the learned topic number distribution on synthetic dataset between single IBP-based sparse NMF (sIBP-NMF) and doubly IBP-based sparse NMF (dIBP-NMF). Note that dIBP-NMF here is based on bivariate beta distribution. ................. 74

3.6 Results on synthetic data to show the flexibility of the different models. The x-axis denotes the trial IDs (order is irrelevant). ........... 75

3.7 Document-word co-clustering (document side) comparisons on Citeseer dataset between GP-dIBP-NMF, BB-dIBP-NMF, C-dIBP-NMF, sIBP-NMF, and SNMF. .................................................. 77

3.8 Document-word co-clustering (word side) comparisons on Citeseer dataset between GP-dIBP-NMF, BB-dIBP-NMF, C-dIBP-NMF, sIBP-NMF, and SNMF. .................................................. 78
LIST OF FIGURES

3.9 Document-word co-clustering (document side) comparisons on Cora dataset between GP-dIBP-NMF, BB-dIBP-NMF, C-dIBP-NMF, sIBP-NMF, and SNMF. ................................................. 79

3.10 Document-word co-clustering (word side) comparisons on Cora dataset between GP-dIBP-NMF, BB-dIBP-NMF, C-dIBP-NMF, sIBP-NMF, and SNMF. ................................................. 80

4.1 An example of Topic Hierarchy. The orange eclipses denote 4 topics at the first layer; the pink eclipses denote 7 topics at the second layer; the cyan eclipses denote 9 topics at the third layer. Data: abstracts of 11 papers from the special issue on Brain Encoding in Pattern Recognition Journal. After removing some stopwords, there are total 635 different words and each abstract contains about 100 words. α function: \( y = 5x \) and λ function is \( y = 5x \). ................................................. 89

4.2 Graphical model for the generation of the proposed Deep Topic Model with three level Topic Hierarchy. The circles denotes random measures or random variables and the solid box denotes the repeat. .......................................................... 91

4.3 Illustration of deep random partition and the relations between topics on the simplex................................................. 91

4.4 The effect of the parameter λ for the DMPM. Here, \( G \) is a mixture of finite number (three) of Dirichlet distributions with equal weights, and each subfigure is an illustration of the \( H(\lambda, G) \) under different value of λ. .......................................................... 93

4.5 Illustration of the base measures in Deep Topic Model ................. 99

4.6 Illustration of the adaptive truncation by Slice variables. The whole length of a black line is one. Each green box denotes a stick weight, and the summation of the sticks on the same black line is one. The red bar represents a slice variable which functions as an adaptive truncation. At each iteration, the slices could move to both directions. With the slice variables, there are only finite number of stick weights and topis need to be updated. ................. 103

4.7 Illustration of the influence of the α function. The synthetic data is: 100 documents and 200 different words, and each document has 20 different words. The λ function is \( y = x \). ................. 108

4.8 Illustration of the influence of the λ function. The synthetic data is: 100 documents and 200 different words, and each document has 20 different words. The α function is \( y = x \). ................. 109
4.9 An example of topic hierarchy. The orange eclipses denote 4 topics at the first layer; the pink eclipses denote 9 topics at the second layer; the cyan eclipses denote 20 topics at the third layer. Dataset: ICDM. The $\alpha$ function is $y = 5x$ and the $\lambda$ function is also $x = 5x$. .................................................. 112

4.10 Results on Dataset PR ........................................ 114

4.11 Results on Dataset ICDM ....................................... 115

4.12 Results on Dataset CORA ............................... 115

4.13 Results on Dataset Citeseer ........................... 116

4.14 Results on Dataset KOS ................................. 116

4.15 Results on Dataset NIPS ............................. 117

5.1 Graphical model of relational-topic-Model ............ 122

5.2 Illustration of Gamma process assignments for a document network. Each document is assigned a Gamma process which has infinite components (represented by the fences in the figure). Each fence denotes a hidden topic, and some examples are given in the figure. The length of the fences denote the weights of different topics in a document. .................................................. 127

5.3 Graphical representation for the Nonparametric Relational Topic (NRT) Model by dependent thinned Gamma Processes and Markov Random Field (MRF). The generative procedures for the document 3 and 4 are omitted to make the figure more concise. 130

5.4 The boxplot of the learned topic numbers by truncated inference method given different truncation levels. .................................................. 142

5.5 Learned topic number distribution from NRT with synthetic datasets under different settings. Normally, the expectation of this distribution will be regarded as the learn topic number of a document network. .................................................. 143

5.6 The illustration of topics learning results. Three red/circle nodes denote three benchmark topics that are also given at the top of each subfigure; the blue/cross nodes denote learned topics from NRT. .................................................. 143

5.7 Effectiveness of SMRF in NRT. The first subfigure is for the comparison between average similarity between topic interests of all test linked document pairs from both NRT and NRT without SMRF; The second and third subfigures are for comparison on Link and Document prediction. .................................................. 144

5.8 Results of NRT and RTM under different setting ($K={20, 30, 40, 50, 60}$) on a first 5-fold of cora dataset. .................................................. 148
LIST OF FIGURES

5.9 Results of NRT and RTM under different settings (K={20, 30, 40, 50, 60}) on a second 5-fold of cora dataset. 148
5.10 Results of NRT and RTM under different settings (K={20, 30, 40, 50, 60}) on a third 5-fold of cora dataset. 149
5.11 Results of NRT and RTM under different settings (K={20, 30, 40, 50, 60}) on a fourth 5-fold of cora dataset. 149
5.12 Results of NRT and RTM under different settings (K={20, 30, 40, 50, 60}) on a fifth 5-fold of cora dataset. 150
5.13 Results of NRT and RTM under different settings (K={2, 5, 10, 20, 30, 40, 50, 100}) on a first 5-fold of citeseer dataset. 150
5.14 Results of NRT and RTM under different settings (K={2, 5, 10, 20, 30, 40, 50, 100}) on a second 5-fold of citeseer dataset. 151
5.15 Results of NRT and RTM under different settings (K={2, 5, 10, 20, 30, 40, 50, 100}) on a third 5-fold of citeseer dataset. 151
5.16 Results of NRT and RTM under different settings (K={2, 5, 10, 20, 30, 40, 50, 100}) on a fourth 5-fold of citeseer dataset. 152
5.17 Results of NRT and RTM under different settings (K={2, 5, 10, 20, 30, 40, 50, 100}) on a fifth 5-fold of citeseer dataset. 152

6.1 Two types of hierarchical structures 157
6.2 The comparison between generated random measures from HDP and CHDP. 164
6.3 Comparing the Chinese restaurant franchise process (three-layer) and international restaurant process. There are four restaurants and three chefs in the figure. In CRF, all the customers in a restaurant can only be served by one chef, but the customers in IRP could be served by different chefs. The main difference between HDP and CHDP is due to Cooperation. 175
6.4 The comparisons between empirical and expected factor numbers of CHDP with two representations: Stick and IRP. Since the factor number is parameterized by $\alpha_0$, $\alpha_a$ and $\alpha_d$. 182
6.5 Gamma-negative binomial process topic model. The left subfigure is related to Eq. (6.38) and the right hand part is related to Eq. (6.39). 185
6.6 Gamma-Gamma-negative binomial process topic model (left one) and mixed Gamma-negative binomial process topic model (right one). 187
6.7 The comparisons between expected and empirical factor numbers of MGNBP under different parameters: $\gamma_0$, $c_0$, $e_a$ and $p_d$. Note that the x-axes of $c_0$ and $e_a$ are in negative (base-10) log space. 200
6.8 The comparisons between the ATM (fixed dimensional model) with CHDP and MGNBP (nonparametric models) on the author prediction. ‘A\(X\)’ denotes the ATM with \(X\) number of topics. Each box expresses the statistics of results from the corresponding model on 5-fold cross-validation. The larger the value of \(AP\), the better the performance. ..........................202

6.9 The comparisons between the ATM (fixed dimensional model) with CHDP and MGNBP (nonparametric models) on the data likelihood. Since CHDP and MGNBP do not need the prefixed topic number, two lines are plotted in this figure. .................202

6.10 The comparisons between LIFT, LEAD, CHDP and MGNBP on the multi-label classification. Each subfigure demonstrates the results on a corresponding metric. ......................206
LIST OF TABLES

2.1 Properties of stochastic processes. ‘Stick-breaking’ denotes if the process has a stick-breaking constructive representation; ‘CRM’ denotes if the process is a completely random measure; ‘NRM’ denotes if the process is a normalized random measure; ‘Power-law’ denotes if there is a version with power-law (cluster number to data number) phenomenon. ........................................ 24

3.1 Notations in this chapter ........................................ 52
3.2 Evaluation metrics for clustering ............................... 76

4.1 The learned topic numbers at three layers from all datasets . . . . 113

5.1 Important notations in this chapter ............................... 125

6.1 Important notations for this chapter ............................... 159