Bayesian Non-parametric Models for Time Segmentation and Regression

Ava Bargi

MEng Baha'i Institute for Higher Education (BIHE)

Faculty of Engineering and Information Technology

University of Technology, Sydney

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

November 2015

*To **Nima**,*

a wonderful companion,

a patient and gentle support,

an unconditional lover,

my husband and most precious friend,

to whom I am willing to dedicate this thesis and all my future accomplishments !

Certificate of Authorship and Originality

Title: **Bayesian Non-parametric Modelling for Infinite-modal Segmentation and Prediction**

Author: **Ava Bargi**

Date: **November 27, 2015**

Degree: **PhD**

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

Signiture of author

Acknowledgements

I would like to sincerely thank the following wonderful people, without whose attention and support this thesis could not have been completed:

My principal supervisor, *Professor Massimo Piccardi*, for gently holding my hands in the toddling stages of a PhD; for constantly caring for my progress and happiness; for teaching me the real essence of research, far beyond publish or perish: the never changing virtue of *'modest learning'*. In addition to a smart supervisor, he has been a great friend and company. It has been both a joy and honour to be his student and I am utterly grateful for all the learning during this PhD.

Dr Richard Xu, my co-supervisor, has been a wonderful technical adviser and support. His timely advice has impacted the most valuable accomplishments. He has been another role model for genuine spirit of research and hard work.

Dr Emily B. Fox whose research has largely inspired this work, offering me an example of noble ideas, elegant presentation and helpful and friendly attitude.

Professor Zoubin Ghahramani for organising a talk in his center, offering me an opportunity to communicate the current research and receive invaluable feedback. His modest and open attitude to collaboration has marked an important point in my research career.

Professor Yee Whye Teh, Dr Christoph Lampert, Dr Fernando De la Torre and *Dr Sebastian Nowozin* for offering their valuable time for discussing the current research and providing feedback.

Dr Dario Garcia Garcia for technical advice and his understanding and flexibility that enabled me to follow this research along with a full-time internship in the Commonwealth Bank.

Professor Didar Zowghi, my former instructor and dear friend, for being there when I most needed personal and career support. I will always cherish her company on enjoyable tea breaks!

Doing a PhD includes days of excitement, exhausting work and sometimes confusion and that is where peers come to rescue! I would like to thank my dear friends *Dr Ehsan Zare Borzeshi, Dr Jaime Andres Garcia Marin* and *Dr Muneera Bano* who have created me unforgettable and joyful moments of chat over coffee!

I would like to acknowledge the staff members of *Baha'i Institute for Higher Education (BIHE)*, the university that hosted my tertiary education when all the other university gates were locked for me and my fellow Baha'i friends in Iran, due to governmental persecutions. BIHE staff have established and run this academic shelter for about 30 years now, sacrificing their time, comfort and freedom. I owe my engieering knowledge, passion for learning and this PhD thesis to all those exceptional people, currently imprisoned for education!

I truly owe my life and achievements to *Rouha and Rouhollah Bargi*, my beloved parents. No matter how geographically dispersed we live, they have always known me better than myself, infusing my life with love, wisdom, support and sacrifice.

Last but never least, I would like to whole-heartedly thank my husband

Dr Nima Khorsandnia, who has always encouraged and reassured me in the critical moments. I feel really prosperous to have his company in life.

Ava Bargi,

March 2015, Sydney

Abstract

Non-parametric Bayesian modelling offers a principled way for avoiding model selection such as pre-defining the number of modes in a mixture model or the optimal number of factors in factor analysis. Instead, Bayesian non-parametric methods allow the data to determine the complexity of model. In particular, the hierarchical Dirichlet process (HDP) is used in a variety of applications to infer an arbitrary number of classes from a set of samples. Within the temporal modelling paradigm, Bayesian non-parametrics is used to model sequential data by integrating HDP priors into state-space models such as HMM, constructing HDP-HMM. Also in latent factor modelling and dimensionality reduction, Indian buffet process (IBP) is a well-known method capable of sparse modelling and selecting an arbitrary number of factors among the often high-dimensional features.

In this PhD thesis, we have applied the above methods to propose novel solutions to two prominent problems. The first model, named as 'ADON HDP-HMM', is an *adaptive online* system based on HDP-HMM. 'ADON HDP-HMM' is capable of segmenting and classifying the sequential data over unlimited number of classes, while meeting the memory and delay constraints of streaming contexts. The model is further enhanced by a number of *learning rate*s, responsible for tuning the adaptability by determining the extent to which the model sustains its previous parameters or adapts to the new data. Empirical results on several variants of synthetic

and action recognition data, show remarkable performance, particularly using adaptive learning rates for evolutionary sequences.

The second proposed solution is an elaborate factor regression model, named as *non-parametric conditional factor regression* (NCFR), to cater for multivariate prediction, preserving the correlations in the response layer. NCFR enhances factor regression by integrating IBP to infer the optimal number of latent factors, in a sparse model. Thanks to this data-driven approach, NCFR can significantly avoid over-fitting even in cases where the ratio between the number of available samples and dimensions is very low. Experimental results on three diverse datasets give evidence of its remarkable predictive performance, resilience to over-fitting, good mixing and computational efficiency.

Contents

Contents **viii**

List of Figures **xii**

List of Tables **xviii**

1 Introduction **1**

1.1 Overview 1

1.2 Research questions 3

1.3 Thesis outline 4

1.4 Related publications 7

2 Machine Learning Background **9**

2.1 Linear regression 9

2.2 Bayesian modelling 11

2.2.1 Inference and learning 11

2.2.2 Generative models 14

2.2.3 Probabilistic graphical models 15

2.2.4 Mixture models 16

2.2.5 Sequential models 18

2.2.6 Exponential family of distributions 22

2.3 Bayesian non-parametric modelling 24

2.3.1 Dirichlet process 26

2.3.2 Hierarchical Dirichlet process 29

2.3.3 Indian buffet process 31

2.4 Graphical models with DP and HDP priors 33

2.4.1 Dirichlet process mixture model 34

2.4.2 HDP prior for HMM 35

2.4.3 HDP prior for SLDS and SAR 36

2.5 Inference methods for Bayesian non-parametrics 37

2.5.1 Markov chain Monte Carlo 38

2.5.2 Variational inference 41

2.6 Online and adaptive learning 42

2.7 Linear latent subspaces 44

2.7.1 Principal component analysis 44

2.7.2 Probabilistic principal component analysis 44

2.7.3 Factor analysis 45

2.7.4 Infinite factor analysis 46

2.7.5 Factor regression 47

3 Action Recognition Background 49

3.1 Human action recognition 50

3.1.1 Feature extraction 51

3.1.2 Action classification and segmentation 57

3.2 Conclusion 58

4 Infinite Adaptive Online Segmentation and Classification for Sequential Data 60

4.1 Introduction 60

4.2 Literature review 62

4.3 Inference and learning 65
4.4 The proposed Adaptive Online HDP-HMM 66
4.4.1 Adaptive learning rate 68
4.4.2 Discussions on the learning rate 76
4.5 Experiments 77
4.5.1 Synthetic data 79
4.5.2 Action recognition datasets 83
4.5.3 Sampling efficiency and computational time 93
4.6 Conclusion 94

5 Online Predictive Batching for Sequential Data 97

5.1 Predictive batching 98
5.2 Experiments 102
5.2.1 The collated Weizmann dataset 103
5.2.2 TUM kitchen dataset 104
5.3 Conclusion 105

6 Multivariate Non-parametric Conditional Factor Regression 107

6.1 Introduction 107
6.2 Literature review 109
6.3 The Non-parametric Conditional Factor Regression 110
6.4 Inference 111
6.4.1 Binary mask 112
6.4.2 Factor loading matrices and latent factor weights 115
6.4.3 Noise and factor loading covariances and IBP parameters . . . 117
6.5 Experiments 117
6.5.1 Synthetic data 119
6.5.2 Human pose prediction 121

6.5.3 Natural gas consumption prediction 122
6.5.4 Sampling efficiency and computational costs 124
6.6 Conclusion 125

7 Conclusions 127

7.1 AdOn HDP-HMM 127
7.1.1 Future studies 129
7.2 NCFR 129
7.2.1 Future studies 130

Appendix A: The balancing effect of τ 131

Appendix B: Conjugacy for τ_Σ 132

Appendix C: Conjugacy for τ_A 134

Appendix D: Impacts of τ on parameter distributions 136

Appendix E: EM for inference of CFR 139

List of Figures

2.1 Bayes ball algorithm, illustrating conditional dependencies of x and z, given y. The top row shows cases where y is a random variable like x and z, whereas in the bottom row y is an observation (Fox, 2009a). . 17

2.2 Mixture model. z_i is the indicator variable and y_i is the observation. The plate notation is used to show replication. 17

2.3 HMM graphical model. 19

2.4 SAR graphical model with $r = 1$. For $r = 2$ there will be extra dependencies (arrows) linking each y_t also to y_{t-2}, stemming from y_{t-2}. A similar pattern holds for higher order models, i.e. auto-regressive models with greater r. 21

2.5 SLDS graphical model. 22

2.6 The Dirichlet process construction; the parameter space has been simplified to one dimension for the sake of visualisation. 27

2.7 The HDP construction. 30

2.8 Binary matrices generated by Indian Buffet Process(IBP) with various αs. The rows and columns are mapped to Indian buffet analyogy of customers and dishes, showing direct propotionality between the number of sampled dishes and α. Figure is reproduced from (Knowles, 2012). 33

2.9 Dirichlet Process Mixture Model (DPMM) graphical model. The *plate* notation implies repetition. 34

2.10 HDP-HMM graphical model. The box notation is used to show replication. 36

2.11 The switching dynamical graphical models with HDP priors. (a) HDP-SLDS, (b) HDP-SAR with $r = 1$. 37

3.1 Pli'e movement in ballet: the kinematic constrains for knee and ankle angles are used for feature detection. 52

3.2 Movement recognition based on a blurred selection of video frames, representing sit down motion (Bobick, 1997, p. 4). 53

3.3 Motion-based features: (a) Representation of Motion-energy images of a sequence of images for sit down action (b) Illustration of Motion-history images in sit-down, arm-wave and crouch-down actions (Bobick, 1997). 53

3.4 (a) Space-time volume of stacked silhouettes (b) Motion history volumes. Even though the representations appear similar, (a) is viewed from a single camera, whereas (b) shows a recency function over reconstructed 3D voxel models. Figures reprinted from (Poppe, 2010). . 54

3.5 Kinect motion capture data. 55

3.6 Spatio-temporal interest points in walking action. (a) 3D plot of leg pattern shown upside down for ease of interpretation. (b) spatio-temporal interest points on walking legs (figures reprinted from Laptev 2005, pp. 115) 57

4.1 Adaptive online learning flowchart: Initialised by a supervised bootstrap, the learning continues completely unsupervised for streaming data split into batches. The figure shows a general case with batches of variable size. For simplicity we assume all batches to have the same size (see (Bargi et al., 2014a) for the variable alternative). The posterior ϕ in each batch is passed to the next batch as prior parameters. 67

4.2 Graphical model for the proposed adaptive online model. ϕ can be any of the parameters in Figure 2.10 (θ, π, β), τ is the respective learning rate (a positive continuous variable) and n represents the batch number. 68

4.3 (a) The distribution of states in the noisy synthetic data set. Note that due to the large standard deviation ($\sigma = 50$) there is a significant overlap between the states, making recognition a challenging task.(b) The sum of density functions for all states. 80

4.4 Segmentation and classification results for *evolutionary* synthetic data, using fixed ($\tau = 1$) and adaptive (*ada* τ) learning rates. (a,d) Shifting means: Without the adapting effects of the learning rate, shifting means can be misclassified as new classes (yellow) in d. (b,e) New class: is shown in yellow in the ground truth. The new class has been recognised and learnt in both cases, assigned a random colour (orange in (b) and red in (e)). (c,f) Combined: Adding both challenges causes slight decrease in accuracy and an extra new class. Nevertheless, the considerable performance gap in utilising τ is still visible. 81

4.5 Weizmann action recognition dataset: Snapshots and action labels. Photos extracted from (Weizmann, Accessed Feb. 2015). 83

4.6 Estimated states for Weizmann dataset. Action labels are represented by colours. 84

4.7 Snapshots of Kitchen TUM dataset (TUM, Accessed Feb. 2015). . . . 86

4.8 Estimated states for the TUM kitchen dataset using the Adaptive Online HDP-HMM. LH and RH stand for left and right hand. The first two columns show *robotic* sequences, whereas the third column includes *complex* ones. (c) is a *human* sequence with altered orders of actions performed spontaneously and (d) contains a new action shown in lavender-blue in the ground-truth and recognised by the model in a random brown colour. In most cases, using adaptive τ causes noticeable improvements on both the performance and cardinality of inferred states. 88

4.9 Snapshots from the Honey bee dataset. 90

4.10 Trajectories of honey bees in each sequence. Movements are colour coded as waggle (red), turn right (blue), and turn left (green). Image extracted from Fox (2009a). 91

4.11 Estimated states for honey bee dataset. 92

4.12 Sampling efficiency: (a-d) Sample mixing for all the 4 proposed learning rates across three online batches and (e) Log likelihood plot for first batch of a Weizmann run, show well mixing and convergence both for the learning rates and generally for all other parameters involved in likelihood calculation. 95

4.13 Computation time per frame (seconds) for Weizmann and TUM kitchen datasets, over the online and offline variants, with fixed and adaptive learning rates. 96

5.1 Two batching alternatives on a single instance of streaming data. Each color represents a ground-truth segment: (a) Fixed size batching, versus (b) predictive batching. 99

5.2 Adaptive online learning flowchart with the *predictive batching* plug-in. 100

5.3 Predictive batching by conditional factor regression (CFR): (a) The graphical model (b) The schematic diagram. 101

5.4 Estimated states with ADON HDP-HMM ($PredBtch$) for Weizmann dataset. The respective quantitative figures are shown in Table 5.1. . . 103

5.5 Estimated states for the TUM kitchen dataset using the ADON HDP-HMM, with predictive batching ($PredBtch$). LH and RH stand for left and right hand. 105

6.1 NCFR Graphical model. The node labeled as $s \odot z$ is deterministic, illustrated with double-ring oval notation (Koller and Friedman, 2009). The plate notation is used to clarify the dimensions of matrices introduced with capital letters in the model. 112

6.2 *IsoN* NCFR performance on synthetic data: (a) Heat map for visual comparison of real $Y_{(D_q \times N_{test})}$ matrix vs. predicted Y. (b) Scatter plot of real Y vs. predicted Y with all dimensions superimposed on the same plot. The plot includes $D_q \times N_{test}$ points all aligned around $y = x$, depicting a significant predictive performance over all dimensions. The figure is best viewed in colour. 120

6.3 Boxplot of prediction performance on TUM kitchen data, sequence 1-2: (a) Normalised prediction errors are presented over 148 frames of Y and sorted across the 3 categories of models. The top 6 variants are magnified in the plot on the right (b), showing noticeably better performance for NCFR variants appearing on the first 3 boxes. 122

6.4 Sampling efficiency and computational cost: (a) Log likelihood plot of NCFR gas results with the last 200 iterations magnified, showing convergence and well mixing. (b) Inference time per iteration for TUM kitchen pose prediction (c) Inference time per iteration per active feature for the same kitchen experiment. Please note that CPU time has been divided by $K \times S_{density}$ to provide a precise metric with respect to number of active features and their average sparsity. Combined analysis of (b) and (c) creates a thorough understanding of inference speed. 124

List of Tables

2.1 A few instances of conjugate priors utilised in this thesis. 14

4.1 Average values for frame-level accuracy, segmentation recall, precision and state cardinality error for the synthetic experiments. Each table section includes the respective results of experiment secions, comparing the performance with and without the learning rate: i) Stationary with high noise reporting average results on 3 sequences, ii) Evolutionary with shifting means, iii) Evolutionary with new states, iv) Evolutionary with combined shifts and new states. 80

4.2 Frame-level accuracy, segmentation F1 score, and difference in decoded state cardinality for ADON HDP-HMM variants and state-of-the-art studies on the collated Weizmann dataset. 84

4.3 Frame-level accuracy and state cardinality error for Adaptive Online HDP-HMM on all TUM kitchen sequences. The comparison between similar sequences with/without learning rate (τ) shows incremental improvement on frame-level accuracy and significant decrease on state cardinality error. Note that the figures under cardinality show differences between inferred vs. actual number of states, considering the sign. However the absolute values are utilised to calculate the average cardinality error. 89

4.4 Adaptability experiment: frame-level accuracy and state cardinality error for Adaptive Online HDP-HMM trained with the *robotic* sequences and tested on the *complex* ones. The comparison between similar sequences with/without learning rate (τ) shows noticeable improvement on frame-level accuracy and significant decrease on state cardinality error. 90

4.5 Frame-level accuracy and state cardinality error for honey bee experiment, with fixed ($\tau = 1$) and adaptive learning rate (*ada* τ). 93

5.1 Frame-level accuracy, segmentation precision, and difference in decoded state cardinality for the adaptive online HDP-HMM variants on the collated Weizmann dataset. The segmentation recall is not reported due to similarity across the variants. 104

5.2 Frame-level accuracy and difference in decoded state cardinality for ADON HDP-HMM, comparing $PredBtch$ and $FixBtch$ on TUM kitchen dataset. 106

6.1 Models and variants exploited and compared in the experiments. Variants are denoted with their italic abbreviations. 119

6.2 Prediction error percentage ($\epsilon = 100\times \parallel y_n - \tilde{y}_n \parallel / \parallel y_n \parallel$) for synthetic prediction. 120

6.3 Prediction error percentage ($\epsilon = 100\times \parallel y_n - \tilde{y}_n \parallel / \parallel y_n \parallel$) for gas consumption forecast. 123