

Bayesian Learning with Dependency Structures via  
Latent Factors, Mixtures, and Copulas

by

Shaobo Han

Department of Electrical and Computer Engineering  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Lawrence Carin, Supervisor

\_\_\_\_\_  
David B. Dunson

\_\_\_\_\_  
Guillermo Sapiro

\_\_\_\_\_  
Ingrid Daubechies

\_\_\_\_\_  
Katherine Heller

Dissertation submitted in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy in the Department of Electrical and Computer Engineering  
in the Graduate School of Duke University

2016

ABSTRACT

Bayesian Learning with Dependency Structures via Latent  
Factors, Mixtures, and Copulas

by

Shaobo Han

Department of Electrical and Computer Engineering  
Duke University

Date: \_\_\_\_\_

Approved:

---

Lawrence Carin, Supervisor

---

David B. Dunson

---

Guillermo Sapiro

---

Ingrid Daubechies

---

Katherine Heller

An abstract of a dissertation submitted in partial fulfillment of the requirements for  
the degree of Doctor of Philosophy in the Department of Electrical and Computer  
Engineering  
in the Graduate School of Duke University  
2016

Copyright © 2016 by Shaobo Han  
All rights reserved except the rights granted by the  
Creative Commons Attribution-Noncommercial Licence

# Abstract

Bayesian methods offer a flexible and convenient probabilistic learning framework to extract interpretable knowledge from complex and structured data. More specifically, such methods can characterize dependencies among multiple levels of hidden variables and share statistical strength across heterogeneous sources. In the first part of this dissertation, we develop two dependent variational inference methods for full posterior approximation in non-conjugate Bayesian models through hierarchical mixture- and copula-based variational proposals, respectively. The proposed methods move beyond the widely used factorized approximation to the posterior and provide generic applicability to a broad class of probabilistic models with minimal model-specific derivations. In the second part of this dissertation, we design probabilistic graphical models to accommodate multimodal data, describe dynamical behaviors and account for task heterogeneity. In particular, the sparse latent factor model is able to reveal common low-dimensional structures from high-dimensional data. We demonstrate the effectiveness of the proposed statistical learning methods on both synthetic and real-world data.

# Contents

<b>Abstract</b>	<b>iv</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Figures</b>	<b>x</b>
<b>Acknowledgements</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Dependent Variational Inference for Non-Conjugate Models . . . . .	1
1.2 Bayesian Modeling of Latent Dependency Structures . . . . .	4
1.3 Learning with Task Dependencies across Heterogeneous Domains . . .	5
<b>2 Mixture-based Dependent Variational Inference</b>	<b>8</b>
2.1 Overview of Variational Inference . . . . .	8
2.2 Parameter Estimation and Marginal Approximation . . . . .	9
2.2.1 Conjugate-Exponential Models . . . . .	10
2.2.2 Variational Expectation Maximization . . . . .	10
2.2.3 Variational Bayesian EM . . . . .	12
2.3 Integrated Nested Laplace Approximation (INLA) . . . . .	13
2.4 Integrated Non-Factorized Variational Inference . . . . .	14
2.4.1 Hybrid Continuous and Discrete Variational Approximations .	15
2.4.2 Variational Optimization . . . . .	16
2.4.3 Links between INF-VB and INLA . . . . .	18

2.5	Variational Bayesian Lasso . . . . .	20
2.5.1	Variational Gaussian Approximation . . . . .	21
2.5.2	Upper Bounds of KL divergence . . . . .	22
2.5.3	Theoretical Analysis . . . . .	25
2.6	Other Inference Methods for Bayesian Lasso . . . . .	27
2.6.1	Data-Augmentation Gibbs Sampler . . . . .	28
2.6.2	Mean-Field Variational Bayes . . . . .	29
2.7	Experiments . . . . .	31
2.7.1	Synthetic Dataset . . . . .	32
2.7.2	Diabetes Dataset . . . . .	33
2.7.3	Comparison: Accuracy and Speed . . . . .	35
2.8	Discussion . . . . .	36
<b>3</b>	<b>Copula-based Dependent Variational Inference</b>	<b>40</b>
3.1	Preliminaries . . . . .	41
3.2	Variational Copula Inference Framework . . . . .	43
3.2.1	KL Additive Decomposition . . . . .	45
3.2.2	Special Case 1: Mean-field VB . . . . .	46
3.2.3	Special Case 2: VG Approximation . . . . .	47
3.3	Variational Gaussian Copula Approximation . . . . .	48
3.3.1	Equivalent Variational Proposals . . . . .	49
3.3.2	VGC with Fixed-form Margins . . . . .	50
3.4	Bernstein Polynomials based Monotone Transformations . . . . .	51
3.4.1	Continuous Margins Constructed via Bernstein Polynomials . . . . .	51
3.4.2	Variational Inverse Transform . . . . .	52
3.5	Stochastic VGC . . . . .	53

3.5.1	Coordinate Transformations . . . . .	54
3.5.2	Update the BP Weights . . . . .	56
3.6	Experiments . . . . .	57
3.6.1	Flexible Margins . . . . .	57
3.6.2	Bivariate Log-Normal . . . . .	58
3.6.3	Horseshoe Shrinkage . . . . .	61
3.6.4	Poisson Log-Linear Regression . . . . .	64
3.7	Discussion . . . . .	67
<b>4</b>	<b>Learning Time-evolving Dependencies via Dynamic Rank Factor Model</b>	<b>68</b>
4.1	Dynamic Rank Factor Model . . . . .	69
4.1.1	Latent Sparse Dynamic Factor Model . . . . .	70
4.1.2	Extension to Handle Multiple Documents . . . . .	72
4.1.3	Comparison with Admixture Topic Models . . . . .	73
4.2	Conjugate Posterior Inference . . . . .	74
4.2.1	Time-evolving Topic and Word Dependencies . . . . .	76
4.2.2	Gibbs Sampling for the Basic Model . . . . .	76
4.2.3	Gibbs Sampling for the Extended Model . . . . .	80
4.2.4	Accelerated MCMC via Document Subsampling . . . . .	85
4.3	Experiments . . . . .	86
4.3.1	Simulation Study: DRFM with Different Innovations . . . . .	86
4.3.2	Case Study I: State of the Union Dataset . . . . .	88
4.3.3	Case Study II: Analysis of <i>Science</i> Dataset . . . . .	90
4.4	Discussion . . . . .	93
<b>5</b>	<b>Leveraging Dependencies in Heterogeneous Multitask Learning</b>	<b>95</b>
5.1	The Latent Probit Model . . . . .	96

5.2	Sparse Oracle Inequalities . . . . .	99
5.3	Parameter Estimation . . . . .	104
5.3.1	Update of Latent Features Distribution . . . . .	106
5.3.2	Update of Domain Transforms . . . . .	107
5.3.3	Update of Probit Classifier . . . . .	107
5.4	Experimental Results . . . . .	107
5.4.1	Cancer Diagnosis . . . . .	107
5.4.2	Mine Detection . . . . .	109
5.5	Discussion . . . . .	111
<b>6</b>	<b>Conclusions</b>	<b>115</b>
6.1	Summary of Contributions . . . . .	115
6.2	Future Directions . . . . .	117
	<b>Bibliography</b>	<b>119</b>
	<b>Biography</b>	<b>129</b>

# List of Tables

3.1	Equivalent Representations of VGC Proposals . . . . .	49
4.1	The 25 Topics of the State of the Union Dataset . . . . .	92
4.2	Selected 20 topics associated with the analysis of the <i>Science</i> dataset and top 10 most probable words. . . . .	92
5.1	The performance of the LPM on Wisconsin breast cancer data with $\alpha = \eta\sqrt{\gamma}$ and $\vartheta = \sqrt{\lambda}$ taking various values. The numbers shown are the improvements in AUC (%) relative to STL, averaged over 50 independent runs. . . . .	108

# List of Figures

2.1	Contour plots for joint posteriors of hyperparameters $q(\sigma^2, \lambda^2   \mathbf{y})$ . . .	33
2.2	Marginal posterior of hyperparameters and coefficients: (a) $q(\sigma^2   \mathbf{y})$ , (b) $q(\lambda^2   \mathbf{y})$ ; (c) $q(x_1   \mathbf{y})$ , (d) $q(x_2   \mathbf{y})$ . . . . .	34
2.3	Posterior marginals of hyperparameters: (a) $q(\sigma^2   \mathbf{y})$ and (b) $q(\lambda^2   \mathbf{y})$ ; posterior marginals of coefficients: (c)-(f) $q(x_j   \mathbf{y})$ ( $j = 1, \dots, 4$ ) . . . .	37
2.4	Posterior marginals of coefficients: (a)-(f) $q(x_j   \mathbf{y})$ ( $j = 5, \dots, 10$ ) . . .	38
2.5	Negative evidence lower bound (ELBO) and elapsed time v.s. grid size; (a), (b) for the Diabetes dataset ( $n = 442, p = 10$ ). (c), (d) for the Prostate cancer dataset ( $n = 97, p = 8$ ) . . . . .	39
3.1	Marginal Adaptation: VIT-BP v.s. VGC-BP . . . . .	59
3.2	Approximate Posteriors via VGC methods . . . . .	60
3.3	RMSE( $\rho$ ) of VGC-LN and VGC-BP v.s. Iterations; Left two: $\rho = 0.4$ ; Right two: $\rho = -0.4$ . . . . .	61
3.4	(Left Panel) Approximated Posteriors (Shown in Log Space for Vi- sualization Purpose); (Right Panel) comparison of ELBO of different variational methods . . . . .	64
3.5	Univariate Margins and Pairwise Posteriors . . . . .	66
4.1	Estimated posterior mean of factor score $s$ with 95% confidence in- terval for $P = 10, K = 2$ and $T = 150$ . Left column: Gaussian innovation with fixed variance $\tau = 1$ ; middle column: Gaussian in- novation with unknown variance $\tau^{-1} \sim \mathcal{G}(0.01, 0.01)$ ; right column: heavy-tailed innovation. . . . .	87

4.2	Estimated posterior mean of latent variable $\mathbf{z}$ vs. the observed data $\mathbf{y}$ for $P = 10$ , $K = 2$ and $T = 150$ . Left: Gaussian innovation with fixed variance $\tau = 1$ ; middle: Gaussian innovation with unknown variance $\tau^{-1} \sim \mathcal{G}(0.01, 0.01)$ ; right: heavy-tailed innovation. . . . .	88
4.3	Above: Time evolving from 1790 to 2014 in the State of the Union dataset for six selected topics. The plotted values represent the posterior means. Below: Top 12 most probable words associated with the above topics. . . . .	89
4.4	First row: Inferred correlations between topics for some specific years associated with some meaningful historical events. Green edges indicate positive correlations and red edges indicate negative correlations. Second row: Learned dendrogram based upon the correlation matrix between the top 10 words associated with each topic (we display 80 unique words in total). . . . .	91
4.5	Time evolving topics from 1790 to 2014. Left up panel: Topics 1 to 7. Right up panel: Topics 8 to 13. Left bottom panel: Topics 14 to 19. Right bottom panel: Topics 20 to 25. The plotted values represent the posterior means. . . . .	93
4.6	The inferred latent trend for variable $\hat{z}_{p,1:T}$ associated with words. . . . .	93
4.7	Inferred correlations between topics for some specific years. Green edges indicate positive correlations and red edges indicate negative correlations. . . . .	94
5.1	A graphic representation of the proposed latent probit model, where solid circles denote data, hollow circles denote unknown parameters and latent variables, and diamonds denote input parameters (including hyper-parameters and fixed model parameters). . . . .	96
5.2	A comparison of performance on the Wisconsin breast cancer data; (a) multitask learning; (b) transfer learning with the original data as the source domain and the diagnostic data as the target domain; (c) transfer learning with the diagnostic data as the source domain and the original data as the target domain. . . . .	112
5.3	A comparison of performance on the the land-mine/underwater mine detection problem; (a) multitask learning; (b) transfer learning with land-mine data as the source domain and underwater mine data as the target domain; (c) transfer learning with underwater mine data as the source domain and land-mine data as the target domain. . . . .	113

5.4	Average AUC on 19 tasks in the landmine detection problem (20 independent runs) . . . . .	114
-----	---	-----

# Acknowledgements

First and foremost I wish to would like to express my sincere gratitude to my advisor, Professor Lawrence Carin, for his excellent guidance, support, and inspiration. Thank you for always being supportive, encouraging me to think critically and independently, and giving me the freedom to explore a diverse set of topics.

I would also like to show my deep appreciation for my dissertation committee members, Professors David B. Dunson, Guillermo Sapiro, Ingrid Daubechies and Katherine Heller for their time and effort to serve on my committee. Special thanks to Professor David B. Dunson for his insightful suggestions on my variational copula work. I am also very grateful to Professors Rebecca Willet, Loren W. Nolte, and Barbara Engelhardt for kindly providing me with guidance and help through classes and reading groups.

I deeply appreciate Drs. Xuejun Liao, Esther Salazar and Yan Kaganovsky for their enlightening discussions and valuable collaborations on my research. Besides, I want to thank my colleagues and friends for providing an excellent research atmosphere and stimulating learning environment within the group. Special thanks go to Bo Chen, Minhua Chen, Mingyuan Zhou, Zhengming Xing, Lingbo Li, Miao Liu, Haojun Chen, Lu Ren, Eric Wang, Yingjian Wang, Xianxing Zhang, Wei Lu, David Carlson, Wenzhao Lian, Changwei Hu, Kyle Ulrich, Zhe Gan, Yunchen Pu, Zhao Song, Yizhe Zhang, Chunyuan Li, Jianbo Yang, Liming Wang, Ricardo Henao, and Piyush Rai for their suggestions, assistance and discussions. I want to thank-

fully acknowledge IBM Thomas J. Watson Research Center for hosting my summer internship in 2014, as well as the many organizations that have provided funding support on my doctoral studies, including NSF, DARPA, ONR, DOE, NGA and ARO.

Over the years, I have fortunately encountered many great people at Duke. Many of them have positively influenced me and inspired me in different ways. Among many others, I am especially thankful to Peng Guan and his wife Shenyu Zhai, for always being great friends, always willing to help and give their best suggestions. The moments we have spent together at Wilson recreation center and during our way home are unforgettable.

Finally and most importantly, I would like to dedicate this dissertation to my wife Kun Li , and my parents Jie Han and Yuling Zhang, for their love and support.

# Introduction

In this dissertation, we study several Bayesian statistical methods for learning, modeling and inference with dependency structures. We exploit the tremendous flexibility provided by Bayesian hierarchical modeling, to recover meaningful dependency structures from data and to allow information sharing and knowledge transferring in statistical learning across heterogeneous domains. We develop two dependent variational inference methods, which are capable of preserve full posterior dependencies among hidden variables in Bayesian (non-conjugate) hierarchical models, aiming to both achieve the accuracy of MCMC and retain its inferential flexibility.

## 1.1 Dependent Variational Inference for Non-Conjugate Models

Markov chain Monte Carlo (MCMC) methods (Gamerman and Lopes, 2006) have been dominant tools for posterior analysis in Bayesian inference. Although MCMC can provide numerical representations of the exact posterior in principle, they usually require intensive runs and are therefore time consuming. Moreover, assessment of a chain's convergence is a well-known challenge (Robert and Casella, 2005). There have been many efforts dedicated to developing deterministic alternatives, including

the Laplace approximation (Kass and Steffey, 1989), variational methods (Jordan *et al.*, 1999a), and expectation propagation (EP) (Minka, 2001). These methods each have their merits and drawbacks (Ormerod, 2011).

More recently, the integrated nested Laplace approximation (INLA) (Rue *et al.*, 2009) has emerged as an attractive method for full posterior inference, which achieves computational accuracy and speed by taking advantage of a (typically) low-dimensional hyper-parameter space, to perform efficient numerical integration and parallel computation on a discrete grid. INLA considers a subclass of structured additive regression models, named latent Gaussian models (LGMs).

In the machine learning community, variational inference has received significant use as an efficient alternative to MCMC. It is also attractive because it provides a closed-form lower bound to the model evidence.

To make inference tractable, mean-field variational Bayes (MFVB) methods (Jordan *et al.*, 1999b; Wainwright and Jordan, 2008) assume  $q(\mathbf{x})$  is factorized over a certain partition of the latent variables  $\mathbf{x} \equiv [\mathbf{x}_1, \dots, \mathbf{x}_J]$ ,  $q_{\text{VB}}(\mathbf{x}) = \prod_j q_{\text{VB}}(\mathbf{x}_j)$ , with marginal densities  $q_{\text{VB}}(\mathbf{x}_j)$  in free-form and correlations between partitions neglected. The structured mean-field approaches (Saul and Jordan, 1996; Hoffman and Blei, 2015) preserve partial correlations and apply only to models with readily identified substructures. The variational Gaussian (VG) approximation (Barber and Bishop, 1998; Opper and Archambeau, 2009a) allows incorporation of correlations by postulating a multivariate Gaussian parametric form  $q_{\text{VG}}(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . The VG approximation, with continuous margins of real variables, are not suitable for variables that are inherently positive or constrained, skewed, or heavy tailed. For multi-modal posteriors, a mixture of MFVB (Jaakkola and Jordan, 1998) or a mixture of uniformly-weighted Gaussians (Gershman *et al.*, 2012) may be employed, which usually requires a further lower bound on the average over the logarithm of the mixture distribution. An active area of research has been focused on developing more

efficient and accurate variational inference algorithms, for example, collapsed inference (Hensman *et al.*, 2012; Foulds *et al.*, 2013), non-conjugate models (Paisley *et al.*, 2012; Wang and Blei, 2012), multimodal posteriors (Gershman *et al.*, 2012), and fast convergent methods (Challis and Barber, 2011; Khan *et al.*, 2012; Kaganovsky *et al.*, 2015a,b).

In Chapter 2, we present a hierarchical mixture-based approach (Han *et al.*, 2013) without requiring the widely used factorized approximation to the posterior. Inspired by INLA, we propose a hybrid continuous-discrete variational approximation, which enables us to preserve full posterior dependencies and is therefore more accurate than the mean-field variational Bayes (MFVB) method (Beal, 2003). The continuous variational approximation is flexible enough for various kinds of latent fields, which makes our method applicable to more general settings than assumed by INLA. The discretization of the low-dimensional hyper-parameter space can overcome the potential non-conjugacy and multimodal posterior problems in variational inference.

In Chapter 3, to address the limitations of current variational methods in failing to simultaneously characterize the posterior dependencies among latent variables while allowing skewness, multimodality, and other characteristics, we propose a new variational copula framework (Han *et al.*, 2016). Our approach decouples the overall inference task into two subtasks: (i) inference of the copula function, which captures the multivariate posterior dependencies; (ii) inference of a set of univariate margins, which are allowed to take essentially any form. Motivated by the work on automated (black-box) variational inference (Ranganath *et al.*, 2014; Mnih and Gregor, 2014; Titsias and Lázaro-Gredilla, 2014; Nguyen and Bonilla, 2014; Kingma and Welling, 2014), we present a stochastic optimization algorithm for *generic* hierarchical Bayesian models with continuous variables, which (i) requires minimal model-specific derivations, (ii) reproduces peculiarities of the true marginal posteriors, and (iii) identifies interpretable dependency structure among latent variables.

## 1.2 Bayesian Modeling of Latent Dependency Structures

Multivariate longitudinal ordinal or count data arise in many areas, including economics, opinion polls, text mining, and social science research. Due to the lack of discrete multivariate distributions supporting a rich enough correlation structure, one popular choice in modeling correlated categorical data employs the multivariate normal mixture of independent exponential family distributions, after appropriate transformations. Examples include the logistic-normal model for compositional data (Aitchison, 1982), the Poisson log-normal model for correlated count data (Chib and Winkelmann, 2001), and the ordered probit model for multivariate ordinal data (Lawrence *et al.*, 2008). Moreover, a dynamic Bayesian extension of the generalized linear model (West *et al.*, 1985) may be considered, for capturing the temporal dependencies of non-Gaussian data (such as ordinal data). In this general framework, the observations are assumed to follow an exponential family distribution, with natural parameter related to a conditionally Gaussian dynamic model (Cargnoni *et al.*, 1997), via a nonlinear transformation. However, these model specifications may still be too restrictive in practice, for the following reasons: (i) Observations are usually discrete, non-negative and with a massive number of zero values and, unfortunately, far from any standard parametric distributions (*e.g.*, multinomial, Poisson, negative binomial and even their zero-inflated variants). (ii) The number of contemporaneous series can be large, bringing difficulties in sharing/learning statistical strength and in performing efficient computations. (iii) The linear state evolution is not truly manifested after a nonlinear transformation, where positive shocks (such as outliers and jumps) are magnified and negative shocks are suppressed; hence, handling temporal jumps (up and down) is a challenge for the above models.

In Chapter 4, we present a flexible semi-parametric Bayesian model, termed *dynamic rank factor model* (DRFM) (Han *et al.*, 2014), that does not suffer these

drawbacks. We first reduce the effect of model misspecification by modeling the sampling distribution non-parametrically. To do so, we fit the observed data only after some implicit monotone transformation, learned automatically via the extended rank likelihood (Hoff, 2007). Second, instead of treating panels of time series as independent collections of variables, we analyze them jointly, with the high-dimensional cross-sectional dependencies estimated via a latent factor model. Finally, by avoiding nonlinear transformations, both smooth transitions and sudden changes (“jumps”) are better preserved in the state-space model, using heavy-tailed innovations.

The proposed model offers an *alternative* to both dynamic and correlated topic models (Blei and Lafferty, 2006b,a; Ahmed and Xing, 2010), with additional modeling facility of word dependencies, and improved ability to handle jumps. It also provides a semi-parametric Bayesian treatment of dynamic sparse factor model. Further, our proposed framework is applicable in the analysis of multiple ordinal time series, where the innovations follow either stationary Gaussian or heavy-tailed distributions.

### 1.3 Learning with Task Dependencies across Heterogeneous Domains

There are two basic approaches for analysis of data from two or more tasks, single-task learning (STL) and multi-task learning (MTL). Whereas STL solves each task in isolation, with possible relations between the tasks ignored, MTL solves the tasks jointly, exploiting between-task relations to reduce the hypothesis space and improve generalization (Baxter, 2000). The advantage of MTL is known to be manifested when the tasks are truly related and the task relations are appropriately employed. For supervised learning, in particular, MTL can achieve the same level of generalization performance as STL, and yet uses significantly fewer labeled examples per task (Baxter, 2000). The reduced sample complexity in each task is achieved by transferring labeling information from related tasks.

While the MTL literature has primarily assumed that the tasks have the same input and output domains and differ only in data distributions (Baxter, 2000; Bakker and Heskes, 2003; Argyriou *et al.*, 2007; Ben-David and Borbely, 2008), a number of recent publications are beginning to break the limit of this assumption, in an attempt of extending MTL to a wider range of applications (He and Rick, 2011; Maayan and Mannor, 2011; Kulis *et al.*, 2011; Wang and Mahadevan, 2011).

In these recent publications, different tasks are permitted to have different feature spaces. In particular, He and Rick (2011) simultaneously performs multi-view learning in each task and multi-task learning in shared views, assuming each task has its own features but may also share features with other tasks. The method in Maayan and Mannor (2011) allows tasks to have different feature representations, learning rotations between the feature representations by matching the tasks' empirical means and covariance matrices. The work in Kulis *et al.* (2011) considers a source task and a target task, assumed to have different feature dimensions, and learns a nonlinear transformation between the source feature domain and the target feature domain using kernel techniques. Finally, Wang and Mahadevan (2011) employs a manifold alignment technique to map each task's input domain to a common latent space, with the task-specific maps achieving the goal of simultaneously clustering examples with the same label, separating examples with different labels, and preserving the topology of each task's manifold.

In Chapter 5, we address the problem of multi-task learning across heterogeneous domains (Han *et al.*, 2012), assuming that each task is a binary classification with a task-specific feature representation. The approach we take differs from Maayan and Mannor (2011); Kulis *et al.* (2011); Wang and Mahadevan (2011) in several important aspects. First, while these previous methods all learn domain transforms and classification in two separate steps, we integrate the two steps by learning domain transforms and classification jointly. Secondly, the domain transforms in our

approach are represented by sparse matrices, with the sparsity enforced by a Laplacian prior on the transform matrices (corresponding to an  $\ell_1$  penalty to the log-likelihood). By contrast, all previous methods do not impose sparsity on domain transforms. The third difference is that the overall model in our approach consists of a factor model for the observed features, which can be used to synthesize new data unseen during training. Finally, our approach is semi-supervised, using labeled as well as unlabeled examples to jointly find the domain transforms and the classification. By contrast, the methods in Maayan and Mannor (2011); Kulis *et al.* (2011) are supervised, and the method in Wang and Mahadevan (2011) is semi-supervised in learning domain transforms, but supervised in learning classification. While full supervision can be challenged by the scarcity of labeled examples (typically assumed in MTL), semi-supervision is doubly beneficial to a joint learning approach, in which unlabeled examples help to perform classification, while labeled examples help to find the domain transforms.

The proposed approach is based on a sparse hierarchical Bayesian model, referred to as the *latent probit model* (LPM), which jointly represents the sparse domain transforms and a common sparse probit classifier (Albert and Chib, 1993) in the latent feature space, with the sparsity imposed by a hierarchical Laplacian prior (Figueiredo, 2003). We employ expectation-maximization (EM) to find the maximum *a posteriori* (MAP) solution to the domain transforms and probit parameters.

Finally, we conclude by summarizing our contributions and outlining some directions for future research in Chapter 6.

## Mixture-based Dependent Variational Inference

In this chapter, we present a hierarchical mixture based approach with hybrid discrete-continuous variational proposals. With the goal of capturing the posterior variable dependencies via efficient and possibly parallel computation, our approach unifies the integrated nested Laplace approximation (INLA) under the variational framework. The proposed method is applicable in more challenging scenarios than typically assumed by INLA, such as Bayesian Lasso, which is characterized by the non-differentiability arising from Laplace priors. We derive an upper bound for the Kullback-Leibler divergence, which yields a fast closed-form solution via decoupled optimization. Our method is a reliable analytic alternative to Markov chain Monte Carlo (MCMC), and it results in a tighter evidence lower bound than that of mean-field variational Bayes (MFVB) method.

### 2.1 Overview of Variational Inference

The problem of Bayesian posterior inference

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{\int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}}$$

can be equivalently transformed into an variational optimization problem by solving

$$\min_{q(\mathbf{x})} \text{KL}[q(\mathbf{x})||p(\mathbf{x})] - \mathbb{E}_{q(\mathbf{x})}[\ln p(\mathbf{y}|\mathbf{x})], \quad s.t. \int q(\mathbf{x})d\mathbf{x} = 1$$

where prior information and/or domain knowledge of parameter encapsulated in  $p(\mathbf{x})$ , the observed data is modeled as generated from the distribution  $p(\mathbf{y}|\mathbf{x})$ . A crucial component of Bayesian inference is approximating the posterior distribution, which represents the current state of knowledge about the latent variables  $\mathbf{x}$  after data  $\mathbf{y}$  have been observed. When intractable integrals are involved, variational inference methods find an approximation  $q(\mathbf{x})$  within some tractable family to the posterior distribution  $p(\mathbf{x}|\mathbf{y})$  by minimizing the Kullback-Leibler (KL) divergence

$$\text{KL}\{q(\mathbf{x})||p(\mathbf{x}|\mathbf{y})\} = \int q(\mathbf{x})\log [q(\mathbf{x})/p(\mathbf{x}|\mathbf{y})] d\mathbf{x}.$$

Minimizing the Kullback-Leibler divergence  $\text{KL}\{q(\mathbf{x})||p(\mathbf{x}|\mathbf{y})\}$  can be interpreted as maximizing a lower bound on the log marginal data likelihood (model evidence)  $\mathcal{L}[q(\mathbf{x})]$ , since

$$\begin{aligned} \ln [p(\mathbf{y})] &= \ln [p(\mathbf{y})] \int q(\mathbf{x})d\mathbf{x} = \int q(\mathbf{x}) \ln [p(\mathbf{y})]d\mathbf{x} = \int q(\mathbf{x}) \ln \left( \frac{p(\mathbf{y}, \mathbf{x}) q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y}) q(\mathbf{x})} \right) d\mathbf{x} \\ &= \underbrace{\int q(\mathbf{x}) \ln \left( \frac{p(\mathbf{y}, \mathbf{x})}{q(\mathbf{x})} \right) d\mathbf{x}}_{\mathcal{L}[q(\mathbf{x})]} + \underbrace{\int q(\mathbf{x}) \ln \left( \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})} \right) d\mathbf{x}}_{\text{KL}\{q(\mathbf{x})||p(\mathbf{x}|\mathbf{y})\}} \end{aligned}$$

## 2.2 Parameter Estimation and Marginal Approximation

Variational methods offer different levels of Bayesian inference (MacKay, 1992) from point parameter estimation, such as maximum a posteriori (MAP) estimation, to marginal posterior approximation and full posterior approximation. Although we focus on full posterior inference for non-conjugate Bayesian models in this dissertation, we first introduce several variational algorithms for partial Bayesian inference within the special conjugate exponential family (Ghahramani and Beal, 2001).

### 2.2.1 Conjugate-Exponential Models

Conjugate-exponential (CE) models (Ghahramani and Beal, 2001) satisfy two conditions:

1. The *complete data likelihood* is in the exponential family:

$$p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{x}, \mathbf{y})g(\boldsymbol{\theta}) \exp[\boldsymbol{\phi}(\boldsymbol{\theta})^T \mathbf{u}(\mathbf{x}, \mathbf{y})] \quad (2.1)$$

where  $\boldsymbol{\phi}(\boldsymbol{\theta})$  is the vector of natural parameters, and  $\mathbf{u}$  and  $f$  and  $g$  are the functions that define the exponential family.

2. The *parameter prior* is conjugate to the complete data likelihood:

$$p(\boldsymbol{\theta}|\eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu})g(\boldsymbol{\theta})^\eta \exp[\boldsymbol{\phi}(\boldsymbol{\theta})^T \boldsymbol{\nu}] \quad (2.2)$$

where  $\eta$  and  $\boldsymbol{\nu}$  are hyperparameters of the prior.

### 2.2.2 Variational Expectation Maximization

We assume  $\mathbf{y}$  is the observed variable,  $\mathbf{x}$  is the latent variable not observed, and the joint probability for  $\mathbf{y}$  and  $\mathbf{x}$  is parameterized by  $\boldsymbol{\theta}$ . We take the maximum likelihood estimation (MLE) computation  $\hat{\boldsymbol{\theta}}$  as an example. Similar properties can easily extend to MAP estimation as well.

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}), \quad p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{p(\mathbf{y})} \int p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) d\mathbf{x} := \mathcal{L}(\boldsymbol{\theta})$$

The standard Expectation Maximization (EM) algorithm (A1) for MLE is as follows,

- E Step: Compute  $q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{(t)})$
- M Step:  $\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \int q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) d\mathbf{x}$ .

Following Neal and Hinton (1998), define a function  $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}, \mathbf{y})$  as follows:

$$\begin{aligned} \mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}, \mathbf{y}) &= -\text{KL}[q_{\mathbf{x}}(\mathbf{x})||p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})] + \mathcal{L}(\boldsymbol{\theta}) \\ &= \mathbb{E}_{q_{\mathbf{x}}(\mathbf{x})}[\log p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})] + H[q_{\mathbf{x}}(\mathbf{x})], \quad H[q_{\mathbf{x}}(\mathbf{x})] = -\mathbb{E}_{q_{\mathbf{x}}(\mathbf{x})}[\log q_{\mathbf{x}}(\mathbf{x})] \end{aligned}$$

The standard EM algorithm (A1) can be expressed in an equivalent algorithm (A2) in terms of function  $\mathcal{F}$  as follows,

- E Step: Set  $q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{(t)})$  to maximize  $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}^{(t-1)}, \mathbf{y})$
- M Step: Set

$$\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{q_{\mathbf{x}}^{(t+1)}(\mathbf{x})} [\ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})] = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}), \boldsymbol{\theta}, \mathbf{y}).$$

The key of the EM algorithm is to compute  $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ . However, in some cases out of the scope of conjugate exponential family, it is not possible to compute it analytically. According to Jordan *et al.* (1999b), one can approximate the exact  $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$  by assuming an appropriate  $q_{\mathbf{x}}(\mathbf{x})$  within certain tractable family  $\mathcal{Q}$ . The variational EM algorithm (A3) for the MLE is as follows,

- E Step: Compute  $q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = \operatorname{argmax}_{q_{\mathbf{x}}(\mathbf{x}) \in \mathcal{Q}} \mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}^{(t)}, \mathbf{y})$
- M Step: Set

$$\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{q_{\mathbf{x}}^{(t+1)}(\mathbf{x})} [\ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})] = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}), \boldsymbol{\theta}, \mathbf{y}).$$

Variational EM applies to more broad cases than standard EM algorithm. For example, for a non-conjugate Poisson log-linear regression model, one can constrain  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$  to be multivariate Gaussian in the E-step. Further, one can optimize  $\boldsymbol{\theta}$  in the M-step to allow automatic relevance determination (ARD) (Tipping, 2001) under the variational EM, or more broadly, alternating minimization (AM) framework (Kaganovsky *et al.*, 2015a,b). The asymptotic distributional behavior of variational Gaussian approximation as a point parameter estimators in a single predictor Poisson mixed model is derived in Hall *et al.* (2011).

### 2.2.3 Variational Bayesian EM

According to (Beal and Ghahramani, 2002), for the conjugate exponential family model, given an i.i.d. data set  $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ , at every iteration of the variational Bayesian EM algorithm and at the maxima of  $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y})$ :

$$\ln p(\mathbf{y}) \geq \int q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} d\mathbf{x} d\boldsymbol{\theta} := \mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y})$$

1.  $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$  is conjugate with parameters  $\tilde{\eta} = \eta + n$ ,  $\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu} + \sum_{i=1}^n \bar{\mathbf{u}}(\mathbf{y}_i)$ :

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = h(\tilde{\eta}, \tilde{\boldsymbol{\nu}}) g(\boldsymbol{\theta})^{\tilde{\eta}} \exp[\boldsymbol{\phi}(\boldsymbol{\theta})^T \tilde{\boldsymbol{\nu}}] \quad (2.3)$$

where  $\bar{\mathbf{u}}(\mathbf{y}_i) = \mathbb{E}_{q_{\mathbf{x}_i}}[\mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)]$ .

2.  $q_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^n q_{\mathbf{x}_i}(\mathbf{x}_i)$  with

$$q_{\mathbf{x}_i}(\mathbf{x}_i) = p(\mathbf{x}_i | \mathbf{y}_i, \bar{\boldsymbol{\phi}}) \propto f(\mathbf{x}_i, \mathbf{y}_i) \exp[\bar{\boldsymbol{\phi}}^T \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)] \quad (2.4)$$

where  $\bar{\boldsymbol{\phi}} = \mathbb{E}_{q_{\boldsymbol{\theta}}}[\boldsymbol{\phi}(\boldsymbol{\theta})]$ , the expectation of the natural parameter.

In summary, the Variational Bayesian EM (VB-EM) algorithm for MAP estimation (A4) is as follows,

- VB-E Step: Compute  $q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x} | \mathbf{y}, \bar{\boldsymbol{\phi}}^{(t)})$
- VB-M Step:  $q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \propto \exp \int q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) d\mathbf{x}$ .

The EM algorithm (A5) for the MAP estimation  $\hat{\boldsymbol{\theta}}^* = \operatorname{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta} | \mathbf{y})$  is as follows,

- E Step: Compute  $q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x} | \mathbf{y}, \boldsymbol{\theta}^{(t)})$
- M Step:  $\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \int q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) d\mathbf{x}$ .

VB-EM (A4) reduces to the EM algorithm (A5) if we degenerate the parameter density to a point estimate (Dirac delta function),  $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ , in which the M step involves re-estimating  $\boldsymbol{\theta}^*$ . In this dissertation, we are interested in full posterior inference, i.e., inferring the joint posterior distribution  $q(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y})$ , which is applicable to models outside the scope of conditionally conjugate models.

### 2.3 Integrated Nested Laplace Approximation (INLA)

In structured additive regression models, observations  $\mathbf{y} \sim p(\mathbf{y}|\boldsymbol{\eta}) = \prod_i p(y_i|\eta_i)$ , where  $\eta_i$  are linear predictors,  $\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(\mu_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i$ . If we assign Gaussian priors on  $\alpha$ ,  $\{f^{(j)}(\cdot)\}$ ,  $\{\beta_k\}$  and  $\{\epsilon_i\}$ , let  $\mathbf{x}$  denote the vector of all the latent Gaussian variables and  $\boldsymbol{\theta}$  the vector of hyperparameters we will have a Bayesian hierarchical model with observation model  $\mathbf{y} \sim p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_i p(y_i|\eta_i, \boldsymbol{\theta})$ , latent variables model  $\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ , and hyperprior  $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ .

In latent Gaussian models (LGMs),  $y_i$  could be non-Gaussian (Poisson, binomial, negative binomial, exponential etc.), the dimension of the latent Gaussian field could be large (e.g.  $10^2 \sim 10^5$ ), and the dimension of parameter  $\boldsymbol{\theta}$  should be no more than 5 or 6.

The exact posterior distribution

$$p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) = \frac{1}{p(\mathbf{y})} p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = \frac{1}{p(\mathbf{y})} p(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) \prod_i p(y_i|x_i, \boldsymbol{\theta}) \quad (2.5)$$

can be difficult to evaluate, since usually the normalization

$$p(\mathbf{y}) = \iint p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} d\boldsymbol{\theta} \quad (2.6)$$

is intractable. The main idea of INLA is discretizing the low-dimensional space  $\boldsymbol{\theta}$  using a grid  $\mathcal{G}$ . INLA includes three steps:

1. Local Laplace approximation (Kass and Steffey, 1989):

$$q_G(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k) = \mathcal{N}(\mathbf{x}; \mathbf{x}^*(\boldsymbol{\theta}_k), \mathbf{H}(\mathbf{x}^*(\boldsymbol{\theta}_k))^{-1}), \quad \forall \boldsymbol{\theta}_k \in \mathcal{G},$$

where  $\mathbf{x}^*(\boldsymbol{\theta}_k) = \underset{\mathbf{x}}{\operatorname{argmax}} p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)$  is the mode, and  $\mathbf{H}(\mathbf{x}^*(\boldsymbol{\theta}_k))$  is the Hessian evaluated at the mode  $\mathbf{x}^*(\boldsymbol{\theta}_k)$

2. Laplace's method of integration (Tierney and Kadane, 1986):

$$q_{LA}(\boldsymbol{\theta}_k|\mathbf{y}) = \frac{p(\mathbf{x}, \boldsymbol{\theta}_k|\mathbf{y})}{q_G(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)} \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta}_k)} \quad (2.7)$$

3. Numerical integration:  $q(\mathbf{x}|\mathbf{y}) = \sum_k q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)q(\boldsymbol{\theta}_k|\mathbf{y})\Delta_k$

INLA provides full posterior inference (i.e. joint density  $q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ ). It is computationally efficient; in some case MCMC may takes hours or days to run, INLA only needs seconds or minutes. INLA offers approximate Bayesian inference for additive regression models, where the latent field is Gaussian. This methodology is particularly attractive if the latent Gaussian model is a GMRF. However, the Gaussian assumption for the latent process prevents INLA from being applied to more general models outside of the family of latent Gaussian models (LGMs). Besides, there is no quantization for the accuracy of approximation  $q_G(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})q_{LA}(\boldsymbol{\theta}|\mathbf{y})$ .

## 2.4 Integrated Non-Factorized Variational Inference

Consider a general Bayesian hierarchical model with observation  $\mathbf{y}$ , latent variables  $\mathbf{x}$ , and hyperparameters  $\boldsymbol{\theta}$ . The exact joint posterior

$$p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{\iint p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})d\mathbf{x}d\boldsymbol{\theta}}$$

can be difficult to evaluate, since usually the normalization  $p(\mathbf{y})$  on the denominator is intractable and numerical integration of  $\mathbf{x}$  is too expensive.

To address this problem, we find a variational approximation to the exact posterior by minimizing the Kullback-Leibler (KL) divergence  $\text{KL}[q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})||p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})]$ . Applying Jensen's inequality to the log-marginal data likelihood, one obtains

$$\begin{aligned} \ln p(\mathbf{y}) &= \ln \iint q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})} d\mathbf{x}d\boldsymbol{\theta} \\ &\geq \iint q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})} d\mathbf{x}d\boldsymbol{\theta} := \mathcal{L} \end{aligned} \quad (2.8)$$

which holds for any proposed approximating distributions  $q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ .  $\mathcal{L}$  is termed the evidence lower bound (ELBO)(Jordan *et al.*, 1999a). The gap in the Jensen's

inequality is exactly the KL divergence. Therefore minimizing the Kullback-Leibler (KL) divergence is equivalent to maximizing the ELBO.

To make the variational problem tractable, the variational distribution  $q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$  is commonly required to take a restricted form. For example, mean-field variational Bayes (VB) method assumes the distribution factorizes into a product of marginals (Beal, 2003),  $q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) = q(\mathbf{x})q(\boldsymbol{\theta})$ , which ignores the posterior dependencies among different latent variables (including hyperparameters) and therefore impairs the accuracy of the approximate posterior distribution.

#### 2.4.1 Hybrid Continuous and Discrete Variational Approximations

We consider a non-factorized approximation to the posterior via a hierarchical model,

$$q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) = q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})q(\boldsymbol{\theta}|\mathbf{y}),$$

to preserve the posterior dependency structure. Unfortunately, this generally leads to a nontrivial optimization problem,

$$\begin{aligned} q^*(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) &= \arg \min_{q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})} \text{KL}(q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})||p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})), \\ &= \arg \min_{q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})} \iint q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \ln \frac{q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}{p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})} d\mathbf{x}d\boldsymbol{\theta}, \\ &= \arg \min_{q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}), q(\boldsymbol{\theta}|\mathbf{y})} \int q(\boldsymbol{\theta}|\mathbf{y}) \left[ \int q(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \ln \frac{q(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})}{p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})} d\mathbf{x} + \ln q(\boldsymbol{\theta}|\mathbf{y}) \right] d\boldsymbol{\theta} \end{aligned} \tag{2.9}$$

We propose a hybrid continuous-discrete variational distribution  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})q_d(\boldsymbol{\theta}|\mathbf{y})$ , where  $q_d(\boldsymbol{\theta}|\mathbf{y})$  is a finite mixture of Dirac-delta distributions,

$$q_d(\boldsymbol{\theta}|\mathbf{y}) = \sum_k \omega_k \delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}), \quad \omega_k = q_d(\boldsymbol{\theta}_k|\mathbf{y}), \quad \sum_k \omega_k = 1.$$

Clearly,  $q_d(\boldsymbol{\theta}|\mathbf{y})$  is an approximation of  $q(\boldsymbol{\theta}|\mathbf{y})$  by discretizing the continuous (typically) low-dimensional parameter space of  $\boldsymbol{\theta}$  using a grid  $\mathcal{G}$  with finite grid

points. The grid points need not to be uniformly spaced, one may put more grid points to potentially high mass regions if credible prior information is available. One can always reduce the discretization error by increasing the number of points in  $\mathcal{G}$ . To obtain a useful discretization at a manageable number of grid points, the dimension of  $\boldsymbol{\theta}$  cannot be too large; this is also the same assumption in INLA (Rue *et al.*, 2009), but we remove here the Gaussian prior assumption of INLA on latent effects  $\boldsymbol{x}$ .

The hybrid variational approximation is found by minimizing the KL divergence, i.e.,

$$\text{KL} [q(\boldsymbol{x}, \boldsymbol{\theta}|\mathbf{y})||p(\boldsymbol{x}, \boldsymbol{\theta}|\mathbf{y})] = \sum_k q_d(\boldsymbol{\theta}_k|\mathbf{y}) \left[ \int q(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathbf{y}) \ln \frac{q(\boldsymbol{x}|\mathbf{y}, \boldsymbol{\theta}_k)}{p(\boldsymbol{x}, \boldsymbol{\theta}_k|\mathbf{y})} d\boldsymbol{x} + \ln q_d(\boldsymbol{\theta}_k|\mathbf{y}) \right] \quad (2.10)$$

which leads to a hierarchical mixture proposal to approximate the marginal posterior,

$$q(\boldsymbol{x}|\mathbf{y}) = \sum_k q(\boldsymbol{x}|\mathbf{y}, \boldsymbol{\theta}_k)q_d(\boldsymbol{\theta}_k|\mathbf{y}).$$

As will be clearer shortly, the problem in (2.10) can be much easier to solve than that in (2.9).

We give the name *integrated non-factorized variational Bayes* (INF-VB) to the method of approximating  $p(\boldsymbol{x}, \boldsymbol{\theta}|\mathbf{y})$  with  $q(\boldsymbol{x}|\mathbf{y}, \boldsymbol{\theta})q_d(\boldsymbol{\theta}|\mathbf{y})$  by solving the optimization problem in (2.10). The use of  $q_d(\boldsymbol{\theta})$  is equivalent to numerical integration, which is a key idea of INLA (Rue *et al.*, 2009), see Section 2.4.3 for details. It has also been used in sampling methods when samples are not easy to obtain directly (Ritter and Tanner, 1992). Here we use this idea in variational inference to overcome the potential non-conjugacy and multimodal posterior problems in  $\boldsymbol{\theta}$ .

#### 2.4.2 Variational Optimization

The proposed INF-VB method consists of two algorithmic steps:

- Step 1: Solving multiple independent optimization problems, each for a grid point in  $\mathcal{G}$ , to obtain the optimal  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)$ ,  $\forall \boldsymbol{\theta}_k \in \mathcal{G}$ , i.e.,

$$\begin{aligned}
q^*(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k) &= \arg \min_{q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)} \sum_k q(\boldsymbol{\theta}_k|\mathbf{y}) \left[ \int q(\mathbf{x}|\boldsymbol{\theta}_k, \mathbf{y}) \ln \frac{q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)}{p(\mathbf{x}, \boldsymbol{\theta}_k|\mathbf{y})} d\mathbf{x} + \ln q(\boldsymbol{\theta}_k|\mathbf{y}) \right] \\
&= \arg \min_{q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)} \int q(\mathbf{x}|\boldsymbol{\theta}_k, \mathbf{y}) \ln \frac{q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)}{p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)} d\mathbf{x} \\
&= \arg \min_{q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)} \text{KL}[q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)||p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)]
\end{aligned}$$

The optimal variational distribution  $q^*(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)$  is the exact posterior  $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)$ . In case it is not available, we may further constrain  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)$  to a parametric form, examples including: (i) multivariate Gaussian (Opper and Archambeau, 2009b), if the posterior asymptotic normality holds; (ii) skew-normal densities (Ormerod, 2011; Challis and Barber, 2012); or (iii) an inducing factorization assumption (see Ch.10.2.5 in (Bishop, 2006)), if the latent variables  $\mathbf{x}$  are conditionally independent or their dependencies are negligible.

- Step 2: Given  $\{q^*(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k) : \boldsymbol{\theta}_k \in \mathcal{G}\}$  obtained in Step 1, one solves

$$\{q^*(\boldsymbol{\theta}_k|\mathbf{y})\} = \arg \min_{\{q(\boldsymbol{\theta}_k|\mathbf{y})\}} \sum_k q_d(\boldsymbol{\theta}_k|\mathbf{y}) \underbrace{\left[ \int q^*(\mathbf{x}|\boldsymbol{\theta}_k, \mathbf{y}) \ln \frac{q^*(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)}{p(\mathbf{x}, \boldsymbol{\theta}_k|\mathbf{y})} d\mathbf{x} + \ln q_d(\boldsymbol{\theta}_k|\mathbf{y}) \right]}_{l(q_d(\boldsymbol{\theta}_k|\mathbf{y}))=l(\omega_k)}$$

Setting  $\partial l(\omega_k)/\partial \omega_k = 0$  (also  $\partial^2 l(\omega_k)/\partial \omega_k^2 > 0$ ), which is solved to give

$$q_d^*(\boldsymbol{\theta}_k|\mathbf{y}) \propto \exp \left( \int q^*(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k) \ln \frac{p(\mathbf{x}, \boldsymbol{\theta}_k|\mathbf{y})}{q^*(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)} d\mathbf{x} \right). \quad (2.11)$$

Note that  $q_d(\boldsymbol{\theta}|\mathbf{y})$  is evaluated at a grid of points  $\boldsymbol{\theta}_k \in \mathcal{G}$ , it needs to be known only up to a multiplicative constant, which can be identified from the normalization constraint  $\sum_k q_d^*(\boldsymbol{\theta}_k|\mathbf{y}) = 1$ . The integral in (2.11) can be analytically evaluated in the application considered in Section 2.5.

### 2.4.3 Links between INF-VB and INLA

The INF-VB is a variational extension of the integrated nested Laplace approximations (INLA) (Rue *et al.*, 2009), a deterministic Bayesian inference method for latent Gaussian models (LGMs), to the case when  $p(\mathbf{x}|\boldsymbol{\theta})$  exhibits strong non-Gaussianity and hence  $p(\boldsymbol{\theta}|\mathbf{y})$  may not be approximated accurately by the Laplace’s method of integration (Tierney and Kadane, 1986). To see the connection, we review briefly the three computation steps of INLA and compare them with INF-VB in below:

1. Based on the Laplace approximation (Kass and Steffey, 1989), INLA seeks a Gaussian distribution

$$q_G(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k) = \mathcal{N}(\mathbf{x}; \mathbf{x}^*(\boldsymbol{\theta}_k), \mathbf{H}(\mathbf{x}^*(\boldsymbol{\theta}_k))^{-1}), \quad \forall \boldsymbol{\theta}_k \in \mathcal{G}$$

that captures most of the probabilistic mass locally, where

$$\mathbf{x}^*(\boldsymbol{\theta}_k) = \underset{\mathbf{x}}{\operatorname{argmax}} p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)$$

is the posterior mode, and  $\mathbf{H}(\mathbf{x}^*(\boldsymbol{\theta}_k))$  is the Hessian matrix of the log posterior evaluated at the mode. By contrast, INF-VB with the Gaussian parametric constraint on  $q^*(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)$  provides a global variational Gaussian approximation  $q_{VG}(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)$  in the sense that the conditions of the Laplace approximation hold on average (Opper and Archambeau, 2009b). As we will see next, the averaging operator plays a crucial role in handling the non-differentiable  $\ell_1$  norm arising from the double-exponential priors.

2. INLA computes the marginal posteriors of  $\boldsymbol{\theta}$  based on the Laplace’s method of integration (Tierney and Kadane, 1986),

$$q_{LA}(\boldsymbol{\theta}|\mathbf{y}) = \left. \frac{p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}{q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})} \right|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})} \quad (2.12)$$

The quality of this approximation depends on the accuracy of  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ . When  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ , one has  $q_{LA}(\boldsymbol{\theta}|\mathbf{y})$  equal to  $p(\boldsymbol{\theta}|\mathbf{y})$ , according to the Bayes

rule. It has been shown in (Rue *et al.*, 2009) that (2.12) is accurate enough for latent Gaussian models with  $q_G(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ . Alternatively, the variational optimal posterior  $q_d^*(\boldsymbol{\theta}|\mathbf{y})$  by INF-VB (2.11) can be derived as a lower bound of the true posterior  $p(\boldsymbol{\theta}|\mathbf{y})$  by Jensen’s inequality.

$$\begin{aligned} \ln p(\boldsymbol{\theta}|\mathbf{y}) &= \ln \left[ \int \frac{p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}{q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})} q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) d\mathbf{x} \right] \\ &\geq \int \ln \left[ \frac{p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}{q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})} \right] q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) d\mathbf{x} = \ln q_d^*(\boldsymbol{\theta}|\mathbf{y}) \end{aligned} \quad (2.13)$$

Its optimality justifications in Section 2.4.2 also explain the often observed empirical successes of hyperparameter selection based on the ELBO of  $\ln p(\mathbf{y}|\boldsymbol{\theta})$  (Challis and Barber, 2011), when the first level of Bayesian inference is performed, i.e. only the conditional posterior  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$  with fixed  $\boldsymbol{\theta}$  is of interest. In Section 4.3 we compare the accuracies of both (2.11) and (2.12) for hyperparameter learning.

- INLA obtains the marginal distributions of interest, e.g.,  $q(\mathbf{x}|\mathbf{y})$  via numerically integrating out  $\boldsymbol{\theta}$ :  $q(\mathbf{x}|\mathbf{y}) = \sum_k q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k) q(\boldsymbol{\theta}_k|\mathbf{y}) \Delta_k$  with area weights  $\Delta_k$ . In INF-VB, we have  $q_d(\boldsymbol{\theta}|\mathbf{y}) = \sum_k \omega_k \delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})$ . Let  $\omega_k = q(\boldsymbol{\theta}_k|\mathbf{y}) \Delta_k$ , we immediately have

$$\begin{aligned} q(\mathbf{x}|\mathbf{y}) &= \int q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) q_d(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &= \sum_k q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k) q_d(\boldsymbol{\theta}_k|\mathbf{y}) = \sum_k q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k) q(\boldsymbol{\theta}_k|\mathbf{y}) \Delta_k \end{aligned}$$

This Dirac-delta mixture interpretation of numerical integration also enables us to quantitize the accuracy of INLA approximation  $q_G(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) q_{LA}(\boldsymbol{\theta}|\mathbf{y})$  using the KL divergence to  $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$  under the variational framework.

In contrast to INLA, INF-VB provides  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$  and  $q_d(\boldsymbol{\theta}|\mathbf{y})$ , both are optimal in a sense of the minimum Kullback-Leibler divergence, within the proposed hybrid

distribution family. In this chapter we focus on the full posterior inference of Bayesian Lasso (Park and Casella, 2008) where the local Laplace approximation in INLA cannot be applied, as the non-differentiability of the  $\ell_1$  norm prevents one from computing the Hessian matrix. Besides, if we do not exploit the scale mixture of normals representation (Andrews and Mallows, 1974) of Laplace priors (i.e., no data-augmentation), we are actually dealing with a non-conjugate variational inference problem in Bayesian Lasso.

## 2.5 Variational Bayesian Lasso

Consider the Bayesian Lasso regression model (Park and Casella, 2008),  $\mathbf{y} = \Phi \mathbf{x} + \mathbf{e}$ , where  $\Phi \in \mathbb{R}^{n \times p}$  is the design matrix containing predictors,  $\mathbf{y} \in \mathbb{R}^n$  are responses. We assume that both  $\mathbf{y}$  and the columns of  $\Phi$  have been mean-centered to remove the intercept term, and  $\mathbf{e} \in \mathbb{R}^n$  contain independent zero-mean Gaussian noise  $\mathbf{e} \sim \mathcal{N}(\mathbf{e}; \mathbf{0}, \sigma^2 \mathbf{I}_n)$ . (Park and Casella, 2008) suggested using scaled double-exponential priors under which they showed that  $p(\mathbf{x}, \sigma^2 | \mathbf{y}, \lambda)$  is unimodal, further, the unimodality helps to accelerate convergence of the data-augmentation Gibbs sampler and makes the posterior mode more meaningful. Gamma prior is put on  $\lambda^2$  for conjugacy. Following (Park and Casella, 2008) we assume,

$$\begin{aligned} x_j | \sigma^2, \lambda^2, &\sim \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}} \|x_j\|_1\right), \\ \sigma^2 &\sim \text{InvGamma}(\sigma^2; a, b), \\ \lambda^2 &\sim \text{Gamma}(\lambda^2; r, s) \end{aligned}$$

While the Lasso estimates (Tibshirani, 1996) provide only the posterior modes of the regression parameters  $\mathbf{x} \in \mathbb{R}^p$ , Bayesian Lasso (Park and Casella, 2008) provides the complete posterior distribution  $p(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y})$ , from which one may obtain whatever statistical properties are desired of  $\mathbf{x}$  and  $\boldsymbol{\theta}$ , including the posterior mode, mean, median, and credible intervals.

Since in our approach variational Gaussian approximation is performed separately (see Section 2.5.1) for each hyperparameter  $\{\lambda, \sigma^2\}$  considered, the efficiency of approximating  $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$  is particularly important. The upper bound of the KL divergence derived in Section 2.5.2 provides an approximate closed-form solution, that is often accurate enough or requires a small number of gradient iterations to converge to optimality. The tightness of the upper bound is analyzed using spectral-norm bounds (See Section 2.5.3), which also provide insights on the connection between the deterministic Lasso (Tibshirani, 1996) and the Bayesian Lasso (Park and Casella, 2008).

### 2.5.1 Variational Gaussian Approximation

The conditional distribution of  $\mathbf{y}$  and  $\mathbf{x}$  given  $\boldsymbol{\theta}$  is

$$p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) = \frac{\lambda^p/(2\sigma)^p}{\sqrt{(2\pi\sigma^2)^n}} \exp \left\{ -\frac{\|\mathbf{y} - \Phi\mathbf{x}\|^2}{2\sigma^2} - \frac{\lambda}{\sigma} \|\mathbf{x}\|_1 \right\}.$$

The postulated approximation,  $q(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{D})$ , is a multivariate Gaussian density (dropping dependencies of variational parameters  $(\boldsymbol{\mu}, \mathbf{D})$  on  $(\boldsymbol{\theta}, \mathbf{y})$  for brevity), whose parameters  $(\boldsymbol{\mu}, \mathbf{D})$  are found by minimizing the KL divergence to  $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ ,

$$\begin{aligned} g(\boldsymbol{\mu}, \mathbf{D}) &\stackrel{Def.}{=} \text{KL}(q(\mathbf{x}; \boldsymbol{\mu}, \mathbf{D}) \| p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})) = \int q(\mathbf{x}; \boldsymbol{\mu}, \mathbf{D}) \ln \frac{q(\mathbf{x}; \boldsymbol{\mu}, \mathbf{D})}{p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})} d\mathbf{x} \\ &= \int q(\mathbf{x}; \boldsymbol{\mu}, \mathbf{D}) \ln \frac{q(\mathbf{x}; \boldsymbol{\mu}, \mathbf{D})}{p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})} d\mathbf{x} + \ln p(\mathbf{y}|\boldsymbol{\theta}), \\ &= -\frac{1}{2} \ln |\mathbf{D}| + \frac{\|\mathbf{y} - \Phi\boldsymbol{\mu}\|^2 + \text{tr}(\Phi'\Phi\mathbf{D})}{2\sigma^2} + \frac{\lambda}{\sigma} \mathbb{E}_q(\|\mathbf{x}\|_1) + \ln p(\mathbf{y}|\boldsymbol{\theta}) - \ln \psi(\sigma^2, \lambda) \end{aligned} \tag{2.14}$$

with

$$\mathbb{E}_q(\|\mathbf{x}\|_1) = \sum_{j=1}^p \left[ \mu_j - 2\mu_j \Psi(h_j) + 2\sqrt{d_j} \psi(h_j) \right], \quad h_j = -\mu_j \sqrt{d_j}, \quad d_j = \mathbf{D}_{jj}$$

where  $\psi(\sigma^2, \lambda) = (4\pi e\lambda^2\sigma^{-2})^{p/2}(2\pi\sigma^2)^{-n/2}$ ,  $\Psi(\cdot)$  and  $\psi(\cdot)$  corresponds to the standard normal cumulative distribution function and probability density function, respectively. Expectation is taken with respect to  $q(\mathbf{x}; \boldsymbol{\mu}, \mathbf{D})$ . Define  $\mathbf{D} = \mathbf{C}\mathbf{C}^T$ , where  $\mathbf{C}$  is the Cholesky factorization of the covariance matrix  $\mathbf{D}$ . Since  $g(\boldsymbol{\mu}, \mathbf{D})$  is convex in the parameter space  $(\boldsymbol{\mu}, \mathbf{C})$ , a global optimal variational Gaussian approximation  $q^*(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$  is guaranteed, which achieves the minimum KL divergence to  $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$  within the family of multivariate Gaussian densities specified (Challis and Barber, 2011).

As a first step, one finds  $q^*(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$  using gradient based procedures independently for each hyperparameter combinations  $\{\lambda, \sigma^2\}$ . Second,  $q^*(\boldsymbol{\theta}|\mathbf{y})$  can be evaluated analytically using either (2.11) or (2.12); both will yield a finite mixture of Gaussian distribution for the marginal posterior  $q(\mathbf{x}|\mathbf{y})$  via numerical integration, which is highly efficient since we only have two hyperparameters in Bayesian Lasso. Finally, the evidence lower bound (ELBO) in (2.8) can also be evaluated analytically after simple algebra. We will show in Section 2.7.3 a comparison with the mean-field variational Bayesian (MFVB) approach, derived based on a scale normal mixture representation (Andrews and Mallows, 1974) of the Laplace prior.

### 2.5.2 Upper Bounds of KL divergence

We provide an approximate solution  $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{D}})$  via minimizing an upper bound of KL divergence (2.14). This solution solves a Lasso problem in  $\boldsymbol{\mu}$ , and has a closed-form expression for  $\mathbf{D}$ , making this computationally efficient. In practice, it could serve as an initialization for gradient procedures.

**Lemma 1.** (*Triangle Inequality*)  $\mathbb{E}_q\|\mathbf{x}\|_1 \leq \mathbb{E}_q\|\mathbf{x} - \boldsymbol{\mu}\|_1 + \|\boldsymbol{\mu}\|_1$ , where  $\mathbb{E}_q\|\mathbf{x} - \boldsymbol{\mu}\|_1 = \sqrt{2/\pi} \sum_{j=1}^p \sqrt{d_j}$ , with the expectation taken with respect to  $q(\mathbf{x}; \boldsymbol{\mu}, \mathbf{D})$ .

**Lemma 2.** For any  $\{d_j \geq 0\}_{j=1}^p$ , it holds  $\sqrt{\sum_{j=1}^p d_j^2} \leq \sum_{j=1}^p d_j \leq \sqrt{p \sum_{j=1}^p d_j^2}$ .

**Lemma 3.** (Golub and Loan, 1996) For any  $\mathbf{A} \in \mathbb{S}_{++}^p$ ,  $\text{tr}(\mathbf{A}^2) \leq \text{tr}(\mathbf{A}) \leq \sqrt{p} \text{tr}(\mathbf{A}^2)$ .

**Theorem 4.** (Upper and Lower bound) For any  $\mathbf{A}, \mathbf{D} \in \mathbb{S}_{++}^p$ ,  $\mathbf{A} = \sqrt{\mathbf{D}}$ ,  $d_j = \mathbf{D}_{jj}$  holds  $\frac{1}{\sqrt{p}} \text{tr}(\mathbf{A}) \leq \sum_{j=1}^p \sqrt{d_j} \leq \sqrt{p} \text{tr}(\mathbf{A})$ .

*Proof.* Each inequality is obtained by first applying Lemma 2 and then Lemma 3,

$$\begin{aligned} \sum_{j=1}^p \sqrt{d_j} &\leq \sqrt{p \sum_{j=1}^p \sqrt{d_j}} = \sqrt{p} \text{tr}(\mathbf{D}) = \sqrt{p} \text{tr}(\mathbf{A}^2) \leq \sqrt{p} \text{tr}(\mathbf{A}) = \sqrt{p} \text{tr}(\sqrt{\mathbf{D}}), \\ \sum_{j=1}^p \sqrt{d_j} &\geq \sqrt{\sum_{j=1}^p \sqrt{d_j}} = \text{tr}(\mathbf{D}) = \text{tr}(\mathbf{A}^2) \geq \frac{1}{\sqrt{p}} \text{tr}(\mathbf{A}) = \frac{1}{\sqrt{p}} \text{tr}(\sqrt{\mathbf{D}}). \end{aligned}$$

□

Since  $\mathbf{D}$  is positive definite, it has a unique symmetric square root  $\mathbf{A} = \sqrt{\mathbf{D}}$ , which can be obtained from  $\mathbf{D}$  by taking square root of the eigenvalues.

Applying Lemma 1 and Theorem 4 in (2.14), one obtains an upper bound for KL divergence,

$$\begin{aligned} f(\boldsymbol{\mu}, \mathbf{D}) &= \underbrace{\frac{\|\mathbf{y} - \Phi \boldsymbol{\mu}\|_2^2}{2\sigma^2} + \frac{\lambda}{\sigma} \|\boldsymbol{\mu}\|_1}_{f_1(\boldsymbol{\mu})} + \underbrace{-\frac{1}{2} \ln |\mathbf{D}| + \frac{\text{tr}(\Phi' \Phi \mathbf{D})}{2\sigma^2} + \frac{\lambda}{\sigma} \sqrt{\frac{2p}{\pi}} \text{tr}(\sqrt{\mathbf{D}})}_{f_2(\mathbf{D})} \\ &+ \ln \frac{p(\mathbf{y}|\boldsymbol{\theta})}{\psi(\sigma^2, \lambda)} \geq g(\boldsymbol{\mu}, \mathbf{D}) = \text{KL}(q(\mathbf{x}; \boldsymbol{\mu}, \mathbf{D}) \| p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})) \end{aligned} \quad (2.15)$$

In the problem of minimizing the KL divergence  $g(\boldsymbol{\mu}, \mathbf{C}\mathbf{C}^T)$ , one needs to iteratively update  $\boldsymbol{\mu}$  and  $\mathbf{C}$ , since they are coupled. However, the upper bound  $f(\boldsymbol{\mu}, \mathbf{D})$  decouples into two additive terms:  $f_1$  is a function of  $\boldsymbol{\mu}$  while  $f_2$  is a function of  $\mathbf{D}$ , which greatly simplifies the minimization.

- The minimization of  $f_1(\boldsymbol{\mu})$  is a convex Lasso problem. Using path-following algorithms (e.g., a modified least angle regression algorithm (LARS) (Efron *et al.*,

2004)), one can efficiently compute the entire solution path of Lasso estimates as a function of  $\lambda_0 = 2\lambda\sigma$  in one shot. Global optimal solutions for  $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_k)$  on each grid point  $\boldsymbol{\theta}_k \in \mathcal{G}$  can be recovered using the piece-wise linear property.

- The function  $f_2(\mathbf{D})$  is convex in the parameter space  $\mathbf{A} = \sqrt{\mathbf{D}}$ , whose minimizer is in closed-form and can be found by setting the gradient to zero and solving the resulting equation,

$$\begin{aligned} \nabla_{\mathbf{A}} f_2 &= -\mathbf{A}^{-1} + \frac{\boldsymbol{\Phi}'\boldsymbol{\Phi}\mathbf{A}}{\sigma^2} + \lambda\sqrt{\frac{2p}{\pi}}\mathbf{I} = \mathbf{0}, \\ \hat{\mathbf{A}} &= \left( \sqrt{\frac{\lambda^2 p}{2\pi\sigma^2}}\mathbf{I} + \sqrt{\frac{\lambda^2 p}{2\pi\sigma^2}}\mathbf{I} + \frac{\boldsymbol{\Phi}'\boldsymbol{\Phi}}{\sigma^2} \right)^{-1}, \end{aligned} \quad (2.16)$$

We have  $\hat{\mathbf{D}} = \hat{\mathbf{A}}^2$ , which is guaranteed to be a positive definite matrix. Note that the global optimum  $\hat{\mathbf{D}}(\boldsymbol{\theta}_k)$  for each grid point  $\boldsymbol{\theta}_k \in \mathcal{G}$  have the same eigenvectors as the Gram matrix  $\boldsymbol{\Phi}'\boldsymbol{\Phi}$  and differ only in eigenvalues. For  $j = 1, \dots, p$ , denote the eigenvalues of  $\mathbf{D}$  and  $\boldsymbol{\Phi}'\boldsymbol{\Phi}$  as  $\alpha_j$  and  $\beta_j$ , respectively. By (2.16), we have

$$\alpha_j = \lambda\sqrt{p/(2\pi\sigma^2)} + \sqrt{\lambda^2 p/(2\pi\sigma^2) + \beta_j/\sigma^2}.$$

Therefore, one can pre-compute the eigenvectors once, and only update the eigenvalues as a function of  $\boldsymbol{\theta}_k$ . This will make the computation efficient both in time and memory.

The solutions  $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{D}})$  which minimize the KL upper bound  $f(\hat{\boldsymbol{\mu}}, \hat{\mathbf{D}})$  in (2.15) achieves its global optimum. Meanwhile, it is also accurate in the sense of the KL divergence  $g(\hat{\boldsymbol{\mu}}, \hat{\mathbf{D}})$  in (2.14), as we will show next. Tightness analysis of the upper bound is also provided, using trace norm bounds.

### 2.5.3 Theoretical Analysis

**Theorem 5.** (*KL Divergence Upper Bound*) Let  $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{D}})$  be the minimizer of the KL upper bound (2.15), i.e.,  $\hat{\boldsymbol{\mu}}$  solves the Lasso and  $\hat{\mathbf{D}}$  is given in (2.16). Then

$$g(\hat{\boldsymbol{\mu}}, \hat{\mathbf{D}}) \leq \min_{\boldsymbol{\mu}, \mathbf{D}} f(\boldsymbol{\mu}, \mathbf{D}) = f_1(\hat{\boldsymbol{\mu}}) + f_2(\hat{\mathbf{D}}) + \ln \frac{p(\mathbf{y}|\boldsymbol{\theta})}{\psi(\sigma^2, \lambda)}$$

where

$$f_1(\hat{\boldsymbol{\mu}}) = \min_{\boldsymbol{\mu}} \left( \frac{\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\mu}\|_2^2}{2\sigma^2} + \frac{\lambda}{\sigma} \|\boldsymbol{\mu}\|_1 \right),$$

$$f_2(\hat{\mathbf{D}}) = \sum_j \ln \alpha_j + \sum_j \frac{\beta_j \alpha_j^{-2}}{2\sigma^2} + \sum_j \sqrt{\frac{2\lambda^2 n}{\pi}} (\alpha_j)^{-1}$$

*Proof.* Theorem 5 holds according to the upper bound of KL divergence and the proof is straightforward.  $\square$

Thus the KL divergence for  $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{D}})$  is upper bounded by the minimum achievable  $\ell_1$ -penalized least square error  $\epsilon_1 = f_1(\hat{\boldsymbol{\mu}})$  and terms in  $f_2(\hat{\mathbf{D}})$  which are ultimately related to the eigenvalues  $\{\beta_j\}$  ( $j = 1, \dots, p$ ) of the Gram matrix  $\boldsymbol{\Phi}'\boldsymbol{\Phi}$ .

Let  $(\boldsymbol{\mu}^*, \mathbf{D}^*)$  be the minimizer of the original KL divergence  $g(\boldsymbol{\mu}, \mathbf{D})$ , and  $g_1(\boldsymbol{\mu}|\mathbf{D})$  collect the terms of  $g(\boldsymbol{\mu}, \mathbf{D})$  that are related to  $\boldsymbol{\mu}$ . Then the Bayesian posterior mean obtained via VG, i.e.,

$$\boldsymbol{\mu}^* = \arg \min_{\boldsymbol{\mu}} g_1(\boldsymbol{\mu}|\mathbf{D}^*) = \arg \min_{\boldsymbol{\mu}} \mathbb{E}_{q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})} (\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2 + 2\lambda\sigma\|\mathbf{x}\|_1),$$

is a counterpart of the deterministic Lasso (Tibshirani, 1996), which appears naturally in the upper bound,

$$\hat{\boldsymbol{\mu}} = \arg \min_{\boldsymbol{\mu}} f_1(\boldsymbol{\mu}) = \arg \min_{\boldsymbol{\mu}} (\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\mu}\|_2^2 + 2\lambda\sigma\|\boldsymbol{\mu}\|_1)$$

Note that the Lasso solution cannot be found by gradient methods due to non-differentiability. By taking the expectation, the objective function is smoothed

around  $\mathbf{0}$  and thus differentiable. This connection indicates that in VG for Bayesian Lasso, the conditions of deterministic Lasso hold on average, with respect to the variational distribution  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ , in the parameter space of  $\boldsymbol{\mu}$ .

The following theorem (with proof sketches) provides quantitative measures of the closeness of the upper bounds,  $f_1(\boldsymbol{\mu})$  and  $f(\boldsymbol{\mu}, \mathbf{D})$ , to their respective true counterparts.

**Theorem 6.** *The tightness of  $f_1(\boldsymbol{\mu})$  and  $f(\boldsymbol{\mu}, \mathbf{D})$  is given by*

$$\begin{aligned} g_1(\boldsymbol{\mu}|\mathbf{D}) - f_1(\boldsymbol{\mu}) &\leq \frac{\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}\mathbf{D})}{2\sigma^2} + \frac{\lambda}{\sigma} \sqrt{\frac{2p}{\pi}} \text{tr}(\sqrt{\mathbf{D}}) \\ f(\boldsymbol{\mu}, \mathbf{D}) - g(\boldsymbol{\mu}, \mathbf{D}) &\leq \frac{2\lambda}{\sigma} \sqrt{\frac{2p}{\pi}} \text{tr}(\sqrt{\mathbf{D}}) \end{aligned} \quad (2.17)$$

which holds for any  $(\boldsymbol{\mu}, \mathbf{D}) \in \mathbb{R}^p \times \mathbb{S}_{++}^p$ . Further assume  $g(\boldsymbol{\mu}^*, \mathbf{D}^*) = \epsilon_2$  (minimum achievable KL divergence, or information gap), we have

$$f_1(\boldsymbol{\mu}^*) \leq g_1(\boldsymbol{\mu}^*) \leq g_1(\hat{\boldsymbol{\mu}}) \leq \epsilon_1 + \text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}\mathbf{D})/(2\sigma^2) + \lambda\sqrt{2p/(\sigma^2\pi)}\text{tr}(\sqrt{\hat{\mathbf{D}}}) \quad (2.18a)$$

$$g(\hat{\boldsymbol{\mu}}, \hat{\mathbf{D}}) \leq f(\hat{\boldsymbol{\mu}}, \hat{\mathbf{D}}) \leq f(\boldsymbol{\mu}^*, \mathbf{D}^*) \leq \epsilon_2 + 2\lambda\sqrt{2p/(\sigma^2\pi)}\text{tr}(\sqrt{\mathbf{D}^*}) \quad (2.18b)$$

*Proof.* To see the first inequality in (2.17), we have

$$\begin{aligned} g_1(\boldsymbol{\mu}) - f_1(\boldsymbol{\mu}) &= \frac{\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}\mathbf{D})}{2\sigma^2} + \frac{\lambda}{\sigma} \mathbb{E}_q(\|\mathbf{x}\|_1 - \|\boldsymbol{\mu}\|_1) \\ &\leq \frac{\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}\mathbf{D})}{2\sigma^2} + \frac{\lambda}{\sigma} \mathbb{E}_q(\|\mathbf{x} - \boldsymbol{\mu}\|_1) \\ &= \frac{\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}\mathbf{D})}{2\sigma^2} + \frac{\lambda}{\sigma} \sqrt{\frac{2}{\pi}} \sum_j \sqrt{d_j} \leq \frac{\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}\mathbf{D})}{2\sigma^2} + \frac{\lambda}{\sigma} \sqrt{\frac{2p}{\pi}} \text{tr}(\sqrt{\mathbf{D}}) \end{aligned}$$

holds for any  $\boldsymbol{\mu} \in \mathbb{R}^p$ . Note that  $f(\boldsymbol{\mu}, \mathbf{D})$  is an upper bound of  $g(\boldsymbol{\mu}, \mathbf{D})$ ,  $f_1(\boldsymbol{\mu}^*) \leq g_1(\boldsymbol{\mu}^*)$  (the proof is straightforward). Thus the first inequality in (2.18a) holds. The second inequality holds since  $\boldsymbol{\mu}^*$  is the global minimum of  $g_1(\boldsymbol{\mu})$ . To see the second

inequality in (2.17), we have that

$$\begin{aligned}
f(\boldsymbol{\mu}, \mathbf{D}) - g(\boldsymbol{\mu}, \mathbf{D}) &= \frac{\lambda}{\sigma} \sqrt{\frac{2p}{\pi}} \text{tr}(\sqrt{\mathbf{D}}) + \frac{\lambda}{\sigma} \mathbb{E}_q(\|\boldsymbol{\mu}\|_1 - \|\mathbf{x}\|_1) \\
&\leq \frac{\lambda}{\sigma} \sqrt{\frac{2p}{\pi}} \text{tr}(\sqrt{\mathbf{D}}) + \frac{\lambda}{\sigma} \mathbb{E}_q(\|\boldsymbol{\mu} - \mathbf{x}\|_1) \\
&= \frac{\lambda}{\sigma} \sqrt{\frac{2p}{\pi}} \text{tr}(\sqrt{\mathbf{D}}) + \frac{\lambda}{\sigma} \sqrt{\frac{2}{\pi}} \sum_j \sqrt{d_j} \leq 2 \frac{\lambda}{\sigma} \sqrt{\frac{2p}{\pi}} \text{tr}(\sqrt{\mathbf{D}})
\end{aligned}$$

holds for any  $(\boldsymbol{\mu}, \mathbf{D}) \in \mathbb{R}^p \times \mathbb{S}_{++}^p$ . The first inequality in (2.18b) holds since  $f(\boldsymbol{\mu}, \mathbf{D})$  is an upper bound of  $g(\boldsymbol{\mu}, \mathbf{D})$ ; the second inequality holds since  $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{D}})$  is the global minimum of  $f(\boldsymbol{\mu}, \mathbf{D})$ .  $\square$

## 2.6 Other Inference Methods for Bayesian Lasso

According to (Park and Casella, 2008), the Bayesian Lasso model (scaled case) with the scale-mixture of normal representation is as follows,

$$\begin{aligned}
\mathbf{y} | \mathbf{x}, \sigma^2 &\sim \mathcal{N}_n(\mathbf{y}; \boldsymbol{\Phi} \mathbf{x}, \sigma^2 \mathbf{I}_n) \\
\mathbf{x} | \sigma^2, \tau_1^2, \dots, \tau_p^2 &\sim \mathcal{N}_p(\mathbf{x}; \mathbf{0}_p, \sigma^2 \mathbf{D}_\tau), \quad \mathbf{D}_\tau = \text{diag}(\tau_1^2, \dots, \tau_p^2) \\
\tau_1^2, \dots, \tau_p^2 &\sim \prod_{j=1}^p \frac{\lambda^2}{2} \exp(-\lambda^2 \tau_j^2 / 2) d\tau_j^2, \quad \tau_1^2, \dots, \tau_p^2 > 0, \quad j = 1, \dots, p \\
\gamma_j &\sim \frac{\lambda^2}{2} \exp(-\lambda^2 / 2\gamma_j) \gamma_j^{-2}, \quad \gamma_j = 1/\tau_j^2, \quad j = 1, \dots, p \\
\sigma^2 &\sim \text{InvGamma}(\sigma^2; a, b) \\
\lambda^2 &\sim \text{Gamma}(\lambda^2; r, s)
\end{aligned}$$

where representation of Laplace distribution as a scale mixture of normals (with an exponential mixing density) is exploited,

$$\frac{t}{2} \exp(-t|z|) = \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp(-z^2/(2s)) \frac{t^2}{2} \exp(-t^2 s/2) ds$$

where  $t > 0$ ,  $t = \lambda/\sigma$ ,  $s = \sigma^2 \tau_j^2 = \sigma^2/\gamma_j$ .

### 2.6.1 Data-Augmentation Gibbs Sampler

The full likelihood can be written as follows,

$$\begin{aligned}
& p(\mathbf{y}|\mathbf{x}, \sigma^2) \times p(\mathbf{x}|\sigma^2, \boldsymbol{\gamma}) \times p(\boldsymbol{\gamma}) \times p(\sigma^2) \times p(\lambda^2) \\
&= \mathcal{N}_n(\mathbf{y}; \boldsymbol{\Phi}\mathbf{x}, \sigma^2\mathbf{I}_n) \times \mathcal{N}_p(\mathbf{x}; \mathbf{0}_p, \sigma^2\mathbf{D}_\tau) \times \left( \prod_{i=1}^p p(\gamma_j) \right) \times p(\sigma^2) \times p(\lambda) \\
&= \frac{1}{(2\pi)^{n/2} |\sigma^2\mathbf{I}_n|^{1/2}} \exp\left(-\frac{(\mathbf{y} - \boldsymbol{\Phi}\mathbf{x})^T(\mathbf{y} - \boldsymbol{\Phi}\mathbf{x})}{2\sigma^2}\right) \\
&\times \frac{1}{(2\pi)^{p/2} [\prod_{j=1}^p \sigma^2/\gamma_j]^{1/2}} \exp\left(-\frac{\mathbf{x}^T\mathbf{D}_\tau^{-1}\mathbf{x}}{2\sigma^2}\right) \\
&\times \prod_{j=1}^p \left( \frac{\lambda^2}{2} \exp(-\lambda^2/(2\gamma_j)) (\gamma_j^{-2}) \right) \\
&\times \frac{b^a}{\Gamma(a)} (\sigma^2)^{-(a+1)} \exp(-b/\sigma^2) \times \frac{s^r}{\Gamma(r)} \lambda^{2(r-1)} \exp(-s\lambda^2)
\end{aligned}$$

where  $\mathbf{D}_\tau = \text{diag}(1/\gamma_1, \dots, 1/\gamma_p)$ .

- Full Conditional distribution of  $\mathbf{x}$ :

$$(\mathbf{x}|\mathbf{y}, \sigma^2, \tau_1^2, \dots, \tau_p^2) \sim \mathcal{N}_p(\mathbf{x}; (\mathbf{D}_\tau^{-1} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{y}, \sigma^2(\mathbf{D}_\tau^{-1} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1})$$

- Full Conditional distribution of  $\sigma^2$ :

$$(\sigma^2|\mathbf{y}, \tau_1^2, \dots, \tau_p^2) \sim \text{InvGamma}(\sigma^2; \tilde{a}, \tilde{b})$$

where

$$\tilde{a} = \frac{n+p-1}{2} + a,$$

$$\tilde{b} = \frac{(\mathbf{y} - \boldsymbol{\Phi}\mathbf{x})^T(\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}) + \mathbf{x}^T\mathbf{D}_\tau^{-1}\mathbf{x}}{2} + b$$

- Full Conditional distribution of  $\gamma_j = 1/\tau_j^2$ :

$$p(1/\tau_j^2|\lambda^2, \sigma^2, x_j) \propto (1/\tau_j^2)^{-\frac{3}{2}} \exp\left\{-\left(\frac{(x_j/\tau_j^2 - \lambda\sigma)^2}{2\sigma^2(1/\tau_j^2)}\right)\right\}$$

$$\sim \text{InvGaussian}(1/\tau_j^2; g, h)$$

where  $g = \sqrt{\lambda^2 \sigma^2 / x_j^2}$  and  $h = \lambda^2$ .

- Full Conditional distribution of  $\lambda^2$ :

$$(\lambda^2 | \tau_j^2) \sim \text{Gamma}(\lambda^2; p + r, s + \sum_{j=1}^p \frac{\tau_j^2}{2})$$

### 2.6.2 Mean-Field Variational Bayes

We seek a variational distribution  $q(\Theta; \Gamma)$  to approximate the exact posterior  $p(\Theta; \Gamma)$ , where  $\Theta \equiv \{\mathbf{x}, \gamma, \sigma^2, \lambda^2\}$ ,  $\Gamma$  are the variational parameters. Consider the variational expression,

$$\tilde{F}(\Gamma) = \int d\Theta q(\Theta; \Gamma) \ln \frac{q(\Theta; \Gamma)}{p(\mathbf{y})p(\Theta|\mathbf{y})} = -\ln p(\mathbf{y}) + \text{KL}[q(\Theta; \Gamma) || p(\Theta|\mathbf{y})]$$

Note that the term  $p(\mathbf{y})$  is a constant with respect to  $\Gamma$ , and therefore the evidence lower bound  $\tilde{F}(\Gamma)$  is maximized when  $\text{KL}[q(\Theta; \Gamma) || p(\Theta|\mathbf{y})]$  is minimized. To make the computation of  $\tilde{F}(\Gamma)$  tractable, we assume  $q(\Theta; \Gamma)$  has a factorized form,

$$q(\Theta; \Gamma) = \prod_{i=1}^k q_i(\Theta_i; \Gamma_i)$$

With appropriate choice of  $q_i$ , the variational expression  $\tilde{F}(\Gamma)$  may be evaluated analytically. Maximizing the lower bound  $\tilde{F}(\Gamma)$  with respect to  $q_i^*(\Theta_i; \Gamma_i)$  yields

$$q_i^*(\Theta_i; \Gamma_i) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{y}, \Theta)])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{y}, \Theta)]) d\Theta_i}$$

The update equations are as follows,

- Update for  $\mathbf{x}$ :

$$q^*(\mathbf{x} | -) \sim \mathcal{N}(\mathbf{x}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$$

$$\hat{\boldsymbol{\mu}} = (\langle \mathbf{D}_\tau^{-1} \rangle + \boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{y},$$

$$\hat{\boldsymbol{\Sigma}} = [\langle \sigma^{-2} \rangle (\langle \mathbf{D}_\tau^{-1} \rangle + \boldsymbol{\Phi}^T \boldsymbol{\Phi})]^{-1}$$

where  $\langle \sigma^{-2} \rangle = \hat{a}/\hat{b}$

- Update for  $\sigma^{-2}$ :

$$q^*(\sigma^{-2}|-) \sim \text{Gamma}(\sigma^{-2}; \hat{a}, \hat{b})$$

$$\hat{a} = \frac{n+p-1}{2} + a$$

$$\hat{b} = \frac{1}{2} \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \Phi \langle \mathbf{x} \rangle + \frac{1}{2} \text{trace} [(\Phi^T \Phi + \langle \mathbf{D}_\tau^{-1} \rangle) \langle \mathbf{x} \mathbf{x}^T \rangle] + b$$

where  $\langle \mathbf{x} \rangle = \hat{\boldsymbol{\mu}}$ ,  $\langle \mathbf{x} \mathbf{x}^T \rangle = \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^T + \hat{\boldsymbol{\Sigma}}$ .

- Update for  $\lambda^2$ :

$$q^*(\lambda^2|-) \sim \text{Gamma}(\lambda^2; \hat{r}, \hat{s}), \quad \hat{r} = p + r, \quad \hat{s} = \sum_{j=1}^p \left\langle \frac{1}{2\gamma_j} \right\rangle + s$$

- Update for  $\gamma_j$ ,  $j = 1, \dots, p$ :

$$q^*(\gamma_j|-) \sim \text{InvGaussian}(\gamma_j; \hat{g}_j, \hat{h}_j), \quad \hat{g}_j = \sqrt{\frac{\langle \lambda^2 \rangle}{\langle \sigma^{-2} \rangle \langle x_j^2 \rangle}}, \quad \hat{h}_j = \langle \lambda^2 \rangle$$

where  $\text{InvGaussian}(x; g, h) = \sqrt{h/(2\pi x^3)} \exp(-h(x-g)^2/(2g^2x))$

( $x > 0$ ) denotes the inverse Gaussian distribution with mean  $\langle x \rangle = g$  and  $\langle x^{-1} \rangle = g^{-1} + h^{-1}$ . We have

$$\langle \lambda^2 \rangle = \hat{r}/\hat{s}, \quad \langle x_j^2 \rangle = \hat{\mu}_j^2 + \hat{\Sigma}_{jj},$$

$$\langle \gamma_j^{-1} \rangle = \hat{g}_j^{-1} + \hat{h}_j^{-1}, \quad \langle \mathbf{D}_\tau^{-1} \rangle = \text{diag} [\hat{g}_j]_{j=1:p}$$

The lower bound  $\tilde{F}(\boldsymbol{\Gamma})$  can be calculated very straightforwardly both for tracking the monotonic increase and for possibly setting a convergence criterion.

$$\begin{aligned} \tilde{F}(\boldsymbol{\Gamma}) = & \langle \ln p(\mathbf{y}|\mathbf{x}, \sigma^2) \rangle + \langle \ln p(\mathbf{x}|\sigma^2, \boldsymbol{\gamma}) \rangle + \langle \ln p(\boldsymbol{\gamma}) \rangle + \langle \ln p(\sigma^2) \rangle + \langle \ln p(\lambda^2) \rangle \\ & - \langle \ln q^*(\mathbf{x}|-) \rangle - \langle \ln q^*(\boldsymbol{\gamma}|-) \rangle - \langle \ln q^*(\sigma^{-2}|-) \rangle - \langle \ln q^*(\lambda^2|-) \rangle \end{aligned}$$

where

$$\langle \ln p(\mathbf{y}|\mathbf{x}, \sigma^2) \rangle = -\frac{n}{2} \ln 2\pi + \frac{n}{2} \langle \ln \sigma^{-2} \rangle - \frac{1}{2} \langle \sigma^{-2} \rangle \left( \|\mathbf{y} - \Phi \hat{\boldsymbol{\mu}}\|_2^2 + \text{trace}(\Phi^T \Phi \hat{\boldsymbol{\Sigma}}) \right)$$

$$\langle \ln \sigma^{-2} \rangle = \psi(\hat{a}) - \ln(\hat{b}), \quad \langle \sigma^{-2} \rangle = \hat{a}/\hat{b}$$

and  $\psi(\cdot)$  is the digamma function.

$$\langle \ln p(\mathbf{x}|\sigma^2, \boldsymbol{\gamma}) \rangle = -\frac{p}{2} \ln 2\pi + \frac{p}{2} \langle \ln \sigma^{-2} \rangle + \frac{1}{2} \sum_{j=1}^p \langle \ln \gamma_j \rangle - \frac{1}{2} \langle \sigma^{-2} \rangle \text{trace}(\langle \mathbf{D}_\tau^{-1} \rangle \langle \mathbf{x}\mathbf{x}^T \rangle)$$

and those  $\langle \ln \gamma_j \rangle$  terms canceled out.

$$\langle \ln p(\boldsymbol{\gamma}) \rangle = \sum_{j=1}^p \langle \log p(\gamma_j) \rangle = \sum_{j=1}^p \left( \langle \ln \frac{\lambda^2}{2} \rangle - \langle \gamma_j^{-1} \rangle \langle \frac{\lambda^2}{2} \rangle - 2 \langle \ln \gamma_j \rangle \right)$$

$$\langle \ln \lambda^2 \rangle = \psi(\hat{r}) - \ln(\hat{s})$$

$$\langle \ln p(\sigma^{-2}) \rangle = \ln \left( \frac{b^a}{\Gamma(a)} \right) + (a-1) \langle \ln \sigma^{-2} \rangle - b \langle \sigma^{-2} \rangle$$

$$\langle \ln p(\lambda^2) \rangle = \ln \left( \frac{s^r}{\Gamma(r)} \right) + (r-1) \langle \ln \lambda^2 \rangle - s \langle \lambda^2 \rangle$$

$$-\langle \ln q^*(\mathbf{x}|-) \rangle = \frac{1}{2} \ln |2\pi e \hat{\boldsymbol{\Sigma}}|$$

$$-\langle \ln q^*(\boldsymbol{\gamma}|-) \rangle = -\sum_{j=1}^p \langle \ln q^*(\gamma_j|-) \rangle$$

$$= \sum_{j=1}^p \left( -\frac{1}{2} \ln \hat{h}_j + \frac{1}{2} \ln 2\pi + \frac{3}{2} \langle \ln \gamma_j \rangle + 0.5 \right)$$

$$-\langle \ln q^*(\sigma^{-2}|-) \rangle = -\hat{a} \ln \hat{b} + \ln \Gamma(\hat{a}) - (\hat{a}-1) \langle \ln \sigma^{-2} \rangle + \hat{b} \langle \sigma^{-2} \rangle$$

$$-\langle \ln q^*(\lambda^2|-) \rangle = -\hat{r} \ln \hat{s} + \ln \Gamma(\hat{r}) - (\hat{r}-1) \langle \ln \lambda^2 \rangle + \hat{s} \langle \lambda^2 \rangle$$

## 2.7 Experiments

We consider long runs of MCMC as reference solutions, and consider two types of INF-VB: INF-VB-1 calculates hyperparameter posteriors using (2.11); while INF-VB-2 uses (2.12) and evaluates it at the posterior mode of  $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ . We also compare INF-VB-1 and INF-VB-2 to VB, a mean-field variational Bayes (VB) solution (See Section 2.6 for update equations). The results show that the INF-VB method is more accurate than VB, and is a promising alternative to MCMC for Bayesian Lasso. In

all experiments shown here, we take intensive MCMC runs as the gold standard (with  $5 \times 10^3$  burn-ins and  $5 \times 10^5$  samples collected). We use data-augmentation Gibbs sampler introduced in (Park and Casella, 2008). Ground truth for latent variables and hyper-parameter are also compared to whenever possible. The hyperparameters for Gamma distributions are set to  $a = b = r = s = 0.001$  through all these experiments. If not mentioned, the grid size is  $50 \times 50$ , which is uniformly created around the ordinary least square (OLS) estimates of hyper-parameters.

### 2.7.1 Synthetic Dataset

We compare the proposed INF-VB methods with VB and intensive MCMC runs, in terms of the joint posterior  $q(\lambda^2, \sigma^2 | \mathbf{y})$ , the marginal posteriors of hyper-parameters  $q(\sigma^2 | \mathbf{y})$  and  $q(\lambda^2 | \mathbf{y})$ , and the marginal posteriors of regression coefficients  $q(x_j | \mathbf{y})$  (see Figure 2.1 and Figure 2.2 ). The observations are generated from

$$y_i = \boldsymbol{\phi}_i^T \mathbf{x} + \epsilon_i, \quad i = 1, \dots, 600,$$

where  $\phi_{ij}$  are drawn from an *i.i.d.* normal distribution, where the pairwise correlation between the  $j$ th and the  $k$ th columns of  $\boldsymbol{\Phi}$  is  $0.5^{|j-k|}$ . The responses  $\mathbf{y}$  and the columns of  $\boldsymbol{\Phi}$  are centered; the columns of  $\boldsymbol{\Phi}$  are also scaled to have unit variance. We assume  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ ,  $x_j | \lambda, \sigma \sim \text{Laplace}(\lambda/\sigma)$ ,  $j = 1, \dots, 300$ , and set  $\sigma^2 = 0.5$ ,  $\lambda = 0.5$ .

See Figure 2.1(a)-(d), both MCMC and INF-VB preserve the strong posterior dependence among hyperparameters, while mean-field VB cannot. While mean-field VB approximates the posterior mode well, the posterior variance can be (sometimes severely) underestimated, see Figure 2.2(a), (b). Since we have analytically approximated  $p(\mathbf{x} | \mathbf{y})$  by a finite mixture of normal distribution  $q(\mathbf{x} | \mathbf{y}, \boldsymbol{\theta})$  with mixing weights  $q(\boldsymbol{\theta} | \mathbf{y})$ , the posterior marginals for the latent variables:  $q(x_j | \mathbf{y})$  are easily accessible from this analytical representation. Perhaps surprisingly, both INF-VB and mean-field VB provide quite accurate marginal distributions  $q(x_j | \mathbf{y})$ , see Figure 2.2(c)-(d)

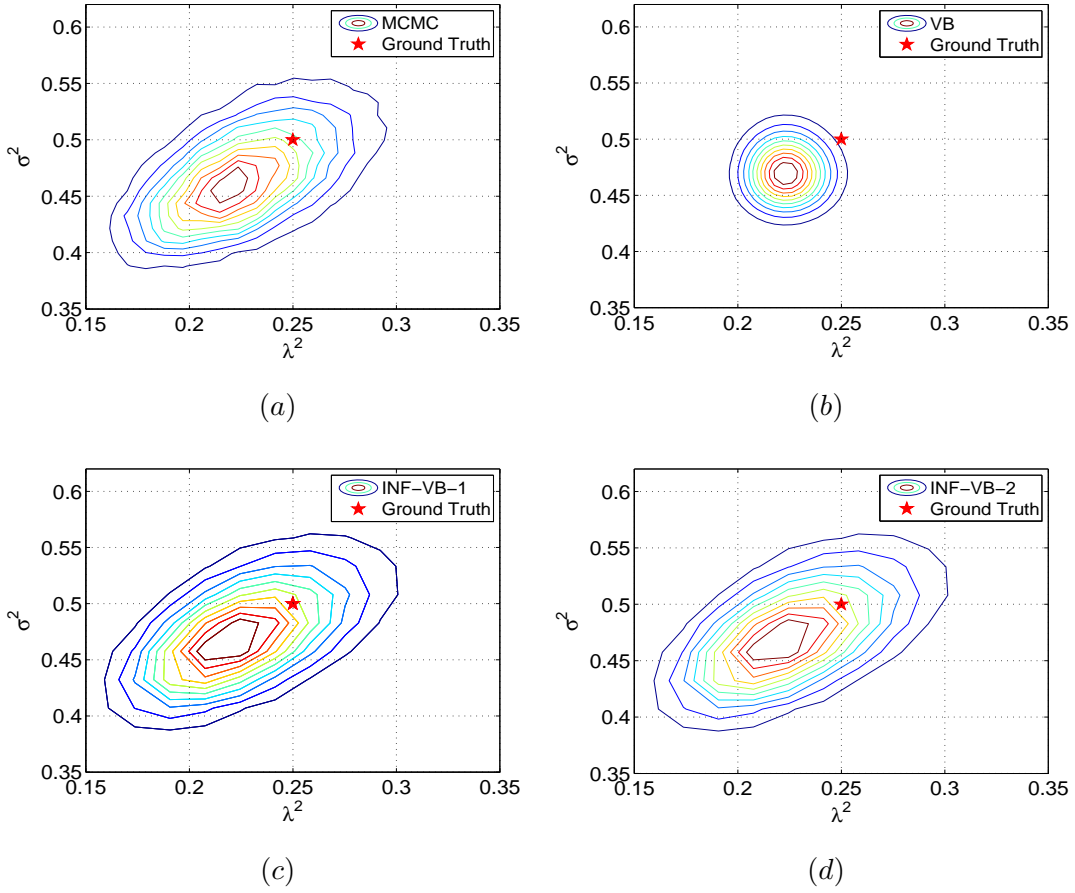


FIGURE 2.1: Contour plots for joint posteriors of hyperparameters  $q(\sigma^2, \lambda^2|\mathbf{y})$

for examples. The differences in the tails of  $q(\boldsymbol{\theta}|\mathbf{y})$  between INF-VB and mean-field VB yield negligible differences in the marginal distributions  $q(x_j|\mathbf{y})$ , when  $\boldsymbol{\theta}$  is integrated out.

### 2.7.2 Diabetes Dataset

We consider the benchmark diabetes dataset (Efron *et al.*, 2004) frequently used in previous studies of Bayesian Lasso; see (Park and Casella, 2008; Hans, 2009), for example. The goal of this diagnostic study, as suggested in (Efron *et al.*, 2004), is

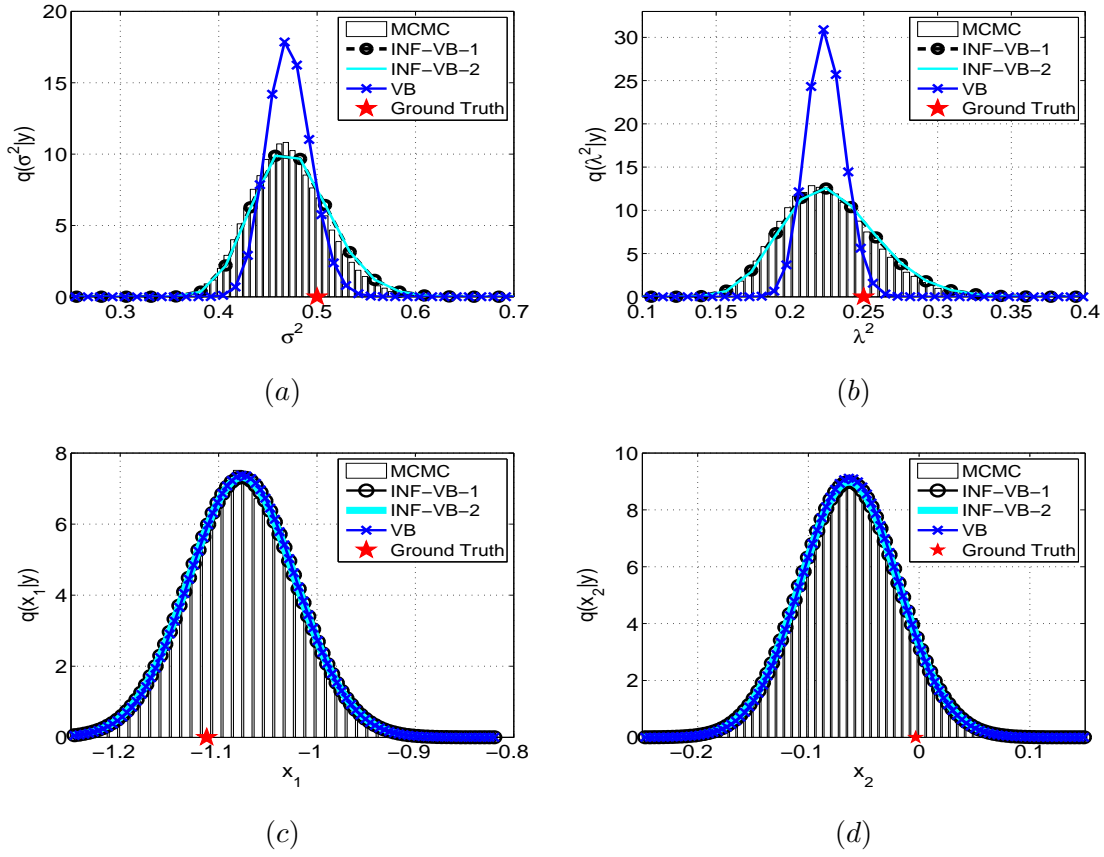


FIGURE 2.2: Marginal posterior of hyperparameters and coefficients: (a)  $q(\sigma^2|\mathbf{y})$ , (b)  $q(\lambda^2|\mathbf{y})$ ; (c)  $q(x_1|\mathbf{y})$ , (d)  $q(x_2|\mathbf{y})$

to construct a linear regression model ( $n = 442$ ,  $p = 10$ ) to reveal the important determinants of the response, and to provide interpretable results to guide disease progression. In Figure 2.3 and Figure 2.4, we show accurate marginal posteriors of hyperparameters  $q(\sigma^2|\mathbf{y})$  and  $q(\lambda^2|\mathbf{y})$  as well as marginals of coefficients  $q(x_j|\mathbf{y})$ ,  $j = 1, \dots, 10$ , which indicate the relevance of each predictor. We also compared them to the ordinary least square (OLS) estimates.

### 2.7.3 Comparison: Accuracy and Speed

We quantitatively measure the quality of the approximate joint probability  $q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$  provided by our non-factorized variational methods, and compare them to mean-field VB under factorization assumptions. The KL divergence  $\text{KL}(q(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})|p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}))$  is not directly available; instead, we compare the negative evidence lower bound (2.8), which can be evaluated analytically in our case and differs from the KL divergence only up to a constant. We also measure the computational time of different algorithms by elapsed times (seconds). In INF-VB, different grids of sizes  $m \times m$  are considered, where  $m = 1, 5, 10, 30, 50$ . We consider two real world datasets: the above Diabetes dataset, and the Prostate cancer dataset (Stamey *et al.*, 1989). Here, INF-VB-3 and INF-VB-4 refer to the methods that use the approximate solution in Section 2.5.2 with no gradient steps for  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ , and use (2.11) or (2.12) for  $q(\boldsymbol{\theta}|\mathbf{y})$ .

The quality of variational methods depends on the flexibility of variational distributions. In INF-VB for Bayesian Lasso, we constrain  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$  to be parametric and  $q(\boldsymbol{\theta}|\mathbf{y})$  to be still in free form. See from Figure 2.5, the accuracy of INF-VB method with a  $1 \times 1$  grid is worse than mean-field VB, which corresponds to the partial Bayesian learning of  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$  with a fixed  $\boldsymbol{\theta}$ . As the grid size increases, the accuracies of INF-VB (even those without gradient steps) also increase and are in general of better quality than mean-field VB, in the sense of negative ELBO (KL divergence up to a constant).

The computational complexities of INF-VB, mean-field VB, and MCMC methods are proportional to the grid size, number of iterations toward local optimum, and the number of runs, respectively. Since the computations on the grid are independent, INF-VB is highly parallelizable, which is an important feature as more multiprocessor computational power becomes available. Besides, one may further reduce its computational load by choosing grid points more economically, which will

be pursued in our next step. Even the small datasets we show here for illustration enjoy good speed-ups. A significant speed-up for INF-VB can be achieved via parallel computing.

## 2.8 Discussion

We have provided a flexible framework for approximate inference of the full posterior  $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$  based on a hybrid continuous-discrete variational distribution, which is optimal in the sense of the KL divergence. As a reliable and efficient alternative to MCMC, our method generalizes INLA to non-Gaussian priors and MFVB to non-factorization settings. While we have used Bayesian Lasso as an example, our inference method is generically applicable. One can also approximate  $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$  using other methods, such as scalable variational methods (Seeger and Nickisch, 2011), or improved EP (Cseke and Heskes, 2011).

The posterior  $p(\boldsymbol{\theta}|\mathbf{y})$ , which is analyzed based on a grid approximation, enables users to do both model averaging and model selection, depending on specific purposes. The discretized approximation of  $p(\boldsymbol{\theta}|\mathbf{y})$  overcomes the potential non-conjugacy or multimodal issues in the  $\boldsymbol{\theta}$  space in variational inference, and it also allows parallel implementation of the hybrid continuous-discrete variational approximation with the dominant computational load (approximating the continuous high dimensional  $q(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ ) distributed on each grid point, which is particularly important when applying INF-VB to large-scale Bayesian inference. INF-VB has limitations. The number of hyperparameters  $\boldsymbol{\theta}$  should be no more than 5 to 6, which is the same fundamental limitation of INLA.

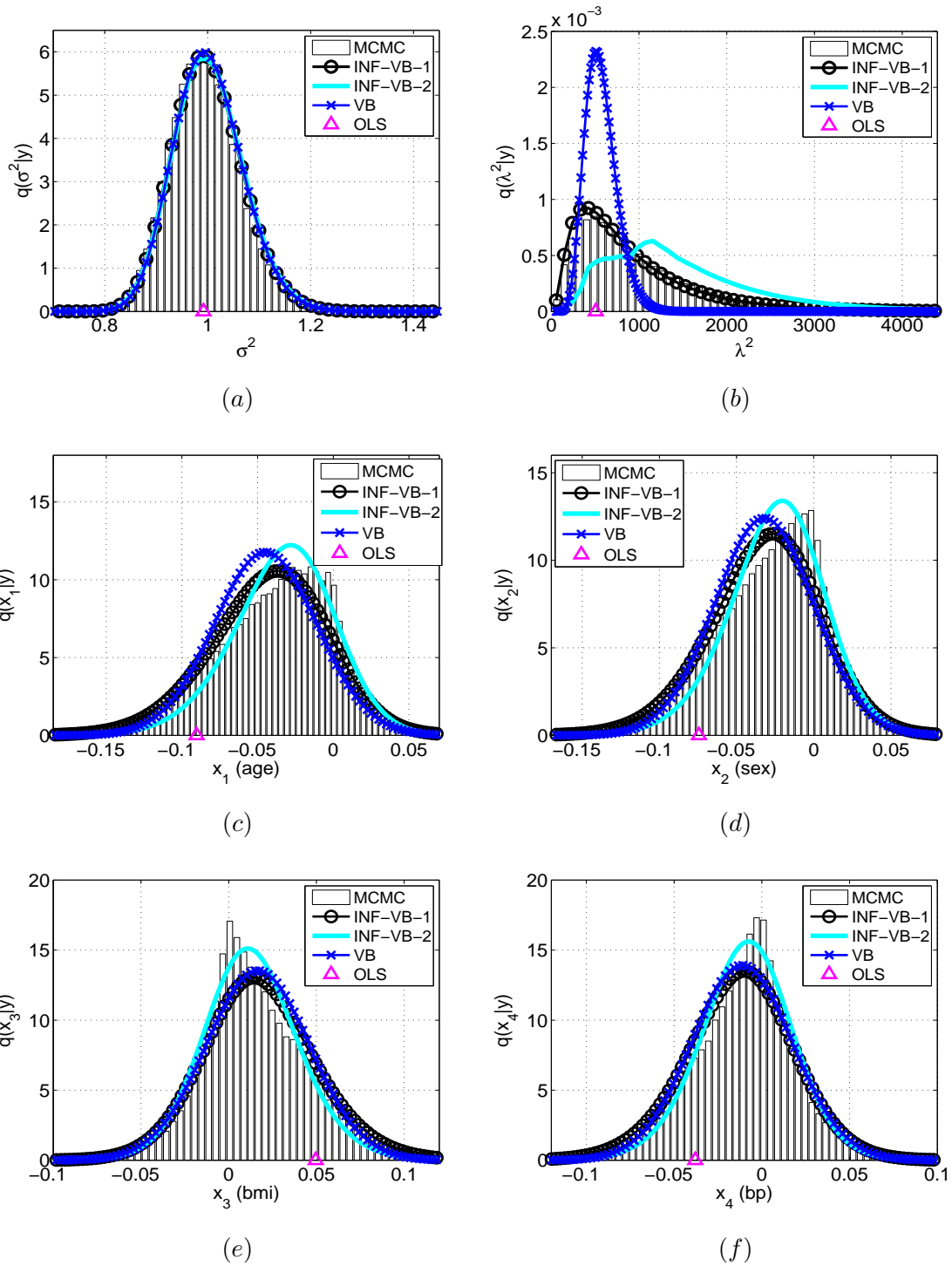


FIGURE 2.3: Posterior marginals of hyperparameters: (a)  $q(\sigma^2|\mathbf{y})$  and (b)  $q(\lambda^2|\mathbf{y})$ ; posterior marginals of coefficients: (c)-(f)  $q(x_j|\mathbf{y})$  ( $j = 1, \dots, 4$ )

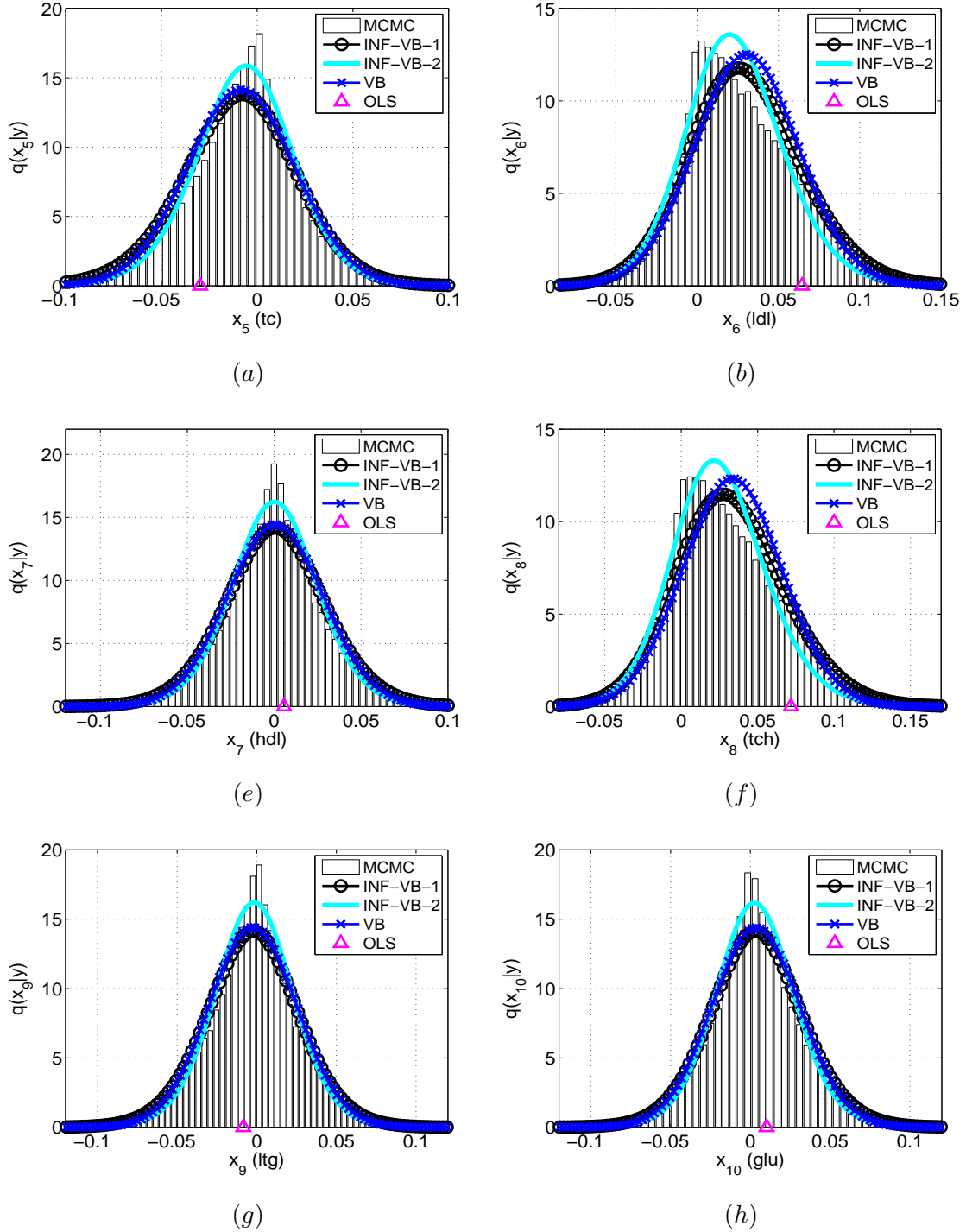


FIGURE 2.4: Posterior marginals of coefficients: (a)-(f)  $q(x_j|\mathbf{y})$  ( $j = 5, \dots, 10$ )

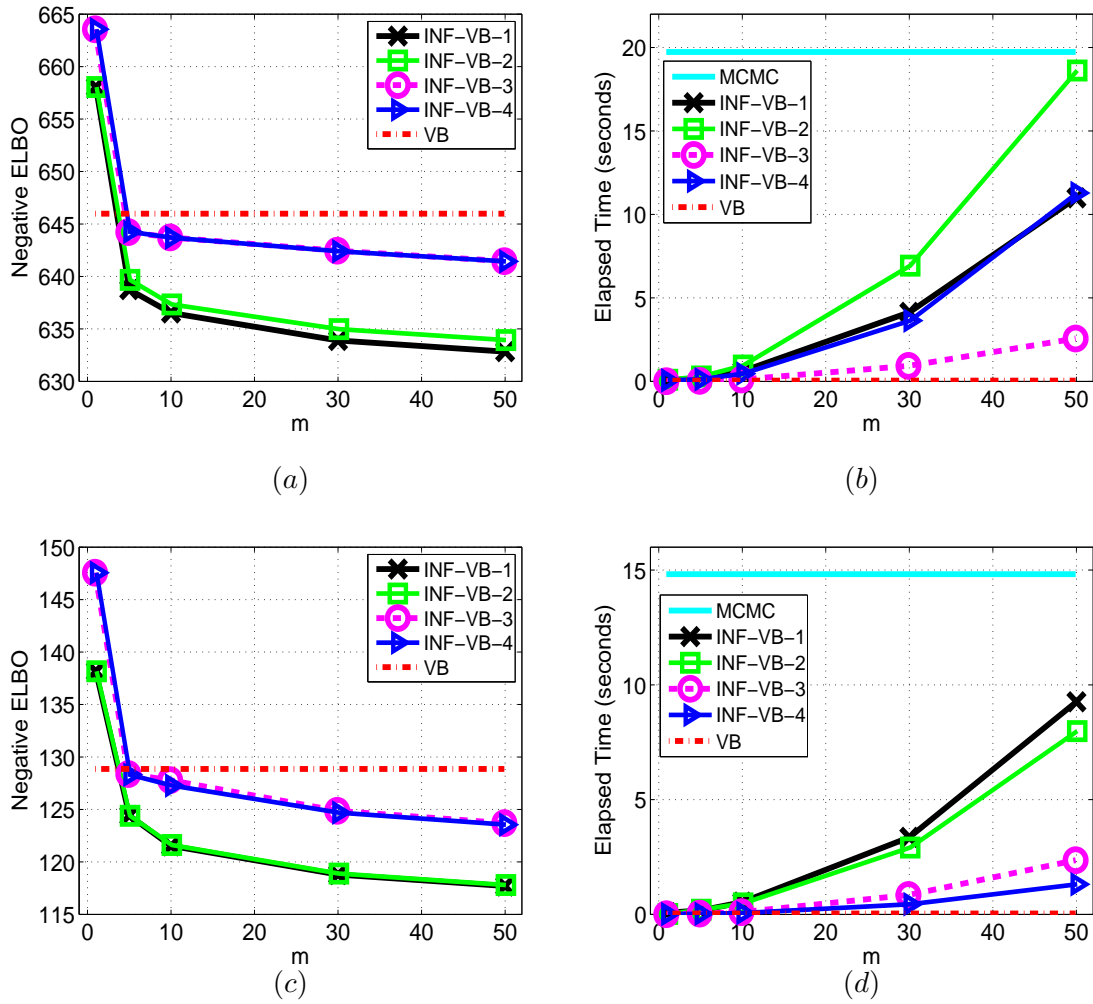


FIGURE 2.5: Negative evidence lower bound (ELBO) and elapsed time v.s. grid size; (a), (b) for the Diabetes dataset ( $n = 442, p = 10$ ). (c), (d) for the Prostate cancer dataset ( $n = 97, p = 8$ )

## Copula-based Dependent Variational Inference

In this chapter, we utilize copulas to constitute a unified variational copula (VC) inference framework for constructing and optimizing variational proposals in hierarchical Bayesian models. The optimal variational posterior under Sklar’s representation is found by minimizing the KL divergence to the true posterior. For models with continuous and non-Gaussian hidden variables, we propose a semiparametric and automated variational Gaussian copula approach, in which the parametric Gaussian copula family is able to preserve multivariate posterior dependence, and the nonparametric transformations based on Bernstein polynomials provide ample flexibility in characterizing the univariate marginal posteriors. Compared with the integrated nested Laplace approximation (INLA) (Rue *et al.*, 2009) and integrated non-factorized variational inference (Han *et al.*, 2013), our approach does not need to discretize the space for non-Gaussian variables and thus does not suffer from the limits on the number of hyperparameters.

### 3.1 Preliminaries

We first provide a brief introduction to some preliminary concepts and theory on copula (Nelsen, 2007) needed for subsequent developments.

- **(Sklar's Theorem)** (Sklar, 1959) Let  $H$  be a joint distribution function with margins  $F$  and  $G$ . Then there exists a copula  $C$  such that for all  $x, y$  in  $\overline{\mathbf{R}}$ ,

$$H(x, y) = C(F(x), G(y)) \quad (3.1)$$

If  $F$  and  $G$  are continuous, then  $C$  is unique; otherwise,  $C$  is uniquely determined on  $\text{Ran}F \times \text{Ran}G$  (the uniqueness property only holds on  $\text{Ran}F \times \text{Ran}G$ ). Conversely, if  $C$  is a copula and  $F$  and  $G$  are distribution functions, then the function  $H$  defined by (3.1) is a joint distribution function with margins  $F$  and  $G$

- Let  $H$  be a joint distribution function with margins  $F$  and  $G$ . Then there exists a unique subcopula  $C'$  such that
  1.  $\text{Dom}C' = \text{Ran}F \times \text{Ran}G$
  2. For all  $x, y$  in  $\overline{\mathbf{R}}$ ,  $H(x, y) = C'(F(x), G(y))$
- **(The inversion method)** With  $C(u, v) = H(F^{(-1)}(u), G^{(-1)}(v))$ , new distribution  $H'(x, y) = C(F'(x), G'(y))$
- **(Quasi-inverses of Distribution Functions)** A quasi-inverse of distribution function  $F$  is any function  $F^{(-1)}$  with domain  $\mathbf{I}$  such that
  1. If  $t$  is in  $\text{Ran}F$ , then  $F^{(-1)}(t)$  is any number  $x$  in  $\overline{\mathbf{R}}$  such that  $F(x) = t$ , i.e. for all  $t$  in  $\text{Ran}F$ ,  $F(F^{(-1)}(t)) = t$
  2. If  $t$  is not in  $\text{Ran}F$ , then  $F^{(-1)}(t) = \inf\{x|F(x) \geq t\} = \sup\{x|F(x) \leq t\}$

If  $F$  is strictly increasing, it has but a single quasi-inverse, which is the ordinary inverse  $F^{-1}$

- If  $\alpha$  and  $\beta$  are strictly increasing on  $\text{Ran}X$  and  $\text{Ran}Y$ , let  $F_1, G_1, F_2, G_2$  denote the distribution functions of  $X, Y, \alpha(X), \beta(Y)$ , respectively, then

$$\begin{aligned}
F_2(X) &= P[\alpha(X) \leq x] = P[X \leq \alpha^{-1}(x)] = F_1(\alpha^{-1}(x)) \\
G_2(Y) &= G_1(\beta^{-1}(y)) \\
C_{\alpha(X)\beta(Y)}(F_2(x), G_2(y)) &= P[\alpha(X) \leq x, \beta(Y) \leq y] \\
&= P[X \leq \alpha^{-1}(x), Y \leq \beta^{-1}(y)] \\
&= C_{XY}(F_1(\alpha^{-1}(x)), G_1(\beta^{-1}(y))) \\
&= C_{XY}(F_2(x), G_2(y))
\end{aligned} \tag{3.2}$$

Thus  $C_{XY}$  is **invariant under strictly increasing transformations** of  $X$  and  $Y$

- **(Independence Copula)**  $X$  and  $Y$  are independent *if and only if*  $C_{XY} = \Pi$ , i.e.  $C(u, v) = uv$ . It is **absolutely continuous**, because for all  $(u, v)$  in  $\mathbf{I}^2$ ,

$$A_{\Pi}(u, v) = \int_0^u \int_0^v \frac{\partial^2}{\partial s \partial t} \Pi(s, t) dt ds = \int_0^u \int_0^v 1 dt ds = uv = \Pi(u, v) \tag{3.3}$$

- **(Impossibility theorem)** Let  $m$  and  $n$  be positive integers such that  $m + n \leq 3$ , and suppose that  $C$  is a 2-copula such that  $H(\mathbf{x}, \mathbf{y}) = C(F(\mathbf{x}), G(\mathbf{y}))$  is an  $(m + n)$  dimensional distribution function with margins  $H(\mathbf{x}, \infty) = F(\mathbf{x})$  and  $H(\infty, \mathbf{y}) = G(\mathbf{y})$  for all  $m$ -dimensional distribution functions  $F(\mathbf{x})$  and  $n$ -dimensional distribution functions  $G(\mathbf{y})$ . Then  $C = \Pi$

- **Gaussian copula** with  $p \times p$  correlation matrix  $\Upsilon$

$$C(u_1, \dots, u_p | \Upsilon) = \Phi_p(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p) | \Upsilon) : [0, 1]^p \rightarrow [0, 1] \tag{3.4}$$

where  $\Phi(\cdot)$  represents the CDF of the standard normal distribution and  $\Phi_p(\cdot|\boldsymbol{\Upsilon})$  is the CDF of  $N_p(\mathbf{0}, \boldsymbol{\Upsilon})$ . In particular, if the correlation matrix is an identity matrix, then Gaussian copula is the independence copula. Furthermore, if the marginals of  $X$  are also normal distributed, then Gaussian copula is corresponds to the multivariate normal distribution.

- **Gaussian copula density**

$$c(u_1, \dots, u_p|\boldsymbol{\Upsilon}) \propto (\det \boldsymbol{\Upsilon})^{-1/2} \exp \left\{ -\frac{1}{2} \langle \boldsymbol{\Upsilon}^{-1} - \mathbf{I}_p, \mathbf{z}^T \mathbf{z} \rangle \right\} \quad (3.5)$$

where  $\mathbf{z} = (\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p))^T$ . The derivation of the copula density is straightforward by the following differentiation

$$c(F_1(X_1), \dots, F_p(X_p)) = \frac{\partial^p C(F_1(X_1), \dots, F_p(X_p))}{\partial F_1(X_1), \dots, \partial F_p(X_p)} \quad (3.6)$$

Assume  $\mathbf{z} = (z_1, \dots, z_p) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Upsilon})$ , with normal scores  $u_j = F(x_j) = \Phi(z_j)$ ,

$$\begin{aligned} c(u_1, \dots, u_p) &= \frac{\partial^p C(u_1, \dots, u_p)}{\partial u_1, \dots, \partial u_p} = \frac{\phi_p(z_1, \dots, z_p)}{\prod_{j=1}^p \phi(z_j)} \Bigg|_{z_j = \Phi^{-1}(u_j)} \\ &= |\boldsymbol{\Upsilon}|^{-1/2} \exp \left( -\frac{1}{2} \mathbf{z}^T (\boldsymbol{\Upsilon}^{-1} - \mathbf{I}_p) \mathbf{z} \right) \end{aligned} \quad (3.7)$$

and  $\mathbf{z} = (\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p))^T$ .

## 3.2 Variational Copula Inference Framework

Sklar's theorem (Sklar, 1959) ensures that any multivariate joint distribution  $Q$  can be written in terms of univariate marginal distributions  $F_j(x) = P(X_j \leq x)$ ,  $j = 1, \dots, p$  and a copula which describes the dependence structures between variables, such that

$$Q(x_1, \dots, x_p) = C[F_1(x_1), \dots, F_p(x_p)]. \quad (3.8)$$

Conversely, if  $C$  is a copula and  $\{F_j\}_{j=1:p}$  are distribution functions, then the function  $Q$  defined by (3.8) is a  $p$ -dimensional joint distribution function with marginal distributions  $F_1, F_2, \dots, F_p$ , owing to the marginally closed property (Song, 2000). Assuming  $Q(x_1, \dots, x_p)$  has  $p$ -order partial derivatives, the joint probability density function (PDF) is

$$q(x_1, \dots, x_p) = c_{\Theta}[F_1(x_1), \dots, F_p(x_p)] \prod_{j=1}^p f_j(x_j),$$

where  $f_j(x_j)$  is the PDF of the  $j$ th variable and it is related to the corresponding cumulative distribution function (CDF) by

$$F_j(x_j) = \int_{-\infty}^x f_j(t) dt,$$

where  $c_{\Theta}$  is the copula density with parameter  $\Theta$ .

Sklar's theorem allows separation of the marginal distributions  $F_j(x_j)$  from the dependence structure, which is appropriately expressed in the copula function  $C$ . As a modeling tool, the specified copula function and margins can be directly fitted to the observed data  $\mathbf{y}$  (Liu *et al.*, 2009a; Wauthier and Jordan, 2010; Lopez-Paz *et al.*, 2013) with their parameters optimized via Bayesian or maximum likelihood estimators (see Smith (2013) and the references therein). In contrast, our goal is to use a copula as an *inference engine* for full posterior approximation. All the unknowns (variables/parameters) in the user-specified hierarchical model are encapsulated into a vector  $\mathbf{x}$ , and the optimal variational approximation  $q_{\text{VC}}(\mathbf{x})$  to the true posterior  $p(\mathbf{x}|\mathbf{y})$  is found under the Sklar's representation. This approach provides users with full modeling freedom and does not require conditional conjugacy between latent variables; thus the approach is applicable to general models. Within some tractable copula family  $C \in \mathcal{C}$ , and assuming  $F(\cdot)$  and  $C(\cdot)$  to be differentiable, we construct

the variational proposal as

$$q_{\mathbf{C}}(\mathbf{x}) = c(\mathbf{u}) \prod_{j=1}^p f_j(x_j),$$

where  $\mathbf{u} = F(\mathbf{x}) = [F_1(x_1), \dots, F_p(x_p)]$ , such that the approximation satisfies

$$q_{\mathbf{C}}^*(\mathbf{x}) = \arg \min_{q_{\mathbf{C}}(\mathbf{x})} \text{KL}\{q_{\mathbf{C}}(\mathbf{x})||p(\mathbf{x}|\mathbf{y})\} = \arg \min_{q_{\mathbf{C}}(\mathbf{x})} \text{KL}\{q_{\mathbf{C}}(\mathbf{x})||p(\mathbf{x})\} - \mathbb{E}_{q_{\mathbf{C}}(\mathbf{x})}[\ln p(\mathbf{y}|\mathbf{x})],$$

where  $p(\mathbf{y}|\mathbf{x})$  is the likelihood and  $p(\mathbf{x})$  is the prior. Letting the true posterior  $p(\mathbf{x}|\mathbf{y})$  in Sklar's representation be  $p(\mathbf{x}|\mathbf{y}) = c^*(\mathbf{v}) \prod_j f_j^*(x_j)$ , where  $\mathbf{v} = [F_1^*(x_1), \dots, F_p^*(x_p)]$ ,  $c^*(\mathbf{v})$  and  $\{f_j^*(x_j)\}_{j=1:p}$  are the true underlying copula density and marginal posterior densities, respectively, the KL divergence decomposes into additive terms (derivations are provided below),

$$\text{KL}\{q_{\mathbf{C}}(\mathbf{x})||p(\mathbf{x}|\mathbf{y})\} = \text{KL}\{c[F(\mathbf{x})]||c^*[F^*(\mathbf{x})]\} + \sum_j \text{KL}\{f_j(x_j)||f_j^*(x_j)\}. \quad (3.9)$$

### 3.2.1 KL Additive Decomposition

Letting the variational proposal in Sklar's representation be

$$q_{\mathbf{V}\mathbf{C}}(\mathbf{x}) = c(\mathbf{u}) \prod_{j=1}^p f_j(x_j),$$

and the true posterior be  $p(\mathbf{x}|\mathbf{y}) = c^*(\mathbf{v}) \prod_j f_j^*(x_j)$ , where

$$\mathbf{u} = F(\mathbf{x}) = [F_1(x_1), \dots, F_p(x_p)], \quad \mathbf{v} = F^*(\mathbf{x}) = [F_1^*(x_1), \dots, F_p^*(x_p)].$$

The KL divergence decomposes into additive terms,

$$\begin{aligned} \text{KL}\{q(\mathbf{x})||p(\mathbf{x}|\mathbf{y})\} &= \int q(\mathbf{x}) \left( \log \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})} \right) d\mathbf{x} \\ &= \int c[F(\mathbf{x})] \prod_j f_j(x_j) \left( \log \frac{c[F(\mathbf{x})] \prod_j f_j(x_j)}{c^*[F^*(\mathbf{x})] \prod_j f_j^*(x_j)} \right) d\mathbf{x} \\ &= \int c[F(\mathbf{x})] \left( \log \frac{c[F(\mathbf{x})]}{c^*[F^*(\mathbf{x})]} \right) \prod_j dF_j(x_j) \\ &+ \int c[F(\mathbf{x})] \prod_j f_j(x_j) \left( \log \frac{\prod_j f_j(x_j)}{\prod_j f_j^*(x_j)} \right) \prod_j dx_j. \end{aligned} \quad (3.10)$$

The first term in (3.10)

$$\begin{aligned} \int c[F(\mathbf{x})] \left( \log \frac{c[F(\mathbf{x})]}{c^*[F^*(\mathbf{x})]} \right) \prod_j dF_j(x_j) &= \int c(\mathbf{u}) \left( \log \frac{c(\mathbf{u})}{c^*(F^*(F^{-1}(\mathbf{u})))} \right) d\mathbf{u} \\ &= \text{KL}\{c(\mathbf{u}) \| c^*[F^*(F^{-1}(\mathbf{u}))]\}, \end{aligned}$$

The second term in (3.10)

$$\begin{aligned} &\int c[F(\mathbf{x})] \prod_j f_j(x_j) \left( \log \frac{\prod_j f_j(x_j)}{\prod_j f_j^*(x_j)} \right) \prod_j dx_j \\ &= \sum_j \int c[F(\mathbf{x})] \prod_j f_j(x_j) \left( \log \frac{f_j(x_j)}{f_j^*(x_j)} \right) \prod_j dx_j \\ &= \sum_j \int f_j(x_j) \left( \log \frac{f_j(x_j)}{f_j^*(x_j)} \right) dx_j \quad (\text{Marginal Closed Property}) \\ &= \sum_j \text{KL}\{f_j(x_j) \| f_j^*(x_j)\}, \end{aligned}$$

Therefore

$$\text{KL}\{q(\mathbf{x}) \| p(\mathbf{x}|\mathbf{y})\} = \text{KL}\{c[F(\mathbf{x})] \| c^*[F^*(\mathbf{x})]\} + \sum_j \text{KL}\{f_j(x_j) \| f_j^*(x_j)\}.$$

Classical methods, such as MFVB and the VG approximation are special cases of the proposed VC inference framework. We next compare their KL divergence under Sklar's representation and offer a reinterpretation of them under the proposed framework.

### 3.2.2 Special Case 1: Mean-field VB

The mean-field proposal corresponds to the independence copula  $C_{\Pi}(\mathbf{u}) = \prod_{j=1}^J u_j$  with free-form marginal densities  $f_j(\mathbf{x}_j)$ . Given  $c_{\Pi}(\mathbf{u}) = 1$  we have

$$q_{\Pi}(\mathbf{x}) = c_{\Pi}(\mathbf{u}) \prod_j f_j(\mathbf{x}_j) = \prod_j f_j(\mathbf{x}_j) = q_{\text{VB}}(\mathbf{x}).$$

If MFVB is not fully factorized, i.e.  $J < p$ , the independence copula is the only copula satisfying the marginal closed property, according to the impossibility theorem (Nelsen, 2007). MFVB assumes an independence copula and only optimizes the free-form margins,

$$\text{KL}\{q_{\text{VB}}(\mathbf{x})||p(\mathbf{x}|\mathbf{y})\} = \text{KL}\{c_{\Pi}[F(\mathbf{x})]||c^*[F^*(\mathbf{x})]\} + \sum_j \text{KL}\{f_j(x_j)||f_j^*(x_j)\}. \quad (3.11)$$

The lowest achievable KL divergence in MFVB is

$$\text{KL}\{q_{\text{VB}}(\mathbf{x})||p(\mathbf{x}|\mathbf{y})\} = \text{KL}\{c_{\Pi}(\mathbf{u})||c^*(\mathbf{u})\},$$

which is achieved when the true posterior marginals are found, i.e.  $F_j \equiv F_j^*, \forall j$ , in which case the overall KL divergence is reduced to the KL divergence between the independence copula and the true copula. As is shown in (3.11), the objective function contains two terms, both involving marginal CDFs  $\{F_j\}_{j=1:p}$ . Since in general  $c^* \neq c_{\Pi}$ , the optimal  $F$  minimizing the first term will not be equal to  $F^*$ . Therefore, minimizing (3.11) will not lead to the correct marginals and this partially explains the reason why MFVB usually cannot find the true marginal posteriors in practice (e.g., variances can be severely underestimated (Neville *et al.*, 2014)), even though it allows for free-form margins.

### 3.2.3 Special Case 2: VG Approximation

In fixed-form variational Bayes (Honkela *et al.*, 2010), such as VG approximation, the multivariate Gaussian proposal  $q_{\text{VG}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  can be written as  $q_{\text{VG}}(\mathbf{x}) = c_{\text{G}}(\mathbf{u}|\boldsymbol{\Upsilon}) \prod_{j=1}^p \phi_j(x_j; \mu_j, \sigma_j^2)$ . VG not only assumes the true copula function is a Gaussian copula (Song, 2000) with parameter  $\boldsymbol{\Upsilon} = \mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2}$ ,  $\mathbf{D} = \text{diag}(\boldsymbol{\Sigma})$ , but is also restricted to univariate Gaussian marginal densities  $\{\phi_j(x_j; \mu_j, \sigma_j^2)\}_{j=1:p}$ ,

$$\text{KL}\{q_{\text{VG}}(\mathbf{x})||p(\mathbf{x}|\mathbf{y})\} = \text{KL}\{c_{\text{G}}[\Phi(\mathbf{x})]||c^*[F^*(\mathbf{x})]\} + \sum_j \text{KL}\{\phi_j(x_j)||f_j^*(x_j)\}. \quad (3.12)$$

We can see in (3.12) that if the margins are misspecified, even if the true underlying copula is a Gaussian copula, there could still be a discrepancy  $\sum_j \text{KL}\{\phi_j(x_j)||f_j^*(x_j)\}$

between margins, and  $\text{KL}\{c_G[\Phi(\mathbf{x})]||c^*[F^*(\mathbf{x})]\}$  is not zero.

Concerning analytical tractability and simplicity, in the sequel we concentrate on variational Gaussian copula (VGC) proposals constructed via Gaussian copula with continuous margins, i.e.  $q_{\text{VGC}}(\mathbf{x}) = c_G(\mathbf{u}|\mathbf{\Upsilon}) \prod_{j=1}^p f_j(x_j)$ , where  $\mathbf{u} = [F_1(x_1), \dots, F_p(x_p)]$ . Our VGC method extends MFVB and VG, and improves upon both by allowing simultaneous updates of the Gaussian copula parameter  $\mathbf{\Upsilon}$  and the adaptation of marginal densities  $\{f_j(x_j)\}_{j=1:p}$ . First, the univariate margins in VGC is not restricted to be Gaussian. Second, the Gaussian copula in VGC is more resistant to local optima than the independence copula assumed in MFVB and alleviates its variance underestimation pitfall, as is demonstrated in Section 3.6.3.

### 3.3 Variational Gaussian Copula Approximation

A Gaussian copula function with  $p \times p$  correlation matrix  $\mathbf{\Upsilon}$  is defined as

$$C_G(u_1, \dots, u_p|\mathbf{\Upsilon}) = \Phi_p(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p)|\mathbf{\Upsilon}) : [0, 1]^p \rightarrow [0, 1] \quad (3.13)$$

where  $\Phi(\cdot)$  is a shorthand notation of the CDF of  $\mathcal{N}(0, 1)$ , and  $\Phi_p(\cdot|\mathbf{\Upsilon})$  is the CDF of  $N_p(\mathbf{0}, \mathbf{\Upsilon})$ . The Gaussian copula density is

$$c_G(u_1, \dots, u_p|\mathbf{\Upsilon}) = \frac{1}{\sqrt{|\mathbf{\Upsilon}|}} \exp \left\{ -\frac{\mathbf{z}^T (\mathbf{\Upsilon}^{-1} - \mathbf{I}_p) \mathbf{z}}{2} \right\},$$

where  $\mathbf{z} = [\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p)]^T$ .

In the proposed VGC approximation, the variational proposal  $q_{\text{VGC}}(\mathbf{x})$  is constructed as a product of Gaussian copula density and continuous marginal densities.

The evidence lower bound (ELBO) of VGC approximation is

$$\mathcal{L}_C[q_{\text{VGC}}(\mathbf{x})] = \int \left[ c_G[F(\mathbf{x})] \times \prod_{j=1}^p f_j(x_j) \right] \ln p(\mathbf{y}, \mathbf{x}) d\mathbf{x} + H[c_G(\mathbf{u})] + \sum_{j=1}^p H[f_j(x_j)], \quad (3.14)$$

where  $u_j = F_j(x_j)$  and  $H[f(x)] = -\int f(x) \ln f(x) dx$ .

However, directly optimizing the ELBO in (3.14) w.r.t. the Gaussian copula parameter  $\Upsilon$  and the univariate marginals  $\{f_j(x_j)\}_{j=1:p}$  often leads to a non-trivial variational calculus problem. For computational convenience, we present several equivalent proposal constructions based on Jacobian transformation and reparameterization.

Table 3.1: Equivalent Representations of VGC Proposals

Posterior Formulation	Optimization Space	
Original	Multivariate (non-Gaussian) density $q(\mathbf{x})$	
Sklar's Representation	Copula density $c_G(\mathbf{u} \Upsilon)$	Univariate marginals $\{f_j\}_{j=1:p}$
Jacobian Transform	Gaussian density $q(\tilde{\mathbf{z}}) = \mathcal{N}(\mathbf{0}, \Upsilon)$	Monotone functions $\{g_j\}_{j=1:p}$
Parameter Expansion	Gaussian density $q(\tilde{\mathbf{z}}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}\mathbf{C}^T)$	Monotone functions $\{h_j\}_{j=1:p}$

### 3.3.1 Equivalent Variational Proposals

We incorporate auxiliary variables  $\mathbf{z}$  by exploiting the latent variable representation of the Gaussian copula:  $x_j = F_j^{-1}(u_j)$ ,  $u_j = \Phi(z_j)$ ,  $\mathbf{z} \sim N_p(\mathbf{0}, \Upsilon)$ . Letting  $g_j(\cdot) = F_j^{-1}(\Phi(\cdot))$  be bijective monotonic non-decreasing functions,  $x_j = g_j(z_j)$ ,  $\forall j$ , the Jacobian transformation gives

$$q_{\text{VGC}}(\mathbf{x}) = \int \left[ \prod_{j=1}^p \delta(x_j - g_j(z_j)) \right] q_G(\mathbf{z}; \mathbf{0}, \Upsilon) d\mathbf{z} = q_G(g^{-1}(\mathbf{x}); \mathbf{0}, \Upsilon) \left[ \prod_{j=1}^p \frac{d}{dx_j} g_j^{-1}(x_j) \right],$$

where  $\delta(\cdot)$  is the Dirac delta function.

It is inconvenient to directly optimize the correlation matrix  $\Upsilon$  of interest, since  $\Upsilon$  is a positive semi-definite matrix with ones on the diagonal and off-diagonal elements between  $[-1, 1]$ . We adopt the parameter expansion (PX) technique (Liu *et al.*, 1998; Liu and Wu, 1999), which has been applied in accelerating variational Bayes (Qi and Jaakkola, 2006) and the sampling of correlation matrix (Talhouk *et al.*, 2012). Further considering

$$\tilde{z}_j = t_j^{-1}(z_j) = \mu_j + \sigma_{jj} z_j, \quad \tilde{\mathbf{z}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \mathbf{D}\Upsilon\mathbf{D}^T, \quad (3.15)$$

$\mathbf{D} = [\text{diag}(\sigma_{jj})]_{j=1:p}$ , thus  $x_j = g(z_j) = g(t(\tilde{z}_j)) := h(\tilde{z}_j)$ , where  $h_j(\cdot) = g_j \circ t_j(\cdot)$  are also bijective monotonic non-decreasing functions, the variational proposal is further written as

$$q_{\text{VGC}}(\mathbf{x}) = \int \left[ \prod_{j=1}^p \delta(x_j - h_j(\tilde{z}_j)) \right] q_{\text{G}}(\tilde{\mathbf{z}}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\tilde{\mathbf{z}} = q_{\text{G}}(h^{-1}(\mathbf{x}); \boldsymbol{\mu}, \boldsymbol{\Sigma}) \left[ \prod_{j=1}^p \frac{d}{dx_j} h_j^{-1}(x_j) \right].$$

Given the transformations  $\{h_j\}_{j=1:p}$ ,  $q_{\text{G}}(\tilde{\mathbf{z}}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  can be further reparameterized by the Cholesky decomposition  $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^T$  (Challis and Barber, 2013; Titsias and Lázaro-Gredilla, 2014), where  $\mathbf{C}$  is a square lower triangular matrix. Table 3.1 summarizes four translatable representations of variational proposals.

### 3.3.2 VGC with Fixed-form Margins

The ELBO under Sklar's representation (3.14) is therefore translated into the Jacobian representation

$$\begin{aligned} \mathcal{L}_{\text{C}}[q_{\text{VGC}}(\mathbf{x})] &= \mathbb{E}_{\mathcal{N}(\tilde{\mathbf{z}}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}[\ell_s(\tilde{\mathbf{z}}) - \ln q_{\text{G}}(\tilde{\mathbf{z}})], \\ \ell_s(\tilde{\mathbf{z}}, h) &= \ln p(\mathbf{y}, h(\tilde{\mathbf{z}})) + \sum_{j=1}^p \ln h'_j(\tilde{z}_j). \end{aligned} \quad (3.16)$$

The monotonic transformations  $h_j(\cdot) = F_j^{-1}[\Phi(t(\cdot))]$  can be specified according to the desired parametric form of marginal posterior, if the inverse CDF  $F_j^{-1}$  is tractable. For example, the multivariate log-normal posterior can be constructed via a Gaussian copula with log-normal (LN) margins,

$$q_{\text{VGC-LN}}(\mathbf{x}) = C_{\text{G}}(\mathbf{u} | \boldsymbol{\Upsilon}) \prod_{j=1}^p \text{LN}(x_j; \mu_j, \sigma_j^2). \quad (3.17)$$

This also corresponds to imposing exponential transform on Gaussian variables,  $\mathbf{x} = h(\tilde{\mathbf{z}}) = \exp(\tilde{\mathbf{z}})$ ,  $\tilde{\mathbf{z}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . In this case,  $\{\mu_j, \sigma_j^2\}_{j=1:p}$  controls the location and dispersion of the marginal density;  $h(\cdot)$  does not have any additional parameters

to control the shape and  $\ln h'(\tilde{z}_j) = \tilde{z}_j$  takes a simple form. VGC-LN is further discussed in Section 3.6.2 and Section 3.6.3.

Given the copula function  $C$ , we only need to find  $p$  one-dimensional margins. However, without knowing characteristics of the latent variables, specifying appropriate parametric form for margins is a difficult task in general cases. First, the marginals might exhibit multi-modality, high skewness or kurtosis, which are troublesome for particular parametric marginals to capture. Second, a tractable inverse CDF with optimizable arguments/parameters, as required here, are available only in a handful of cases. Instead of using some arbitrary parametric form, we construct bijective transform functions via kernel mixtures, which lead to highly flexible (ideally free-form) marginal proposals.

### 3.4 Bernstein Polynomials based Monotone Transformations

The marginal densities in VGC can be recovered through Jacobian transformation,

$$f_j(x_j) = q_G(h_j^{-1}(x_j); \mu_j, \sigma_2^2) \frac{d}{dx_j} h_j^{-1}(x_j) = q_G(h_j^{-1}(x_j); \mu_j, \sigma_2^2) \frac{1}{h_j'(h_j^{-1}(x_j))}, \quad (3.18)$$

where the  $[h_j'(h_j^{-1}(x_j))]^{-1}$  term is interpreted as a marginal-correction term. To guarantee analytical tractability, we require  $h(\cdot)$  to be (i) bijective; (ii) monotonic non-decreasing; (iii) having unbounded/constrained range; (iv) differentiable with respect to both its argument and parameters; and (v) sufficiently flexible. We propose a class of continuous and smooth transformations  $h(\cdot)$  constructed via kernel mixtures that automatically have these desirable properties.

#### 3.4.1 Continuous Margins Constructed via Bernstein Polynomials

The Bernstein polynomials (BPs) have a uniform convergence property for continuous functions on unit interval  $[0, 1]$  and have been used for nonparametric density estimation (Petrone, 1999). It seems more natural to use kernel mixtures directly as

the variational proposal. However, the difficulty lies in tackling the term  $f(F^{-1}(\cdot))$  involving the inverse CDF of mixtures (not analytical) and the need of a further lower bound on the entropy of mixtures. In this chapter, we overcome this issue by using a sandwich-type construction of the transform  $h(\tilde{z})^1$  which maps from  $(-\infty, \infty)$  to some target range building upon BP,

$$h(\tilde{z}) = \Psi^{-1}[B(\Phi(\tilde{z}); k, \boldsymbol{\omega})],$$

$$B(u; k, \boldsymbol{\omega}) = \sum_{r=1}^k \omega_{r,k} I_u(r, k - r + 1), \quad (3.19)$$

where  $I_u(r, k - r + 1)$  is the regularized incomplete beta function.  $\Phi(\cdot)$  is the standard normal CDF mapping from  $(-\infty, \infty)$  to  $[0, 1]$ , and  $\Psi^{-1}(\cdot)$  is some predefined tractable inverse CDF with fixed parameters; for example, the inverse CDF of the exponential distribution helps map from  $[0, 1]$  to  $(0, \infty)$  for positive variables.  $B(u; k, \boldsymbol{\omega})$  relocates the probability mass on the unit interval  $[0, 1]$ . The degree  $k$  is an unknown smoothing parameter, and  $\boldsymbol{\omega}$  is the unknown mixture weights on the probability simplex  $\Delta_k = \{(\omega_1, \dots, \omega_k) : \omega_i \geq 0, \sum_i \omega_i = 1\}$ . The proposed sandwich-type transformation avoids the difficulty of specifying any particular types of marginals, while still leads to tractable derivations presented in Section 3.5.

### 3.4.2 Variational Inverse Transform

Considering a 1-d variational approximation problem ( $x$  is a scalar, the true posterior  $f(x)$  is known up to the normalizing constant), fix  $q(\tilde{z}) = \mathcal{N}(0, 1)$ , thus  $u = \Phi(\tilde{z}) \sim \mathcal{U}[0, 1]$ , we can learn the monotonic transformation  $\xi(\cdot) = Q^{-1}(\cdot)$  on the base uniform distribution  $q_0(u)$  by solving a variational problem,

$$\xi^*(\cdot) = \arg \min_{\xi} \text{KL}\{q(x) || f(x)\}, \quad x = \xi(u) = Q^{-1}(u),$$

---

<sup>1</sup> The index  $j$  on  $\tilde{z}$  is temporarily omitted for simplicity, and is added back when necessary.

i.e., if we generate  $u \sim \mathcal{U}[0, 1]$ , then  $x = \xi^*(u) \sim Q^*$ .  $Q^*$  is closest to the true distribution  $F$  with the minimum KL divergence. This can be interpreted as the variational counterpart of the inverse transform sampling (Devroye, 1986), termed as variational inverse transform (VIT). Our BP-based construction  $\xi(\cdot) = Q^{-1}(\cdot) = \Psi^{-1}(B(u; k, \boldsymbol{\omega}))$  is one appropriate parameterization scheme for the inverse probability transformation  $Q^{-1}(\cdot)$ . VIT-BP offers two clear advantages. First, as opposed to fixed-form variational Bayes, it does not require any specification of parametric form for  $q(x)$ . Second, the difficult task of calculating the general inverse CDFs  $Q^{-1}(\cdot)$  is lessened to the much easier task of calculating the predefined tractable inverse CDF  $\Psi^{-1}(\cdot)$ . Some choices of  $\Psi(\cdot)$  include CDF of  $\mathcal{N}(0, 1)$  for variables in  $(-\infty, \infty)$ , Beta(2, 2) for truncated variables in  $(0, 1)$ .

To be consistent with VIT, we shall set  $\Phi(\cdot)$  in (3.19) to be  $\Phi(\cdot|\mu, \sigma^2)$ , instead of  $\Phi(\cdot|0, 1)$ , such that  $u$  is always uniformly distributed. Ideally, BP itself suffices to represent arbitrary continuous distribution function on the unit interval. However, it might require a higher order  $k$ . As is demonstrated in Section 3.6.1, this requirement can be alleviated by incorporating auxiliary parameters  $\{\mu, \sigma^2\}$  in VGC-BP, which potentially help in changing location and dispersion of the probability mass.

### 3.5 Stochastic VGC

The derivations of deterministic VGC updates are highly model-dependent. First, due to the cross terms often involved in the log likelihood/prior, the corresponding Gaussian expectations and their derivatives may not be analytically tractable. Second, owing to the non-convex nature of many problems, only locally optimal solutions can be guaranteed. In contrast, stochastic implementation of VGC only requires the evaluation of the log-likelihood and log-prior along with their derivatives, eliminating most model-specific derivations, and it provides a chance of escaping local optima by introducing randomness in gradients.

---

**Algorithm 1** (VGC-BP) Stochastic Variational Gaussian Copula Inference with Bernstein Polynomials
 

---

**Input:** observed data  $\mathbf{y}$ , user specified model  $\ln p(\mathbf{y}, \mathbf{x})$  and first-order derivatives  $\nabla_{\mathbf{x}} \ln p(\mathbf{y}, \mathbf{x})$ , Bernstein polynomials degree  $k$ , predefined  $\Psi(\cdot)$  and  $\Phi(\cdot)$

**Initialize** variational parameter  $\Theta_0 = (\boldsymbol{\mu}_0, \mathbf{C}_0, \{\boldsymbol{\omega}_0^{(j)}\}_{j=1:p})$ ,  $t = 0$ .

**repeat**

$t = t + 1$ ,

Sample  $\tilde{\boldsymbol{\epsilon}} \sim q_G(\tilde{\boldsymbol{\epsilon}}, \mathbf{0}, \mathbf{I}_p)$ , and set  $\tilde{\mathbf{z}} = \boldsymbol{\mu}_{t-1} + \mathbf{C}_{t-1}\boldsymbol{\epsilon}$ ,

$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \lambda_t [\nabla_{\tilde{\mathbf{z}}} \ell_s(\tilde{\mathbf{z}}, h) - \nabla_{\tilde{\mathbf{z}}} \ln q_G(\tilde{\mathbf{z}})]$ , % Update  $\boldsymbol{\mu}_{t-1}$  with stepsize  $\lambda_t$

$\mathbf{C}_t = \mathbf{C}_{t-1} + \eta_t [\nabla_{\tilde{\mathbf{z}}} \ell_s(\tilde{\mathbf{z}}, h) - \nabla_{\tilde{\mathbf{z}}} \ln q_G(\tilde{\mathbf{z}})] \boldsymbol{\epsilon}^T$ , % Update  $\mathbf{C}_{t-1}$  with stepsize  $\eta_t$

**for**  $j = 1$  **to**  $p$  **do**

$\boldsymbol{\omega}_t^{(j)} = \mathcal{P}(\boldsymbol{\omega}_{t-1}^{(j)} + \xi_t^{(j)} \nabla_{\boldsymbol{\omega}^{(j)}} \ell_s(\tilde{\mathbf{z}}, h))$ , % Update  $\boldsymbol{\omega}_{t-1}^{(j)}$  with stepsize  $\xi_t^{(j)}$  and gradient projection  $\mathcal{P}$

**end for**

**until** convergence criterion is satisfied

**Output:** marginal parameters  $(\{\boldsymbol{\omega}^{(j)}\}_{j=1:p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$  and copula parameters  $\Upsilon$

---

### 3.5.1 Coordinate Transformations

Applying the coordinate transformations of stochastic updates,  $\tilde{\mathbf{z}} = \boldsymbol{\mu} + \mathbf{C}\boldsymbol{\epsilon}$ ,  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , introduced in (Rezende *et al.*, 2014; Titsias and Lázaro-Gredilla, 2014), the gradient of the ELBO w.r.t. variational parameter  $(\boldsymbol{\mu}, \mathbf{C})$  can be written as

$$\begin{aligned} \nabla_{\boldsymbol{\mu}} \mathcal{L}_C &= \mathbb{E}_{q_G(\tilde{\mathbf{z}})} [\nabla_{\tilde{\mathbf{z}}} \ell_s(\tilde{\mathbf{z}}, h) - \nabla_{\tilde{\mathbf{z}}} \ln q_G(\tilde{\mathbf{z}})], \\ \nabla_{\mathbf{C}} \mathcal{L}_C &= \mathbb{E}_{q_G(\tilde{\mathbf{z}})} [\nabla_{\tilde{\mathbf{z}}} (\ell_s(\tilde{\mathbf{z}}, h) - \ln q_G(\tilde{\mathbf{z}})) \boldsymbol{\epsilon}^T], \end{aligned} \quad (3.20)$$

where the stochastic gradient terms

$$\nabla_{\tilde{z}_j} \ell_s(\tilde{\mathbf{z}}) = \nabla_{\tilde{z}_j} \ln p(\mathbf{y}, h(\tilde{\mathbf{z}})) + \nabla_{\tilde{z}_j} \ln h'_j(\tilde{z}_j) = \frac{\partial \ln p(\mathbf{y}, \mathbf{x})}{\partial x_j} h'_j(\tilde{z}_j) + \nabla_{\tilde{z}_j} \ln h'_j(\tilde{z}_j).$$

If necessary, the Gaussian copula can be replaced with other appropriate parametric forms. The coordinate transformation supports many other distributions as well, for example, those described in Appendix C.2. of Rezende *et al.* (2014).

According to the chain rule, the first derivative of  $h(\cdot)$  w.r.t  $\tilde{z}$  is,

$$h'(\tilde{z}) = \frac{d\Psi^{-1}[B(\Phi(\tilde{z}); k, \boldsymbol{\omega})]}{dB(\Phi(\tilde{z}); k, \boldsymbol{\omega})} \frac{dB(\Phi(\tilde{z}); k, \boldsymbol{\omega})}{d\Phi(\tilde{z})} \frac{d\Phi(\tilde{z})}{d\tilde{z}} = \frac{b(\Phi(\tilde{z}); k, \boldsymbol{\omega})\phi(\tilde{z})}{\psi(h(\tilde{z}))}, \quad (3.21)$$

where  $b(u; k, \boldsymbol{\omega}) = \sum_{r=1}^k \omega_{r,k} \beta(u; r, k-r+1)$ ,  $\beta(x; a, b)$  is the beta density  $\beta(x; a, b) = \Gamma(a+b)/(\Gamma(a)\Gamma(b))x^{a-1}(1-x)^{b-1}$ . Therefore,  $\ln h'(\tilde{\mathbf{z}}) = \ln b(\Phi(\tilde{\mathbf{z}}); k, \boldsymbol{\omega}) + \ln \phi(\tilde{\mathbf{z}}) - \ln \psi(h(\tilde{\mathbf{z}}))$  and  $\nabla_{\tilde{\mathbf{z}}_j} \ln h'_j(\tilde{z}_j) = h''_j(\tilde{z}_j)/h'_j(\tilde{z}_j)$  all take analytical expressions, where

$$h''_j(\tilde{z}_j) = [\rho'_1(\tilde{z}_j)\rho_2(\tilde{z}_j)\rho_3(\tilde{z}_j) + \rho_1(\tilde{z}_j)\rho'_2(\tilde{z}_j)\rho_3(\tilde{z}_j) - \rho_1(\tilde{z}_j)\rho_2(\tilde{z}_j)\rho'_3(\tilde{z}_j)]/[\rho_3(\tilde{z}_j)]^2,$$

where

$$\rho_1(\tilde{z}_j) = b(u_j; k, \boldsymbol{\omega}^{(j)}), \quad \rho_2(\tilde{z}_j) = \phi(\tilde{z}_j), \quad \rho_3(\tilde{z}_j) = \psi(h_j(\tilde{z}_j))$$

$$\rho'_1(\tilde{z}_j) = \phi(\tilde{z}_j) \sum_{r=1}^k \omega_{r,k}^{(j)} \beta'(u_j; r, k-r+1),$$

$$\rho'_2(\tilde{z}_j) = -\tilde{z}_j \phi(\tilde{z}_j), \quad \rho'_3(\tilde{z}_j) = \psi'(h_j(\tilde{z}_j))h'_j(\tilde{z}_j), \quad u_j = \Phi(\tilde{z}_j),$$

$\phi(\cdot)$  is the PDF of  $\mathcal{N}(0, 1)$ ,  $\psi(\cdot)$  and  $\psi'(\cdot)$  are the predefined PDF and its derivative respectively. Defining  $\beta(x; a, 0) = \beta(x; 0, b) = 0$ , the derivative is written as a combination of two polynomials of lower degree

$$\beta'(x; a, b) = (a+b-1)[\beta(x; a-1, b) - \beta(x; a, b-1)]. \quad (3.22)$$

In stochastic optimization, the gradients expressed in terms of expectations are approximated using Monte Carlo integration with finite samples. The gradients contain expectations on additive terms. Note that Rezende *et al.* (2014) and Titsias and Lázaro-Gredilla (2014) ignore the stochasticity in the entropy term  $\mathbb{E}_{q_G(\tilde{\mathbf{z}})}[-\ln q_G(\tilde{\mathbf{z}})]$  and assume  $\nabla_{\boldsymbol{\mu}} \mathbb{E}_{q_G(\tilde{\mathbf{z}})}[-\ln q_G(\tilde{\mathbf{z}})] = 0$  and  $\nabla_{\mathbf{C}} \mathbb{E}_{q_G(\tilde{\mathbf{z}})}[-\ln q_G(\tilde{\mathbf{z}})] = \text{diag}[1/C_{jj}]_{j=1:p}$ . This creates an inconsistency as we only take finite samples in approximating the term  $\mathbb{E}_{q_G(\tilde{\mathbf{z}})}[\nabla_{\tilde{\mathbf{z}}} \ell_s(\tilde{\mathbf{z}})]$ , and perhaps surprisingly, this also results in an increase of the gradient variance and the sensitivity to the learning rates. Our method is inherently more stable, as the difference between the gradients,  $\nabla_{\tilde{\mathbf{z}}}[\ell_s(h(\tilde{\mathbf{z}})) - q_G(\tilde{\mathbf{z}})]$ ,  $\forall \tilde{\mathbf{z}}$ , tends to zero when the convergent point is approached. In contrast, the gradients in previous method diffuses with a constant variance even around the global maximum. This phenomenon is illustrated in Section 3.6.2.

The alternative log derivative approach are also applicable to VGC inference and other types of copulas, see Paisley *et al.* (2012); Mnih and Gregor (2014); Rezende *et al.* (2014) for references. We leave this exploration open for future investigation.

### 3.5.2 Update the BP Weights

Under a given computational budget, we prefer a higher degree  $k$ , as there is no over-fitting issue in this variational density approximation task. Given  $k$ , the basis functions are completely known, depending only on index  $r$ . The only parameter left to be optimized in the Bernstein polynomials is the mixture weights. Therefore, this construction is relatively simpler than Gaussian mixture proposals (Gershman *et al.*, 2012; Nguyen and Bonilla, 2014). Assuming permissibility of interchange of integration and differentiation holds, we have  $\nabla_{\omega^{(j)}} \mathcal{L}_C = \mathbb{E}_{q_G(\tilde{\mathbf{z}})} [\nabla_{\omega^{(j)}} \ell_s(\tilde{\mathbf{z}}, h, \mathbf{y})]$ , with the stochastic gradients

$$\begin{aligned} \nabla_{\omega^{(j)}} \ell_s(\tilde{\mathbf{z}}, h, \mathbf{y}) &= \nabla_{\omega^{(j)}} \ln p(\mathbf{y}, h(\tilde{\mathbf{z}})) + \nabla_{\omega^{(j)}} \ln h'_j(\tilde{z}_j) \\ &= \frac{\partial \ln p(\mathbf{y}, \mathbf{x})}{\partial x_j} \left[ \frac{\partial h_j(\tilde{z}_j)}{\partial \omega_{r,k}^{(j)}} \right]_{r=1:k} + \left[ \frac{\partial \ln h'_j(\tilde{z}_j)}{\partial \omega_{r,k}^{(j)}} \right]_{r=1:k}, \end{aligned}$$

where

$$\begin{aligned} \frac{\partial h_j(\tilde{z}_j)}{\partial \omega_{r,k}^{(j)}} &= \frac{\partial \Psi^{-1}[B(u_j; k, \omega^{(j)})]}{\partial \omega_{r,k}^{(j)}} = \frac{I_{u_j}(r, k - r + 1)}{\psi(h_j(\tilde{z}_j))}, \\ \frac{\partial \ln h'_j(\tilde{z}_j)}{\partial \omega_{r,k}^{(j)}} &= \beta(u_j; r, k - r + 1) / b(u_j; k, \omega^{(j)}) - \frac{\psi'(h_j(\tilde{z}_j))}{\{\psi(h_j(\tilde{z}_j))\}^2} I_{u_j}(r, k - r + 1). \end{aligned}$$

The gradients w.r.t  $\omega^{(j)}$  turn into expectation straightforwardly, to enable stochastic optimization of the ELBO. To satisfy the constraints of  $\omega^{(j)}$  on the probability simplex, we apply the gradient projection operation  $\mathcal{P}$  introduced in Duchi *et al.* (2008) with complexity  $\mathcal{O}(k \log k)$ . The above derivations related to BPs together with those in Section 3.5.1 are all analytic and model-independent. The only two

model-specific terms are  $\ln p(\mathbf{y}, \mathbf{x})$  and  $\partial \ln p(\mathbf{y}, \mathbf{x})/\partial \mathbf{x}$ . The stochastic optimization algorithm is summarized in Algorithm 2, with little computational overhead added relative to stochastic VG. The stability and efficiency of the stochastic optimization algorithm can be further improved by embedding adaptive subroutines (Duchi *et al.*, 2011) and considering second-order optimization method (Fan *et al.*, 2015).

### 3.6 Experiments

We use Gaussian copulas with fixed/free-form margins as automated *inference engines* for posterior approximation in generic hierarchical Bayesian models. We evaluate the peculiarities reproduced in the univariate margins and the posterior dependence captured broadly across latent variables. This is done by comparing VGC methods to the ground truth and other baseline methods such as MCMC, MFVB, and VG. Matlab code for VGC is available from the GitHub repository: <https://github.com/shaobohan/VariationalGaussianCopula>

#### 3.6.1 Flexible Margins

We first assess the marginal approximation accuracy of our BP-based constructions in Section 3.4.2, i.e.,  $h(\cdot) = \Psi^{-1}(B(\Phi(\tilde{z}); k, \boldsymbol{\omega}))$  via 1-d variational optimization, where  $\tilde{z} \sim \mathcal{N}(0, 1)$  in VIT-BP, and  $\tilde{z} \sim \mathcal{N}(\mu, \sigma^2)$  in VGC-BP. For fixed BP order  $k$ , the shape of  $q(x)$  is adjusted solely by updating  $\boldsymbol{\omega}$ , according to the variational rule. In VGC-BP, the additional marginal parameters  $\{\mu, \sigma^2\}$  also contribute in changing location and dispersion of  $q(x)$ . Examining Figure 3.1, VGC-BP produces more accurate densities than VIT-BP under the same order  $k$ . Hereafter, the predefined  $\Psi(\cdot)$  for real variables, positive real variable, and truncated  $[0, 1]$  variables are chosen to be the CDF of  $\mathcal{N}(0, 1)$ ,  $\text{Exp}(1)$  and  $\text{Beta}(2, 2)$ , respectively. The model-specific derivations in these univariate cases are detailed as follows.

### *Skew Normal Distribution*

1.  $\ln p(x) \propto \ln \phi(x) + \ln \Phi(\alpha x)$  and  $\partial \ln p(x) / \partial x = -x + \alpha \phi(\alpha x) / \Phi(\alpha x)$ ,  $\alpha$  is the shape parameter
2.  $\Psi(x)$  is predefined as CDF of  $\mathcal{N}(0, 1)$

### *Student's t Distribution*

1.  $\ln p(x) \propto -(\nu + 1)/2 \ln(1 + x^2/\nu)$  and  $\partial \ln p(x) / \partial x = -(\nu + 1)x / (\nu + x^2)$ ,  $\nu > 0$  is the degrees of freedom
2.  $\Psi(x)$  is predefined as CDF of  $\mathcal{N}(0, 1)$

### *Gamma Distribution*

1.  $\ln p(x) \propto (\alpha - 1) \ln x - \beta x$  and  $\partial \ln p(x) / \partial x = (\alpha - 1)/x - \beta$ ,  $\alpha$  is the shape parameter,  $\beta$  is the rate parameter
2.  $\Psi(x)$  is predefined as CDF of  $\text{Exp}(1)$

### *Beta Distribution*

1.  $\ln p(x) \propto (a - 1) \ln x + (b - 1) \ln(1 - x)$  and
$$\partial \ln p(x) / \partial x = (a - 1)/x - (b - 1)/(1 - x), \quad \text{both } a, b > 0$$
2.  $\Psi(x)$  is predefined as CDF of  $\text{Beta}(2, 2)$

### *3.6.2 Bivariate Log-Normal*

The bivariate log-normal PDF  $p(x_1, x_2)$  (Aitchison and Brown, 1957) is given by

$$p(x_1, x_2) = \exp(-\zeta/2) / [2\pi x_1 x_2 \sigma_1 \sigma_2 \sqrt{1 - \rho^2}],$$
$$\zeta = \frac{1}{1 - \rho^2} \left[ \alpha_1^2(x_1) - 2\rho\alpha_1(x_1)\alpha_2(x_2) + \alpha_2^2(x_2) \right],$$

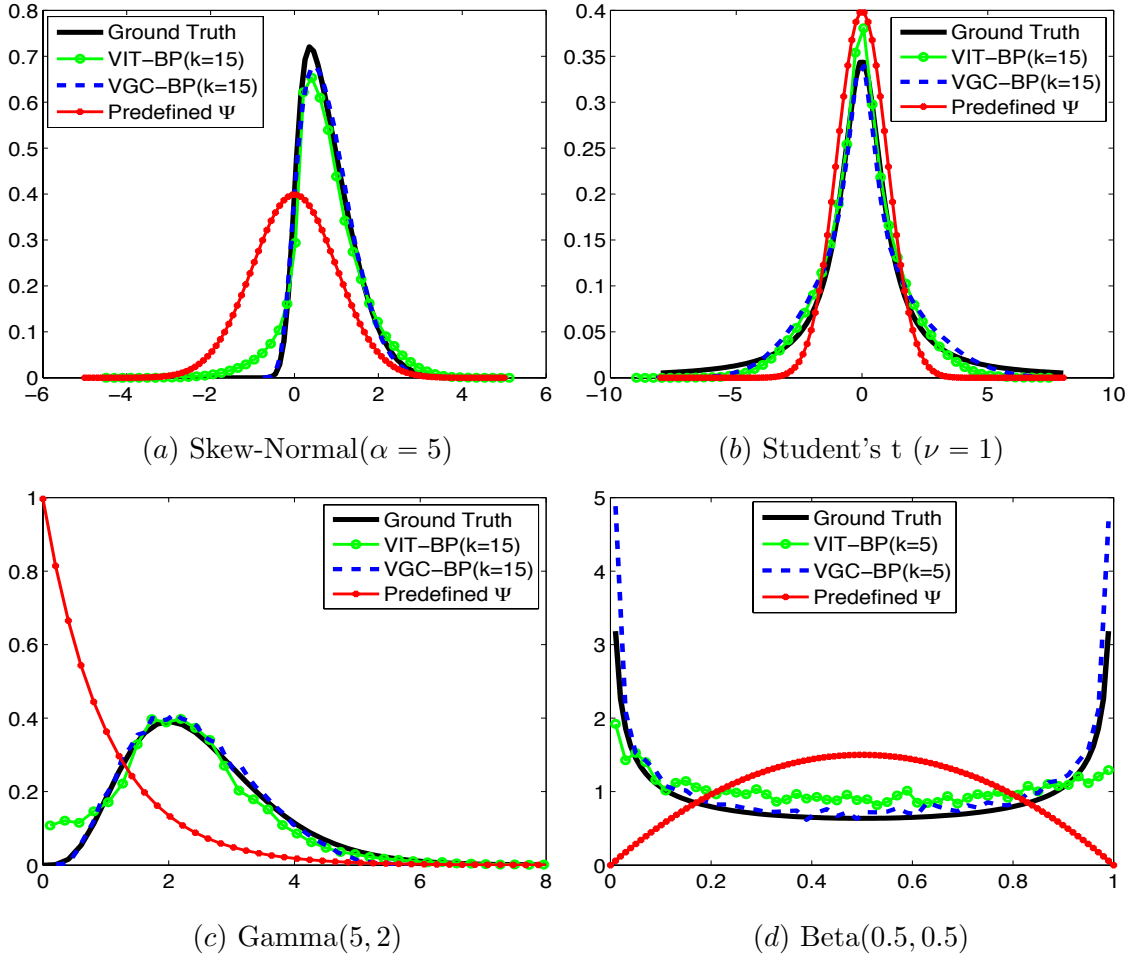


FIGURE 3.1: Marginal Adaptation: VIT-BP v.s. VGC-BP

where  $\alpha_i(x_i) = (\ln x_i - \mu_i)/\sigma_i$ ,  $i = 1, 2$ ,  $-1 < \rho < 1$ .

The model-specific derivations for this model are detailed in below.

1.  $\ln p(x_1, x_2) \propto -\ln x_1 - \ln x_2 - \zeta/2$  and

$$\frac{\partial \ln f(x_1, x_2)}{\partial x_1} = -\frac{1}{x_1} - \frac{\alpha_1(x_1) - \rho\alpha_2(x_2)}{(1 - \rho^2)x_1\sigma_1}$$

$$\frac{\partial \ln f(x_1, x_2)}{\partial x_2} = -\frac{1}{x_2} - \frac{\alpha_2(x_2) - \rho\alpha_1(x_1)}{(1 - \rho^2)x_2\sigma_2}$$

2.  $\Psi(x)$  is predefined as CDF of  $\text{Exp}(1)$

We construct a bivariate Gaussian copula with (i) Log-normal margins (VGC-LN) and (ii) BP-based margins (VGC-BP). We set  $\mu_1 = \mu_2 = 0.1$  and  $\sigma_1 = \sigma_2 = 0.5$ ,  $\rho = 0.4$  or  $-0.4$  (first and second row in Figure 3.2). Both VGC-LN and VGC-BP methods presume the correct form of the underlying copula (bivariate Gaussian) and learn the copula parameters  $\rho$ . VGC-LN further assumes exactly the true form of the univariate margins (log-normal) while VGC-BP is without any particular assumptions on parametric form of margins. Figure 3.2 shows that VGC-BP find as accurate joint posteriors as VGC-LN, even though the former assumes less knowledge about the true margins.

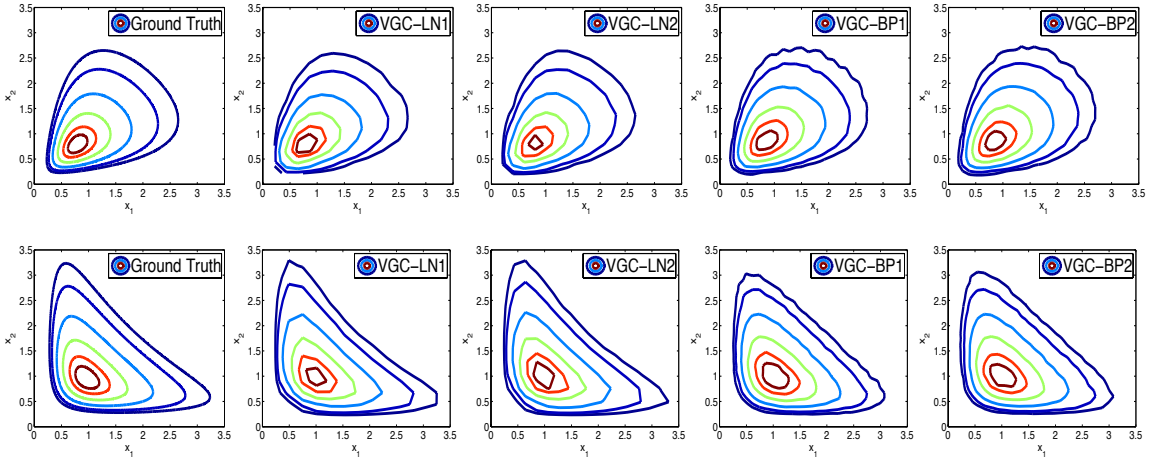


FIGURE 3.2: Approximate Posteriors via VGC methods

In updating  $(\boldsymbol{\mu}, \mathbf{C})$ , VGC-LN1 and VGC-BP1 follow the scheme in (Titsias and Lázaro-Gredilla, 2014) and neglect the stochasticity in the entropy term; while VGC-LN2 and VGC-BP2 are based on our scheme in (3.20). Under the same learning rates, we define the relative mean square error (RMSE) of the copula parameter as  $R(\rho) = \frac{(\hat{\rho} - \rho)^2}{\rho^2}$ ; both VGC-LN and VGC-BP results in Figure 3.3 consistently show that our method leads to less noisy gradients and converges faster.

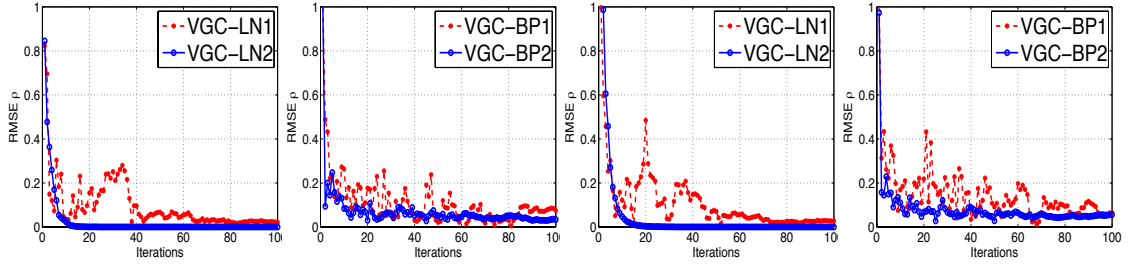


FIGURE 3.3: RMSE( $\rho$ ) of VGC-LN and VGC-BP v.s. Iterations; Left two:  $\rho = 0.4$ ; Right two:  $\rho = -0.4$

### 3.6.3 Horseshoe Shrinkage

The horseshoe distribution (Carvalho *et al.*, 2010) can be represented in equivalent conjugate hierarchies (Neville *et al.*, 2014)  $y|\tau \sim \mathcal{N}(0, \tau)$ ,  $\tau|\lambda \sim \text{InvGa}(0.5, \lambda)$ ,  $\lambda \sim \text{InvGa}(0.5, 1)$ . Here we assume  $y = 0.01$  is the (single) observation. Denoting  $\mathbf{x} = (x_1, x_2) = (\tau, \gamma = 1/\lambda)$ , we implemented the VGC-BP algorithm ( $k = 10$ ) and VGC-LN algorithms (deterministic implementations are available in this special case. For gradient updates, we use a quasi-Newton strategy implemented in Schmidt (2012)). We compared them with two baselines: (i) Gibbs sampler ( $1 \times 10^6$  samples), and (ii) MFVB. The equivalent hierarchical model is

$$y|\tau \sim \mathcal{N}(0, \tau), \quad \tau|\gamma \sim \text{InvGa}(0.5, \gamma), \quad \gamma \sim \text{Ga}(0.5, 1)$$

The inference updates for this model is detailed as follows.

#### *Gibbs Sampler*

The full conditional posterior distributions are

$$p(\tau|y, \gamma) = \text{InvGa}(1, y^2/2 + \gamma), \quad p(\gamma|\tau) = \text{Ga}(1, \tau^{-1} + 1)$$

#### *Mean-field Variational Bayes*

The ELBO under MFVB is

$$\mathcal{L}_{\text{MFVB}}[q_{\text{VB}}(\tau, \gamma)] = \mathbb{E}_{q(\tau)q(\gamma)}[\ln p(y, \tau, \gamma)] + H_1[q(\tau; \alpha_1, \beta_1)] + H_2[q(\gamma; \alpha_2, \beta_2)]$$

where

$$\begin{aligned}\mathbb{E}_{q(\tau)q(\gamma)}[\ln p(y, \tau, \gamma)] &= -0.5 \ln(2\pi) - 2 \ln \Gamma(0.5) - 2\langle \ln \tau \rangle \\ &\quad - y^2 \langle \tau^{-1} \rangle / 2 - \langle \gamma \rangle \langle \tau^{-1} \rangle - \langle \gamma \rangle \\ H_1[q(\tau; \alpha_1, \beta_1)] &= \alpha_1 + \ln \beta_1 + \ln [\Gamma(\alpha_1)] - (1 + \alpha_1)\psi(\alpha_1) \\ H_2[q(\gamma; \alpha_2, \beta_2)] &= \alpha_2 - \ln \beta_2 + \ln [\Gamma(\alpha_2)] + (1 - \alpha_2)\psi(\alpha_2)\end{aligned}$$

The variational distribution

$$\begin{aligned}q(\tau) &= \mathcal{IG}(\tau; \alpha_1, \beta_1) = \mathcal{IG}(\tau; 1, y^2/2 + \langle \gamma \rangle), \\ q(\gamma) &= \mathcal{G}(\gamma; \alpha_2, \beta_2) = \mathcal{G}(\gamma; 1, \langle \tau^{-1} \rangle + 1)\end{aligned}$$

where

$$\begin{aligned}\langle \ln \tau \rangle &= \ln \beta_1 - \psi(\alpha_1) = \ln(y^2/2 + \langle \gamma \rangle) - \psi(1), \\ \langle \tau^{-1} \rangle &= \frac{\alpha_1}{\beta_1} = \frac{1}{(y^2/2 + \langle \gamma \rangle)}, \quad \langle \gamma \rangle = \frac{\alpha_2}{\beta_2} = \frac{1}{\langle \tau^{-1} \rangle + 1}\end{aligned}$$

*Deterministic VGC-LN*

Denoting  $\mathbf{x} = (x_1, x_2) = (\tau, \gamma)$ , we construct a variational Gaussian copula proposal with (1) a bivariate Gaussian copula, and (2) fixed-form margin for both  $x_1 = \tau \in (0, \infty)$  and  $x_2 = \gamma \in (0, \infty)$ ; we employ  $f_j(x_j; \mu_j, \sigma_{jj}^2) = \mathcal{LN}(x_j; \mu_j, \sigma_{jj}^2)$ ,  $x_j = h_j(\tilde{z}_j) = \exp(\tilde{z}_j) = g(z_j) = \exp(\sigma_{jj}z_j + \mu_j)$ ,  $j = 1, 2$ . The ELBO of VGC-LN is

$$\begin{aligned}\mathcal{L}_{\text{VGC}}(\boldsymbol{\mu}, \mathbf{C}) &= c_1 - \mu_1 + \mu_2 - \frac{y^2 \exp\left(-\mu_1 + \frac{C_{11}^2}{2}\right)}{2} \\ &\quad - \ell_0 - \exp\left(\mu_2 + \frac{C_{21}^2 + C_{22}^2}{2}\right) + \ln |\mathbf{C}| \\ \ell_0 &= \exp\left((\mu_2 - \mu_1) + \frac{C_{11}^2 - 2C_{11}C_{21} + C_{21}^2 + C_{22}^2}{2}\right)\end{aligned}$$

where  $c_0 = -0.5 \ln(2\pi) - 2 \ln \Gamma(0.5)$ ,  $c_1 = c_0 + \ln(2\pi e)$ .

The gradients are

$$\frac{\partial \mathcal{L}_{\text{VGC}}(\boldsymbol{\mu}, \mathbf{C})}{\partial \mu_1} = -1 + \frac{y^2}{2} \exp\left(\frac{C_{11}^2}{2} - \mu_1\right) + \ell_0$$

$$\frac{\partial \mathcal{L}_{\text{VGC}}(\boldsymbol{\mu}, \mathbf{C})}{\partial \mu_2} = 1 - \ell_0 - \exp\left(\mu_2 + \frac{C_{21}^2 + C_{22}^2}{2}\right)$$

$$\frac{\partial \mathcal{L}_{\text{VGC}}(\boldsymbol{\mu}, \mathbf{C})}{\partial C_{11}} = -\frac{y^2}{2} C_{11} \exp\left(\frac{C_{11}^2}{2} - \mu_1\right) - (C_{11} - C_{21})\ell_0 + \frac{1}{C_{11}}$$

$$\frac{\partial \mathcal{L}_{\text{VGC}}(\boldsymbol{\mu}, \mathbf{C})}{\partial C_{21}} = (C_{11} - C_{21})\ell_0 - C_{21} \exp\left(\mu_2 + \frac{C_{21}^2 + C_{22}^2}{2}\right)$$

$$\frac{\partial \mathcal{L}_{\text{VGC}}(\boldsymbol{\mu}, \mathbf{C})}{\partial C_{22}} = -C_{22}\ell_0 - C_{22} \exp\left(\mu_2 + \frac{C_{21}^2 + C_{22}^2}{2}\right) + \frac{1}{C_{22}}$$

### *Stochastic VGC-LN*

The stochastic part of the ELBO is,

$$\ell_s(\tilde{\mathbf{z}}) = c_0 + \tilde{z}_2 - \tilde{z}_1 - \frac{y^2 \exp(-\tilde{z}_1)}{2} - \exp(\tilde{z}_2 - \tilde{z}_1) - \exp(\tilde{z}_2)$$

and

$$\nabla_{\tilde{z}_1} \ell_s(\tilde{\mathbf{z}}) = -1 + \frac{y^2 \exp(-\tilde{z}_1)}{2} + \exp(\tilde{z}_2 - \tilde{z}_1)$$

$$\nabla_{\tilde{z}_2} \ell_s(\tilde{\mathbf{z}}) = 1 - \exp(\tilde{z}_2 - \tilde{z}_1) - \exp(\tilde{z}_2)$$

### *Stochastic VGC-BP*

1.  $\ln p(y, x_1, x_2) = c_0 - 2 \ln x_1 - y^2/(2x_1) - x_2/x_1 - x_2$ , and

$$\partial \ln p(y, x_1, x_2)/\partial x_1 = -2/x_1 + y^2/(2x_1^2) + x_2/x_1^2,$$

$$\partial \ln p(y, x_1, x_2)/\partial x_2 = -1/x_1 - 1$$

2.  $\Psi(x)$  is predefined as CDF of  $\text{Exp}(0.01)$ .

From Figure 3.4, it is noted that the VGC methods with full correlation matrix (VGC-LN-full, VGC-BP-full) are able to preserve the posterior dependence and alleviate the under-estimation of the posterior variance. VGC-LN-full lead to higher

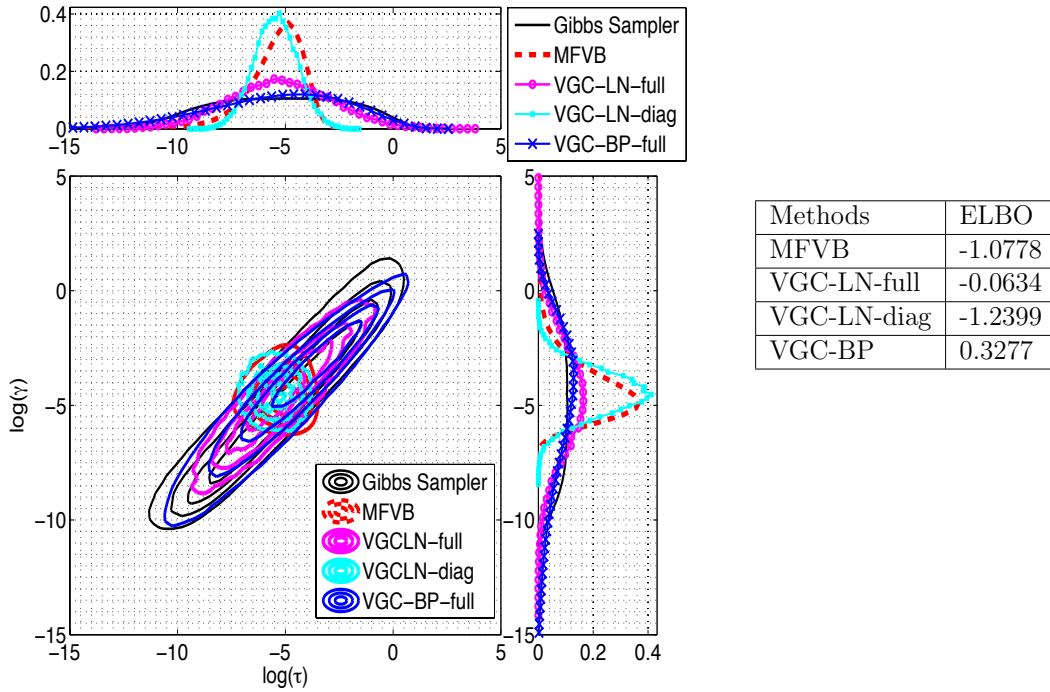


FIGURE 3.4: (Left Panel) Approximated Posteriors (Shown in Log Space for Visualization Purpose); (Right Panel) comparison of ELBO of different variational methods

ELBO than MFVB, and the gain is lost with factorized assumption  $\Upsilon = \mathbf{I}$  (VGC-LN-diag) in which case the Gaussian copula reduces to the independence copula. The restriction of parametric margins is relaxed in VGC-BP. With refinement of the mixture weights, VGC-BP leads to higher ELBO than VGC-LN. Since the Gaussian copula admits neither lower nor upper tail dependence, the posterior dependence it is able to preserve can be restrictive. It is a future research topic to explore other copula families that allow more complex posterior dependencies in variational copula inference.

### 3.6.4 Poisson Log-Linear Regression

We consider the tropical rain forest dataset (Møller and Waagepetersen, 2007), a point pattern giving the locations of 3605 trees accompanied by covariate data giving

the elevation. Resampling the data into a grid of  $50 \times 50\text{m}$  ( $u_i$  locates the  $i$ -th grid), the number of trees  $y_i$  per unit area is modeled as,  $y_i \sim \text{Poisson}(\mu_i)$ ,  $i = 1, \dots, n$ ,  $\log(\mu_i) = \beta_0 + \beta_1 u_i + \beta_2 u_i^2$ ,  $\beta_0 \sim N(0, \tau)$ ,  $\beta_1 \sim N(0, \tau)$ ,  $\beta_2 \sim N(0, \tau)$ ,  $\tau \sim \text{Ga}(1, 1)$ . We denote  $\boldsymbol{x} = (\beta_0, \beta_1, \beta_2, \tau)$ , and choosing  $\Psi^{-1}(\cdot)$  to be the CDF of  $\mathcal{N}(0, 1)$  or  $\text{Exp}(1)$  accordingly. The implementation of VGC-BP leads to highly accurate marginal and pairwise posteriors (See Figure 3.5), as compared to the MCMC sampler ( $1 \times 10^6$  runs) implemented in JAGS (<http://mcmc-jags.sourceforge.net/>) as reference solutions.

### *Model Specific Derivations in Poisson Log Linear Regression*

For  $i = 1, \dots, n$ , the hierarchical model is

$$\begin{aligned} y_i &\sim \text{Poisson}(\mu_i), \quad \log(\mu_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2, \\ \beta_0 &\sim N(0, \tau), \quad \beta_1 \sim N(0, \tau), \quad \beta_2 \sim N(0, \tau), \quad \tau \sim \text{Ga}(1, 1) \end{aligned}$$

The log likelihood and prior,

$$\begin{aligned} \ln p(\mathbf{y}, \boldsymbol{\beta}, \tau) &= \sum_{i=1}^n \ln p(y_i | \boldsymbol{\beta}) + \ln \mathcal{N}(\beta_0; 0, \tau) + \ln \mathcal{N}(\beta_1; 0, \tau) + \ln \mathcal{N}(\beta_2; 0, \tau) \\ &\quad + \ln \text{Ga}(\tau; 1, 1) \end{aligned}$$

where  $\ln p(y_i | \boldsymbol{\beta}) = y_i \ln \mu_i - \mu_i - \ln y_i!$ , and  $\mu_i = \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2)$ .

The derivatives are

$$\begin{aligned} \frac{\partial \ln p(\mathbf{y}, \boldsymbol{\beta}, \tau)}{\partial \beta_0} &= \left[ \sum_{i=1}^n (y_i - \mu_i) \right] - \tau^{-1} \beta_0 \\ \frac{\partial \ln p(\mathbf{y}, \boldsymbol{\beta}, \tau)}{\partial \beta_1} &= \left[ \sum_{i=1}^n x_i (y_i - \mu_i) \right] - \tau^{-1} \beta_1 \\ \frac{\partial \ln p(\mathbf{y}, \boldsymbol{\beta}, \tau)}{\partial \beta_2} &= \left[ \sum_{i=1}^n x_i^2 (y_i - \mu_i) \right] - \tau^{-1} \beta_2 \\ \frac{\partial \ln p(\mathbf{y}, \boldsymbol{\beta}, \tau)}{\partial \tau} &= -\frac{3}{2\tau} + \frac{\beta_0^2 + \beta_1^2 + \beta_2^2}{2\tau^2} + \frac{a_0 - 1}{\tau} - b_0 \end{aligned}$$

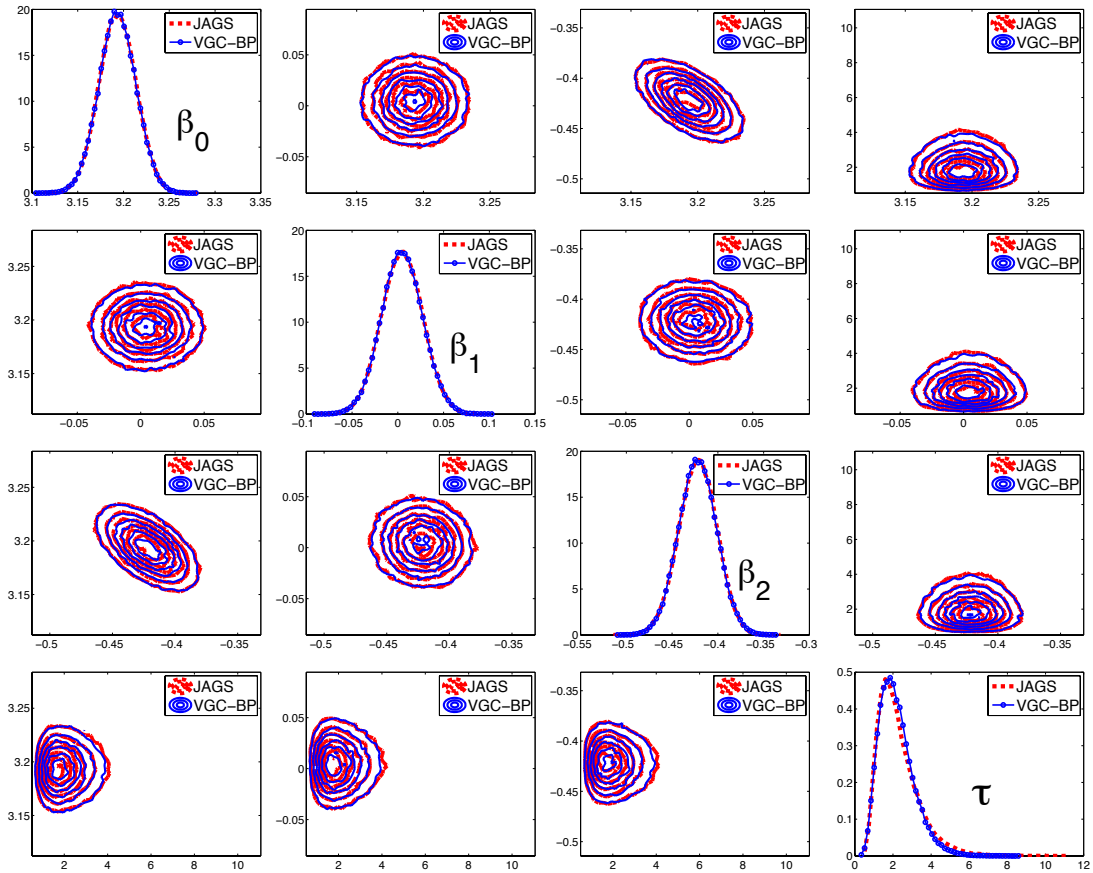


FIGURE 3.5: Univariate Margins and Pairwise Posteriors

Interestingly, for non-conjugate models with unknown exact joint posteriors, VGC still provides a Sklar's representation of the approximated posterior, including an analytical Gaussian copula, and a number of univariate margins (summarized as univariate histograms if not in closed-form). For further uses such as calculating sample quantiles, simulating samples from  $q_{\text{VGC}}(\mathbf{x})$  is independent and faster, as compared to MCMC. The obtained posterior approximation could possibly improve the efficiency of Metropolis-Hastings (MH) samplers by replacing the MCMC prerun as a reasonable proposal (Schmidl *et al.*, 2013). The proposed method is an automated approach of approximating full posteriors. It is readily applicable to a broad scope

of latent Gaussian models with non-conjugate likelihoods.

Using copulas to improve approximate Bayesian inference is a natural idea that has also been explored recently in other contexts (Li *et al.*, 2015; Ferkingstad and Rue, 2015). Independently from our work, Tran *et al.* (2015) presented a copula augmented variational method with fixed-form marginals, and utilizes regular vines to decompose the multivariate dependency structure into bivariate copulas and a nest of trees. Our method provides complementary perspectives on nonparametric treatment of univariate marginals.

### 3.7 Discussion

This chapter proposes a unified variational copula inference framework. In VGC, we have focused on Gaussian copula family for simplicity, however, other more flexible forms such as Gaussian mixture copula can be considered as well. To avoid the difficulty of specifying marginals for hidden variables, a nonparametric procedure based on Bernstein polynomials indirectly induces highly flexible univariate margins. Tran *et al.* (2015) and Kucukelbir *et al.* (2015) could potentially benefit from our flexible margins, while our approach is likely to benefit from the vine copula decomposition (Tran *et al.*, 2015) to allow richer or more complex dependencies and the automatic differentiation techniques applied in Kucukelbir *et al.* (2015).

## Learning Time-evolving Dependencies via Dynamic Rank Factor Model

In this chapter, we propose a semi-parametric and dynamic rank factor model for topic modeling, capable of (i) discovering topic prevalence over time, and (ii) learning contemporary multi-scale dependence structures, providing topic and word correlations as a byproduct. The high-dimensional and time-evolving ordinal/rank observations (such as word counts), after an arbitrary monotone transformation, are well accommodated through an underlying dynamic sparse factor model. The framework naturally admits heavy-tailed innovations, capable of inferring abrupt temporal jumps in the importance of topics. Posterior inference is performed through straightforward Gibbs sampling, based on the forward-filtering backward-sampling algorithm. Moreover, an efficient data subsampling scheme is leveraged to speed up inference on massive datasets. The modeling framework is illustrated on two real datasets: the US State of the Union Address and the JSTOR collection from *Science*.

## 4.1 Dynamic Rank Factor Model

We perform analysis of multivariate ordinal time series. In the most general sense, such ordinal variables indicate a ranking of responses in the sample space, rather than a cardinal measure (Hoff, 2009). Examples include real continuous variables, discrete ordered variables with or without numerical scales or, more specially, counts, which can be viewed as discrete variables with integer numeric scales. Our goal is twofold: (i) discover the common trends that govern variations in observations, and (ii) extract interpretable patterns from the cross-sectional dependencies.

Dependencies among multivariate non-normal variables may be induced through normally distributed latent variables. Suppose we have  $P$  ordinal-valued time series  $y_{p,t}$ ,  $p = 1, \dots, P$ ,  $t = 1, \dots, T$ . The general framework contains three components:

$$y_{p,t} \sim g(z_{p,t}), \quad z_{p,t} \sim p(\boldsymbol{\theta}_t), \quad \boldsymbol{\theta}_t \sim q(\boldsymbol{\theta}_{t-1}), \quad (4.1)$$

where  $g(\cdot)$  is the sampling distribution, or marginal likelihood for the observations, the latent variable  $z_{p,t}$  is modeled by  $p(\cdot)$  (assumed to be Gaussian) with underlying system parameters  $\boldsymbol{\theta}_t$ , and  $q(\cdot)$  is the system equation representing Markovian dynamics for the time-evolving parameter  $\boldsymbol{\theta}_t$ .

In order to gain more model flexibility and robustness against misspecification, we propose a semi-parametric Bayesian dynamic factor model for multiple ordinal time series analysis. The model is based on the *extended rank likelihood* (Hoff, 2007), allowing the transformation from the latent conditionally Gaussian dynamic model to the multivariate observations, treated non-parametrically.

Extended Rank Likelihood (ERL): There exist many approaches for dealing with ordinal data, however, they all have some restrictions. For continuous variables, the underlying normality assumption could be easily violated without a carefully chosen deterministic transformation. For discrete ordinal variables, an ordered probit model, with cut points, becomes computationally expensive if the number of categories

is large. For count variables, a multinomial model requires finite support on the integer values. Poisson and negative binomial models lack flexibility from a practical viewpoint, and often lead to non-conjugacy when employing log-normal priors.

Being aware of these issues, a natural candidate for consideration is the ERL (Hoff, 2007). With appropriate monotone transformations learned automatically from data, it offers a unified framework for handling both continuous (Pettitt, 1982) and discrete ordinal variables. The ERL depends only on the ranks of the observations (zero values in observations are further restricted to have negative latent variables),

$$z_{p,t} \in D(\mathbf{Y}) \equiv \{z_{p,t} \in \mathbb{R} : y_{p,t} < y_{p',t'} \Rightarrow z_{p,t} < z_{p',t'}, \text{ and } z_{p,t} \leq 0 \text{ if } y_{p,t} = 0\}. \quad (4.2)$$

In particular, this offers a distribution-free approach, with relaxed assumptions compared to parametric models, such as Poisson log-normal (Aitchison and Ho, 1989). It also avoids the burden of computing nuisance parameters in the ordered probit model (cut points). The ERL has been utilized in Bayesian Gaussian copula modeling, to characterize the dependence of mixed data (Hoff, 2007). In (Murray *et al.*, 2013) a low-rank decomposition of the covariance matrix is further employed and efficient posterior sampling is developed in (Kalaitzis and Silva, 2013). The proposed work herein can be viewed as a dynamic extension of that framework.

#### 4.1.1 Latent Sparse Dynamic Factor Model

In the forthcoming text,  $\mathcal{G}(\alpha, \beta)$  denotes a gamma distribution with shape parameter  $\alpha$  and rate parameter  $\beta$ ,  $\text{TN}_{(l,u)}(\mu, \sigma^2)$  denotes a univariate truncated normal distribution within the interval  $(l, u)$ , and  $\mathcal{N}_+(0, \sigma^2)$  is the half-normal distribution that only has non-negative support.

Assume  $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}_t)$ , where  $\mathbf{\Omega}_t$  is usually a high-dimensional ( $P \times P$ ) covariance matrix. To reduce the number of parameters, we assume a low rank factor model

decomposition of the covariance matrix  $\mathbf{\Omega}_t = \mathbf{\Lambda}\mathbf{V}_t\mathbf{\Lambda}^T + \mathbf{R}$  such that

$$\mathbf{z}_t = \mathbf{\Lambda}\mathbf{s}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \quad \mathbf{R} = \mathbf{I}_P. \quad (4.3)$$

Common trends (importance of topics) are captured by a low-dimensional factor score parameter  $\mathbf{s}_t$ . We assume autoregressive dynamics on  $s_{k,t} \leftarrow \text{AR}(1|(\rho_k, \delta_{k,t}))$  with heavy-tailed innovations,

$$s_{k,t} = \rho_k s_{k,t-1} + \delta_{k,t}, \quad 0 < \rho_k < 1, \quad \delta_{k,t} \sim \text{TPBN}(e, f, \nu), \quad \nu^{1/2} \sim \mathcal{C}^+(0, h), \quad (4.4)$$

where  $\delta_{k,t}$  follows the *three-parameter beta mixture of normal* TPBN( $e, f, \nu$ ) distribution (Armagan *et al.*, 2011). Parameter  $e$  controls the peak around zero,  $f$  controls the heaviness on the tails, and  $\nu$  controls the global sparsity with a half-Cauchy prior (Polson and Scott, 2012). This prior encourages smooth transitions in general, while jumps are captured by the heavy tails. The conjugate hierarchy may be equivalently represented as

$$\delta_{k,t} \sim \mathcal{N}(0, \tau_{k,t}), \quad \tau_{k,t} \sim \mathcal{G}(e, \eta_{k,t}), \quad \eta_{k,t} \sim \mathcal{G}(f, \nu) \quad \nu \sim \mathcal{G}(1/2, \zeta), \quad \zeta \sim \mathcal{G}(1/2, h^2).$$

Truncated normal priors are employed on  $\rho_k$ ,  $\rho_k \sim \text{TN}_{(0,1)}(\mu_0, \sigma_0^2)$ , and assume  $s_{0,k} \sim \mathcal{N}(0, \sigma_s^2)$ . Note that the extended rank likelihood is scale-free; therefore, we do not need to include a redundant intercept parameter in (4.3). For the same reason, we set  $\mathbf{R} = \mathbf{I}_P$ .

**Model Identifiability Issues:** Although the covariance matrix  $\mathbf{\Omega}_t$  is not identifiable (Hoff, 2009), the related correlation matrix  $\mathbf{C}_t = \Omega_{[i,j],t} / \sqrt{\Omega_{[i,i],t}\Omega_{[j,j],t}}$ , ( $i, j = 1, \dots, P$ ) may be identified, using the parameter expansion technique (Lawrence *et al.*, 2008; Murray *et al.*, 2013). Further, the rank  $K$  in the low-rank decomposition of  $\mathbf{\Omega}_t$  is also not unique. For the Purpose of brevity, we do not explore this uncertainty here, but the tools developed in the Bayesian factor analysis literature (Lopes and West, 2004; Ghosh and Dunson, 2009; Bhattacharya and Dunson, 2011) can be easily adopted.

Identifiability is a key concern for factor analysis. Conventionally, for fixed  $K$ , a full-rank, lower-triangular structure in  $\mathbf{\Lambda}$  ensures identifiability (Geweke and Zhou, 1996). Unfortunately, this assumption depends on the ordering of variables. As a solution, we add nonnegative and sparseness constraints on the factor loadings, to alleviate the inherent ambiguity, while also improving interpretability. Also, we add a Procrustes post-processing step (Christian *et al.*, 2014) on the posterior samples, to reduce this indeterminacy.

The nonnegative and (near) sparseness constraints are imposed by the following hierarchy,

$$\lambda_{p,k} \sim \mathcal{N}_+(0, l_{p,k}) \quad l_{p,k} \sim \mathcal{G}(a, u_{p,k}), \quad u_{p,k} \sim \mathcal{G}(b, \phi_k), \quad \phi_k^{1/2} \sim \mathcal{C}^+(0, d). \quad (4.5)$$

Integrating out  $l_{p,k}$  and  $u_{p,k}$ , we obtain a half-TPBN prior  $\lambda_{p,k} \sim \text{TPBN}_+(a, b, \phi_k)$ . The column-wise shrinkage parameters  $\phi_k$  enable factors to be of different sparsity levels (Gao and Engelhardt, 2012). We set hyperparameters  $a = b = e = f = 0.5$ ,  $d = P$ ,  $h = 1$ ,  $\sigma_s^2 = 1$ . For weakly informative priors, we set  $\alpha = \beta = 0.01$ ;  $\mu_0 = 0.5$ ,  $\sigma_0^2 = 10$ .

#### 4.1.2 Extension to Handle Multiple Documents

At each time point  $t$  we may have a corpus of documents  $\{\mathbf{y}_t^{n_t}\}_{n_t=1}^{N_t}$ , where  $\mathbf{y}_t^{n_t}$  is a  $P$ -dimensional observation vector, and  $N_t$  denotes the number of documents at time  $t$ . The model presented in Section 4.1.1 is readily extended to handle this situation. Specifically, at each time point  $t$ , for each document  $n_t$ , the ERL representation for word count  $p$ , denoted by  $y_{p,t}^{n_t}$ , is

$$y_{p,t}^{n_t} = g(z_{p,t}^{n_t}), \quad p = 1, \dots, P, \quad t = 1, \dots, T, \quad n_t = 1, \dots, N_t,$$

where  $\mathbf{z}_t^{n_t} \in \mathbb{R}^P$  and  $P$  is the vocabulary size. We assume a latent factor model for  $\mathbf{z}_t^{n_t}$  such that

$$\mathbf{z}_t^{n_t} = \mathbf{\Lambda} \mathbf{b}_t^{n_t} + \boldsymbol{\epsilon}_t^{n_t}, \quad \boldsymbol{\epsilon}_t^{n_t} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_P), \quad \mathbf{b}_t^{n_t} \sim \mathcal{N}(\mathbf{s}_t, \mathbf{\Gamma}), \quad \mathbf{\Gamma} = \text{diag}(\boldsymbol{\gamma}), \quad \boldsymbol{\gamma}_k^{-1} \sim \mathcal{G}(\alpha, \beta),$$

where  $\mathbf{\Lambda} \in \mathbb{R}_+^{P \times K}$  is the topic-word loading matrix, representing the  $K$  topics as columns of  $\mathbf{\Lambda}$ . The factor score vector  $\mathbf{b}_t^{n_t} \in \mathbb{R}^K$  is the topic usage for each document  $\mathbf{y}_t^{n_t}$ , corresponding to locations in a low-dimensional  $\mathbb{R}^K$  space. The other parts of the model remain unchanged. The latent trajectory  $\mathbf{s}_{1:T}$  represents the common trends for the  $K$  topics. Moreover, through the forward filtering backward sampling (FFBS) algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994), we also obtain time-evolving topic correlation matrices  $\mathbf{\Phi}_t \in \mathbb{R}^{K \times K}$  and word dependencies matrices  $\mathbf{C}_t \in \mathbb{R}^{P \times P}$ , offering a multi-scale graph representation, a useful tool for document visualization.

#### 4.1.3 Comparison with Admixture Topic Models

Many topic models are unified in the admixture framework (Inouye *et al.*, 2014),

$$\mathbb{P}_{\text{Admix}}(\mathbf{y}_n | \mathbf{w}, \mathbf{\Phi}) = \mathbb{P}_{\text{Base}} \left( \mathbf{y}_n \mid \bar{\phi}_n = \sum_{k=1}^K w_{k,n} \phi_k \right), \quad (4.6)$$

where  $\mathbf{y}_n$  is the  $P$ -dimensional observation vector of word counts in the  $n$  th document, and  $P$  denotes the vocabulary size. Traditionally,  $\mathbf{y}_n$  is generated from an admixture of base distributions,  $\mathbf{w}_n$  is the admixture weight (topic proportion for document  $n$ ), and  $\phi_k$  is the canonical parameter (word distribution for topic  $k$ ), which denotes the location of the  $k$ th topic on the  $P-1$  dimensional simplex. For example, latent Dirichlet allocation (LDA) (Blei *et al.*, 2003) assumes the base distribution to be multinomial, with  $\phi_k \sim \text{Dir}(\boldsymbol{\alpha}_0)$ ,  $\mathbf{w}_n \sim \text{Dir}(\boldsymbol{\beta}_0)$ . The correlated topic model (CTM) (Blei and Lafferty, 2006a) modifies the topic distribution, with  $\mathbf{w}_n \sim \text{Logistic Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . The dynamic topic model (DTM) (Blei and Lafferty, 2006b) analyzes document collections in a known chronological order. In order to incorporate the state space model, both the topic proportion and the word distribution are changed to logistic normal, with isotropic covariance matrices  $\mathbf{w}_t \sim \text{Logistic Normal}(\mathbf{w}_{t-1}, \sigma^2 \mathbf{I}_K)$  and  $\phi_{k,t} \sim \text{Logistic Normal}(\phi_{k,t-1}, v \mathbf{I}_P)$ ,

respectively. To overcome the drawbacks of multinomial base, spherical topic models (Reisinger *et al.*, 2010) assume the von Mises-Fisher (vMF) distribution as its base distribution, with  $\phi_k \sim \text{vMF}(\boldsymbol{\mu}, \xi)$  lying on a unit  $P-1$  dimensional sphere. Recently in (Inouye *et al.*, 2014) the base and word distribution are both replaced with Poisson Markov random fields (MRFs), which characterizes word dependencies.

We present here a semi-parametric factor model formulation,

$$P(\mathbf{y}_n | \mathbf{s}, \boldsymbol{\Lambda}) \triangleq P \left( z_n \in D(\mathbf{Y}) \mid \bar{\boldsymbol{\lambda}}_n = \sum_{k=1}^K s_{k,n} \boldsymbol{\lambda}_k \right), \quad (4.7)$$

with  $\mathbf{y}_n$  defined as above,  $\boldsymbol{\lambda}_k \in \mathbb{R}_+^P$  is a vector of nonnegative weights, indicating the  $P$  vocabulary usage in each individual topics  $k$ , and  $\mathbf{s}_n \in \mathbb{R}^K$  is the topic usage. Note that the extended rank likelihood does not depend on any assumptions about the data marginal distribution, making it appropriate for a broad class of ordinal-valued observations, *e.g.*, term frequency-inverse document frequency (tf-idf) or rankings, beyond word counts. However, the proposed model here is not an admixture model, as the topic usage is allowed to be either positive or negative.

The DRFM framework has some appealing advantages: (i) It is more natural and convenient to incorporate with sparsity, rank selection, and state-space model; (ii) it provides topic-correlations and word-dependences as a byproduct; and (iii) computationally, this model is tractable and often leads to locally conjugate posterior inference. DRFM has limitations. Since the marginal distributions are of unspecified types, objective criteria (*e.g.* perplexity) is not directly computable. This makes quantitative comparisons to other parametric baselines developed in the literature very difficult.

## 4.2 Conjugate Posterior Inference

Let  $\Theta = \{\boldsymbol{\Lambda}, \mathbf{S}, \mathbf{L}, \mathbf{U}, \phi, \boldsymbol{\omega}, \boldsymbol{\rho}, \boldsymbol{\tau}, \boldsymbol{\eta}, \nu, \zeta\}$  denote the set of parameters in basic model, and let  $\mathbf{Z}$  be the augmented data (from the ERL). We use Gibbs sampling to approx-

imate the joint posterior distribution  $p(\mathbf{Z}, \Theta | \mathbf{Z} \in R(\mathbf{Y}))$ . The algorithm alternates between sampling  $p(\mathbf{Z} | \Theta, \mathbf{Z} \in R(\mathbf{Y}))$  and  $p(\Theta | \mathbf{Z}, \mathbf{Z} \in R(\mathbf{Y}))$  (reduced to  $p(\Theta | \mathbf{Z})$ ). The derivation of the Gibbs sampler is straightforward, and for brevity here we first highlight the sampling steps for  $\mathbf{Z}$ , and the forward filtering backward sampling (FFBS) steps for the trajectory  $\mathbf{s}_{1:T}$ . Section 4.2.2 and section 4.2.3 contain full details of the inference.

- Sampling  $z_{p,t}$ :  $p(z_{p,t} | \Theta, \mathbf{Z} \in R(\mathbf{Y}), \mathbf{Z}_{-p,-t}) \sim \text{TN}_{[z_{p,t}, \bar{z}_{p,t}]}(\sum_{k=1}^K \lambda_{p,k} s_{k,t}, 1)$ , where  $\underline{z}_{p,t} = \max\{z_{p',t'} : y_{p',t'} < y_{p,t}\}$  and  $\bar{z}_{p,t} = \min\{z_{p',t'} : y_{p',t'} > y_{p,t}\}$ .

This conditional sampling scheme is widely used in (Hoff, 2007, 2009; Murray *et al.*, 2013). In (Kalaitzis and Silva, 2013) a novel Hamiltonian Monte Carlo (HMC) approach has been developed recently, for a Gaussian copula extended rank likelihood model, where ranking is only within each row of  $\mathbf{Z}$ . This method simultaneously samples a column vector of  $\mathbf{z}_i$  conditioned on other columns  $\mathbf{Z}_{-i}$ , with higher computation but better mixing.

- Sampling  $\mathbf{s}_t$ : we have the state model  $\mathbf{s}_t | \mathbf{s}_{t-1} \sim \mathcal{N}(\mathbf{A}\mathbf{s}_{t-1}, \mathbf{Q}_t)$ , and the observation model  $\mathbf{z}_t | \mathbf{s}_t \sim \mathcal{N}(\mathbf{\Lambda}\mathbf{s}_t, \mathbf{R})$ , where  $\mathbf{A} = \text{diag}(\boldsymbol{\rho})$ ,  $\mathbf{Q}_t = \text{diag}(\boldsymbol{\tau}_t)$ ,  $\mathbf{R} = \mathbf{I}_P$ , for  $t = 1, \dots, T$ . For brevity, we omit the dependencies on  $\Theta$  in notation.

1. Forward Filtering: beginning at  $t = 0$  with  $\mathbf{s}_0 \sim \mathcal{N}(\mathbf{0}, \sigma_s^2 \mathbf{I}_K)$ , for all  $t = 1, \dots, T$ , we find the on-line posteriors at  $t$ ,  $p(\mathbf{s}_t | \mathbf{z}_{1:t}) = \mathcal{N}(\mathbf{m}_t, \mathbf{V}_t)$ , where  $\mathbf{m}_t = \mathbf{V}_t \{\mathbf{\Lambda}^T \mathbf{R}^{-1} \mathbf{z}_t + \mathbf{H}_t^{-1} \mathbf{A} \mathbf{m}_{t-1}\}$ ,  $\mathbf{V}_t = [\mathbf{H}_t^{-1} + \mathbf{\Lambda}^T \mathbf{R}^{-1} \mathbf{\Lambda}]^{-1}$ , and  $\mathbf{H}_t = \mathbf{Q}_t + \mathbf{A} \mathbf{V}_{t-1} \mathbf{A}^T$ .
2. Backward Sampling: starting from  $\mathcal{N}(\tilde{\mathbf{m}}_t, \tilde{\mathbf{V}}_t)$ , the backward smoothing density, *i.e.*, the conditional distribution of  $\mathbf{s}_{t-1}$  given  $\mathbf{s}_t$ , is  $p(\mathbf{s}_{t-1} | \mathbf{s}_t, \mathbf{z}_{1:(t-1)}) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_{t-1}, \tilde{\boldsymbol{\Sigma}}_{t-1})$ , where  $\tilde{\boldsymbol{\mu}}_{t-1} = \tilde{\boldsymbol{\Sigma}}_{t-1} \{\mathbf{A}^T \mathbf{Q}_t^{-1} \mathbf{s}_t + \mathbf{V}_{t-1}^{-1} \mathbf{m}_{t-1}\}$ ,  $\tilde{\boldsymbol{\Sigma}}_{t-1} = (\mathbf{V}_{t-1}^{-1} + \mathbf{A}^T \mathbf{Q}_t^{-1} \mathbf{A})^{-1}$ .

There exist different variants of FFBS schemes (see (Reis *et al.*, 2006) for a detailed comparison); the method we choose here enjoys fast decay in autocorrelation and reduced computation time.

#### 4.2.1 Time-evolving Topic and Word Dependencies

We also have the backward recursion density at  $t - 1$ ,  $p(\mathbf{s}_{t-1}|\mathbf{z}_{1:T}) = \mathcal{N}(\tilde{\mathbf{m}}_{t-1}, \tilde{\mathbf{V}}_{t-1})$ , where

$$\tilde{\mathbf{m}}_{t-1} = \tilde{\Sigma}_{t-1}(\mathbf{A}^T \mathbf{Q}_t^{-1} \tilde{\mathbf{m}}_t + \mathbf{V}_{t-1}^{-1} \mathbf{m}_{t-1})$$

and

$$\tilde{\mathbf{V}}_{t-1} = \tilde{\Sigma}_{t-1} + \tilde{\Sigma}_{t-1} \mathbf{A}^T \mathbf{Q}_t^{-1} \tilde{\mathbf{V}}_t \mathbf{Q}_t^{-1} \mathbf{A} \tilde{\Sigma}_{t-1}.$$

We perform inference on the  $K \times K$  time-evolving topic dependencies in  $\mathbf{s}_{1:T}$ , using the posterior covariances  $\{\tilde{\mathbf{V}}_{1:T}\}$  (with topic correlation matrices  $\Phi_{1:T}$ ,  $\Phi_{[r,s],t} = V_{[r,s],t}/\sqrt{V_{[r,r],t}V_{[s,s],t}}$ ,  $r, s = 1, \dots, K$ ), and further obtain the  $P \times P$  time-evolving word dependencies capsuled in  $\{\Omega_{1:T}\}$  with  $\Omega_t = \Lambda \tilde{\mathbf{V}}_t \Lambda^T + \mathbf{I}_P$ . Essentially, this can be viewed as a dynamic Gaussian copula model,  $y_{p,t} = g(\tilde{z}_{p,t})$ ,  $\tilde{\mathbf{z}}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_t)$ , where  $g(\cdot)$  is a non-decreasing function of a univariate marginal likelihood and  $\mathbf{C}_t$  ( $t = 1, \dots, T$ ) is the correlation matrix capturing the multivariate dependence. We obtain a posterior distribution for  $\mathbf{C}_{1:T}$  as a byproduct, without having to estimate the nuisance parameters in marginal likelihoods  $g(\cdot)$ . This decoupling strategy resembles the idea of copula models.

#### 4.2.2 Gibbs Sampling for the Basic Model

Given the basic model,

$$y_{p,t} = g(z_{p,t})$$

$$\mathbf{z}_t = \Lambda \mathbf{s}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \quad \mathbf{R} = \mathbf{I}_P \quad (4.8)$$

$$\lambda_{p,k} \sim \text{TPBN}_+(a, b, \phi_k), \quad \phi_k^{1/2} \sim \mathcal{C}^+(0, d), \quad s_{k,t} = \rho_k s_{k,t-1} + \delta_{k,t}, \quad 0 < \rho_k < 1$$

$$\delta_{k,t} \sim \text{TPBN}(e, f, \nu), \quad \nu^{1/2} \sim \mathcal{C}^+(0, h), \quad \rho_k \sim \text{TN}_{(0,1)}(\mu_0, \sigma_0^2), \quad s_{0,k} \sim \mathcal{N}(0, \sigma_s^2),$$

and its conjugate hierarchy

$$\lambda_{p,k} \sim \mathcal{N}_+(0, l_{p,k}) \quad l_{p,k} \sim \mathcal{G}(a, u_{p,k}), \quad u_{p,k} \sim \mathcal{G}(b, \phi_k), \quad \phi_k \sim \mathcal{G}\left(\frac{1}{2}, \omega_k\right), \quad \omega_k \sim \mathcal{G}\left(\frac{1}{2}, d^2\right)$$

$$\delta_{k,t} \sim \mathcal{N}(0, \tau_{k,t}), \quad \tau_{k,t} \sim \mathcal{G}(e, \eta_{k,t}), \quad \eta_{k,t} \sim \mathcal{G}(f, \nu), \quad \nu \sim \mathcal{G}\left(\frac{1}{2}, \zeta\right), \quad \zeta \sim \mathcal{G}\left(\frac{1}{2}, h^2\right)$$

### Gibbs Sampling Updates

Denote  $\Theta = \{\mathbf{\Lambda}, \mathbf{S}, \mathbf{L}, \mathbf{U}, \phi, \omega, \rho, \tau, \eta, \nu, \zeta\}$ , we use Gibbs sampling to approximate the joint posterior distribution of  $(\mathbf{Z}, \Theta)$ ,

1. Given  $\Theta$ , find  $p(z_{p,t} | \Theta, \mathbf{Z} \in R(\mathbf{Y}), \mathbf{Z}_{-p,-t})$ , for  $p = 1, \dots, P, t = 1, \dots, T$ .
2. Given  $\mathbf{Z}$ , find  $p(\Theta | \mathbf{Z}, \mathbf{Z} \in R(\mathbf{Y}))$  reduce to  $p(\Theta | \mathbf{Z})$

Treat  $\mathbf{Z}$  as augmented data, the full likelihood for  $(\mathbf{Z}, \Theta)$  is

$$p(\mathbf{Z}, \Theta) = \left( \prod_{t=1}^T \mathcal{N}(z_t; \mathbf{\Lambda} s_t, \mathbf{R}) \right) \times \left( \prod_{k=1}^K \mathcal{G}(\phi_k; 1/2, \omega_k) \mathcal{G}(\omega_k; 1/2, d^2) \right)$$

$$\times \prod_{k=1}^K \left[ \mathcal{N}(s_{0,k}; 0, \sigma_s^2) \left( \prod_{t=1}^T \mathcal{N}(s_{k,t}; \rho_k s_{k,t-1}, \tau_{k,t}) \mathcal{G}(\tau_{k,t}; e, \eta_{k,t}) \mathcal{G}(\eta_{k,t}; f, \nu) \right) \right]$$

$$\times \left( \prod_{p=1}^P \prod_{k=1}^K \mathcal{N}_+(\lambda_{p,k}; 0, l_{p,k}) \mathcal{G}(l_{p,k}; a, u_{p,k}) \mathcal{G}(u_{p,k}; b, \phi_k) \right)$$

$$\times \prod_{k=1}^K \text{TN}_{(0,1)}(\rho_k; \mu_0, \sigma_0^2) \times \mathcal{G}(\nu; 1/2, \zeta) \times \mathcal{G}(\zeta; 1/2, h^2)$$

- Sampling  $z_{p,t}$

$$p(z_{p,t} | \Theta, \mathbf{Z} \in R(\mathbf{Y}), \mathbf{Z}_{-p,-t}) \sim \text{TN}_{[z_{p,t}, \overline{z_{p,t}}]} \left( \sum_{k=1}^K \lambda_{p,k} s_{k,t}, 1 \right)$$

where  $\underline{z_{p,t}} = \max\{z_{p',t'} : y_{p',t'} < y_{p,t}\}$  and  $\overline{z_{p,t}} = \min\{z_{p',t'} : y_{p',t'} > y_{p,t}\}$

- Sampling  $\lambda_{p,k}$

$$\begin{aligned}
p(\lambda_{p,k}|-) &\propto \left( \prod_{t=1}^T \mathcal{N}(z_{p,t}; \lambda_{p,k} s_{k,t} + \sum_{j \neq k} s_{j,t} \lambda_{p,j}, 1) \right) \mathcal{N}_+(\lambda_{p,k}; 0, l_{p,k}) \\
&= \mathcal{N}_+ \left( \lambda_{p,k}; v_{\lambda_{p,k}} \sum_{t=1}^T \left[ s_{k,t} z_{p,t} - s_{k,t} \sum_{j \neq k} s_{j,t} \lambda_{p,j} \right], v_{\lambda_{p,k}} \right) \\
v_{\lambda_{p,k}} &= (l_{p,k}^{-1} + \sum_{t=1}^T s_{k,t}^2)^{-1}
\end{aligned}$$

- Sampling  $l_{p,k}, u_{p,k}$

$$p(l_{p,k}|-) = \text{GIG}(a - 1/2, 2u_{p,k}, (\lambda_{p,k})^2), \quad p(u_{p,k}|-) = \mathcal{G}(a + b, u_{p,k} + \phi_k)$$

The Generalized Inverse Gaussian (GIG) distribution can be expressed as

$$\text{GIG}(x; p, a, b) = \frac{(a/b)^{\frac{p}{2}}}{2K_p(\sqrt{ab})} x^{p-1} \exp\left(-\frac{1}{2}(ax + \frac{b}{x})\right) \quad (x > 0)$$

where  $K_p(\theta)$  is the modified Bessel function of the second kind

$$K_p(\theta) = \int_0^\infty \frac{1}{2} \theta^{-p} t^{p-1} \exp\left(-\frac{1}{2}(t + \frac{\theta^2}{t})\right) dt$$

with property  $K_{-\frac{1}{2}}(\theta) = \frac{1}{2}\sqrt{2\pi}\theta^{-\frac{1}{2}} \exp(-\theta)$  and  $K_{p+1}(\theta) = K_{p-1}(\theta) + \frac{2p}{\theta}K_p(\theta)$

- Sampling  $\phi_k, \omega_k$

$$p(\phi_k|-) = \mathcal{G}(1/2 + bP, \omega_k + \sum_{p=1}^P u_{p,k}), \quad p(\omega_k|-) = \mathcal{G}(1, \phi_k + d^2)$$

- Sampling  $\tau_{k,t}, \eta_{k,t}$

$$p(\tau_{k,t}|-) = \text{GIG}(e - \frac{1}{2}, 2\eta_{k,t}, (s_{k,t} - \rho_k s_{k,t-1})^2), \quad p(\eta_{k,t}|-) = \mathcal{G}(e + f, \tau_{k,t} + \nu)$$

- Sampling  $\nu, \zeta$

$$p(\nu|-) = \mathcal{G}(\frac{1}{2} + fTK, \zeta + \sum_{k=1}^K \sum_{t=1}^T \eta_{k,t}), \quad p(\zeta|-) = \mathcal{G}(1, \nu + h^2)$$

- Sampling  $\rho_k$

$$p(\rho_k | -) = \text{TN}_{(0,1)} \left( \sigma_{\rho_k}^2 (\sigma_0^{-2} \mu_0 + \sum_{t=1}^T \tau_{k,t}^{-1} s_{k,t-1} s_{k,t}), \sigma_{\rho_k}^2 \right)$$

where  $\sigma_{\rho_k}^2 = 1 / (\sigma_0^{-2} + \sum_{t=1}^T \tau_{k,t}^{-1} s_{k,t-1}^2)$ .

- Sampling  $s_{k,t}$ : we have the state model and the observation model. Again, we omit the dependencies on  $\Theta$  in notation for brevity.

$$\mathbf{s}_t | \mathbf{s}_{t-1} \sim \mathcal{N}(\mathbf{A} \mathbf{s}_{t-1}, \mathbf{Q}_t), \quad \mathbf{A} = \text{diag}(\boldsymbol{\rho}), \quad \mathbf{Q}_t = \text{diag}(\boldsymbol{\tau}_t), \quad (4.9)$$

$$\mathbf{z}_t | \mathbf{s}_t \sim \mathcal{N}(\boldsymbol{\Lambda} \mathbf{s}_t, \mathbf{R}), \quad \mathbf{R} = \mathbf{I}_P \quad (4.10)$$

for  $t = 1, \dots, T$

1. **Forward Filtering**: beginning at  $t = 0$  with  $\mathbf{s}_0 \sim \mathcal{N}(\mathbf{0}, \sigma_s^2 \mathbf{I}_K)$ , we have, for all  $t = 1, \dots, T$ , the on-line posteriors  $p(\mathbf{s}_t | \mathbf{z}_{1:t}) = \mathcal{N}(\mathbf{m}_t, \mathbf{V}_t)$ . Start from

$$p(\mathbf{s}_{t-1} | \mathbf{z}_{1:(t-1)}) = \mathcal{N}(\mathbf{m}_{t-1}, \mathbf{V}_{t-1}) \quad (4.11)$$

Combine (4.9) with (4.11), integrate out  $\mathbf{s}_{t-1}$ , we have the predictive density at  $t$ ,

$$p(\mathbf{s}_t | \mathbf{z}_{1:(t-1)}) = \mathcal{N}(\mathbf{A} \mathbf{m}_{t-1}, \mathbf{Q}_t + \mathbf{A} \mathbf{V}_{t-1} \mathbf{A}^T) \quad (4.12)$$

Further combine (4.10) with (4.12), we have the on-line posteriors at  $t$ ,  $p(\mathbf{s}_t | \mathbf{z}_{1:t}) = \mathcal{N}(\mathbf{m}_t, \mathbf{V}_t)$ , where  $\mathbf{m}_t = \mathbf{V}_t \{ \boldsymbol{\Lambda}^T \mathbf{R}^{-1} \mathbf{z}_t + \mathbf{H}_t^{-1} \mathbf{A} \mathbf{m}_{t-1} \}$ ,  $\mathbf{V}_t = [\mathbf{H}_t^{-1} + \boldsymbol{\Lambda}^T \mathbf{R}^{-1} \boldsymbol{\Lambda}]^{-1}$ , and  $\mathbf{H}_t = \mathbf{Q}_t + \mathbf{A} \mathbf{V}_{t-1} \mathbf{A}^T$ .

2. **Backward Sampling**: define the backward smoothing density

$$p(\mathbf{s}_t | \mathbf{z}_{1:T}) = \mathcal{N}(\tilde{\mathbf{m}}_t, \tilde{\mathbf{V}}_t) \quad (4.13)$$

At  $t = T$ , we have the initialization condition  $p(\mathbf{s}_T | \mathbf{z}_{1:T}) = \mathcal{N}(\tilde{\mathbf{m}}_T, \tilde{\mathbf{V}}_T) = \mathcal{N}(\mathbf{m}_T, \mathbf{V}_T)$ . Combine (4.9) with (4.11), we have the conditional distribution of  $\mathbf{s}_{t-1}$  given  $\mathbf{s}_t$ ,

$$p(\mathbf{s}_{t-1} | \mathbf{s}_t, \mathbf{z}_{1:(t-1)}) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_{t-1}, \tilde{\boldsymbol{\Sigma}}_{t-1})$$

where  $\tilde{\boldsymbol{\mu}}_{t-1} = \tilde{\boldsymbol{\Sigma}}_{t-1} \{ \mathbf{A}^T \mathbf{Q}_t^{-1} \mathbf{s}_t + \mathbf{V}_{t-1}^{-1} \mathbf{m}_{t-1} \}$ ,  $\tilde{\boldsymbol{\Sigma}}_{t-1} = (\mathbf{V}_{t-1}^{-1} + \mathbf{A}^T \mathbf{Q}_t^{-1} \mathbf{A})^{-1}$ .

3. **Backward Recursion:** For each  $t = T - 1, T - 2, \dots, 0$ , start from (4.13), we are able to find  $p(\mathbf{s}_{t-1}|\mathbf{z}_{1:T}) = \mathcal{N}(\tilde{\mathbf{m}}_{t-1}, \tilde{\mathbf{V}}_{t-1})$  via backward recursion.

According to the Markov property,

$$p(\mathbf{s}_{t-1}|\mathbf{s}_{t:T}, \mathbf{z}_{1:T}) \equiv p(\mathbf{s}_{t-1}|\mathbf{s}_t, \mathbf{z}_{1:T}) \equiv p(\mathbf{s}_{t-1}|\mathbf{s}_t, \mathbf{z}_{1:(t-1)}) \quad (4.14)$$

using the second equality in (4.14), we obtain

$$p(\mathbf{s}_{t-1}|\mathbf{s}_t, \mathbf{z}_{1:T}) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_{t-1}, \tilde{\boldsymbol{\Sigma}}_{t-1}) \quad (4.15)$$

Combine (4.13) with (4.15), integrated out  $\mathbf{s}_t$ , we have the backward smoothing density at  $t - 1$ ,

$$\begin{aligned} p(\mathbf{s}_{t-1}|\mathbf{z}_{1:T}) &= \mathcal{N}(\tilde{\mathbf{m}}_{t-1}, \tilde{\mathbf{V}}_{t-1}) \\ \tilde{\mathbf{m}}_{t-1} &= \tilde{\boldsymbol{\Sigma}}_{t-1}(\mathbf{A}^T \mathbf{Q}_t^{-1} \tilde{\mathbf{m}}_t + \mathbf{V}_{t-1}^{-1} \mathbf{m}_{t-1}) \\ \tilde{\mathbf{V}}_{t-1} &= \tilde{\boldsymbol{\Sigma}}_{t-1} + \tilde{\boldsymbol{\Sigma}}_{t-1} \mathbf{A}^T \mathbf{Q}_t^{-1} \tilde{\mathbf{V}}_t \mathbf{Q}_t^{-1} \mathbf{A} \tilde{\boldsymbol{\Sigma}}_{t-1} \end{aligned} \quad (4.16)$$

#### 4.2.3 Gibbs Sampling for the Extended Model

At each time point  $t$ , for each document  $n_t$ , the likelihood is

$$y_{p,t}^{n_t} = g(z_{p,t}^{n_t})$$

To consider  $N_t$  documents per time point, add additional layer,

$$\begin{aligned} \mathbf{z}_t^{n_t} &= \mathbf{A} \mathbf{b}_t^{n_t} + \boldsymbol{\epsilon}_t^{n_t}, \quad \boldsymbol{\epsilon}_t^{n_t} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \quad \mathbf{R} = \mathbf{I}_P, \quad n_t = 1, \dots, N_t \\ \mathbf{b}_t^{n_t} &\sim \mathcal{N}(\mathbf{s}_t, \boldsymbol{\Gamma}), \quad \boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma}), \quad \gamma_k^{-1} \sim \mathcal{G}(\alpha, \beta), \quad k = 1, \dots, K \end{aligned}$$

#### Gibbs Sampler

Denote  $\boldsymbol{\Theta} = \{\mathbf{A}, \mathbf{S}, \mathbf{B}_{1:T}, \mathbf{L}, \mathbf{U}, \boldsymbol{\Gamma}, \boldsymbol{\phi}, \boldsymbol{\omega}, \boldsymbol{\rho}, \boldsymbol{\tau}, \boldsymbol{\eta}, \nu, \zeta\}$ , we use Gibbs sampling to approximate the joint posterior distribution of  $(\mathbf{Z}, \boldsymbol{\Theta})$ ,

1. Given  $\boldsymbol{\Theta}$ , find  $p(z_{p,t}^{n_t}|\boldsymbol{\Theta}, \mathbf{Z} \in R(\mathbf{Y}), \mathbf{Z}_{-p,-t,-n_t})$ , for  $p = 1, \dots, P$ ,  $t = 1, \dots, T$ ,  $n_t = 1, \dots, N_t$

2. Given  $\mathbf{Z}$ , find  $p(\Theta|\mathbf{Z}, \mathbf{Z} \in R(\mathbf{Y}))$  reduce to  $p(\Theta|\mathbf{Z})$

Treat  $\mathbf{Z}$  as augmented data, the full likelihood for  $(\mathbf{Z}, \Theta)$  is

$$\begin{aligned}
p(\mathbf{Z}, \Theta) &= \left( \prod_{t=1}^T \prod_{n_t=1}^{N_t} \mathcal{N}(\mathbf{z}_t^{n_t}; \Lambda \mathbf{b}_t^{n_t}, \mathbf{R}) \right) \times \left( \prod_{k=1}^K \mathcal{G}(\phi_k; 1/2, \omega_k) \mathcal{G}(\omega_k; 1/2, d^2) \right) \\
&\times \left( \prod_{t=1}^T \prod_{n_t=1}^{N_t} \mathcal{N}(\mathbf{b}_t^{n_t}, \mathbf{s}_t, \Gamma) \right) \times \prod_{k=1}^K \mathcal{G}(\gamma_k^{-1}; \alpha, \beta) \\
&\times \prod_{k=1}^K \left[ \mathcal{N}(s_{0,k}; 0, \sigma_s^2) \left( \prod_{t=1}^T \mathcal{N}(s_{k,t}; \rho_k s_{k,t-1}, \tau_{k,t}) \mathcal{G}(\tau_{k,t}; e, \eta_{k,t}) \mathcal{G}(\eta_{k,t}; f, \nu) \right) \right] \\
&\times \left( \prod_{p=1}^P \prod_{k=1}^K \mathcal{N}_+(\lambda_{p,k}; 0, l_{p,k}) \mathcal{G}(l_{p,k}; a, u_{p,k}) \mathcal{G}(u_{p,k}; b, \phi_k) \right) \\
&\times \prod_{k=1}^K \text{TN}_{(0,1)}(\rho_k; \mu_0, \sigma_0^2) \times \mathcal{G}(\nu; 1/2, \zeta) \times \mathcal{G}(\zeta; 1/2, h^2)
\end{aligned}$$

- Sampling  $z_{p,t}^{n_t}$

$$p(z_{p,t}^{n_t} | \Theta, \mathbf{Z} \in R(\mathbf{Y}), \mathbf{Z}_{-p,-t,-n_t}) \sim \text{TN}_{[\underline{z}_{p,t}^{n_t}, \overline{z}_{p,t}^{n_t}]} \left( \sum_{k=1}^K \lambda_{p,k} b_{k,t}^{n_t}, 1 \right)$$

where  $\underline{z}_{p,t}^{n_t} = \max\{z_{p',t'}^{n_t} : y_{p',t'}^{n_t} < y_{p,t}^{n_t}\}$  and  $\overline{z}_{p,t}^{n_t} = \min\{z_{p',t'}^{n_t} : y_{p',t'}^{n_t} > y_{p,t}^{n_t}\}$

- Sampling  $\lambda_{p,k}$

$$\begin{aligned}
p(\lambda_{p,k} | -) &\propto \left( \prod_{t=1}^T \prod_{n_t=1}^{N_t} \mathcal{N}(z_{p,t}^{n_t}; \lambda_{p,k} b_{k,t}^{n_t} + \sum_{j \neq k} b_{j,t}^{n_t} \lambda_{p,j}, 1) \right) \mathcal{N}_+(\lambda_{p,k}; 0, l_{p,k}) \\
&= \mathcal{N}_+ \left( \lambda_{p,k}; v_{\lambda_{p,k}} \sum_{t=1}^T \sum_{n_t=1}^{N_t} b_{k,t}^{n_t} [z_{p,t}^{n_t} - \lambda_p \mathbf{b}_t^{n_t} + b_{k,t}^{n_t} \lambda_{p,k}], v_{\lambda_{p,k}} \right) \\
v_{\lambda_{p,k}} &= (l_{p,k}^{-1} + \sum_{t=1}^T \sum_{n_t=1}^{N_t} (b_{k,t}^{n_t})^2)^{-1}
\end{aligned}$$

- Sampling  $\mathbf{b}_t^{n_t}$

$$p(\mathbf{b}_t^{n_t} | -) = \mathcal{N}(\Sigma_{\mathbf{b}_t^{n_t}} (\Lambda^T \mathbf{R}^{-1} \mathbf{z}_t^{n_t} + \Gamma^{-1} \mathbf{s}_t), \Sigma_{\mathbf{b}_t^{n_t}}), \quad \Sigma_{\mathbf{b}_t^{n_t}} = (\Gamma^{-1} + \Lambda^T \mathbf{R}^{-1} \Lambda)^{-1}$$

- Sampling  $\gamma_k^{-1}$

$$p(\gamma_k^{-1}|-) \sim \mathcal{G}\left(\alpha + \frac{1}{2} \sum_{t=1}^T N_t, \beta + \frac{1}{2} \sum_{t=1}^T \sum_{n_t=1}^{N_t} (b_{k,t}^{n_t} - s_{k,t})^2\right)$$

- Sampling  $l_{p,k}, u_{p,k}$

$$p(l_{p,k}|-) = \text{GIG}(a - 1/2, 2u_{p,k}, (\lambda_{p,k})^2), \quad p(u_{p,k}|-) = \mathcal{G}(a + b, l_{p,k} + \phi_k)$$

- Sampling  $\phi_k, \omega_k$

$$p(\phi_k|-) = \mathcal{G}(1/2 + bP, \omega_k + \sum_{p=1}^P u_{p,k}), \quad p(\omega_k|-) = \mathcal{G}(1, \phi_k + d^2)$$

- Sampling  $\tau_{k,t}, \eta_{k,t}$

$$p(\tau_{k,t}|-) = \text{GIG}(e - \frac{1}{2}, 2\eta_{k,t}, (s_{k,t} - \rho_k s_{k,t-1})^2), \quad p(\eta_{k,t}|-) = \mathcal{G}(e + f, \tau_{k,t} + \nu)$$

- Sampling  $\nu, \zeta$

$$p(\nu|-) = \mathcal{G}\left(\frac{1}{2} + fTK, \zeta + \sum_{k=1}^K \sum_{t=1}^T \eta_{k,t}\right), \quad p(\zeta|-) = \mathcal{G}(1, \nu + h^2)$$

- Sampling  $\rho_k$

$$p(\rho_k|-) = \text{TN}_{(0,1)}\left(\sigma_{\rho_k}^2 (\sigma_0^{-2} \mu_0 + \sum_{t=1}^T \tau_{k,t}^{-1} s_{k,t-1} s_{k,t}), \sigma_{\rho_k}^2\right)$$

where  $\sigma_{\rho_k}^2 = 1/(\sigma_0^{-2} + \sum_{t=1}^T \tau_{k,t}^{-1} s_{k,t-1}^2)$ .

We have the state model and the observation model. For brevity, we omit the dependencies on  $\Theta$  in notation.

$$\mathbf{s}_t | \mathbf{s}_{t-1} \sim \mathcal{N}(\mathbf{A} \mathbf{s}_{t-1}, \mathbf{Q}_t), \quad \mathbf{A} = \text{diag}(\rho), \quad \mathbf{Q}_t = \text{diag}(\tau_t), \quad (4.17)$$

$$\mathbf{b}_t^{n_t} | \mathbf{s}_t \sim \mathcal{N}(\mathbf{s}_t, \mathbf{\Gamma}), \quad \mathbf{\Gamma} = \text{diag}(\gamma_k) \quad (4.18)$$

for  $n_t = 1, \dots, N_t, t = 1, \dots, T$

1. **Forward Filtering:** beginning at  $t = 0$  with  $\mathbf{s}_0 \sim \mathcal{N}(\mathbf{0}, \sigma_s^2 \mathbf{I}_K)$ , we have, for all  $t = 1, \dots, T$ , the on-line posteriors  $p(\mathbf{s}_t | \mathbf{B}_{1:t}) = \mathcal{N}(\mathbf{m}_t, \mathbf{V}_t)$ . Start from

$$p(\mathbf{s}_{t-1} | \mathbf{B}_{1:(t-1)}) = \mathcal{N}(\mathbf{m}_{t-1}, \mathbf{V}_{t-1}) \quad (4.19)$$

Combine (4.17) with (4.19), integrate out  $\mathbf{s}_{t-1}$ , we have the predictive density at  $t$ ,

$$p(\mathbf{s}_t | \mathbf{B}_{1:(t-1)}) = \mathcal{N}(\mathbf{A}\mathbf{m}_{t-1}, \mathbf{Q}_t + \mathbf{A}\mathbf{V}_{t-1}\mathbf{A}^T) \quad (4.20)$$

Further combine (4.18) with (4.20), we have the on-line posteriors at  $t$ ,

$$\begin{aligned} p(\mathbf{s}_t | \mathbf{B}_{1:t}) &= \mathcal{N}(\mathbf{m}_t, \mathbf{V}_t) \\ \mathbf{m}_t &= \mathbf{V}_t \{ N_t \mathbf{\Gamma}^{-1} \bar{\mathbf{b}}_t + (\mathbf{Q}_t + \mathbf{A}\mathbf{V}_{t-1}\mathbf{A}^T)^{-1} \mathbf{A}\mathbf{m}_{t-1} \}, \\ \mathbf{V}_t &= [(\mathbf{Q}_t + \mathbf{A}\mathbf{V}_{t-1}\mathbf{A}^T)^{-1} + N_t \mathbf{\Gamma}^{-1}]^{-1}, \quad \bar{\mathbf{b}}_t = \frac{1}{N_t} \sum_{n_t=1}^{N_t} \mathbf{b}_t^{n_t} \end{aligned} \quad (4.21)$$

Define  $\tilde{\mathbf{\Omega}}_t = (\mathbf{Q}_t + \mathbf{A}\mathbf{V}_{t-1}\mathbf{A}^T)$ , according to the Woodbury lemma,

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$$

we have

$$(\tilde{\mathbf{\Omega}}_t^{-1} + N_t \mathbf{\Gamma}^{-1})^{-1} = \tilde{\mathbf{\Omega}}_t - \tilde{\mathbf{\Omega}}_t (N_t^{-1} \mathbf{\Gamma} + \tilde{\mathbf{\Omega}}_t)^{-1} \tilde{\mathbf{\Omega}}_t$$

2. **Backward Sampling:** define the backward smoothing density

$$p(\mathbf{s}_t | \mathbf{B}_{1:T}) = \mathcal{N}(\tilde{\mathbf{m}}_t, \tilde{\mathbf{V}}_t) \quad (4.22)$$

At  $t = T$ , we have the initialization condition  $p(\mathbf{s}_T | \mathbf{B}_{1:T}) = \mathcal{N}(\tilde{\mathbf{m}}_T, \tilde{\mathbf{V}}_T) = \mathcal{N}(\mathbf{m}_T, \mathbf{V}_T)$ . Combine (4.17) with (4.19), we have the conditional distribution of  $\mathbf{s}_{t-1}$  given  $\mathbf{s}_t$ ,

$$\begin{aligned} p(\mathbf{s}_{t-1} | \mathbf{s}_t, \mathbf{B}_{1:(t-1)}) &= \mathcal{N}(\tilde{\boldsymbol{\mu}}_{t-1}, \tilde{\boldsymbol{\Sigma}}_{t-1}) \\ \tilde{\boldsymbol{\mu}}_{t-1} &= \tilde{\boldsymbol{\Sigma}}_{t-1} \{ \mathbf{A}^T \mathbf{Q}_t^{-1} \mathbf{s}_t + \mathbf{V}_{t-1}^{-1} \mathbf{m}_{t-1} \} \\ \tilde{\boldsymbol{\Sigma}}_{t-1} &= (\mathbf{V}_{t-1}^{-1} + \mathbf{A}^T \mathbf{Q}_t^{-1} \mathbf{A})^{-1} \end{aligned} \quad (4.23)$$

Similarly, apply Woodbury matrix inversion lemma we have

$$\tilde{\Sigma}_{t-1} = \mathbf{V}_{t-1} - \mathbf{V}_{t-1} \mathbf{A}^T \tilde{\Omega}_t^{-1} \mathbf{A} \mathbf{V}_{t-1}$$

- Sampling  $\tilde{\mathbf{V}}_{0:(T-1)}$

Integrated out  $\mathbf{B}_t$ , we have the observation model

$$\mathbf{z}_t^{n_t} | \mathbf{s}_t \sim \mathcal{N}(\mathbf{\Lambda} \mathbf{s}_t, \tilde{\mathbf{R}}), \quad \tilde{\mathbf{R}} = \mathbf{I}_P + \mathbf{\Lambda} \mathbf{\Gamma} \mathbf{\Lambda}^T, \quad \tilde{\mathbf{R}}^{-1} = \mathbf{I}_P - \mathbf{\Lambda} (\mathbf{\Gamma}^{-1} + \mathbf{\Lambda}^T \mathbf{\Lambda})^{-1} \mathbf{\Lambda}^T \quad (4.24)$$

for  $n_t = 1, \dots, N_t$ ,  $t = 1, \dots, T$ . We have the on-line posteriors at  $t$ ,

$$\begin{aligned} p(\mathbf{s}_t | \mathbf{Z}_{1:t}) &= \mathcal{N}(\mathbf{m}_t, \mathbf{V}_t) \\ \mathbf{m}_t &= \mathbf{V}_t \{ N_t \mathbf{\Lambda}^T \tilde{\mathbf{R}}^{-1} \bar{\mathbf{z}}_t + \tilde{\Omega}_t^{-1} \mathbf{A} \mathbf{m}_{t-1} \}, \\ \mathbf{V}_t &= [\tilde{\Omega}_t^{-1} + N_t \mathbf{\Lambda}^T \tilde{\mathbf{R}}^{-1} \mathbf{\Lambda}]^{-1}, \quad \bar{\mathbf{z}}_t = \frac{1}{N_t} \sum_{n_t=1}^{N_t} \mathbf{z}_t^{n_t} \end{aligned} \quad (4.25)$$

The conditional distribution of  $\mathbf{s}_{t-1}$  given  $\mathbf{s}_t$ ,

$$\begin{aligned} p(\mathbf{s}_{t-1} | \mathbf{s}_t, \mathbf{Z}_{1:(t-1)}) &= \mathcal{N}(\tilde{\boldsymbol{\mu}}_{t-1}, \tilde{\Sigma}_{t-1}) \\ \tilde{\boldsymbol{\mu}}_{t-1} &= \tilde{\Sigma}_{t-1} \{ \mathbf{A}^T \mathbf{Q}_t^{-1} \mathbf{s}_t + \mathbf{V}_{t-1}^{-1} \mathbf{m}_{t-1} \} \\ \tilde{\Sigma}_{t-1} &= (\mathbf{V}_{t-1}^{-1} + \mathbf{A}^T \mathbf{Q}_t^{-1} \mathbf{A})^{-1} \end{aligned} \quad (4.26)$$

Similarly, apply Woodbury matrix inversion lemma we have

$$\tilde{\Sigma}_{t-1} = \mathbf{V}_{t-1} - \mathbf{V}_{t-1} \mathbf{A}^T \tilde{\Omega}_t^{-1} \mathbf{A} \mathbf{V}_{t-1}$$

Further, the backward smoothing density at  $t - 1$ ,

$$\begin{aligned} p(\mathbf{s}_{t-1} | \mathbf{Z}_{1:T}) &= \mathcal{N}(\tilde{\mathbf{m}}_{t-1}, \tilde{\mathbf{V}}_{t-1}) \\ \tilde{\mathbf{m}}_{t-1} &= \tilde{\Sigma}_{t-1} (\mathbf{A}^T \mathbf{Q}_t^{-1} \tilde{\mathbf{m}}_t + \mathbf{V}_{t-1}^{-1} \mathbf{m}_{t-1}) \\ \tilde{\mathbf{V}}_{t-1} &= \tilde{\Sigma}_{t-1} + \tilde{\Sigma}_{t-1} \mathbf{A}^T \mathbf{Q}_t^{-1} \tilde{\mathbf{V}}_t \mathbf{Q}_t^{-1} \mathbf{A} \tilde{\Sigma}_{t-1} \end{aligned} \quad (4.27)$$

#### 4.2.4 Accelerated MCMC via Document Subsampling

For large-scale datasets, recent approaches efficiently reduce the computational load of Monte Carlo Markov chain (MCMC) by data subsampling (Korattikara *et al.*, 2014; Quiroz *et al.*, 2014). We borrow this idea of subsampling documents when considering a large corpora (*e.g.*, in our experiments, we consider analysis of articles in the magazine *Science*, composed of 139379 articles from years 1880 to 2002, and a vocabulary size 5855). In our model, the augmented data  $\mathbf{z}_t^{n_t}$  ( $n_t = 1, \dots, N_t$ ) for each document is relatively expensive to sample. One simple method is random document sampling without replacement. However, by treating all likelihood contributions symmetrically, this method leads to a highly inefficient MCMC chain with poor mixing (Korattikara *et al.*, 2014).

Alternatively, we adopt the probability proportional-to-size (PSS) sampling scheme in (Quiroz *et al.*, 2014), *i.e.*, sampling the documents with inclusion probability proportional to the likelihood contributions. For each MCMC iteration, the subsampling procedure for documents at time  $t$  is designed as follows:

- **Step 1:** Given a small subset  $\mathcal{V}_t \subset \{1, \dots, N_t\}$  of chosen documents, only sample  $\{\mathbf{z}_t^d\}$  for all  $d \in \mathcal{V}_t$  and compute the augment log-likelihood contributions (with  $\mathbf{B}_t$  integrated out)  $\ell_{\mathcal{V}_t}(\mathbf{z}_t^d) = \mathcal{N}(\mathbf{\Lambda}\mathbf{s}_t, \tilde{\mathbf{R}})$ , where  $\tilde{\mathbf{R}} = \mathbf{\Lambda}\mathbf{\Gamma}\mathbf{\Lambda}^T + \mathbf{I}_P$ . Note that, only a  $K$ -dimensional matrix inversion is required, by using the Woodbury matrix inversion formula  $\tilde{\mathbf{R}}^{-1} = \mathbf{I}_P - \mathbf{\Lambda}(\mathbf{\Gamma}^{-1} + \mathbf{\Lambda}^T\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T$ .
- **Step 2:** Similar to (Quiroz *et al.*, 2014), we use a Gaussian process (Rasmussen, 2004) to predict the log-likelihood for the remaining documents

$$\ell_{\mathcal{V}_t^c}(\mathbf{z}_t^d) = \mathcal{K}(\mathcal{V}_t^c, \mathcal{V}_t)\mathcal{K}(\mathcal{V}_t, \mathcal{V}_t)^{-1}\ell_{\mathcal{V}_t}(\mathbf{z}_t^d),$$

where  $\mathcal{K}$  is a  $N_t \times N_t$  squared-exponential kernel, which denotes the similarity of documents:  $\mathcal{K}(\mathbf{y}_t^i, \mathbf{y}_t^j) = \sigma_f^2 \exp(-\|\mathbf{y}_t^i - \mathbf{y}_t^j\|^2 / (2s^2))$ ,  $i, j = 1, \dots, N_t$ ,  $\sigma_f^2 = 1$ ,  $s = 1$ .

- **Step 3:** Calculate the inclusion probability  $w_d \propto \exp[\ell(\mathbf{z}_t^d)]$ ,  $d = 1, \dots, N_t$ ,  $\tilde{w}_d = w_d / \sum_{d'} w_{d'}$ .
- **Step 4:** Sampling the next subset  $\mathcal{V}_t$  of pre-specified size  $|\mathcal{V}_t|$  with inclusion probability  $\tilde{w}_d$ , and store it for the use of the next MCMC iteration.

In practice, this adaptive design allows MCMC to run more efficiently on a full dataset of large scale, often mitigating the need to do parallel MCMC implementation. Future work could also consider nonparametric function estimation subject to monotonicity constraint, e.g. Gaussian process projections recently developed in (Lin and Dunson, 2014).

### 4.3 Experiments

Different from DTM (Blei and Lafferty, 2006b), the proposed model has the jumps directly at the level of the factor scores (no exponentiation or normalization needed), and therefore it proved more effective in uncovering jumps in factor scores over time. Demonstrations of this phenomenon in a synthetic experiment are detailed below.

#### 4.3.1 Simulation Study: DRFM with Different Innovations

We conducted a simulation study to assess the performance of the proposed approach. We first generate the latent continuous variable  $\mathbf{Z}$  from the augmented model with  $e = f = 0.5$ ,  $\nu = 1$ ,  $\rho_k = 0.5$ ,  $l_{p,k} = 1/P$  for  $k = 1, \dots, K$ ,  $p = 1, \dots, P$  and then round it to integer value. Three different approaches are considered here: Gaussian innovation with fixed variance  $\delta \sim \mathcal{N}(0, 1)$ , Gaussian innovation with unknown variance  $\delta \sim \mathcal{N}(0, \tau)$ , and  $\tau^{-1} \sim \mathcal{G}(0.01, 0.01)$ , and heavy-tailed innovation  $\delta \sim \text{TPBN}(0.5, 0.5, \phi)$  with  $\phi^{1/2} \sim \mathcal{C}^+(0, 1)$ . The results are shown in Figure 4.1.

Note that here we are dealing with a simple two factor dynamic model, and we recover the ground truth of the trajectory of factor score. In contrast to other

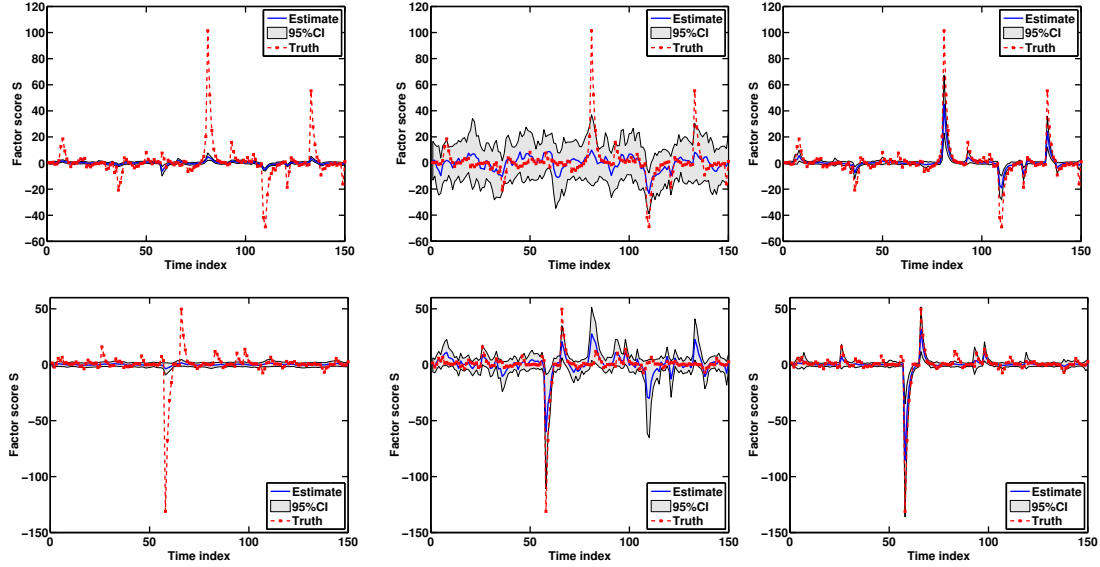


FIGURE 4.1: Estimated posterior mean of factor score  $s$  with 95% confidence interval for  $P = 10$ ,  $K = 2$  and  $T = 150$ . Left column: Gaussian innovation with fixed variance  $\tau = 1$ ; middle column: Gaussian innovation with unknown variance  $\tau^{-1} \sim \mathcal{G}(0.01, 0.01)$ ; right column: heavy-tailed innovation.

models that involve nonlinear transformation, the smooth transition and sudden jumps can be well preserved under the proposed DFRM framework, using heavy-tailed innovation.

Figure 4.2 shows the monotone relationship between observation  $\mathbf{y}$  and the latent variable  $\mathbf{z}$  inferred by extended rank likelihood in our DFRM model. It can be seen that the rank likelihood approach maintains the order information of  $\mathbf{y}$  in  $\mathbf{z}$  and provides a flexible link between  $\mathbf{y}$  and  $\mathbf{z}$ .

In the following, we present exploratory data analysis on two real examples, demonstrating the ability of the proposed model to infer temporal jumps in topic importance, and to infer correlations across topics and words.

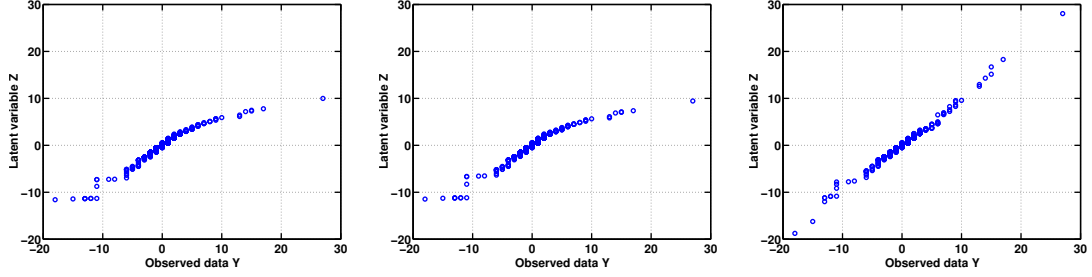


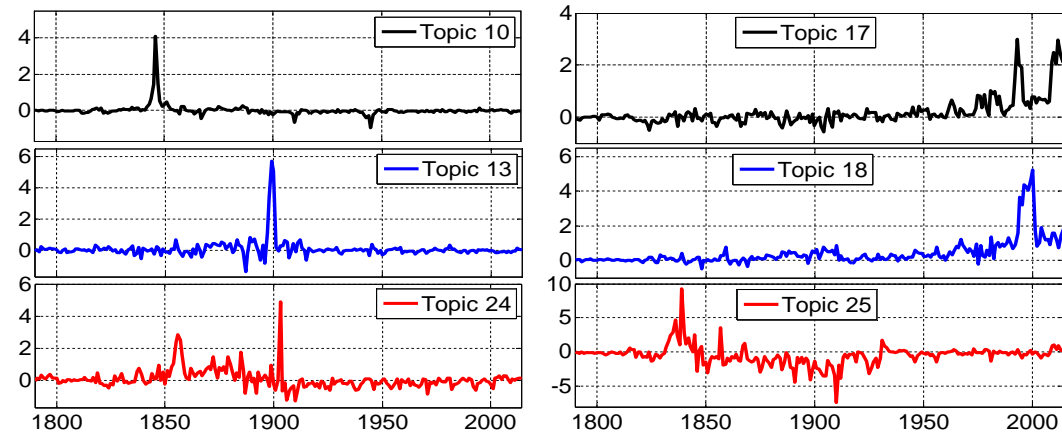
FIGURE 4.2: Estimated posterior mean of latent variable  $z$  vs. the observed data  $y$  for  $P = 10$ ,  $K = 2$  and  $T = 150$ . Left: Gaussian innovation with fixed variance  $\tau = 1$ ; middle: Gaussian innovation with unknown variance  $\tau^{-1} \sim \mathcal{G}(0.01, 0.01)$ ; right: heavy-tailed innovation.

#### 4.3.2 Case Study I: State of the Union Dataset

The State of the Union dataset contains the transcripts of  $T = 225$  US State of the Union addresses, from 1790 to 2014. We take each transcript as a document, *i.e.*, we have one document per year. After removing stop words, and removing terms that occur fewer than 3 times in one document and less than 10 times overall, we have  $P = 7518$  unique words. The observation  $y_{p,t}$  corresponds to the frequency of word  $p$  of the State of the Union transcript from year  $t$ .

We apply the proposed DRFM setting and learned  $K = 25$  topics. To better understand the temporal dynamic per topic, six topics are selected and the posterior mean of their latent trajectories  $s_{k,1:T}$  are shown in Figure 4.3 (with also the top 12 most probable words associated with each of the topics). A complete table with all 25 learned topics and top 12 words is provided in the Table 4.2. The learned trajectory associated with every topic indicates different temporal patterns across all the topics. Figure 4.5 presents the learned trajectory for each topic. Clearly, we can identify jumps associated with some key historical events. For instance, for Topic 10, we observe a positive jump in 1846 associated with the Mexican-American war. Topic 13 is related with the Spanish-American war of 1898, with a positive jump

in that year. In Topic 24, we observe a positive jump in 1914, when the Panama canal was officially opened (words *Panana* and *Canal* are included). In Topic 18, the positive jumps observed from 1997 to 1999 seem to be associated with the creation of the state children’s health insurance program in 1997. We note that the words for this topic are explicitly related with this issue. Topic 25 appears to be related to banking; the significant spike around 1836 appears to correspond to the second bank of the united states, which was allowed to go out of existence, and end national banking that year. In 1863 congress passed the national banking act, which ended the “free-banking” period from 1836-1863; note the spike around 1863 in Topic 25.



Topic#10	Topic#13	Topic#24	Topic#17	Topic#18	Topic#25
Mexico	Government	United	Jobs	Children	Government
Government	United	Treaty	Country	America	Public
Texas	Islands	Isthmus	Tax	Americans	Banks
United	Commission	Public	American	Care	Bank
War	Island	Panama	Economy	Tonight	Currency
Mexican	Cuba	Law	Deficit	Support	Money
Army	Spain	Territory	Americans	Century	United
Territory	Act	America	Energy	Health	Federal
Country	General	Canal	Businesses	Working	American
Peace	Military	Service	Health	Challenge	National
Policy	International	Banks	Plan	Security	Duty
Lands	Officers	Colombia	Care	Families	Institutions

FIGURE 4.3: Above: Time evolving from 1790 to 2014 in the State of the Union dataset for six selected topics. The plotted values represent the posterior means. Below: Top 12 most probable words associated with the above topics.

Our modeling framework is able to capture dynamic patterns of topics and word correlations. To illustrate this, we select three years (associated with some meaningful historical events) and analyze their corresponding topic and word correlations. Figure 4.4 (first row) shows graphs of the topic correlation matrices, in which the nodes represent topics and the edges indicate positive (green) and negative (red) correlations (we show correlations with absolute value larger than 0.01). We notice that Topics 11 and 22 are positively correlated with those years. Some of the most probable words associated with each of them are: *increase, united, law and legislation* (for Topic 11) and *war, Mexico, peace, army, enemy and military* (for Topic 22). We also are interested in understanding the time-varying correlation between words. To do so, and for the same years as before, in Figure 4.4 (second row) we plot the dendrogram associated with the learned correlation matrix for words. In the plots, different colors indicate highly correlated word clusters defined by cutting the branches off the dendrogram. Those figures reveal different sets of highly correlated words for different years. By inspecting all the words correlation, we noticed that the set of words  $\{government, federal, public, power, authority, general, country\}$  are highly correlated across the whole period.

#### 4.3.3 Case Study II: Analysis of Science Dataset

We analyze a collection of scientific documents from the JSTOR *Science* journal (Blei and Lafferty, 2006b). This dataset contains a collection of 139379 documents from 1880 to 2002 ( $T = 123$ ), with approximately 1100 documents per year. After removing terms that occurred fewer than 25 times, the total vocabulary size is  $P = 5855$ . We learn  $K = 50$  topics from the inferred posterior distribution, for brevity and simplicity, we only show 20 of them. We handle about 2700 documents per iteration (subsampling rate: 2%). Table 4.2 shows the 20 selected topics and the top 10 most probable words associated with each of them. By inspection, we notice

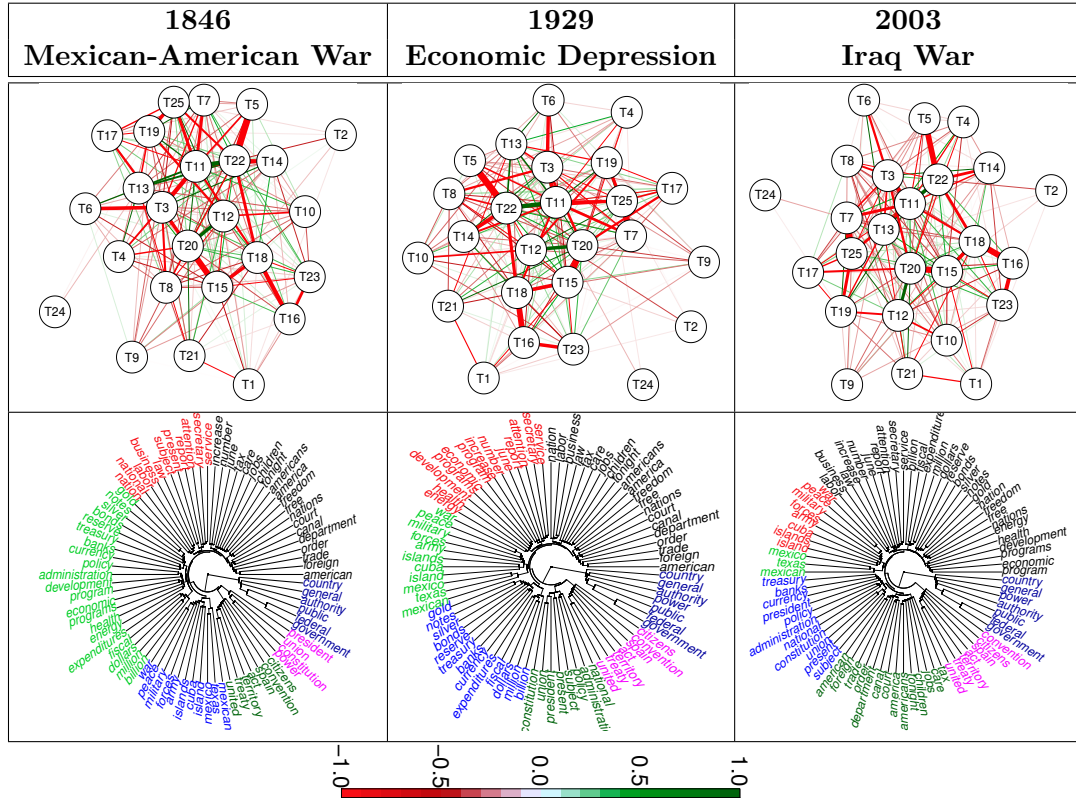


FIGURE 4.4: First row: Inferred correlations between topics for some specific years associated with some meaningful historical events. Green edges indicate positive correlations and red edges indicate negative correlations. Second row: Learned dendrogram based upon the correlation matrix between the top 10 words associated with each topic (we display 80 unique words in total).

that those topics are related with specific fields in science. For instance, Topic 2 is more related to “scientific research”, Topic 10 to “natural resources”, and Topic 15 to “genetics”. Figure 4.6 shows the time-varying trend for some specific words,  $\hat{z}_{p,1:T}$ , which reveals the *importance* of those words across time. Finally, Figure 4.7 shows the correlation between the selected 20 topics. For instance, in 1950 and 2000, topic 9 (related to *mouse, cells, human, transgenic*) and topic 17 (related to *virus, rna, tumor, infection*) are highly correlated.

Table 4.1: The 25 Topics of the State of the Union Dataset

Topic#1	Topic#2	Topic#3	Topic#4	Topic#5	Topic#6	Topic#7
United Act Public Treaty Duties Present Nations Treasury Session Commerce Citizens War	Dollars War Million Fiscal Expenditures Government Billion Program United Federal Estimated Legislation	Administration Federal Program Policy Energy Programs Economic Development Security Nation Major Act	Government American United Foreign Department National Canal Policy Republic Order Administration Banks	Government Service Public Department Report Secretary District Attention Present Fiscal Laws Court	Law Country National Public Business Government Action Control United Interstate Labor Corporations	Government United Department Public Law Court Service Federal Canal Tariff District Lands
Topic#8	Topic#9	Topic#10	Topic#11	Topic#12	Topic#13	
Government General Public Character Interests Subject Country Power Duty Attention Federal Means	Constitution Country War President Power Mexico Public Union California Service House Period	Mexico Government Texas United War Mexican Army Territory Country Peace Policy Lands	Increase United Cent Law Legislation Secretary Free Increased Fiscal American Tariff Products	Government Public Nation American Law Power Conditions Business Islands Service War Laws	Government United Islands Commission Island Cuba Spain Act General Military International Officers	
Topic#14	Topic#15	Topic#16	Topic#17	Topic#18	Topic#19	
Free Nations Freedom Economic Military Defense United Peace Strength Security Program Nation	Government Federal Public National Country Economic Agriculture Banks Present America Reduction Construction	Statute Law Business General American Purpose Court Mexico Service Federal Commission Present	Jobs Country Tax American Economy Deficit AmericanS Energy Businesses Health Plan Care	America America Americans Care Tonight Support Century Health Working Challenge Security Families Budget	America Government Nation American Federal Tonight Peace War Freedom AmericanS Future Budget	
Topic#20	Topic#21	Topic#22	Topic#23	Topic#24	Topic#25	
Gold Government Notes Treasury Silver United Bonds Currency Reserve Circulation Issued Large	Government Constitution United Power Union Federal Duty American Kansas Question Law Present	War Mexico Peace Army Enemy Forces Military Mexican Production Japanese Fighting American	Government United Spain Cuba Spanish War Island Secretary June Duty Department Fiscal	United Isthmus Public Panama Law Territory America Canal Service Banks Colombia	Government Public Banks Bank Currency Money United Federal American National Duty Institutions	

Table 4.2: Selected 20 topics associated with the analysis of the *Science* dataset and top 10 most probable words.

Topic#1	Topic#2	Topic#3	Topic#4	Topic#5	Topic#6	Topic#7	Topic#8	Topic#9	Topic#10
cells cell normal two growth development tissue body egg blood	research National government support federal development new program scientific basic	field magnetic solar Energy spin state electron quantum temperature current	animals brain neurons activity response rats control fig effects days	Energy oil percent production fuel total growth states electricity coal	university professor college president department research institute director society school	science scientific new scientists human men sciences knowledge meeting work	work research scientific laboratory made university results science survey department	mouse type wild fig cells human transgenic animals mutant	water surface temperature soil pressure sea plants solution plant air
Topic#11	Topic#12	Topic#13	Topic#14	Topic#15	Topic#16	Topic#17	Topic#18	Topic#19	Topic#20
system nuclear new systems power cost computer fuel coal plant	energy theory temperature radiation atoms surface atomic mass atom time	association science meeting university American society section president committee secretary	protein proteins cell membrane amino sequence binding acid residues sequences	human genome sequence chromosome gene genes map data sequences genetic	professor university society Department college president director american appointed medical	virus rna viruses particles tumor mice disease viral human infection	energy electron state fig two structure reaction laser high temperature	stars mass star temperature solar gas data density surface galaxies	rna fig mrna protein site sequence splicing synthesis trna rnas

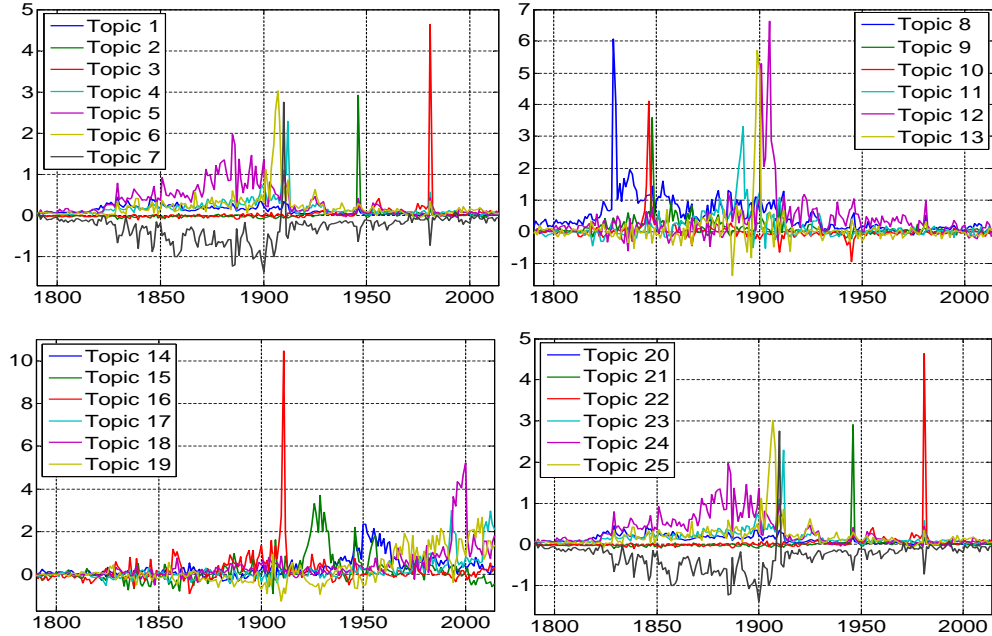


FIGURE 4.5: Time evolving topics from 1790 to 2014. Left up panel: Topics 1 to 7. Right up panel: Topics 8 to 13. Left bottom panel: Topics 14 to 19. Right bottom panel: Topics 20 to 25. The plotted values represent the posterior means.

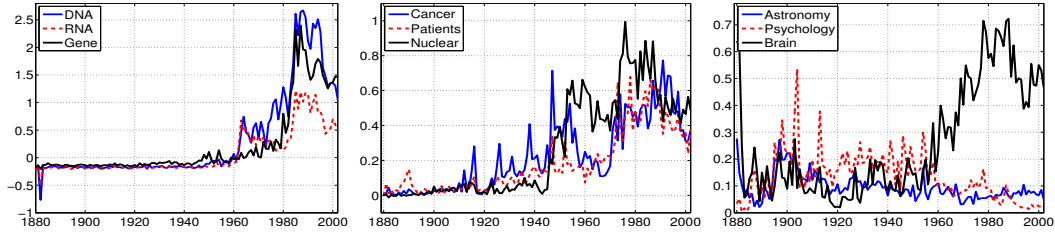


FIGURE 4.6: The inferred latent trend for variable  $\hat{z}_{p,1:T}$  associated with words.

## 4.4 Discussion

We have proposed a DRFM framework that could be applied to a broad class of applications such as: (i) dynamic topic model for the analysis of time-stamped document collections; (ii) joint analysis of multiple time series, with ordinal valued observations; and (iii) multivariate ordinal dynamic factor analysis or dynamic copula analysis for mixed type of data. The proposed model is a semi-parametric methodology, which

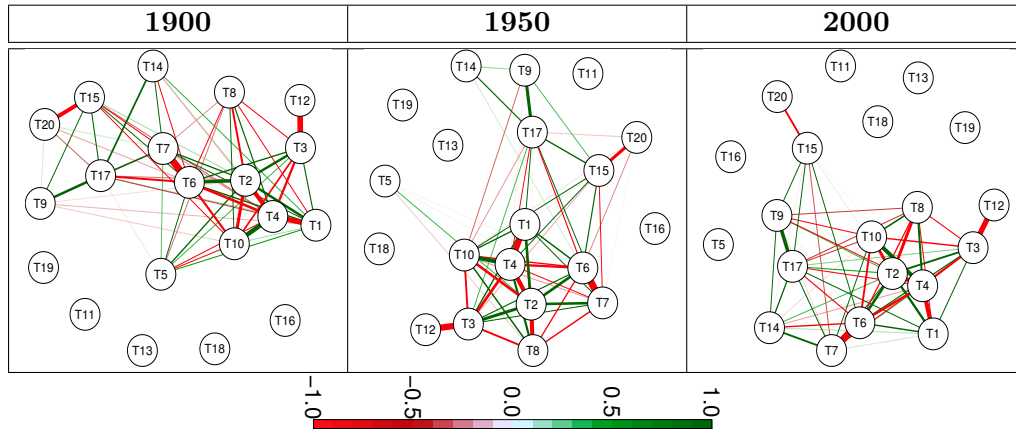


FIGURE 4.7: Inferred correlations between topics for some specific years. Green edges indicate positive correlations and red edges indicate negative correlations.

offers modeling flexibilities and reduces the effect of model misspecification. However, as the marginal likelihood is distribution-free, we could not calculate the model evidence or other evaluation metrics based on it (e.g. held-out likelihood). As a consequence, we are lack of objective evaluation criteria, which allow us to perform formal model comparisons. In our proposed setting, we are able to perform either retrospective analysis or multi-step ahead forecasting (using the recursive equations derived in the FFBS algorithm). Finally, our inference framework is easily adaptable for using sequential Monte Carlo (SMC) methods (Doucet *et al.*, 2001) allowing on-line learning.

# Leveraging Dependencies in Heterogeneous Multitask Learning

Learning multiple tasks across heterogeneous domains is a challenging problem since the feature space may not be the same for different tasks. In this chapter, we assume the data in multiple tasks are generated from a latent common domain via sparse domain transforms and propose a latent probit model (LPM) to jointly learn the domain transforms, and a probit classifier shared in the common domain. To learn meaningful task relatedness and avoid over-fitting in classification, we introduce sparsity in the domain transforms matrices, as well as in the common classifier parameters. We derive theoretical bounds for the estimation error of the classifier parameters in terms of the sparsity of domain transform matrices. An expectation-maximization algorithm is derived for learning the LPM. The effectiveness of the approach is demonstrated on several real datasets.

The contributions are summarized as follows. First, we propose the latent probit model (LPM) (Section 5.1) and analyze the estimation error of the classifier as a function of the sparsity level of domain transforms (Section 5.2). Second, we develop

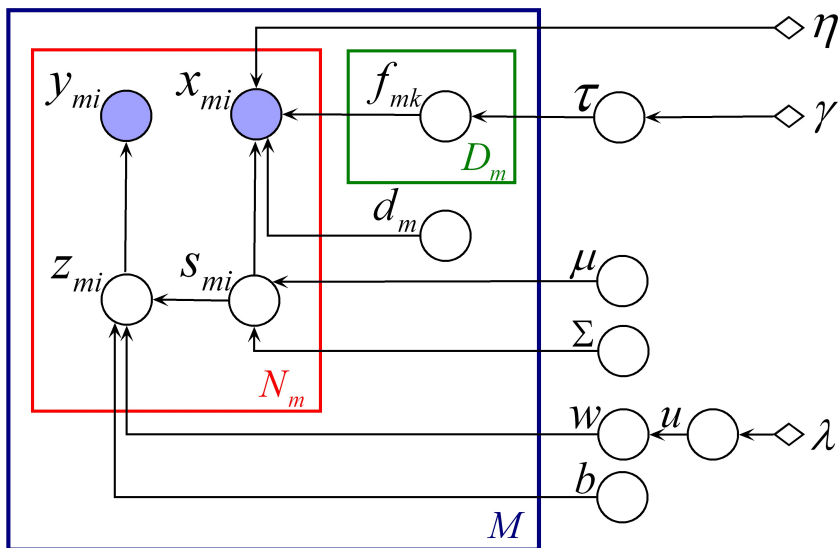


FIGURE 5.1: A graphic representation of the proposed latent probit model, where solid circles denote data, hollow circles denote unknown parameters and latent variables, and diamonds denote input parameters (including hyper-parameters and fixed model parameters).

an EM algorithm for learning the LPM (Section 5.3). Third, we provide extensive experimental results evaluating the LPM’s performance on two real datasets (Section 5.4).

## 5.1 The Latent Probit Model

The latent probit model (LPM) is a generative probabilistic model for  $M \geq 2$  partially labeled sets of feature vectors (data points), assuming each dataset has a different feature representation. The LPM has a hierarchical Bayesian structure, as graphically shown in Figure 5.1, and is parameterized by  $\{\eta, \boldsymbol{\mu}, \boldsymbol{\Sigma}, b, \boldsymbol{w}\}$  and  $\{\boldsymbol{F}_m, \boldsymbol{d}_m\}_{m=1}^M$ . The parameters  $\boldsymbol{w}$  specify the probit classifier shared by the tasks in the latent feature space, and  $\boldsymbol{F}_m$  specifies the domain transform for the  $m$ -th dataset up to a translation (which is specified by  $\boldsymbol{d}_m$ ). The parameters  $\boldsymbol{w}$  and  $\{\boldsymbol{F}_m\}_{m=1}^M$  are given hi-

erarchical Laplacian priors (Figueiredo, 2003) to encourage sparsity, with the priors specified by hyper-parameters  $\{\gamma, \lambda\}$ . The other hollow circles in Figure 5.1 denote latent variables, which include  $\{\boldsymbol{\tau}, \mathbf{u}, \mathbf{s}, z\}$ . The generative process in the LPM is described below, with  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  denoting a normal distribution with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ .

Given hyper-parameters  $\{\gamma, \lambda\}$ , the sparse parameters  $\mathbf{w}$  and  $\{\mathbf{F}_m\}_{m=1}^M$  are generated as follows.

1. Draw  $\mathbf{w} = [w_j]_{F_0 \times 1}$ , the sparse parameters of the probit model shared by the tasks in the latent feature space,

$$w_j \sim \mathcal{N}(0, u_j),$$

$$u_j \sim \frac{\lambda}{2} \exp\left\{-\frac{\lambda}{2}u_j\right\}, \quad u_j \geq 0, \quad j = 1, 2, \dots, F_0,$$

where  $F_0$  is the latent feature dimensionality.

2. For each task  $m = 1, 2, \dots, M$ , draw the sparse domain specific transform matrix  $\mathbf{F}_m = [f_{mkj}]_{D_m \times F_0}$  by

$$f_{mkj} \sim \mathcal{N}(0, \tau_{mkj}),$$

$$\tau_{mkj} \sim \frac{\gamma}{2} \exp\left\{-\frac{\gamma}{2}\tau_{mkj}\right\}, \quad \tau_{mkj} \geq 0,$$

$k = 1, \dots, D_m$  and  $j = 1, \dots, F_0$ , with  $D_m$  the observed feature dimensionality of the  $m$ -th dataset.

Given parameters  $\{\eta, \boldsymbol{\mu}, \boldsymbol{\Sigma}, b, \mathbf{w}\} \cup \{\mathbf{F}_m, \mathbf{d}_m\}_{m=1}^M$ , the data sets are generated as follows.

For  $i = 1, 2, \dots, N_m$  and  $m = 1, 2, \dots, M$ ,

1. Draw a latent feature vector

$$\mathbf{s}_{mi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{5.1}$$

where  $\boldsymbol{\mu} \in \mathbb{R}^{F_0 \times 1}$  and  $\boldsymbol{\Sigma} \in \mathbb{R}^{F_0 \times F_0}$  are the mean and covariance matrix, respectively.

2. Draw an observed feature vector

$$\mathbf{x}_{mi} \sim \mathcal{N}(\mathbf{F}_m \mathbf{s}_{mi} + \mathbf{d}_m, \eta \mathbf{I}), \quad (5.2)$$

where  $\mathbf{d}_m \in \mathbb{R}^{D_m}$ ,  $\eta > 0$  and  $\mathbf{I}$  denotes an identity matrix of appropriate dimensions.

3. If the feature vector  $\mathbf{x}_{mi}$  requires a label, draw the label by

$$y_{mi} = \begin{cases} +1, & \text{if } z_{mi} \geq 0, \\ -1, & \text{otherwise,} \end{cases}$$

$$z_{mi} \sim \mathcal{N}(\mathbf{w}^T \mathbf{s}_{mi} + b, 1), \quad b \in \mathbb{R}. \quad (5.3)$$

Note that the latent normal distribution in (5.1) can be extended to a mixtures of normal distributions to account for more complicated data manifolds.

The sparsity of domain transforms in LPM plays a pivotal role in defining the between-task relations. Roughly speaking, a greater sparsity in domain transforms indicates closer relations between the tasks. In other words, sparser domain transforms imply that different tasks look more similar to each other in the latent feature space, and thus greater performance gain may be achieved by sharing information among the tasks. We give a quantitative analysis of the performance gain by providing an upper bound to the estimation error of the probit classifier, which is shared among the tasks in the latent space. The bound has an analytic functional dependence on the sparsity level of domain transforms, showing that sparsity contributes directly to the error reduction. In addition, the bound also reveals the error's dependency on the number of tasks, the number of labeled examples in each task, and the latent dimensionality.

With  $\{\mathbf{F}_m\}_{m=1}^M$  drawn from sparse prior distributions, most entries of these matrices will be zero; by (5.2) this implies that only a few latent features are responsible for

generating the observed features. Since this is true for any  $m$ , the chance for different datasets to use the same features to generate their observed features is large. However, latent features are identically distributed; thus the shared latent features must have the same statistics across the tasks. Therefore, the datasets (sets of features vectors) generated by the LPM model are encouraged to be closely related.

While the sparsity of  $\{\mathbf{F}_m\}_{m=1}^M$  reflects the relatedness between the sets of features vectors, the sparsity of  $\mathbf{w}$  encourages the classification to be dependent on a few latent features. This is important, because even when the observed features differ among tasks to entail less sparse  $\{\mathbf{F}_m\}_{m=1}^M$ , the tasks may still be able to share information for classification through appropriately selected latent features.

## 5.2 Sparse Oracle Inequalities

The goal of our analysis is to quantify the notion that sparse domain transforms encourage the tasks to be related, and that better generalization can be achieved by sharing information among related tasks to learn the common classifier. The analysis is based on an upper bound for the estimation error of  $\mathbf{w}$ , with the bound represented in terms of the number of nonzero elements of the true  $\{\mathbf{F}_m\}_{m=1}^M$ .

Since we are analyzing the general information-sharing mechanism in the LPM, we expect the results to be insensitive to the choice of estimation method. We therefore employ a simple two-step approach to estimate  $\mathbf{w}$ . The estimation is based on training data generated by the true LPM parameterized by  $\{\eta, \boldsymbol{\mu}, \boldsymbol{\Sigma}, b, \mathbf{w}^*\} \cup \{\mathbf{F}_m, \mathbf{d}_m\}_{m=1}^M$ , with the simplifications  $b = 0$ ,  $\boldsymbol{\Sigma} = \mathbf{I}$ ,  $\boldsymbol{\mu} = \mathbf{0}$ , and  $\mathbf{d}_m = \mathbf{0} \forall m$ , where  $\mathbf{0}$  is a vector of zeros of appropriate dimensions. Note we have used a superscript  $*$  to emphasize  $\mathbf{w}^*$  is the vector of unknown parameters to be estimated.

Let  $\{\mathbf{X}_m\}_{m=1}^M$ , with  $\mathbf{X}_m = [\mathbf{x}_{m1}, \mathbf{x}_{m2}, \dots, \mathbf{x}_{mL_m}]$ , be  $M$  sets of feature vectors,

each corresponding to a task. By the generative process of the LPM,

$$\mathbf{X}_m = \mathbf{F}_m \mathbf{S}_m + [\epsilon_{mij}]_{D_m \times N_m},$$

where  $\{\epsilon_{mij}\}$  are i.i.d. drawn from a zero-mean normal distribution with variance  $\eta$ , and the entries of  $\mathbf{S}_m$  are i.i.d. from the standard normal distribution. Given  $\mathbf{X}_m$ , the maximum-likelihood solutions to  $\{\mathbf{S}_m\}$  are given by

$$\hat{\mathbf{S}}_m = (\mathbf{F}_m^T \mathbf{F}_m)^{-1} \mathbf{F}_m^T \mathbf{X}_m, \quad \forall m, \quad (5.4)$$

which form a global data matrix by pooling data across the tasks,

$$\Psi = [\hat{\mathbf{S}}_1, \hat{\mathbf{S}}_2, \dots, \hat{\mathbf{S}}_M] \in \mathbb{R}^{F_0 \times n_t}, \quad (5.5)$$

where  $n_t = \sum_{m=1}^M L_m$  is the total number of training examples across all  $M$  tasks.

To simplify the analysis, we assume access to the latent responses of  $\mathbf{w}^*$  to  $\Psi$ , i.e.,

$$\mathbf{z} = \Psi^T \mathbf{w}^* + \mathbf{e} \quad (5.6)$$

where  $\mathbf{z} = [\mathbf{z}_1, \dots, \mathbf{z}_M]^T$  with  $\mathbf{z}_m = [z_{m1}, \dots, z_{mL_m}]$ , and the entries in  $\mathbf{e}$  are assumed i.i.d. from the standard normal distribution. These assumptions may be avoided at the price of complicating the bound, which is not pursued here. The estimate of  $\mathbf{w}^*$  is given by

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} (\|\mathbf{z} - \Psi^T \mathbf{w}\|_2^2 + r \|\mathbf{w}\|_1). \quad (5.7)$$

We derive an upper bound to  $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2$ , following similar arguments as in Bickel *et al.* (2009); Lounici *et al.* (2009) and making use of a key result in Byrne (2009) on extreme singular values of Hermitian matrices. Our main results are stated in Theorem 7.

**Theorem 7.** *Let  $\mathbf{w}^*$  have nonzero and zero elements indexed respectively by  $J$  and  $J^c$ . Denote  $s = |J|$  as the cardinality of  $J$ . Let  $\boldsymbol{\delta} = \hat{\mathbf{w}} - \mathbf{w}^*$  with  $\hat{\mathbf{w}}$  given in (5.7),*

and  $c_0$  be the minimum nonnegative number such that  $\|\boldsymbol{\delta}_{J^c}\|_1 \leq c_0 \|\boldsymbol{\delta}_J\|_1$ . Let  $\boldsymbol{\Psi}_j$  be the transpose of the  $j$ -th row of  $\boldsymbol{\Psi}$  and  $\varepsilon_\psi = \max_j \|\boldsymbol{\Psi}_j\|_2$ . For any  $F_0 \geq 2$  and  $a \geq \sqrt{8}$ , it holds with probability of at least  $P_e = 1 - F_0^{1-a^2/8}$  that

$$\|\boldsymbol{\delta}\|_2 \leq \frac{2a\varepsilon_\psi n_t^{-1} \sqrt{s(1+c_0^2 s) \ln F_0}}{\sum_{m=1}^M \frac{\omega_{\min}(\mathbf{X}_m^T \mathbf{X}_m / n_t)}{\max_i \left( \sum_{j=1}^{F_0} \|\mathbf{f}_{m,:j}\|_0 |f_{ij}|^2 \right)}}, \quad (5.8)$$

where  $\mathbf{f}_{m,:j}$  denotes the  $j$ -th column of  $\mathbf{F}_m$  and  $\|\mathbf{f}\|_0$  denotes the number of nonzero elements in vector  $\mathbf{f}$ .

*Proof.* By (5.7), one has

$$\frac{1}{n_t} \|\boldsymbol{\Psi}^T \hat{\mathbf{w}} - \mathbf{z}\|_2^2 + r \|\hat{\mathbf{w}}\|_1 \leq \frac{1}{n_t} \|\boldsymbol{\Psi}^T \mathbf{w}^* - \mathbf{z}\|_2^2 + r \|\mathbf{w}^*\|_1.$$

Substituting  $\mathbf{z} = \boldsymbol{\Psi}^T \mathbf{w}^* + \mathbf{e}$ , one obtains

$$\frac{1}{n_t} \|\boldsymbol{\Psi}^T (\hat{\mathbf{w}} - \mathbf{w}^*) - \mathbf{e}\|_2^2 \leq \frac{1}{n_t} \|\mathbf{e}\|_2^2 + r (\|\mathbf{w}^*\|_1 - \|\hat{\mathbf{w}}\|_1),$$

which, using the notations  $\boldsymbol{\delta} = \hat{\mathbf{w}} - \mathbf{w}^*$  and  $r_e = \|\boldsymbol{\Psi} \mathbf{e}\|_\infty / n_t$ , is expanded to give

$$\begin{aligned} \frac{1}{n_t} \|\boldsymbol{\Psi}^T \boldsymbol{\delta}\|_2^2 &\leq \frac{2}{n_t} \boldsymbol{\delta}^T \boldsymbol{\Psi} \mathbf{e} + r (\|\mathbf{w}^*\|_1 - \|\hat{\mathbf{w}}\|_1), \\ &\leq 2r_e \|\boldsymbol{\delta}\|_1 + r (\|\mathbf{w}^*\|_1 - \|\hat{\mathbf{w}}\|_1), \\ &= 2r_e (\|\boldsymbol{\delta}_J\|_1 + \|\hat{\mathbf{w}}_{J^c}\|_1) + r (\|\mathbf{w}_J^*\|_1 - \|\hat{\mathbf{w}}_J\|_1) - r \|\hat{\mathbf{w}}_{J^c}\|_1, \\ &\stackrel{(a)}{\leq} \|\boldsymbol{\delta}_J\|_1 (2r_e + r) + \|\hat{\mathbf{w}}_{J^c}\|_1 (2r_e - r), \\ &\leq \sqrt{s} \|\boldsymbol{\delta}_J\|_2 (2r_e + r) + \|\hat{\mathbf{w}}_{J^c}\|_1 (2r_e - r), \end{aligned} \quad (5.9)$$

where inequality (a) arises because  $\|\mathbf{w}^*\|_1 - \|\hat{\mathbf{w}}\|_1 \leq \|\mathbf{w}^* - \hat{\mathbf{w}}\|_1 = \|\boldsymbol{\delta}_J\|_1$ . Dividing both sides of (5.9) by  $\|\boldsymbol{\Psi}^T \boldsymbol{\delta}\|_2$  gives

$$\frac{1}{n_t} \|\boldsymbol{\Psi}^T \boldsymbol{\delta}\|_2 \leq \frac{\sqrt{s} \|\boldsymbol{\delta}_J\|_2}{\|\boldsymbol{\Psi}^T \boldsymbol{\delta}\|_2} (2r_e + r) + \frac{\|\hat{\mathbf{w}}_{J^c}\|_1}{\|\boldsymbol{\Psi}^T \boldsymbol{\delta}\|_2} (2r_e - r), \quad (5.10)$$

which is reduced to

$$\frac{1}{n_t} \|\Psi^T \boldsymbol{\delta}\|_2 \leq 2r\sqrt{s} \frac{\|\boldsymbol{\delta}_J\|_2}{\|\Psi^T \boldsymbol{\delta}\|_2}. \quad (5.11)$$

when  $2r_e \leq r$ . Clearly the inequality in (5.11) holds with probability no less than  $P_e = p(2r_e \leq r)$ . We will come back to find the expression of  $P_e$ ; until then we assume  $2r_e \leq r$  is true. We follow Bickel *et al.* (2009); Lounici *et al.* (2009) to similarly define  $\kappa_s = \min_{\boldsymbol{\delta} \neq \mathbf{0}} n_t^{-1/2} \|\boldsymbol{\delta}_J\|_2^{-1} \|\Psi^T \boldsymbol{\delta}\|_2$ , then

$$\|\boldsymbol{\delta}_J\|_2 \leq \kappa_s^{-1} n_t^{-1/2} \|\Psi^T \boldsymbol{\delta}\|_2, \quad (5.12)$$

Substitution of (5.12) into (5.11) yields  $\|\Psi^T \boldsymbol{\delta}\|_2 \leq 2r\sqrt{n_t s}/\kappa_s$ , which is substituted back into (5.12) to give

$$\|\boldsymbol{\delta}_J\|_2 \leq 2r\kappa_s^{-2}\sqrt{s}. \quad (5.13)$$

By the definition of  $\kappa_s$ , one has

$$n_t \kappa_s^2 = \min_{\mathbf{v} \neq \mathbf{0}} \frac{\|\Psi^T \mathbf{v}\|_2^2}{\|\mathbf{v}_J\|_2^2} \geq \min_{\mathbf{v} \neq \mathbf{0}} \frac{\|\Psi^T \mathbf{v}\|_2^2}{\|\mathbf{v}\|_2^2}.$$

Substituting (5.5), alongside (5.4), one gets

$$\begin{aligned} n_t \kappa_s^2 &= \min_{\mathbf{v} \neq \mathbf{0}} \sum_{m=1}^M \frac{\|\mathbf{X}_m^T \mathbf{F}_m (\mathbf{F}_m^T \mathbf{F}_m)^{-1} \mathbf{v}\|_2^2}{\|\mathbf{v}\|_2^2} \\ &\geq \sum_{m=1}^M \min_{\mathbf{v} \neq \mathbf{0}} \frac{\|\mathbf{X}_m^T \mathbf{F}_m (\mathbf{F}_m^T \mathbf{F}_m)^{-1} \mathbf{v}\|_2^2}{\|\mathbf{v}\|_2^2}, \text{ (Weyl's Inequality)} \\ &= \sum_{m=1}^M \min_{\mathbf{v} \neq \mathbf{0}} \frac{\|\mathbf{X}_m^T \mathbf{F}_m (\mathbf{F}_m^T \mathbf{F}_m)^{-1} \mathbf{v}\|_2^2}{\|\mathbf{F}_m (\mathbf{F}_m^T \mathbf{F}_m)^{-1} \mathbf{v}\|_2^2} \frac{\mathbf{v}^T (\mathbf{F}_m^T \mathbf{F}_m)^{-1} \mathbf{v}}{\mathbf{v}^T \mathbf{v}}, \\ &\geq \sum_{m=1}^M \min_{\mathbf{v} \neq \mathbf{0}} \frac{\|\mathbf{X}_m^T \mathbf{F}_m (\mathbf{F}_m^T \mathbf{F}_m)^{-1} \mathbf{v}\|_2^2}{\|\mathbf{F}_m (\mathbf{F}_m^T \mathbf{F}_m)^{-1} \mathbf{v}\|_2^2} \times \min_{\mathbf{v} \neq \mathbf{0}} \frac{\mathbf{v}^T (\mathbf{F}_m^T \mathbf{F}_m)^{-1} \mathbf{v}}{\mathbf{v}^T \mathbf{v}}, \\ &\geq \sum_{m=1}^M \min_{\tilde{\mathbf{v}} \neq \mathbf{0}} \frac{\|\mathbf{X}_m^T \tilde{\mathbf{v}}\|_2^2}{\|\tilde{\mathbf{v}}\|_2^2} \min_{\mathbf{v} \neq \mathbf{0}} \frac{\mathbf{v}^T (\mathbf{F}_m^T \mathbf{F}_m)^{-1} \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \geq \sum_{m=1}^M \frac{\omega_{\min}(\mathbf{X}_m^T \mathbf{X}_m)}{\omega_{\max}(\mathbf{F}_m^T \mathbf{F}_m)}, \end{aligned} \quad (5.14)$$

where  $\omega_{\min}(\cdot)$  and  $\omega_{\max}(\cdot)$  respectively represents the maximum and minimum eigenvalue of a Hermitian matrix. Substitution of (5.14) into (5.13) gives

$$\|\boldsymbol{\delta}_J\|_2 \leq \frac{2r\sqrt{s}}{\sum_{m=1}^M \omega_{\min}(\mathbf{X}_m^T \mathbf{X}_m/n_t) \omega_{\max}^{-1}(\mathbf{F}_m^T \mathbf{F}_m)}. \quad (5.15)$$

By the result in Byrne (2009),

$$\omega_{\max}(\mathbf{F}_m^T \mathbf{F}_m) \leq \max_i \left( \sum_{j=1}^{F_0} \|\mathbf{f}_{m, \cdot, j}\|_0 |f_{ij}|^2 \right), \quad (5.16)$$

$m = 1, 2, \dots, M$ , which is substituted into (5.15) to yield (5.8), using the auxiliary variable defined as  $a = n_t r (\ln F_0)^{-1/2} \varepsilon_\psi^{-1}$  and  $\|\boldsymbol{\delta}_{J^c}\|_2 \leq \|\boldsymbol{\delta}_{J^c}\|_1 \leq c_0 \|\boldsymbol{\delta}_J\|_1 \leq c_0 \sqrt{s} \|\boldsymbol{\delta}_J\|_2$ , where  $\|\boldsymbol{\delta}_{J^c}\|_1 \leq c_0 \|\boldsymbol{\delta}_J\|_1$  by assumption.

Recall that (5.11) holds with probability no less than  $P_e = p(2r_e \leq r)$ . Since (5.8) is implied by (5.11), the probability for (5.8) being true is no less than  $P_e$  also.

To evaluate  $P_e$ , we first plug  $r_e = \|\boldsymbol{\Psi} \mathbf{e}\|_\infty/n_t$  into  $P_e = p(2r_e \leq r)$  and expand the result, yielding

$$\begin{aligned} P_e &= p(2\|\boldsymbol{\Psi} \mathbf{e}\|_\infty/n_t \leq r) = 1 - p(2\|\boldsymbol{\Psi} \mathbf{e}\|_\infty/n_t \geq r), \\ &\geq 1 - \sum_{j=1}^{F_0} p(2|\boldsymbol{\Psi}_j^T \mathbf{e}|/n_t \geq r) = 1 - \sum_{j=1}^{F_0} p(|\boldsymbol{\Psi}_j^T \mathbf{e}| \|\boldsymbol{\Psi}_j\|_2^{-1} \geq n_t r 2^{-1} \|\boldsymbol{\Psi}_j\|_2^{-1}), \\ &\geq 1 - \sum_{j=1}^{F_0} p(|\boldsymbol{\Psi}_j^T \mathbf{e}| \|\boldsymbol{\Psi}_j\|_2^{-1} \geq n_t r / (2\varepsilon_\psi)), \end{aligned}$$

where the first inequality results from the union bound. Since the elements of  $\mathbf{e}$  are i.i.d. from the standard normal distribution, so is  $\boldsymbol{\Psi}_j^T \mathbf{e} / \|\boldsymbol{\Psi}_j\|_2$ . Using the inequality  $\mathbb{P}(|X| > x) \leq 2 \exp(-x^2/2)/(x\sqrt{2\pi})$ ,  $x > 0$ , for any standard normal-distributed random number  $X$ , one obtains

$$P_e \geq 1 - \frac{4F_0 \exp(-\frac{n_t^2 r^2}{8\varepsilon_\psi^2})}{\sqrt{2\pi} n_t r \varepsilon_\psi^{-1}} = 1 - \frac{4}{a\sqrt{2\pi} \ln F_0} F_0^{1-a^2/8} \geq 1 - F_0^{1-a^2/8},$$

where the equation is due to  $a = n_t r (\ln F_0)^{-1/2} \varepsilon_\psi^{-1}$ , and the second inequality arises because  $F_0 \geq 2$  and  $a \geq \sqrt{8}$  by assumption, which ensure  $\frac{4}{a\sqrt{2\pi \ln F_0}} \leq 1$ .  $\square$

The bound in (5.8) establishes the functional dependency of  $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2$  on a number of characteristic parameters of the LPM. Foremost, the term  $\|\mathbf{f}_{m,:j}\|_0$  measures the number of nonzero elements in the  $j$ -th column of  $\mathbf{F}_m$ . A sparse  $\mathbf{F}_m$  has small  $\|\mathbf{f}_{m,:j}\|_0$  for its columns, which decreases the term  $\max_j \left( \|\mathbf{f}_{m,:j}\|_0 \sum_{j=1}^{F_0} |f_{ij}|^2 \right)$  and contributes to the error reduction. Second,  $s$  is the number of nonzero elements in  $\mathbf{w}$ ; a sparse  $\mathbf{w}$  has a small  $s$ , which makes the error small.

Third, recall that  $n_t = \sum_{m=1}^M L_m$ , where  $M$  is the number of tasks, and  $L_m$  is the number of training samples in the  $m$ -th task. The  $n_t$  in the denominator of (5.8) plays the role of normalization with respect to the training examples across all tasks, leaving the  $n_t$  in the numerator to influence the error: large  $n_t$  indicates small error. Note that some tasks may have few examples while other have abundant ones; as long as they add up to a large  $n_t$ , similar error reduction will be achieved. Lastly,  $F_0$  is the dimensionality of latent features shared across the tasks. The error bound decreases as  $F_0$  becomes smaller.

### 5.3 Parameter Estimation

We seek a MAP estimate of the parameters  $\Theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, b, \mathbf{w}\} \cup \{\mathbf{F}_m, \mathbf{d}_m\}_{m=1}^M$ . Taking into account all data (labeled and unlabeled) and the sparse priors, and integrating out the latent variables  $\{\boldsymbol{\tau}, \mathbf{u}, \mathbf{s}, z\}$ , one obtains the logarithmic posterior probability

---

**Algorithm 2** The EM algorithm for learning the LPM
 

---

**Input:**  $\{\mathbf{x}_{mi}\}_{i=1}^{N_m} \cup \{y_{mi}\}_{i \in \mathcal{L}_m}$ ,  $m = 1, 2, \dots, M$ ;  $\{\gamma, \lambda\}$  and  $\{\eta, F_0\}$ .  
**Initialize**  $\Theta$ .  
**repeat**  
   Update  $\Sigma$ ,  $\boldsymbol{\mu}$  using  $\{\mathbf{x}_{mi}\}_{i=1}^{N_m} \cup \{y_{mi}\}_{i \in \mathcal{L}_m}$ ,  $m = 1, 2, \dots, M$ , according to (5.17)  
   **for**  $m = 1$  **to**  $M$  **do**  
     Update  $\mathbf{F}_m$ ,  $\mathbf{d}_m$  using  $\{\mathbf{x}_{mi}\}_{i=1}^{N_m} \cup \{y_{mi}\}_{i \in \mathcal{L}_m}$  according to (5.18)  
   **end for**  
   Estimate  $\mathbf{w}$ ,  $b$  according to (5.19) using  $\{\mathbf{x}_{mi}\}_{i \in \mathcal{L}_m} \cup \{y_{mi}\}_{i \in \mathcal{L}_m}$ ,  $m = 1, 2, \dots, M$ ,  
**until**  $\ell(\Theta)$  Converges

---

of  $\Theta$ ,

$$\begin{aligned}
 \ell(\Theta) &= \sum_{m=1}^M \sum_{i \in \mathcal{U}_m} \ln \int p(\mathbf{x}_{mi}, \mathbf{s}_{mi} | \Theta) d\mathbf{s}_{mi} \\
 &+ \sum_{m=1}^M \sum_{i \in \mathcal{L}_m} \ln \int p(\mathbf{x}_{mi}, y_{mi}, z_{mi}, \mathbf{s}_{mi} | \Theta) dz_{mi} d\mathbf{s}_{mi} \\
 &+ \sum_{m=1}^M \sum_k \sum_j \ln \int p(f_{mkj} | \tau_{mkj}) p(\tau_{mkj} | \gamma) d\tau_{mkj} + \sum_j \ln \int p(w_j | u_j) p(u_j | \lambda) du_j
 \end{aligned}$$

where  $\mathcal{L}_m$  and  $\mathcal{U}_m$  respectively index the labeled and unlabeled feature vectors in the  $m$ -th data set, i.e.,  $\mathcal{L}_m \cup \mathcal{U}_m = \{1, 2, \dots, N_m\}$ .

We employ an expectation-maximization (EM) algorithm to maximize  $\ell(\Theta)$ , with  $\{\eta, F_0\}$  and hyper-parameters  $\{\gamma, \lambda\}$  treated as input parameters to the algorithm, determined separately by cross-validation when necessary. The EM algorithm consists of an iteration of E-step and M-step. In the E-step, one computes the conditional moments of latent variables  $\{z_{mi}, \mathbf{s}_{mi}, \boldsymbol{\tau}, \mathbf{u}\}$  given the data and the most recent parameters  $\Theta$ . In the M-step, one calculates the updated model parameters  $\hat{\Theta}$  using the latent variables' moments obtained in E-step. The complete EM algorithm is given in Algorithm 2, with major update equations summarized below. The algorithm requires  $O\left(F_0 \sum_{m=1}^M D_m (F_m + F_0^2)\right)$  scalar products per iteration.

### 5.3.1 Update of Latent Features Distribution

It can be shown that, under the LPM, the marginal distribution of  $\mathbf{x}_{mi}$  is  $\mathcal{N}(\mathbf{F}_m\boldsymbol{\mu} + \mathbf{d}_m, \eta\mathbf{I} + \mathbf{F}_m\boldsymbol{\Sigma}\mathbf{F}_m^T)$ , with the mean and covariance matrix defined dublicately by  $(\boldsymbol{\mu}, \mathbf{d}_m)$  and  $(\mathbf{F}_m, \boldsymbol{\Sigma})$ , respectively. Similar situations exist for  $z_{mi}$ . To void duplicat-  
edness, one may wish to set  $\boldsymbol{\mu} = \mathbf{0}$ ,  $\boldsymbol{\Sigma} = \mathbf{I}$ , and do not update them during learning.  $g_{cdf}$ ,  $g_{pdf}$  are the *c.d.f* and *p.d.f* of  $\mathcal{N}(0, 1)$ , respectively.

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= \frac{1}{n_a} \sum_{m=1}^M \sum_{i=1}^{N_m} \phi_{mi} \\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{n_a} \sum_{m=1}^M \sum_{i=1}^{N_m} ((\phi_{mi} - \boldsymbol{\mu})(\phi_{mi} - \boldsymbol{\mu})^T + \mathbf{R}_m \mathbf{w} \beta_{mi} \mathbf{w}^T \mathbf{R}_m + \mathbf{R}_m)\end{aligned}\quad (5.17)$$

where  $n_a = \sum_{m=1}^M N_m$ .

$$\phi_{mi} = \mathbf{R}_m (\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \mathbf{w}(\xi_{mi} - b) + \eta^{-1} \mathbf{F}_m^T (\mathbf{x}_{mi} - \mathbf{d}_m)),$$

$$\mathbf{R}_m = (\boldsymbol{\Sigma}^{-1} + \mathbf{w} \mathbf{w}^T + \eta^{-1} \mathbf{F}_m^T \mathbf{F}_m)^{-1},$$

$$\beta_{mi} = \begin{cases} \rho_{mi}, & \text{if } i \in \mathcal{U}_m, \\ (\zeta_{mi}^2 + \rho_{mi}) g_{cdf}(\frac{\zeta_{mi}}{\sqrt{\rho_{mi}}}) \\ + \zeta_{mi} \sqrt{\rho_{mi}} g_{pdf}(\frac{\zeta_{mi}}{\sqrt{\rho_{mi}}}), & \text{if } i \in \mathcal{L}_m, \end{cases}$$

$$\xi_{mi} = \begin{cases} \zeta_{mi}, & \text{if } i \in \mathcal{U}_m, \\ \zeta_{mi} g_{cdf}(\frac{\zeta_{mi}}{\sqrt{\rho_{mi}}}) \\ + \sqrt{\rho_{mi}} g_{pdf}(\frac{\zeta_{mi}}{\sqrt{\rho_{mi}}}), & \text{if } i \in \mathcal{L}_m, \end{cases}$$

$$\rho_{mi} = 1 + \mathbf{w}^T \mathbf{Q}_m \mathbf{w},$$

$$\zeta_{mi} = \mathbf{w}^T \boldsymbol{\mu} + b + \eta^{-1} \mathbf{w}^T \mathbf{Q}_m \mathbf{F}_m^T (\mathbf{x}_{mi} - \mathbf{F}_m \boldsymbol{\mu} - \mathbf{d}_m),$$

$$\mathbf{Q}_m = (\boldsymbol{\Sigma}^{-1} + \eta^{-1} \mathbf{F}_m^T \mathbf{F}_m)^{-1}.$$

### 5.3.2 Update of Domain Transforms

$$\begin{aligned}\hat{\mathbf{d}}_m &= \frac{1}{N_m} \sum_{i=1}^{N_m} (\mathbf{x}_{mi} - \mathbf{F}_m \boldsymbol{\phi}_{mi}), \\ \hat{\mathbf{f}}_{mk} &= \mathbf{V}_{mk} (\alpha \mathbf{I}_{F_0} + \mathbf{V}_{mk} \boldsymbol{\Gamma}_{m1} \mathbf{V}_{mk})^{-1} \mathbf{V}_{mk} \times \sum_{i=1}^{N_m} \boldsymbol{\phi}_{mi}^T (x_{mik} - \hat{d}_{mk}),\end{aligned}\quad (5.18)$$

for  $k = 1, 2, \dots, F_0$  and  $m = 1, 2, \dots, M$ , where  $\alpha = \eta\sqrt{\gamma}$  is a regularization parameter and

$$\begin{aligned}\boldsymbol{\Gamma}_{m1} &= \sum_{i=1}^{N_m} (\boldsymbol{\phi}_{mi} \boldsymbol{\phi}_{mi}^T + \mathbf{R}_m + \beta_{mi} \mathbf{R}_m \mathbf{w} \mathbf{w}^T \mathbf{R}_m), \\ \mathbf{V}_{mk} &= \text{diag}(\sqrt{|f_{mk1}|}, \sqrt{|f_{mk2}|}, \dots, \sqrt{|f_{mkF_0}|}).\end{aligned}$$

### 5.3.3 Update of Probit Classifier

$$\begin{aligned}\hat{\mathbf{w}} &= \mathbf{G} (\vartheta \mathbf{I} + \mathbf{G} \boldsymbol{\Gamma}_1 \mathbf{G})^{-1} \mathbf{G} \sum_{m=1}^M \sum_{i \in \mathcal{L}_m} \boldsymbol{\phi}_{mi} (\xi_{mi} - b), \\ \hat{b} &= \frac{1}{\sum_{m=1}^M N_m} \sum_{m=1}^M \sum_{i \in \mathcal{L}_m} (\xi_{mi} - \boldsymbol{\phi}_{mi}^T \hat{\mathbf{w}}),\end{aligned}\quad (5.19)$$

where  $\vartheta = \sqrt{\lambda}$  is another regularization parameter and

$$\begin{aligned}\boldsymbol{\Gamma}_2 &= \sum_{m=1}^M \sum_{i \in \mathcal{L}_m} (\boldsymbol{\phi}_{mi} \boldsymbol{\phi}_{mi}^T + \mathbf{R}_m + \mathbf{R}_m \mathbf{w} \beta_{mi} \mathbf{w}^T \mathbf{R}_m), \\ \mathbf{G} &= \text{diag}(\sqrt{|w_1|}, \sqrt{|w_2|}, \dots, \sqrt{|w_{F_0}|}).\end{aligned}$$

## 5.4 Experimental Results

### 5.4.1 Cancer Diagnosis

We first consider the two Wisconsin breast cancer datasets (original and diagnostic) from the UCI machine learning repository (<http://archive.ics.uci.edu/ml/>).

Table 5.1: The performance of the LPM on Wisconsin breast cancer data with  $\alpha = \eta\sqrt{\gamma}$  and  $\vartheta = \sqrt{\lambda}$  taking various values. The numbers shown are the improvements in AUC (%) relative to STL, averaged over 50 independent runs.

	LABEL=50				LABEL=100				LABEL=150			
	$\vartheta = 0$	$\vartheta = 0.1$	$\vartheta = 1$	$\vartheta = 10$	$\vartheta = 0$	$\vartheta = 0.1$	$\vartheta = 1$	$\vartheta = 10$	$\vartheta = 0$	$\vartheta = 0.1$	$\vartheta = 1$	$\vartheta = 10$
$\alpha = 0$	1.69	1.69	1.74	1.81	-0.28	-0.27	-0.24	-0.17	-0.68	-0.67	-0.64	-0.58
$\alpha = 0.01$	2.40	2.37	2.40	2.12	0.51	0.52	0.57	0.49	0.20	0.20	0.23	0.19
$\alpha = 0.05$	2.42	2.43	2.51	1.95	0.67	0.68	0.73	0.66	0.36	0.36	0.39	0.38
$\alpha = 0.1$	2.42	2.44	2.55	1.97	0.74	0.75	0.79	0.74	0.40	0.40	0.43	0.44
$\alpha = 0.5$	2.35	2.37	2.49	2.09	0.70	0.71	0.74	0.66	0.42	0.43	0.46	0.48
$\alpha = 1$	1.73	1.78	1.95	1.34	0.40	0.42	0.49	0.55	0.26	0.27	0.29	0.27
$\alpha = 5$	2.51	2.51	2.57	2.00	0.74	0.74	0.79	0.72	0.41	0.41	0.43	0.42
$\alpha = 10$	-1.36	-1.17	-0.90	-0.46	-2.23	-2.13	-2.03	-1.68	-1.71	-1.70	-1.66	-1.56

The objective of both tasks is to identify benign or malignant cells. The feature dimensionality is 9 for the original data and 30 for the diagnostic data. We set  $F_0$  to the smallest dimensionality among the tasks to favor error reduction (as suggested by (5.8)), and  $\eta = 10^{-3}$  to enlarge the role of domain transforms in connecting the tasks, with the regularization parameters  $(\alpha, \vartheta)$  determined via cross-validation (the robustness to these parameters is shown below). We perform both multitask learning and transfer learning experiments, and compare the LPM to STL and the methods in (Wang and Mahadevan, 2011) (abbreviated as HDAMA), (Maayan and Mannor, 2011), and (Kulis *et al.*, 2011), with all competing methods using standard probit classifiers. The method in (Maayan and Mannor, 2011) cannot perform MTL and is excluded in the comparisons on MTL. The performance is measured in terms of the area under ROC curve (AUC), as a function of the number of labeled examples per task in the MTL case, or the number of labeled examples in the target task in the transfer learning case. The results are averaged over 50 independent runs, each constituting an independent split of the data into training sets and test sets.

Figure 5.2(a) shows that, for MTL, the LPM performs comparably as or slightly better than HDAMA and both outperform the other methods, especially when labeled data are scarce. In transfer learning, all data in the source domain are labeled,

and we have only a few labeled data in the target domain. We transfer all the labeled data from the source domain to the target domain. Figure 5.2(b-c) show that the performance of the LPM is slightly better than HDAMA, probably due to the fact that the amount of data (labeled and unlabeled) is balanced between the two tasks.

The regularization parameters  $\alpha$  and  $\vartheta$  control the sparsity of domain transforms and the classifier, respectively. Table 5.1 summarizes the performance of the LPM relative to STL, under a wide range of settings for these parameters. The importance of sparsity is indicated by the diminishing performance improvements as the regularization parameters approach zero. Over a wide range in the middle, the LPM maintains stable performance improvements over STL, indicating the learning is robust to the settings of regularization parameters. The table also shows that the sparsity of domain transforms plays a more prominent role in influencing the performance than the classifier itself, signaling that the benefit of sharing information among the tasks can outweigh the benefit of feature selection.

#### 5.4.2 Mine Detection

The land-mine detection problem<sup>1</sup> (Xue *et al.*, 2007) is based on airborne synthetic-aperture radar (SAR) data and the underwater mine detection problem<sup>2</sup> (Liu *et al.*, 2009b) is based on synthetic-aperture sonar (SAS) data. Here we solve these two problems together, using the proposed cross-domain multitask learning approach. The feature dimensionality of land-mine data is 9 and that of underwater mine data is 13, and the labels do not have the same exact meaning for the two problem domains. There are a total of 19 land-mine tasks and 8 underwater mine tasks. The number of data points in the underwater mine tasks ranges from 756 to 3562, which is much larger than that for the land-mine tasks (ranging from 445 to 454).

---

<sup>1</sup> The land mine dataset is available at <http://www.ee.duke.edu/~lcarin/LandmineData.zip>

<sup>2</sup> The underwater mine dataset is available at <http://www.ece.duke.edu/~lcarin/UnderwaterMines.zip>

This problem can be viewed as a multitask learning across heterogeneous input and output domains (although the labels have known correspondence). We consider 9 land-mine tasks and all 8 underwater tasks, pairing them up to form  $9 \times 8 = 72$  MTL problems. The results are reported as an average over the 72 problems, with the setting of  $F_0$  and regularization parameters based on the same rule as in Section 5.4.1.

The performance comparisons for multi-task learning are shown in Figure 5.3(a) in terms of average AUC. Each curve results from an average of 100 independent runs of independently splitting the data into training and test sets and  $9 \times 8$  combinations of underwater tasks versus land-mine tasks. In the transfer learning case, 50 labeled samples together with all other unlabeled samples are transferred to the target domain. The performance on the target task is shown in Figure 5.3(b-c). It is seen that the LPM outperforms all other methods by significantly large margins, in both multi-task learning and transfer learning from land-mine data to underwater mine data. The competition on transfer learning from underwater mine data to land-mine data is more intense, but the LPM still gives the best overall outperformance.

While the amount of examples is balanced between the two Wisconsin tasks, it is highly unbalanced between the land-mine tasks and the underwater mine tasks (as detailed above). The results indicate that the LPM is more robust to this unbalance than the other methods.

#### *Homogeneous Multitask Learning*

We also conducted symmetric multitask learning (SMTL) experiments on the land-mine datasets (Xue *et al.*, 2007). There are total of 19 tasks in the homogeneous space. We examine the performance of three methods on accuracy of label prediction: (i) cross-domain multitask learning using latent probit models, (ii) single task learning plus separate probit classifiers, (iii) simply pooling the data in all tasks and

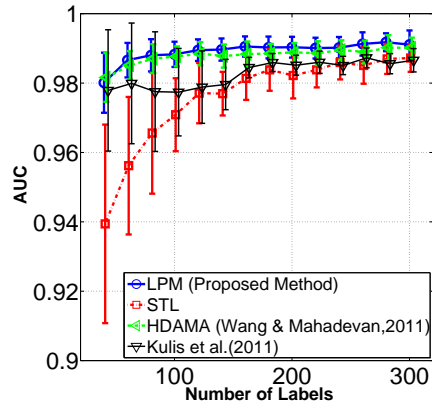
then learning a single probit classifier.

The performance is measured by average AUC (Area under Roc curve) on 19 tasks, the number of training samples for every task is set as 20, 40, ..., 300. For each task, the training samples are randomly chosen from the corresponding data set and the remaining samples are used for testing.

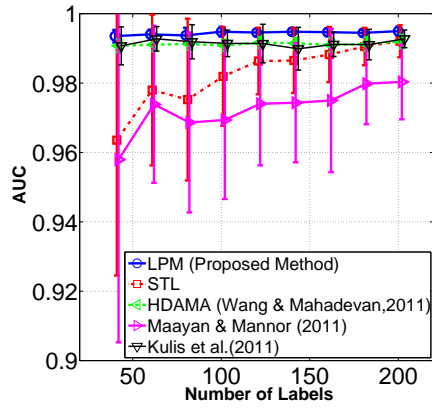
It is a multitask learning problem in homogeneous setting. As is shown in Figure. 5.4, our performance is comparable to (Xue *et al.*, 2007), without using the correspondence information of features among 19 tasks.

## 5.5 Discussion

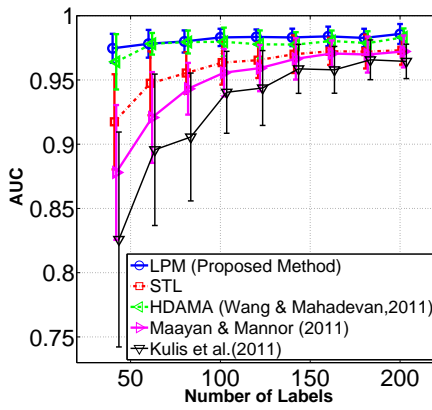
We have proposed the LPM model for cross-domain multi-task learning, assuming heterogenous feature representations across the tasks. The benefit of MTL in the LPM is based on the tasks' relatedness in the latent feature space, which is characterized by the sparse domain transforms. By promoting sparseness of domain transforms and the common classifiers, information sharing is encouraged to the advantage of improving performance in each individual task. The importance of sparsity is demonstrated by both theoretical analysis and experimental results.



(a)

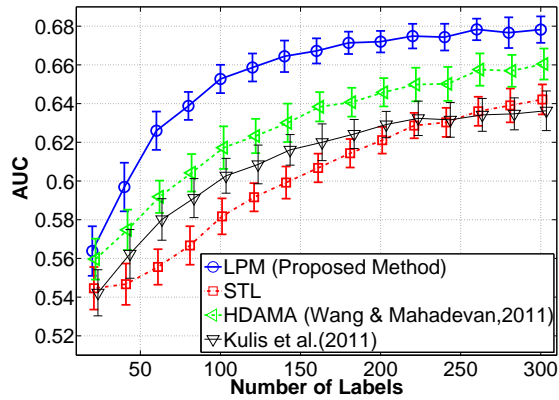


(b)

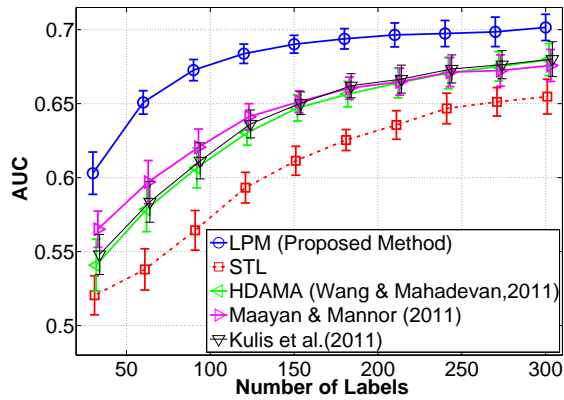


(c)

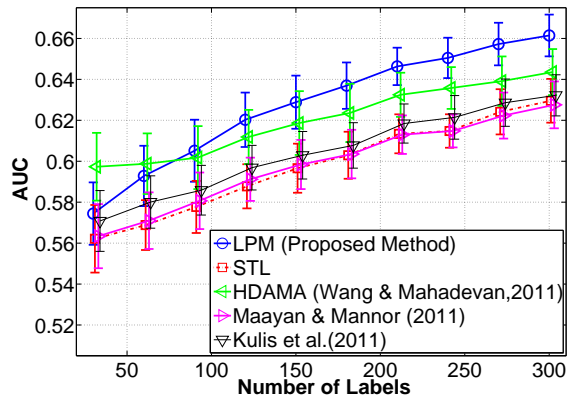
FIGURE 5.2: A comparison of performance on the Wisconsin breast cancer data; (a) multitask learning; (b) transfer learning with the original data as the source domain and the diagnostic data as the target domain; (c) transfer learning with the diagnostic data as the source domain and the original data as the target domain.



(a)



(b)



(c)

FIGURE 5.3: A comparison of performance on the the land-mine/underwater mine detection problem; (a) multitask learning; (b) transfer learning with land-mine data as the source domain and underwater mine data as the target domain; (c) transfer learning with underwater mine data as the source domain and land-mine data as the target domain.

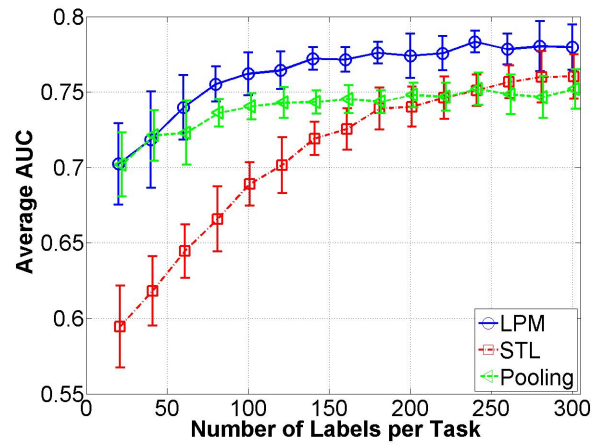


FIGURE 5.4: Average AUC on 19 tasks in the landmine detection problem (20 independent runs)

## Conclusions

### 6.1 Summary of Contributions

This dissertation has addressed the problems of multitask learning, Bayesian hierarchical modeling, and variational inference with dependency structures by providing advances in the following key aspects:

- We have provided a unified variational copula (VC) inference framework on constructing and optimizing variational proposals in hierarchical Bayesian models. The optimal variational posterior under Sklar’s representation is found by minimizing the KL divergence to the true posterior. Classical methods, such as mean-field VB and the VG approximation are special cases of the proposed VC inference framework.
- We have moved away from the common factorization assumption in variational inference and developed dependent new variational inference methods based on hierarchical mixtures or copulas. As a reliable and efficient alternative to MCMC, our INF-VB method generalizes INLA to non-Gaussian priors and mean-field VB to non-factorization settings. The hierarchical mixture-based

proposal overcomes the potential non-conjugacy or multimodal issues in traditional variational inference, and it also allows parallel implementation with the dominant computational load distributed on each grid point, which is particularly important when applying INF-VB to large-scale Bayesian inference.

- For models with continuous and non-Gaussian hidden variables, we propose a semiparametric and automated variational Gaussian copula (VGC) method, in which the parametric Gaussian copula family is able to preserve multivariate posterior dependence. To avoid the difficulty of specifying marginals for hidden variables, the proposed nonparametric transformations based on Bernstein polynomials provide ample flexibility in characterizing the univariate marginal posteriors. Compared with the hierarchical mixture-based approach, VGC does not need to discretize the space for non-Gaussian variables and thus does not suffer from the limits on the number of hyperparameters.
- We have also proposed a dynamic factor model for ordinal time series analysis. For the application of dynamic topic modeling, our model is able to discover topic prevalence over time, and reveal contemporary dependence structures, providing topic and word correlations as a byproduct. The model is semi-parametric with time-evolving ordinal observations well accommodated through the adopted extended rank likelihood, and the latent state space model naturally admits heavy-tailed innovations, capable of inferring abrupt temporal jumps in the importance of topics. The performance of the model is illustrated on simulated dataset and two real datasets.
- We have also studied the problem of learning multiple tasks across heterogeneous domains. This is a challenging problem since the feature space may not be the same for different tasks. We assume the data in multiple tasks are

generated from a latent common domain via sparse domain transforms and propose a hierarchical model to jointly learn the domain transforms, and a probit classifier shared in the common domain. To learn meaningful task relatedness and avoid over-fitting in classification, we introduce sparsity in the domain transforms and the classifier parameters. We derive oracle inequalities for the estimation error of the classifier parameters in terms of the sparsity of domain transform matrices. The effectiveness of the model is demonstrated on real-world applications including cancer diagnosis and mine detection.

## 6.2 Future Directions

This dissertation motivates several possible directions for future research.

1. It is an ongoing challenge to develop scalable deterministic inference methods, to both achieve the accuracy of MCMC methods and retain their inferential flexibility. For example, the inference of nonlinear, non-Gaussian state-space models requires approximating a sequence of probability distributions defined on a sequence of (measurable) spaces, which has relied on sequential Monte Carlo (SMC) methods for a long time. Besides the approaches we proposed in Chapter 2 and Chapter 3, an interesting prospect for future research is to see if we can develop sequential variational approximations, to deal with high dimensionality and complex patterns of dependence with similar sprits of SMC.
2. The semiparametric Bayesian approach alleviates model misspecification and offers more modeling flexibility. It naturally handles data with arbitrary marginal distributions and still enjoys conjugate inference. This dissertation has demonstrated its utility for dynamic topic modeling in Chapter 4. It is interesting to explore other semiparametric models with information-theoretic guarantees, in

which the redundant information is selectively ignored without influence the goals of interest.

3. The increased availability of multiple heterogenous datasets brings opportunities of integrative knowledge discovery through joint analysis of data from heterogeneous sources. For example, omics data may contain different data types including transcriptomics, metabolomics, genomics, etc. Along the directions of research introduced in Chapter 5 on heterogeneous multitask learning and transfer learning, an interesting line of research is to explore if we could build new statistical models with cross-domain dependency structures in life science applications, to better represent the observed data and encapsulate the information for sharing, with the goal of improving the predictive performance.

# Bibliography

- Ahmed, A. and Xing, E. P. (2010). Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Uncertainty in Artificial Intelligence*.
- Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **44**(2), 139–177.
- Aitchison, J. and Brown, J. A. (1957). *The lognormal distribution with special reference to its uses in economics*. Cambridge Univ. Press.
- Aitchison, J. and Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, **76**(4), 643–653.
- Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association (JASA)*, **88**(422), 669–679.
- Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **36**(1), pp. 99–102.
- Argyriou, A., Evgeniou, T., and Pontil, M. (2007). Multi-task feature learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 41–48.
- Armagan, A., Clyde, M., and Dunson, D. B. (2011). Generalized Beta mixtures of Gaussians. In *Advances in Neural Information Processing Systems (NIPS)*.
- Bakker, B. and Heskes, T. (2003). Task clustering and gating for Bayesian multitask learning. *Journal of Machine Learning Research (JMLR)*, **4**, 83–99.
- Barber, D. and Bishop, C. M. (1998). Ensemble learning in Bayesian neural networks. *Neural networks and machine learning*, **168**, 215–238.
- Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research*, **12**(149-198), 3.

- Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London.
- Beal, M. J. and Ghahramani, Z. (2002). The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics 7: Proceedings of the 7th Valencia International Meeting*, page 453.
- Ben-David, S. and Borbely, R. S. (2008). A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine Learning*, **73**(3), 273–287.
- Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, **98**(2), 291–306.
- Bickel, P. J., Ritov, Y., and Tsybakov, B. T. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc.
- Blei, D. M. and Lafferty, J. D. (2006a). Correlated topic models. In *Advances in Neural Information Processing Systems (NIPS)*.
- Blei, D. M. and Lafferty, J. D. (2006b). Dynamic topic models. In *International Conference on Machine Learning (ICML)*.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, **3**, 993–1022.
- Byrne, C. (2009). Bounds on the largest singular value of a matrix and the convergence of simultaneous and block-iterative algorithms for sparse linear systems. *International Transactions in Operational Research*, **16**(4), 465–479.
- Cargnoni, C., Müller, P., and West, M. (1997). Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. *Journal of the American Statistical Association (JASA)*, **92**(438), 640–647.
- Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, **81**, 541–553.
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**(2), 465–480.
- Challis, E. and Barber, D. (2011). Concave Gaussian variational approximations for inference in large-scale Bayesian linear models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 199–207.

- Challis, E. and Barber, D. (2012). Affine independent variational inference. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2186–2194.
- Challis, E. and Barber, D. (2013). Gaussian Kullback-Leibler approximate inference. *Journal of Machine Learning Research (JMLR)*, **14**(1), 2239–2286.
- Chib, S. and Winkelmann, R. (2001). Markov chain Monte Carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, **19**(4).
- Christian, A., Jens, B., and Markus, P. (2014). Bayesian analysis of dynamic factor models: an ex-post approach towards the rotation problem. Kiel Working Papers 1902, Kiel Institute for the World Economy.
- Cseke, B. and Heskes, T. (2011). Approximate marginals in latent Gaussian models. *Journal of Machine Learning Research (JMLR)*, **12**, 417–454.
- Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer-Verlag.
- Doucet, A., Nando, D. F., and Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Springer.
- Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the  $\ell_1$ -ball for learning in high dimensions. In *International Conference on Machine Learning (ICML)*, pages 272–279.
- Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research (JMLR)*, **12**, 2121–2159.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**, 407–499.
- Fan, K., Wang, Z., Beck, J., Kwok, J., and Heller, K. (2015). Fast second-order stochastic backpropagation for variational inference. In *arXiv:1509.02866*.
- Ferkingstad, E. and Rue, H. (2015). Improving the INLA approach for approximate bayesian inference for latent Gaussian models. *arXiv:1503.07307*.
- Figueiredo, M. A. T. (2003). Adaptive sparseness for supervised learning. *IEEE Trans. Pattern Anal. Machine Intelligence*, **25**(9), 1150–1159.
- Foulds, J., Boyles, L., Dubois, C., Smyth, P., and Welling, M. (2013). Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In *19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.
- Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Times Series Analysis*, **15**, 183–202.

- Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman & Hall Texts in Statistical Science Series. Taylor & Francis.
- Gao, C. and Engelhardt, B. E. (2012). A sparse factor analysis model for high dimensional latent spaces. In *NIPS: Workshop on Analysis Operator Learning vs. Dictionary Learning: Fraternal Twins in Sparse Modeling*.
- Gershman, S. J., Hoffman, M. D., and Blei, D. M. (2012). Nonparametric variational inference. In *International Conference on Machine Learning (ICML)*.
- Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies*, **9**(2), 557–587.
- Ghahramani, Z. and Beal, M. J. (2001). Propagation algorithms for variational Bayesian learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 507–513.
- Ghosh, J. and Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, **18**(2), 306–320.
- Golub, G. H. and Loan, C. V. (1996). *Matrix Computations(Third Edition)*. Johns Hopkins University Press.
- Hall, P., Pham, T., Wand, M. P., and Wang, S. S. J. (2011). Asymptotic normality and valid inference for gaussian variational approximation. *Ann. Statist.*, **39**(5), 2502–2532.
- Han, S., Liao, X., and Carin, L. (2012). Cross-domain multitask learning with latent probit models. In *International Conference on Machine Learning (ICML)*, pages 1463–1470.
- Han, S., Liao, X., and Carin, L. (2013). Integrated non-factorized variational inference. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2481–2489.
- Han, S., Du, L., Salazar, E., and Carin, L. (2014). Dynamic rank factor model for text streams. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2663–2671.
- Han, S., Liao, X., Dunson, D. B., and Carin, L. (2016). Variational gaussian copula inference. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Hans, C. (2009). Bayesian Lasso regression. *Biometrika*, **96**(4), 835–845.

- He, J. and Rick, L. (2011). A graph-based framework for multi-task multi-view learning. In *International Conference on Machine Learning (ICML)*, pages 25–32.
- Hensman, J., Rattray, M., and Lawrence, N. D. (2012). Fast variational inference in the conjugate exponential family. In *Advances in Neural Information Processing Systems (NIPS)*.
- Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Statist.*, **1**(1), 265–283.
- Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer.
- Hoffman, M. D. and Blei, D. M. (2015). Structured stochastic variational inference. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Honkela, A., Raiko, T., Kuusela, M., Tornio, M., and Karhunen, J. (2010). Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *Journal of Machine Learning Research (JMLR)*, **11**, 3235–3268.
- Inouye, D., Ravikumar, P., and Dhillon, I. (2014). Admixture of Poisson MRFs: A topic model with word dependencies. In *International Conference on Machine Learning (ICML)*.
- Jaakkola, T. S. and Jordan, M. I. (1998). Improving the mean field approximation via the use of mixture distributions. In *Learning Graphical Models*.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999a). An introduction to variational methods for graphical models. In *Learning in graphical models*, pages 105–161, Cambridge, MA. MIT Press.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999b). An introduction to variational methods for graphical models. *Machine learning*, **37**(2), 183–233.
- Kaganovsky, Y., Han, S., Degirmenci, S., Politte, D. G., Brady, D. J., O’Sullivan, J. A., and Carin, L. (2015a). Alternating minimization algorithm with automatic relevance determination for transmission tomography under Poisson noise. *SIAM Journal on Imaging Sciences*, **8**(3), 2087–2132.
- Kaganovsky, Y., Degirmenci, S., Han, S., Odina, I., Politte, D. G., Brady, D. J., O’Sullivan, J. A., and Carin, L. (2015b). Alternating minimization algorithm with iteratively reweighted quadratic penalties for compressive transmission tomography. In *SPIE Medical Imaging*.
- Kalaitzis, A. and Silva, R. (2013). Flexible sampling of discrete data correlations without the marginal distributions. In *Advances in Neural Information Processing Systems (NIPS)*.

- Kass, R. E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association (JASA)*, **84**(407), 717–726.
- Khan, M. E., Mohamed, S., and Murphy, K. P. (2012). Fast Bayesian inference for non-conjugate Gaussian process regression. In *Advances in Neural Information Processing Systems (NIPS)*.
- Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Korattikara, A., Chen, Y., and Welling, M. (2014). Austerity in MCMC land: cutting the Metropolis-Hastings budget. In *International Conference on Machine Learning (ICML)*, pages 181–189.
- Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. M. (2015). Automatic variational inference in Stan. In *arXiv:1506.03431*.
- Kulis, B., Saenko, K., and Darrell, T. (2011). What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1785–1792.
- Lawrence, E., Bingham, D., Liu, C., and Nair, V. N. (2008). Bayesian inference for multivariate ordinal data using parameter expansion. *Technometrics*, **50**(2).
- Li, J., Nott, D. J., Fan, Y., and Sisson, S. A. (2015). Extending approximate Bayesian computation methods to high dimensions via Gaussian copula. *arXiv:1504.04093*.
- Lin, L. and Dunson, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika*, **101**(2), 303–317.
- Liu, C., Rubin, D. B., and Wu, Y. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, **85**(4), 755–770.
- Liu, H., Lafferty, J., and Wasserman, L. (2009a). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research (JMLR)*, **10**, 2295–2328.
- Liu, J. S. and Wu, Y. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association (JASA)*, **94**(448), 1264–1274.
- Liu, Q., Liao, X., Li, H., Stack, J. R., and Carin, L. (2009b). Semi-supervised multitask learning. *IEEE Trans. Pattern Anal. Machine Intelligence*, **31**(6), 1074–1086.
- Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, **14**(1), 41–68.

- Lopez-Paz, D., Hernandez-Lobato, J. M., and Ghahramani, Z. (2013). Gaussian process vine copulas for multivariate dependence. In *International Conference on Machine Learning (ICML)*, pages 10–18.
- Lounici, K., Pontil, M., Tsybakov, A. B., and de Geer, S. V. (2009). Taking advantage of sparsity in multi-task learning. In *Proceedings of the 22nd Conference on Information Theory*, pages 73–82.
- Maayan, H. and Mannor, S. (2011). Learning from multiple outlooks. In *ICML*, pages 401–408.
- MacKay, D. J. (1992). Bayesian interpolation. *Neural computation*, **4**(3), 415–447.
- Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In J. S. Breese and D. Koller, editors, *Uncertainty in Artificial Intelligence*, pages 362–369.
- Mnih, A. and Gregor, K. (2014). Neural variational inference and learning in belief networks. In *International Conference on Machine Learning (ICML)*, pages 1791–1799.
- Møller, J. and Waagepetersen, R. P. (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, **34**(4), 643–684.
- Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association (JASA)*, **108**(502), 656–665.
- Neal, R. M. and Hinton, G. E. (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer.
- Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.
- Neville, S. E., Ormerod, J. T., and Wand, M. (2014). Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electronic Journal of Statistics*, **8**, 1113–1151.
- Nguyen, T. V. and Bonilla, E. V. (2014). Automated variational inference for Gaussian process models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1404–1412.
- Opper, M. and Archambeau, C. (2009a). The variational Gaussian approximation revisited. *Neural computation*, **21**(3), 786–792.
- Opper, M. and Archambeau, C. (2009b). The variational Gaussian approximation revisited. *Neural Computation*, **21**(3), 786–792.

- Ormerod, J. T. (2011). Skew-normal variational approximations for Bayesian inference. *Technical Report CRG-TR-93-1, School of Mathematics and Statistics, Univeristy of Sydney*.
- Paisley, J. W., Blei, D. M., and Jordan, M. I. (2012). Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning (ICML)*.
- Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association (JASA)*, **103**(482), 681–686.
- Petrone, S. (1999). Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics*, **27**(1), 105–126.
- Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **44**(2), 234–243.
- Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, **7**(4), 887–902.
- Qi, Y. and Jaakkola, T. S. (2006). Parameter expanded variational Bayesian methods. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1104.
- Quiroz, M., Villani, M., and Kohn, R. (2014). Speeding up MCMC by efficient data subsampling. *arXiv:1404.4178*.
- Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Rasmussen, C. E. (2004). *Gaussian processes in machine learning*. Springer.
- Reis, E. A., Salazar, E., and Gamerman, D. (2006). Comparison of sampling schemes for dynamic linear models. *International Statistical Review*, **74**(2), 203–214.
- Reisinger, J., Waters, A., Silverthorn, B., and Mooney, R. J. (2010). Spherical topic models. In *International Conference on Machine Learning (ICML)*.
- Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, pages 1278–1286.
- Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the gridly-Gibbs sampler. *Journal of the American Statistical Association (JASA)*, **87**(419), pp. 861–868.
- Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc.

- Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(2), 319–392.
- Saul, L. K. and Jordan, M. I. (1996). Exploiting tractable substructures in intractable networks. *Advances in Neural Information Processing Systems (NIPS)*, pages 486–492.
- Schmidl, D., Czado, C., Hug, S., and Theis, F. J. (2013). A vine-copula based adaptive MCMC sampler for efficient inference of dynamical systems. *Bayesian Analysis*, **8**(1), 1–22.
- Schmidt, M. (2012). minfunc: unconstrained differentiable multivariate optimization in matlab.
- Seeger, M. W. and Nickisch, H. (2011). Large scale Bayesian inference and experimental design for sparse linear models. *SIAM Journal on Imaging Sciences*, **4**(1), 166–199.
- Sklar, A. (1959). *Fonctions de Répartition à n Dimensions Et Leurs Marges*. Publ. Inst. Statist. Univ. Paris 8.
- Smith, M. S. (2013). Bayesian approaches to copula modelling. *Bayesian Theory and Applications*, page 336.
- Song, P. X. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, **27**(2), 305–320.
- Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freha, F., Redwine, E., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *Journal of Urology*, **16**, pp. 1076–1083.
- Talhouk, A., Doucet, A., and Murphy, K. (2012). Efficient Bayesian inference for multivariate probit models with sparse inverse correlation matrices. *Journal of Computational and Graphical Statistics*, **21**(3), 739–757.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58**, 267–288.
- Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association (JASA)*, **81**, 82–86.
- Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research (JMLR)*, **1**, 211–244.

- Titsias, M. and Lázaro-Gredilla, M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning (ICML)*, pages 1971–1979.
- Tran, D., Blei, D. M., and Airoldi, E. M. (2015). Variational inference with copula augmentation. In *arXiv:1506.03159*.
- Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, **1**(1-2), 1–305.
- Wang, C. and Blei, D. M. (2012). Truncation-free online variational inference for Bayesian nonparametric models. In *Advances in Neural Information Processing Systems (NIPS)*.
- Wang, C. and Mahadevan, S. (2011). Heterogeneous domain adaptation using manifold alignment. In *International Joint Conference on Artificial Intelligence*, pages 1541–1546.
- Wauthier, F. L. and Jordan, M. I. (2010). Heavy-tailed process priors for selective shrinkage. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2406–2414.
- West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association (JASA)*, **80**(389), 73–83.
- Xue, Y., Liao, X., Carin, L., and Krishnapuram, B. (2007). Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research (JMLR)*, **8**, 35–63.

# Biography

Shaobo Han was born on September 27, 1985, and was raised in Lintsing, Shandong Province, China. He received the B.S. degree in Electrical Engineering from Xidian University, Xi' an, China, in 2007, and the M.S. degree in Signal and Information Processing from the Chinese Academy of Sciences, Beijing, China, in 2010. Beginning August 2010, He has been with the Department of Electrical and Computer Engineering at Duke University, Durham, NC, where he is currently studying toward the Ph.D. degree in Electrical and Computer Engineering, under the guidance of Professor Lawrence Carin. His research interests include machine learning and Bayesian statistics.

## List of Publications

- **Shaobo Han**, Xuejun Liao, David B. Dunson and Lawrence Carin. “Variational Gaussian Copula Inference”, In Proceedings of the 19th *International Conference on Artificial Intelligence and Statistics*, (AISTATS 2016), Cadiz, Spain, JMLR: W&CP volume 51, 2016
- Yan Kaganovsky, **Shaobo Han**, Soysal Degirmenci, David G. Politte, David J. Brady, Joseph A. O’Sullivan, and Lawrence Carin. “Alternating Minimization Algorithm with Automatic Relevance Determination for Transmission Tomography under Poisson Noise”, *SIAM Journal on Imaging Sciences (SIIMS)*, Vol. 8, No. 3, (2015), 2087-2132

- **Shaobo Han\***, Lin Du\*, Esther Salazar and Lawrence Carin. “Dynamic Rank Factor Model for Text Streams”, In Advances in *Neural Information Processing Systems* 27 (NIPS 2014), pp. 2663-2671, Montreal, Canada, 2014
- **Shaobo Han**, Xuejun Liao and Lawrence Carin. “Integrated Non-Factorized Variational Inference”, In Advances in *Neural Information Processing Systems* 26 (NIPS 2013), pp. 2481-2489, Lake Tahoe, NV, USA, 2013
- **Shaobo Han**, Xuejun Liao and Lawrence Carin. “Cross-Domain Multitask Learning with Latent Probit Models”, In Proceedings of the 29th *International Conference on Machine Learning* (ICML 2012), pp. 1463-1470, Edinburgh, Scotland, UK, 2012
- Yan Kaganovsky, Soysal Degirmenci, **Shaobo Han**, Ikenna Odinaka, David G. Politte, David J. Brady, Joseph A. O’Sullivan, and Lawrence Carin. “Alternating Minimization Algorithm with Iteratively Reweighted Quadratic Penalties for Compressive Transmission Tomography”, *SPIE Medical Imaging Conference*, Orlando, FL, Feb. 2015
- HyungJu Jeon, Yan Kaganovsky, **Shaobo Han** and Lawrence Carin. “GPU-Based Sparse Bayesian Learning for Adaptive Tomographic Imaging”, *IEEE Nuclear Science Symposium & Medical Imaging Conference* (NSS/MIC 2014), Seattle, WA, Nov. 2014
- Xin Yuan, Vinayak Rao, **Shaobo Han** and Lawrence Carin. “Hierarchical Infinite Divisibility for Multiscale Shrinkage”, *IEEE Transactions on Signal Processing*, Vol. 62, No.17(2014) 4363-4374