Faculty of Engineering and Information Technology

University of Technology Sydney

Non-IID Latent Variable Models

A thesis submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

by

Trong Dinh Thac Do

February 2019

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Trong Dinh Thac Do declare that this thesis, is submitted in fulfilment of the requirements for the award of the Ph.D. degree, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

Signature:

Date:

Acknowledgments

Foremost, I would like to express the deepest appreciation to my supervisor, Professor Longbing Cao for the continuous support of my Ph.D. study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

In addition, I would like to thank all colleagues in Advanced Analytics Institute: Guansong Pang, Shoujin Wang, Chengzhang Zhu, Songlei Jian, Liang Hu and Wei Wang for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last three years and a half.

Last but not least, I would like to thank my family: my parents, my parents in law, my wife, my dear son, my younger sister, her husband and her two children for their unconditional support, both financially and emotionally throughout the Ph.D. study.

Trong Dinh Thac Do

February 2019 @ UTS

Contents

List of Figures

List of Tables

List of Publications

Published Papers

C-1. **Trong Dinh Thac Do**, and Longbing Cao, "Gamma-Poisson Dynamic Matrix Factorization Embedded with Metadata Influence," *The Thirty-second Conference on Neural Information Processing Systems (NIPS-18)*, 2018. Accepted.

C-2. **Trong Dinh Thac Do**, and Longbing Cao, "Coupled Poisson Factorization Integrated with User/Item Metadata for Modeling Popular and Sparse Ratings in Scalable Recommendation," *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

C-3. **Trong Dinh Thac Do**, and Longbing Cao, "Metadata-dependent Infinite Poisson Factorization for Efficiently Modelling Sparse and Large Matrices in Recommendation," *The 27th International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018.

To be Submitted/Under Reviewed Papers

J-1. **Trong Dinh Thac Do**, and Longbing Cao, "Coupled Attributes-dependent Mondrian Process for Both Static and Dynamic Infinite Relational Learning," *Submitted to Machine Learning Journal.* Under Reviewed.

J-2. **Trong Dinh Thac Do**, and Longbing Cao, "Bayesian Nonparametric Metadata-integrated Coupled Poisson Factorization for Scalable Recommendations." To be Submitted.

J-3. **Trong Dinh Thac Do**, and Longbing Cao, "Explicit and Implicit Relations-based Infinite Latent Feature Learning for Link Prediction". To be Submitted.

C-4. **Trong Dinh Thac Do**, and Longbing Cao, "Group-based Gamma-Poisson Matrix Factorization on Multi-source, Large and Sparse Recommendation Data." To be Submitted.

C-5. **Trong Dinh Thac Do**, and Longbing Cao, "HDIM: A Heterogeneous Data-driven Infinite Model for Learning Hierarchical Relations in Time-varying Attributed Networks". To be Submitted.

Abstract

Latent Variable Model (LVM) is the statistical model that aims to uncover hidden information behind data. These models have been widely used for real-world applications such as community detection, link prediction or recommender systems. However, LVM faces significant challenges in modeling complex relations since LVM assumes that the data are independent and identically distributed (IID). However, real-world data are often coupled in terms of object attributes, object relations, or even hidden variable relations. For example, in social networks, users that indicate a similar 'age', 'location' and 'high school' are often friends. To this end, non-IID learning has the potential to describe the above hierarchical relations in real-world data which are typically not independent or identically distributed (non-IID).

In this thesis, we are interested in determining the relations behind observations and hidden variables in LVM. More specifically, we focus on coupling relations in non-IID data in terms of various LVM, including Latent Class Model (LCM), Latent Feature Model (LFM), and Latent Factor Model-Matrix Factorization (LFM-MF). In particular, we aim to model the following relations: (1) relations between attributes in observed data (e.g., user/item metadata such as 'location' of a user or 'genre' of a movie); (2) relations between different sources of observed data (e.g., metadata and user's friendships); and (3) relations between latent variables in LVM. We also apply

Bayesian Nonparametric (BNP) techniques to the proposed LVM models to automatically tune the number of latent variables in LVM for efficient computation. Furthermore, to work with large and sparse data, we introduce several methods for better inference of the proposed LVM models.

The empirical analysis of both proposed models reveals that our models significantly outperform state-of-the-art models in the same family. Together with improved optimization techniques (i.e., BNP and inference methods), our proposed models indicate their potential for online modeling of large, sparse data.