



Doctor of Philosophy

Diversified Probabilistic Graphical
Models

by

Maoying Qiao

supervised

by

Prof. Dacheng Tao

the Centre for Quantum Computation and Intelligent Systems (QCIS)

the Faculty of Engineering and Information Technology (FEIT)

the University of Technology Sydney (UTS)

July, 2016

CERTIFICATE OF AUTHORSHIP / ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the 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

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Dacheng Tao for his continuous support of my Ph.D study. Thanks for his consistent patience and motivation, for his encouraging attitude and expert knowledge for my research. His strict academic attitude and diligent work style have played a model role for me and will continue to benefit me through my life. It is no exaggeration to say without his help steering my research direction, I would not have finished this thesis so smoothly and on time.

Besides my principal supervisor, I would like to thank Dr. Wei Bian. I want to thanks for his never bored discussion with my seemingly endless simple questions from research motivation, model development, algorithm implementation, and paper drafting. With these specific and detailed technical discussion, I have gained practical skills to effectively and efficiently develop and implement my research problems. I also gain deep understandings for my research areas. In addition, these discussion moments between us and with Qiang Li have been an inseparable part of my Ph.D life, and will always be cherished by myself.

I would like to thank Dr. Richard Yi da Xu for his advices on my study of Monte Carlo sampling methods. His help undoubtably has widened my research area. I also want to thank my colleagues in Prof. Tao's group for their academic discussion and help. My special thanks goes to Ms. Jemima Moore for helping improve my English skills in both academic publications and presentations.

My sincere thanks also goes to China Scholarship Council (CSC), Univer-

sity of Technology Sydney (UTS) and the centre for Quantum Computation & Information System (QCIS). With the financial support from CSC-UTS joint scholarship, I could concentrate on my research study and did not have to worry about my living. QCIS has provided a excellent and supportive research environment for academic-related activities.

Last but not the least, my gratitude extends to my family who have been patiently encouraging and waiting for the finish of this thesis.

ABSTRACT

Probabilistic graphical models (PGMs) as diverse as Bayesian networks and Markov random fields have provided a fundamental framework to learn and reason using limited and noisy observations. Examples include, but are not limited to, hidden Markov models (HMMs), sequential graphical models, and probabilistic principal component analysis mixture models (PPCA-MM). PGMs have been used in a wide variety of applications such as speech recognition, natural language processing, web searching, and image understanding. However, one potential drawback of using PGMs with traditional learning and inference methods is that the learned parameters or inferred variables are easily trapped within local, clustered optima rather than distributed evenly across the whole space. Taking mixture models as an example, the learned mixing components might overlap. Consequently, the resulting models might show ambiguity when clustering is performed based on these overlapping mixing components. This phenomenon might limit PGM performance.

Although efforts have been made to explore a variety of priors to alleviate this potential drawback and to enhance PGM performance, diverse priors have yet to be fully explored and utilized. Diversity is a concept that encourages counterpart model parameters and variables to repel as much as possible and, in doing so, spread out model components and decrease overlapping. However, how to explicitly encode these priors into a PGM and how to solve the resulting diversified PGMs are two critical problems that must be solved. This thesis proposes a uni-

fied framework to constrain PGMs with diverse priors. Three different PGMs - HMMs, time-varying determinantal point processes (TV-DPPs), and PCA-MMs - are elaborated to demonstrate the proposed diversified PGM framework. For each PGM, three basic constituent framework elements are examined: which part of the traditional PGM is diversified, how to formulate the diversity, and how to solve the diversified version, e.g., parameter learning and inference. In addition, experiments are conducted using various application scenarios to verify the effectiveness of the proposed diversified PGMs.

Keywords: Probabilistic graphical models (PGMs), diversity prior, determinantal point processes (DPPs), hidden Markov models (HMMs), time-varying DPPs, probabilistic principal component analysis (PPCA).

TABLE OF CONTENT

CERTIFICATE OF AUTHORSHIP/ORGINALITY	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	v
TABLE OF CONTENT	vii
LIST OF FIGURES	x
LIST OF TABLES	xiv
Chapter 1 Introduction	1
1.1 Background	1
1.1.1 Probabilistic graphical models	1
1.1.2 Determinantal point processes (DPPs)	3
1.2 Related works	6
1.2.1 DPP-related developments	7
1.2.2 Diversity-extending models	10
1.2.3 Diversity-requiring applications	12
1.3 Motivations and research significance	16
1.3.1 Motivations and contributions	16
1.3.2 Research significance	17
1.4 Thesis Structure	18
Chapter 2 Diversified Hidden Markov Models for Sequential La- belling	20
2.1 Introduction	21
2.2 Related Work	24
2.3 Diversified Hidden Markov Models	26
2.3.1 Kernels for probability diversity	27
2.3.2 Log-likelihood function of Hidden Markov Models	27
2.3.3 Proposed Diversified HMM	30

TABLE OF CONTENT

2.3.4	Solutions	32
2.4	Experimental results	40
2.4.1	Toy experiments	40
2.4.2	Real-world experiments	44
2.5	Summary	51
Chapter 3 Fast Sampling for Time-Varying Determinantal Point Processes		61
3.1	Introduction	62
3.2	Related Work	64
3.3	Review of Sequential Monte Carlo	66
3.4	Time Varying Determinantal Point Processes	67
3.5	Experimental Results	76
3.5.1	News recommendation	77
3.5.2	Enron corpus	87
3.6	Summary	95
Chapter 4 Diverse Learning for Mixtures of Exponential Family PCA Model		97
4.1	Introduction	98
4.2	Related Work	102
4.2.1	PCA and its Mixture Extensions	102
4.3	Background	104
4.3.1	Simple Exponential Family PCA (SePCA)	104
4.3.2	SePCA Mixture Models (SePCA-MM)	105
4.4	The proposed model	109
4.4.1	Motivation for Matrices-valued DPP	109
4.4.2	Matrices-valued DPP	110
4.4.3	Diversified SePCA-MM	114
4.5	Learning and inference	116
4.5.1	Learning Parameters: M-step	116
4.5.2	Inference: E-step	118
4.5.3	Algorithm Summary and Bernoulli Distributions	122
4.5.4	Algorithm Complexity Analysis	123
4.6	Experimental Results	124
4.6.1	Synthetic Experiments	124
4.6.2	Real-world Dataset experiment	129
4.7	Summary	133
Chapter 5 Conclusion and Further Study		134
5.1	Conclusions	134

TABLE OF CONTENT

5.2 Further study	136
REFERENCES	137

LIST OF FIGURES

1.1	Illustration of diverse subset	4
1.2	Demonstration of diversity captured by DPP.	6
1.3	Graphical representation for Markov DPP	9
1.4	Graphical representation for sequential DPP	9
1.5	Thesis structure	18
2.1	Graphical model of diversified HMM	52
2.2	Parameters of ground-truth, learned by proposed dHMM and by traditional HMM	53
2.3	Diversities of transition matrix of ground-truth, dHMM-learned and HMM-learned with regard to the parameter of variance of the Gaussian emission distributions	54
2.4	Histograms of hidden states inferred from parameters of ground-truth, dHMM-learned and HMM-learned	55
2.5	Number of hidden states inferred by model parameters of dHMM-learned and HMM-learned with regard to the variance of Gaussian emission distributions	56
2.6	Sentence example with PoS tags	56
2.7	Effectiveness of α for PoS tagging	57
2.8	Transition diversity comparison between dHMM and HMM for tag ‘1’ and all other tags	57
2.9	Histogram comparisons among ground-truth, HMM and dHMM	58
2.10	Effectiveness of α for OCR	58
2.11	Test accuracies of different classifiers	59
2.12	Transition diversity comparison between dHMM and HMM	60

- 3.1 Time Varying DPPs: In the first diagram, the first row represents the news updating process along time stamps. Six different news sources are schematically listed, i.e. ‘The Daily Telegraph’, ‘Daily Mail’, ‘ABC NEWS’, ‘The Guardian’, ‘Reuters’ and ‘Indiatimes’. From time to time, only a small portion of the news sources are updated. The arrows make which news sources are updating clear: It starts at a news source with old news and points to the same source with new headlines -bordered in cyan - at the next time stamp. The second row shows the evolution of DPP marginal kernel L along with the news updates. The difference between two successive L -s is highlighted with different colours and is apparently tiny. The third row shows explanatory diverse subsets outputted by TV-DPPs. In the second diagram, the solid circles represent the observations, which correspond to the news dataset shown in the first row of the above figure, and the hollow circles represent the variables obeying the DPP distribution, one example of which can be found in the third row of the above figure. One important truth is that given the observations $\{X_1, X_2, \dots, X_T\}$, the variables $\{Y_1, Y_2, \dots, Y_T\}$ are independent. 69
- 3.2 Illustration of Sequential Monte Carlo: At time stamp $t - 1$, 10 particles in red with equal weights are given, i.e. $\{x_{t-1}^{(i)}\}_{i=1}^{10}$. At this stage, two computations will be done - One is computing the incremental weights; the other is computing the particles for the next time stamp. For the incremental weight of each particle at time $t - 1$, according to Eq. (3.7), it is simply the likelihood ratio between time stamps t and $t - 1$. The corresponding relationship is denoted by the dashed line connecting two neighbour distributions and the weight for each particle is illustrated by size of blue solid circle. For the particle’s location at next time stamp t , usually, a Markov transition kernel is used to qualify the transition job between two slightly different neighbour distributions. The transition relationship is indicated by solid line with arrow. For the particle’s weight at time stamp t , it is gained by multiplying the weight at time stamp $t - 1$ by the incremental weight. To alleviate the degeneracy of the algorithm which is measured by effective sample size (ESS), a re-sampling step is applied when $N_{ESS} < \alpha \cdot N$. High weighted particles will re-birth as several equal weighted particles, while particles with low weights may disappear. To increase the samples’ diversity, a move step is followed. Once particle’ locations and weights at t are prepared, it will recursively carry out the whole above procedure. 75

LIST OF FIGURES

3.3 Analysis for fast DPPs sampling algorithm at $t = 1$ 79

3.4 Diversity comparison of news subsets selected by sep-DPPs and TV-DPPs with regard to both diverse probability and cosine similarity. 82

3.5 Time cost comparison between sep-DPPs and TV-DPPs for news recommendation. X-axis indicates the time stamps, while Y-axis shows the accumulated seconds over time. 83

3.6 Demonstration of diverse subset of news articles sampled by TV-DPPs. 85

3.7 Demonstration of news topic drift: Comparing sequential subsets from TV-DPPs with the subsets from sep-DPPs, it extracts a more smooth news evolvement. 86

3.8 One example of Enron communication network. 87

3.9 Time cost comparison between sep-DPPs and TV-DPPs for Enron communication network. X-axis represents the day stamps, while Y-axis shows the accumulated seconds. 90

3.10 Demonstration of news topic drift: Comparing sequential subsets from TV-DPPs with the subsets from sep-DPPs, it extracts a more smooth news evolvement. 92

3.11 Demonstration of Enron communication drift: No communication patterns or events can be detected from the above figure. However, three different stages are clearly obtained with smoothness at each stage. Two separating points - one is around 2001-Oct-01 and the other around 2001-Nov-10 - coincide with two important turn points for Enron incorporation. 93

3.12 Three Enron communication networks from different stages detected by TV-DPPs. 96

4.1 Graphical representations of models: (a) SePCA mixture model; (b) Diversified SePCA mixture model. The only difference between these two models is obviously the priors over K mixture component parameters, i.e., W s. Traditional SePCA assigns independent isotropic Gaussian distributions, represented by a plate. Comparatively, the proposed diversified SePCA-MM assigns a joint distribution over its component parameter, i.e., $\mathbf{W} = \{W^1, \dots, W^K\}$. The distribution is a DPP, parameterized with ϱ, ξ, λ and represented by a double-struck. 108

4.2	Illustration of proposed similarity formulations with two orthonormal matrices including two PCs. The dashed pairwise perpendicular lines represent PCs of orthonormal matrix Υ^1 , while the solid pairwise lines represent PCs of Υ^2 . θ_1 and θ_2 represent angles of direct-matching pairwise PCs as shown in all subfigures, while θ_3 and θ_4 represent angles of cross-matching pairwise PCs which are only demonstrated in the first subfigure and ignored by other subfigures to keep their symbols less crowded. From left to right, the angles between direct-matching pairwise PCs of the two orthonormal matrices starts from 0° and increases with 45° , while the similarities between these two matrices decreases from the largest value $e^{2\varrho}$ to the smallest value $e^{-2\varrho}$, which are calculated from (4.9)..	112
4.3	Black indicates 0 and white indicate 1.	125
4.4	The first row is the result from traditional PCA-MM, while the second row is the result from the proposed diversified PCA-MM. $d = 12, K = 3$	126
4.5	Hinton diagram of diagonal of $\{\Phi^k\}_{k=1}^K$, where white boxes indicate positive values, and black ones indicate negative values. The magnitudes of Φ^k 's are symbolized by the sizes of boxes.	127
4.6	Illustrate effectiveness of diversity over model redundancy reduction with fixed $d = 4$ and $K = 3$ and $K = 2$ for both traditional PCA-MM (the first rows) and diversified PCA-MM (the second rows).	128
4.7	Samples from USPS digits 2, 3, 4.	130
4.8	The fitted three local PCA transformation matrices of our diversified mixture models with $K = 3, d = 2$. Each row represents one PC.	131
4.9	Dominant PCs of two mixing components with dark black pixels indicating larger absolute values than light grey pixels.	132

LIST OF TABLES

2.1	Comparison of state frequencies and accuracies between dHMM and HMM	43
2.2	Summary of PoS tags of WJS corpus	46
2.3	Examples of OCR dataset	49
3.1	Summary of news categories for different news media sources . . .	77