Bayesian Modeling and Computation for Mixed Data

by

Kai Cui

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
David B. Dunson, Supervisor

_____
Fan Li

_____
Joseph Lucas

_____
Lawrence Carin

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2012

Abstract

Bayesian Modeling and Computation for Mixed Data

by

Kai Cui

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
David B. Dunson, Supervisor

_____
Fan Li

_____
Joseph Lucas

_____
Lawrence Carin

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2012

Abstract

Multivariate or high-dimensional data with mixed types are ubiquitous in many fields of studies, including science, engineering, social science, finance, health and medicine, and joint analysis of such data entails both statistical models flexible enough to accommodate them and novel methodologies for computationally efficient inference. Such joint analysis is potentially advantageous in many statistical and practical aspects, including shared information, dimensional reduction, efficiency gains, increased power and better control of error rates.

This thesis mainly focuses on two types of mixed data: (i) mixed discrete and continuous outcomes, especially in a dynamic setting; and (ii) multivariate or high dimensional continuous data with potential non-normality, where each dimension may have different degrees of skewness and tail-behaviors. Flexible Bayesian models are developed to jointly model these types of data, with a particular interest in exploring and utilizing the factor models framework. Much emphasis has also been placed on the ability to scale the statistical approaches and computation efficiently up to problems with long mixed time series or increasingly high-dimensional heavy-tailed and skewed data.

To this end, in Chapter 1, we start with reviewing the mixed data challenges. We start developing generalized dynamic factor models for mixed-measurement time series in Chapter 2. The framework allows mixed scale measurements in different time series, with the different measurements having distributions in the exponential fam-

ily conditional on time-specific dynamic latent factors. Efficient computational algorithms for Bayesian inference are developed that can be easily extended to long time series. Chapter 3 focuses on the problem of jointly modeling of high-dimensional data with potential non-normality, where the mixed skewness and/or tail-behaviors in different dimensions are accurately captured via the proposed heavy-tailed and skewed factor models. Chapter 4 further explores the properties and efficient Bayesian inference for the generalized semiparametric Gaussian variance-mean mixtures family, and introduce it as a potentially useful family for modeling multivariate heavy-tailed and skewed data.

I dedicate this thesis to

my family, my wife Yuanyuan, my son Eric and my parents

for their constant support and unconditional love.

I love you all dearly.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

DP	Dirichlet Process
DPM	Dirichlet Process Mixtures
FFBS	Forward Filtering Backward Sampling
GDKA	Greedy Density Kernel Approximation
GLTM	Generalized Latent Trait Models
GVMM	Gaussian Variance-Mean Mixtures
KF	Kalman Filters
LARS	Least-Angle Regression
logGN	Generalized log-Normal
MCMC	Markov Chain Monte Carlo
N-(I)FM	Normal Variance-Mean Mixture (Infinite) Factor Models
PCR	Principle Component Regression
PLS	Partial Least Squares Regression
PX	Parameter Expansion
RMSEP	Root of Mean Squared Prediction Error
VAR	Vector Autoregression

Acknowledgements

I would like to express my heartfelt gratitude to my PhD advisor, Professor David Dunson, for his support throughout my doctoral studies. He is someone you will instantly respect as a student and never forget once you meet him. As one of the smartest people I know, I hope that I could be as knowledgeable, energetic and enthusiastic as David and to someday be able to command an audience as well as he can. David has been very supportive and given me the freedom to pursue various projects. I am also very grateful for his in-depth advice, numerous insightful discussions and invaluable suggestions. I really can not think of anything else a student could ask from his advisor, and I feel really lucky and thankful to have David as my PhD advisor. I also have to sincerely thank the members of my PhD committee, Professor Fan Li, Joseph Lucas and Lawrence Carin for their valuable comments and advice in general.

I am deeply thankful to Professor Mike West. Mike was my first research advisor at Duke Statistics and has been really helpful and generous in providing advice many times during my graduate school career. His enthusiasm and love for working with students is contagious.

I thank all the faculty members of the statistical science department for providing such a resourceful and friendly environment. Further thanks go to Professor Fan Li for being my first year advisor, Professor Alan Gelfand, James Berger, Merlise Clyde, Scott Schmidler, Sayan Mukherjee, Jerome Reiter, Ronald Gallant, Rich Durrett

and Michael Brandt for their wonderful lectures, which all have provided me a solid foundation to progress both in my graduate study and future career. I also would like to thank Professor Cliburn Chan in the biostatistics department for providing data and research projects to me to start with in my first year.

A good support system is important to surviving and staying calm in graduate school. I was lucky to be surrounded by many friends and colleagues, who have made my life in graduate school full of joy. Thanks go to Guoxian Zhang as being such a wonderful roommate, and for the numerous discussions. Thanks go to Hao Wang, Zhengzi Li, Junyao Tang, Minghui Shi, Lin Lin, Yajuan Si, Hui Wang, Yanhua Rao and many others as generous and supportive friends, Andrew Cron, Anjishnu Banerjee, Jochi Nakajima for being nice officemates, Fangpo Wang and Yun Yang for being on my last-semester "stress eating"team, and all the graduate students in the statistics department and other friends at Duke for making my life here so colorful.

I am forever grateful to my whole family for their overwhelming love and trust, for always being there for me through thick and thin, and for all the sacrifice and dedication they have made. To them, I dedicate this dissertation.

1

Introduction

1.1 Motivation

Multivariate mixed data occur frequently in many fields of studies. The analysis of such data has created new challenges that have made it necessary to develop new statistical techniques and methodologies. Because correlations between the mixed data are usually of practical interest, particular attention has been given to how such dependence structures can be incorporated in the models. Remarkable advances have been made to meet these challenges over the past two decades, including, among others, extensions of Bayesian methods to mixed data settings, development of factorization and generalized latent variable models, dependence modeling via copula and graphical models, and shared or correlated random effect models to incorporate correlations between mixed outcomes. However, a close scrutiny of the literature reveals that both modeling and computational issues are yet to be resolved in the following mixed data scenarios:

1.1.1 Mixed-Measurement Time Series

These kinds of data streams consisting of either discrete (binary, categorical and counts) or continuous (most likely assumed normal) outcomes are ubiquitous in many applications areas. An effective approach of jointly modeling mixed-measurement time series and thus sharing information is essential to characterizing the dependence among the variables and can improve prediction and interpolation across missing data regions.

As stated previously, in the static setting, there has been a rich history of jointly modeling mixed outcomes involving binary, categorical and continuous variables, using either underlying Gaussian models (Muthen (1984)) or generalized latent trait models (GLTM) that specify GLMs for each response with shared latent factors (Moustaki and Knott (2000) and Dunson (2000)). Underlying Gaussian models assume categorical manifest variables arise by thresholding normal underlying variables. This leads to substantial computational simplifications, as algorithms for Gaussian latent factor models can be easily adapted. However, the price to be paid for computational efficiency is lack of modeling flexibility in that categorical variables will be assigned probit models and count variables cannot be naturally accommodated. Although substantial flexibility can be gained through using mixtures of underlying Gaussian models (Yang and Dunson (2010a)), there is an associated price in terms of computational and modeling complexity.

The GLTM framework represents a compromise in that it is much more flexible than the underlying Gaussian framework, while maintaining a simple parametric form. For example, counts can be characterized via Poisson log-linear models, and categorical variables via logistic regressions. Dunson (2003) applied GLTMs to dynamic modeling of mixed scale multivariate longitudinal data, specifying GLMs for each variable. Time-varying Gaussian latent factors were included, assuming a dis-

crete Markov model. However, substantial computational challenges result outside of the underlying Gaussian framework. In particular in Bayesian analysis, one can no longer specify conditionally-conjugate priors and implement simple Gibbs sampling algorithms. Although adaptive rejection sampling can be used within Gibbs samplers, the sampler chains are commonly poorly behaved due to high posterior dependence in the model parameters leading to slow mixing (Ghosh and Dunson (2009)). Such problems are particularly prevalent in dynamic factor models. Therefore, latent factor models with distributions in the exponential family are much more challenging computationally than the underlying normal model class.

Therefore, a flexible and computationally efficient framework for mixed time series analysis is definitely needed, for which we propose the use of generalized Bayesian dynamic factor model. The framework allows mixed scale measurements in different time series, with the different measurements having distributions in the exponential family conditional on time-specific dynamic latent factors. Efficient computational algorithms for Bayesian inference are also developed, that can be easily extended to long mixed-measurement time series. Since incomplete data are ubiquitous in studies involving mixed data, where little research has been done in this area, adaptation of the models to handle missing data would also be of great importance in practice.

1.1.2 High-Dimensional Data with Potential Non-Normality

High-dimensional continuous data with potential non-normality is another class of mixed data in a broader sense, given that the skewness or tail behaviors may be different in each dimension. In recent years a renewed attention has also been devoted to joint modeling these types of data, due to the increased collection of large volumes of heavy-tailed and/or skewed data in a variety of application such as signal processing, network traffic, gene expression and finance. Several proposals have been put forward in the literature, including skew-symmetric distributions (Wang et al.

(2004)), Gaussian variance-mean mixtures (Barndorff-Nielsen and Sorensen (1982)) and mixtures of multivariate normal or heavy-tailed distributions. Related models have been successfully applied to numerous applications from a wide range of fields, but essentially entirely in univariate or low-dimensional cases.

Extension to high-dimensional problems with potential heavy-tailed and skewed observations is very challenging. On the one hand, it is challenging to directly define multivariate distributions that can flexibly characterize heavy tails and skewness in each dimension while also appropriately characterizing the types of dependence that arise in applications. For example, Gaussian copula models can be unrealistic as there is commonly tail-dependence. Even if a distributional family with sufficient flexibility could be defined, with limited numbers of observations, estimation of high-dimensional model parameters is problematic without dimensionality reduction.

Alternatively, factor models have been widely used to explain the dependence in p-dimensional continuous data via k latent factors with $k \ll p$. Gaussian factor models assume that the latent factor distribution is multivariate normal. This gives rise to a multivariate normal marginal distribution with a sparse decomposition of the $p \times p$ covariance matrix as $\Lambda\Lambda' + \Sigma$, where Λ is a $p \times k$ factor loadings matrix and Σ is a $p \times p$ diagonal matrix with nonnegative diagonal elements. The normal assumption was introduced mainly for convenience and computational efficiency, while Bartholomew (1988) suggested that the assumption is quite robust. However, this robustness was questioned by Hesketh et al. (2003) arguing that inappropriate specifications for latent factor distributions give more errors in the prediction of latent scores, in latent factor regression and lead to misinterpretation of the estimation results. Ma and Genton (2010) also described in the generalized linear latent variable model setting that imposing the normal assumption inappropriately biased the estimates.

Using alternatives to normal distributions for the latent factors to accommodate heavy tails and skewness is needed and has been of growing interest. In signal pro-

cessing, independent component analysis uses a linear combination of non-Gaussian source signals to flexibly characterize a non-Gaussian signal (Comon (1994)). In low-dimensional problems with a small number of factors, Attias (1999) used univariate mixtures of normals for the latent factors. Yung (1997) and later Montanari and Viroli (2010a) also modeled latent factors using mixtures of normals, with Montanari and Viroli (2010b) instead using skew-normals. Motivated by non-Gaussian high-dimensional microarray gene expression data, Carvalho et al. (2008a) modeled the latent factor distributions non-parametrically using Dirichlet process priors. Yang and Dunson (2010b) used a similar idea in structural equation models with latent variable distributions completely unknown and inferred via centered Dirichlet process (CDP) and CDP mixture priors.

Although it is conceptually appealing to use an extremely flexible nonparametric approach, allowing the distribution of the latent factors and data to follow essentially any form, there is a big price to be paid for such flexibility in terms of efficiency. One aspect of efficiency relates to the well known curse of dimensionality in which even the optimal rate of estimation for a p-variate density degrades rapidly with p, necessitating an enormous sample size n for accurate estimation in the absence of constraints. Another aspect is computational efficiency, with computation in mixture models for high-dimensional data quite challenging due in large part to the multimodality that arises and difficulties in adequately exploring possible modes. Also it is often questionable whether such flexibility is needed in most applications in which there may be deviations from normality such as heavy tails and skewness without multimodality.

With this motivation, our focus is on introducing a new framework for Bayesian sparse factor modeling of high-dimensional heavy-tailed and skewed data, with an emphasis on developing an approach that is flexible enough to characterize the majority of non-Gaussian data encountered in practice while also being computationally

tractable to implement and leading to good estimation and prediction performance. The proposed model is a semiparametric Gaussian variance-mean mixture factor model, with the usual Gaussian factor model arising as a limiting case. We generalize the approach of Bhattacharya and Dunson (2011a) for efficient inference under Gaussian factor models with unknown numbers of factors to the heavy-tailed and skewed case.

1.1.3 Low-Dimensional Heavy-tailed and Skewed Data

Low dimensional heavy-tailed and skewed data typically have relative large sample size compared to dimension, and thus require developing new classes of multivariate distributions that flexibly characterize heavy tails and skewness, while accommodating tail dependence. Such tail dependence arises in many applications and is a natural consequence of dependence in outlying events. Such dependence is well known to occur in financial data, communication networks, weather and other settings, but is not adequately characterized by common approaches such as Gaussian copula models. Salmon (2012) provides a compelling commentary on how reliance on a single measure of correlation in two variables based on a Gaussian copula may have played a substantial role in the financial crisis. We need statistical methods based on new classes of distributions that do not rely on such unrealistic assumptions but that are still tractable to apply even in moderate to high-dimensional settings.

There is an existing literature relevant to this topic. Wang et al. (2004) proposed a class of skew-symmetric distributions having probability density functions (pdfs) of the form $2f(\mathbf{x})Q(\mathbf{x})$, where f is a continuous symmetric density and $Q : \Re^n \to [0, 1]$ is a skewing function. Choosing f as normal leads to the skew normal class (Azzalini (1985), Azzalini and Dalla Valle (1996)), with other special cases corresponding to skew-t (Sahu et al. (2003), Gupta (2003)), skew slash (Wang and Genton (2006)) and skew elliptical distributions (Genton and Loperfido (2005)). These parametric models

are useful in providing computationally tractable distributions that have parameters regulating skewness and kurtosis in the data. However, choosing a specific parametric family for f and Q can be challenging in practice, with different choices yielding potentially different results. Although one can potentially conduct model selection or averaging, this adds to the computational burden.

Alternatively, nonparametric approaches have been explored to handle heavy-tailed and skewed observations with more flexibility. For instance, mixtures of normal distributions have been widely used to approximate arbitrary distributions. S. Venturini and Parmigiani (2008) use mixtures of gamma distributions over the shape parameter to model heavy-tailed medical expenditure data. Mixtures of other heavy-tailed distributions have also been proposed. Such nonparametric density estimation approaches face substantial challenges in multivariate cases due to the curse of dimensionality. Fully nonparametric density estimation is almost too flexible in allowing densities that have arbitrary numbers of modes and complex shapes, which are difficult to estimate accurately based on available data in many cases. There has been some attempt to reduce dimensionality in multivariate density estimation using mixtures of factor analyzers (Chen et al. (2010b)) and alternative approaches, but nonetheless the curse is only partly thwarted by such efforts.

We propose Bayesian semiparametric Gaussian variance-mean mixture models, in which the mixing distribution G is modeled nonparametrically to flexibly accommodate heavy-tails and skewness, while letting the data inform about the appropriate distribution choice. Efficient Bayesian computational strategies are developed for reliable inference on parameters.

1.2 Thesis Outline

This thesis is primarily motivated by the mixed data problems presented above. Emphasis is placed on providing both flexible models and computational efficient

algorithms that can be easily scaled up to long time series or high-dimensional data.

Chapter 2 We start developing generalized dynamic factor models for mixed-measurement time series. The framework allows mixed scale measurements in different time series, with the different measurements having distributions in the exponential family conditional on time-specific dynamic latent factors. Efficient computational algorithms based on a Greedy Density Kernel Approximation (GDKA) for Bayesian inference are developed that can be easily extended to long time series.

Chapter 3 focuses on the problem of jointly modeling of high-dimensional data with potential non-normality, where the mixed skewness and/or tail-behaviors in different dimensions are accurately captured via the proposed Gaussian variance-mean mixture factor models. In this framework, the distributions of the independent factors are model separately as coming from the family of Gaussian variance-mean mixtures, many members of which have been widely used in modeling non-Gaussian observations. The degree of flexibility required by many applications of latent factor distributions can be achieved through a semi-parametric specification on the latent factor distributions. Most importantly, we show that even with this additional flexibility, we maintain very efficient Bayesian inference based on Markov Chain Monte Carlo (MCMC) that scales easily up to high-dimensional problems. The efficiency is mainly achieved by the conditionally joint multivariate conjugacy property of the framework, which allows block updating of the loading matrix and latent factors.

Chapter 4 explores the property and efficient Bayesian inference for the more general semiparametric Gaussian variance-mean mixtures family, and introduce it as a potentially useful family for modeling multivariate heavy-tailed and skewed data.

2

Generalized Dynamic Factor Models for Mixed-Measurement Time Series

2.1 Introduction

In this chapter, we consider the modeling and computation of multiple co-evolving time series with mixed-scale measurements within a dynamic factor model framework.

When it comes to multivariate time series analysis, the literature has focused primarily on joint models for multiple time series with normal (or conditionally normal) observations. Latent factor models, among other multiple time series analysis approaches, are useful tools for both dimension reduction and characterizing dependency between multiple time series (Anderson (1963) and Hamann et al. (2005)). In the absence of mixed-measurement data, the latent time-varying factors can be analyzed using either principal component analysis in an approximated dynamic factor framework (see e.g. Stock and Watson (2002), Bai (2003), Bai and Ng (2002) and Mohamed et al. (2008)), frequency domain methods (Boashash (2006)), or filtering and smoothing techniques as in the state-space framework (Aguilar and West (2000)

and Jungbacker and Koopman (2008)). Alternatively, Zhang and Nesselroade (2007) used an underlying Gaussian variable approach to jointly model multiple categorical time series. However, little effort has been made to jointly model mixed-measurement time series.

Motivated by providing a flexible and computationally efficient framework for mixed-measurement time series analysis, we propose the use of generalized Bayesian dynamic factor model. The framework allows mixed scale measurements in different time series, with the different measurements having distributions in the exponential family conditional on time-specific dynamic latent factors. Efficient computational algorithms for Bayesian inference are also developed, which, as shown later, can be easily extended to mixed-measurement time series with missing values and/or mixed-frequency. Briefly, a Metropolis Hastings algorithm with adaptive proposals is developed to efficiently obtain posterior samples of time-varying latent factors. The algorithm is based on a Greedy Density Kernel Approximation (GDKA) via a mixture of normal distributions and further combined with the forward filtering backward sampling algorithm. In addition, parameter expansion with latent factor normalization is proposed to improve the efficiency of sampling model parameters. These steps are further embedded in a higher-level MCMC sampler, which gives efficient updating, improved inference, and prediction.

We introduce the general framework of Bayesian dynamic factor modeling for mixed-measurement time series in Section 2. Section 3 describes computational algorithms developed for Bayesian inference on both latent factors and model parameters. Section 4 applies the model and algorithms to simulated data sets. In section 5 the model is applied to jointly modeling corporate default risk, bond recovery rates and business cycles. Section 6 contains a discussion of the model and results.

2.2 Generalized Dynamic Factor Model Framework

Consider p mixed-measurement time series $\mathbf{y}_t = \{y_{it}\}$ ($i = 1, \ldots, p$, and $t = 1, \ldots, T$), with y_{it} denoting the value of the i^{th} time series at time point t, and the points assumed to be equally-spaced. Each time series can have a different measurement scale, so, for example, y_{1t} may be binary, y_{2t} a count and y_{3t} continuous. We follow the setting of Dunson (2003) and jointly model the mixed-measurement time series $\mathbf{y}_t$ via m common latent dynamic factors $\eta_t = (\eta_{t1}, \ldots, \eta_{tm})'$ ($t = 1, \ldots, T$) and time series-specific observed covariates z_{it}.

Conditional on a factor path $(\eta_1, \ldots, \eta_T)$, the observation y_{it} is modeled as coming from a distribution in the exponential family with canonical parameter θ_{it}:

$$y_{it} \sim f_i(\theta_{it}, \phi_i). \tag{2.1}$$

The canonical parameters are defined by a generalized linear model linked to the latent factors and observed covariates:

$$h_i(\theta_{it}) = \alpha_i + z_{it}'\beta_i + \Lambda_i \eta_t, \tag{2.2}$$

where $h_i(\cdot)$ is the link function for the i^{th} time series, α_i is the intercept term, z_{it} is a $b \times 1$ vector of observed covariates at time t, β_i is a $b \times 1$ parameter vector, η_t is $m \times 1$ latent factor vector at time t, and Λ_i is $1 \times m$ factor loading vector. Let $\Lambda = [\Lambda_1', \ldots, \Lambda_m']'$ denote the factor loading matrix. Considering identifiability of factors, Λ is lower-triangular with positive elements on the diagonal.

For simplicity in exposition, a stationary VAR(1) vector autoregressive process is assumed for the equally spaced m-dimensional factor η_t (2.3). However, the methods are easily modified to place arbitrary Gaussian process priors on continuous time factors to accommodate unequally-spaced and mixed-frequency time series. For example, a simple yet flexible choice corresponds to the Ornstein-Uhlenbeck (O-U) process prior (Barndorff-Nielsen and Shephard (2001)).

Considering the identifiability of latent factors, the covariance matrix of η_1 is set to I_m, and mean is set to 0 (in this paper, N(mean, precision) is used in this paper to represent Normal distributions), and we let

$$\begin{aligned} &\eta_1 \sim N(0, I_m) \\ &\eta_t = A_\eta \eta_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma_\eta^{-1}), \quad t = 2, \ldots, T \\ &\text{where } \Sigma_\eta = \Phi_\eta^{-1} = I_m - A_\eta A_\eta' \end{aligned} \tag{2.3}$$

Conditional on the factor path $(\eta_1, \ldots, \eta_T)$, the covariates $\mathbf{z}_{1:T}$ and all other model parameters Θ, the observations $\mathbf{y}_{1:T}$ are independent of each other, which implies that the likelihood is of the form

$$f(\mathbf{y}_{1:T}|\eta_{1:T}, \mathbf{z}_{1:T}, \Theta) = \prod_{t=1}^{T} \prod_{i=1}^{m} f(y_{it}|\eta_{1:T}, \mathbf{z}_{1:T}, \Theta) \tag{2.4}$$

Further combined with the Bayesian computational algorithms explored in the next section, the model allows missing data on a subset of the time series and/or mixed-frequency.

2.3 Computational Algorithms for Bayesian Inference

For Bayesian computation in dynamic generalized latent trait models, Dunson (2003) proposed to use one-at-a-time Gibbs updates with adaptive rejection sampling (ARS) to sample from the full conditional posterior distributions of the model parameters and latent factors. There are several problems that arise with this approach. First, ARS can be slow as many sampling and updating steps are needed. Second, due to high posterior dependence in the unknowns, one-at-a-time Gibbs samplers tend to be very slow to converge and mix. This is particularly the case as the number of time points increases, with computation becoming completely infeasible for moderate to large T. To address these problems, we propose a Metropolis Hastings algorithm with adaptive proposals, which relies on a Greedy Density Kernel Approximation

(GDKA) via mixture of normal distributions, to efficiently draw samples of latent factors in a block. For efficient updating of model parameters, parameter expansion techniques are proposed and combined with latent factor normalization to improve mixing.

2.3.1 Sampling Latent Factors – Background

Posterior computation in latent factor time series models shares many characteristics with similar issues in state-space models, where sequential filtering and backward sampling (FFBS) using Monte Carlo is pivotal. In the absence of mixed-measurement data, when smoothing distributions of the states do not have closed forms, computational techniques based on sequential Monte Carlo (Liu and West (2001)), particle filtering and using approximated distributions have been proposed for non-linear and non-Gaussian state-space models (see e.g. Harrison and Stevens (1971), Sorenson and Alspach (1971), Alspach (1974) and more recently Ravines and Schmidt (2007)). Stroud et al. (2003) further proposed to jointly update all unobserved states by using a Metropolis Hastings proposal that approximates the exact conditional posterior. Closer to our perspective is the recent work of Niemi and West (2010), who proposed adaptive mixture approximations by matching moments for the smoothing distributions in non-linear state-space models. However, most of the these algorithms are not designed for cases with multidimensional mixed-measurement response data, in which presence of multiple exponential family distributions of data makes sequential approximation of smoothing distributions more computationally challenging, especially when missing data exist.

In this paper, we proposed a fast and efficient Greedy Density Kernel Approximation (GDKA) to the smoothing distributions via mixtures of Gaussian kernels. The GDKA needs to be performed for each filtering step and chooses mixture Gaussian backward sampling distributions. Assuming the approximations are reasonably

accurate, Metropolis Hastings steps based on the approximated mixture of normal distributions as proposals (the actual proposal distributions are based on the mixtures of normals approximation, modified to have heavier tails) jointly update latent factors efficiently.

2.3.2 Sampling Latent Factors – Approximating Filtering Distributions

Let $\alpha_t(\eta_t)$ denote the forward filtering function at time t

$$\alpha_t(\eta_t) = f(\eta_t|\mathbf{y}_{1:t})$$

We have the recursive relationship between time t and t+1

$$\alpha_{t+1}(\eta_{t+1}) \propto \left[\int \alpha_t(\eta_t) f(\eta_{t+1}|\eta_t) d\eta_t\right] \cdot f(y_{i,(t+1)}, \forall i|\eta_{t+1}) \tag{2.5}$$

For simplicity of exposition, we focus on the case η_t is one-dimensional. Clearly, the computational bottleneck is that $\alpha_t(\eta_t)$ does not have an analytic form with mixed-measurement time series, making it infeasible to perform filtering and smoothing. To tackle this, we propose to use a Greedy Density Kernel Approximation (GDKA) to approximate the forward filtering functions $\alpha_t(\eta_t)$ over a compact domain Ω via mixtures of K Gaussian kernels:

$$\alpha_t(\eta_t) \approx \sum_{k=1}^{K} w_{tk} N(\eta_t; \mu_{tk}, \phi_t), \quad w_{tk} \geqslant 0. \tag{2.6}$$

Suppose there are N_t pairs of $\{\alpha_t(\eta_t^{(i)}), \eta_t^{(i)}\}$ $(i = 1, \ldots, N_t)$ over domain Ω available compromising a training set, and let $\Gamma_t = (\Gamma_t^{(1)}, \ldots, \Gamma_t^{(N_t)})^T$ be a $N_t \times 1$ vector with $\Gamma_t^{(i)} \propto \alpha_t(\eta_t^{(i)})$. Then GDKA via a mixture of K Gaussian kernels has the form:

$$\hat{\Gamma}_t^{(i)} = \sum_{k=1}^{K} w_{tk} \exp\left(-\frac{\phi_t}{2}||\eta_t^{(i)} - C_{tk}||^2\right), \quad \text{where } w_j \geqslant 0$$

where C_{tk} $(k = 1, \ldots, K)$ are centers (means) of Gaussian kernels, and ϕ_t is the Gaussian precision, which are set to be the same for all K kernels. Letting H_t denote the distance matrix with $H_{t,i,k} = \exp\left(-\frac{\phi_t}{2}||\eta_t^{(i)} - C_{tk}||^2\right)$, $(i = 1, \ldots, N_t$, $k = 1, \ldots, K)$, and $\mathbf{w_t} = (w_{t1}, \ldots, w_{tK})^T$ be a $K \times 1$ vector of the weights, we can represent the fitting/approximation problem in the regression form:

$$\Gamma_t = H_t \mathbf{w_t} + \epsilon_t \tag{2.7}$$

Given the means and variances of the Gaussian kernels, the best approximation can be achieved by minimizing the residual errors. The optimal weights are:

$$\hat{\mathbf{w}}_\mathbf{t} = arg\min_{\mathbf{w_t}} ||\epsilon_t||^2; \quad \text{subject to: } \mathbf{w_t} \geqslant 0; \tag{2.8}$$

If we cannot be sure that H_t is well conditioned or even a full ranked matrix, a regularized least square solution is:

$$\hat{\mathbf{w}}_\mathbf{t} = arg\min_{\mathbf{w_t}} ||\epsilon_t||^2 + \alpha||\mathbf{w_t}||^2; \quad 0 < \alpha << 1; \quad \text{subject to: } \mathbf{w_t} \geqslant 0; \tag{2.9}$$

It is trivial to prove that with the non-negative constraints, the quadratic optimization problem always has a unique solution.

Clearly, specification of number of kernels used (K), choice of training pairs, Gaussian centers C_{tk} and Gaussian precision ϕ_t is critical for the approximation. In details, to form the training pairs, a grid of N_t equally-spaced α_t are chosen within a compact domain on which the density $\alpha_t(\eta_t)$ is above certain threshold. In our studies, we chose the threshold of $\alpha_t(\eta_t)$ to be $> 10^{-3}$ relative to the mode of the filtering distribution (Note that the filtering distributions have to be uni-modal if the observations are assumed to be coming from distributions in the exponential family). Gaussian centers are chosen from the N_t training points α_t. Given equation (2.7), choosing a greedy set of K Gaussian centers from N_t values of α_t can be

achieved using the forward-selection algorithm as in the variable selection settings in linear regression problems. For the value of K, in our preliminary simulation studies, we found that $K = 20$ provides reasonably good approximations to a variety of distributions based on our algorithm, so in all our studies, we chose K to be 20. A greedy search step is also included within GDKA to find a reasonable value of ϕ_t. A detailed step-by-step procedure for GDKA is shown in Section 2.7.1 (in the appendix), and an example is shown in Figure 2.10 to show the fine approximation.

With the approximated $\alpha_t(\eta_t)$, forward filtering can be conducted sequentially. Let $\psi_t(\eta_{t+1}) = \int \alpha_t(\eta_t) f(\eta_{t+1}|\eta_t) d\eta_t \propto f(\eta_{t+1}|y_{i1}, \ldots, y_{i,t}, \forall i)$ denote the one-step prediction function (the integral part in the filtering function). Then we have:

$$\begin{aligned} &\psi_t(\eta_{t+1}) \propto \sum_k w_k^* N(\eta_{t+1}; \mu_k^*, \phi_k^*), \\ &\text{where } w_k^* = w_{tk}, \quad \mu_k^* = A_\eta \mu_{tk}, \quad \phi_k^* = \frac{\phi_{tk}\Phi_\eta}{\phi_{tk} + A_\eta^2 \Phi_\eta}, \\ &\text{with } A_\eta, \mu_\eta \text{ and } \Phi_\eta \text{ defined in (2.3), and } w_{tk}, \mu_{tk}, \phi_{tk} \text{ from (2.6).} \end{aligned} \tag{2.10}$$

With the approximated closed-form $\psi_t(\eta_{t+1})$, we can evaluate $\alpha_{t+1}(\eta_{t+1})$, generate N_{t+1} training pairs $\{\eta_{t+1}^{(i)}, \alpha_{t+1}(\eta_{t+1}^{(i)})\}(i = 1, \ldots, N_{t+1})$ and approximate $\alpha_{t+1}(\eta_{t+1})$ using the same GDKA algorithm we used to approximate $\alpha_t(\eta_t)$, and obtain the updated filtering function $\alpha_{t+1}(\eta_{t+1})$:

$$\alpha_{t+1}(\eta_{t+1}) \propto \sum_k w_{t+1,k} N(\eta_t; \mu_{t+1,k}, \phi_{t+1,k}) \tag{2.11}$$

The normalizing constant is not involved in the computation, and the nature of the recursion relationship implies that we can start with approximating

$$\alpha_1(\eta_t) \propto \pi_0(\eta_1) L(\eta_1|y_{i1}, \forall i).$$

And then proceed with sequentially approximating each $\alpha_t(\eta_t)$.

2.3.3 Sampling Latent Factors – Backward Sampling Distributions

Given the approximated filtering functions, backward sampling distributions $\pi_t(\cdot)$ $(t = 1, \ldots, T)$ can be approximated straightforwardly via $\hat{\pi}_t(\cdot)$:

$$\hat{\pi}_T(\eta_T|\mathbf{y}_{1:T}) = \alpha_T(\eta_T) \propto \sum_k w_{Tk} N(\eta_T; \mu_{Tk}, \phi_{Tk}),$$

$$\hat{\pi}_{t-1}(\eta_{t-1}|\eta_t, \mathbf{y}_{1:T}) \propto \sum_k \tilde{w}_{t-1,k} N(\eta_{t-1}; \tilde{\mu}_{t-1,k}, \tilde{\phi}_{t-1,k}),$$

$$\text{where } \tilde{w}_{t-1,k} = w_{t-1,k} \exp\left(\frac{A_\eta \Phi_\eta \mu_{t-1,k} \phi_{t-1,k} [2(\eta_t - \mu_\eta) - A_\eta \mu_k^{(t-1)}]}{2(\phi_k^{(t-1)} + A_\eta^2 \Phi_\eta)} \right), \tag{2.12}$$

$$\tilde{\mu}_{t-1,k} = \frac{\mu_{t-1,k} \Phi_\eta + A_\eta \Phi_\eta (\eta_t - \mu_\eta)}{\phi_{t-1,k} + A_\eta^2 \Phi_\eta},$$

$$\phi_{t-1,k} = \phi_{t-1,k} + A_\eta^2 \Phi_\eta,$$

with A_η, Φ_η defined in (2.3), and $w_{t-1,k}, \mu_{t-1,k}, \phi_{t-1,k}$ from (2.6).

Given (2.12) the joint posterior distribution of all latent factors can be approximated by a sequential product of mixture of normals:

$$\hat{\pi}(\eta_{1:T}|\mathbf{y}_{1:T}) \propto \hat{\pi}_T(\eta_T|\mathbf{y}_{1:T}) \prod_{t=2}^{T} \hat{\pi}_t(\eta_{t-1}|\eta_t, \mathbf{y}_{1:T}) \tag{2.13}$$

2.3.4 Sampling Latent Factors – Metropolis Hastings Updating

Using the approximated joint posterior distribution $\hat{\pi}(\eta_{1:T}|\mathbf{y}_{1:T})$ of the latent factors as proposal distributions, sample paths of latent factors can be drawn from a Metropolis Hastings step. For practical reasons, the approximated joint posterior distribution $\hat{\pi}(\eta_{1:T}|\mathbf{y}_{1:T})$ is further combined with a mixture of corresponding heavier-tailed Cauchy distributions $\tilde{\pi}(\eta_{1:T}|\mathbf{y}_{1:T})$, forming the proposal distribution $g(\eta_{1:T}|\mathbf{y}_{1:T})$ (as shown in (2.14)), to avoid irregular behaviors of the Markov chain in

the tail region:

$$g(\eta_{1:T}|\mathbf{y}) = p_\eta \hat{\pi}(\eta_{1:T}|\mathbf{y}_{1:T}) + (1 - p_\eta)\tilde{\pi}(\eta_{1:T}|\mathbf{y}_{1:T}) \tag{2.14}$$

where $0 << p_\eta < 1$ is the weight of the approximated joint distribution. $p_\eta = 0.95$ is used in all the following computations.

The Metropolis Hastings step in the $(m + 1)^{th}$ iteration of the MCMC then proceeds by:

1. Generate the entire block of candidate values γ^{m+1} for $\eta = \eta_{1:T}$ from the proposal distribution $\sim g(\cdot)$, with $g(\cdot)$ following (2.14).

2. Accept the candidate and let $\eta^{(m+1)} = \gamma^{(m+1)}$ with probability

$$\min\left(1, \frac{\pi(\gamma^{(m+1)})g(\eta^{(m)})}{\pi(\eta^{(m)})g(\gamma^{(m+1)})}\right)$$

   where $\pi(\cdot)$ is the unnormalized posterior full conditional joint distribution of latent factors $\eta_{1:T}$. Otherwise, let $\eta^{(m+1)} = \eta^{(m)}$.

2.3.5 Sampling Model Parameters

For models with other parameters Θ in addition to the latent factors, the above latent factor sampler is applied at each stage of an MCMC algorithm that also includes steps for updating Θ. To reduce slow-mixing which is common in MCMC algorithms for latent factor models, we combine parameter expansion with latent factor normalization to induce default priors and improve computational efficiency for model parameters.

Parameter Expansion (PX) has been introduced as a useful approach for both introducing new families of prior distributions and improving computational efficiency (Liu and Wu (1999), Gelman (2004) and Kinney and Dunson (2007)). Ghosh and

Dunson (2009) used parameter expansion in Bayesian Gaussian factor analysis to induce heavy-tailed priors for factor loadings and improve mixing of samplers simultaneously. In this paper, we extend the parameter expansion technique to the Bayesian dynamic factor model setting with mixed-measurement data. Basically, parameter expansion introduces a *working model* that is over-parameterized, and parameters in the working model are then transformed back to parameters in the inferential model.

Considering model (2.1)-(2.3) as our inferential model, we define the following working PX-factor model:

$$\begin{aligned} &h_i(\theta_{it}) = \alpha_i^* + z_{it}'\beta_i + \Lambda_i^*\eta_t^*, \\ &\eta_1^* \sim N(\mu_1^*, (\Psi_1^*)^{-1}), \\ &\eta_t^* \sim N(\mu_\eta^* + A_\eta^*\eta_{t-1}^*, (\Psi_\eta^*)^{-1}), \\ &\text{where } \mu_\eta^* = (I - A_\eta^*)\mu_1^*, \quad \Psi_\eta^* = \Psi_1^* - A_\eta^*\Psi_1^*A_\eta^{*'}, \end{aligned} \tag{2.15}$$

where Λ_i^* is an unconstrained working factor loading matrix having a lower triangular structure, η_t^* is an $m \times 1$ working latent factor vector, $\Psi_1 = \text{diag}(\psi_1, \ldots, \psi_m)$ is a diagonal covariance matrix, and α_i^* is the intercept term, with $\alpha_k^* = 0$, for $k \leqslant m$. $A_\eta^* = \text{diag}(A_1^*, \ldots, A_m^*)$ is a diagonal matrix of AR(1) coefficients, with $|A_i^*| < 1$ to guarantee stationarity. Note that the PX-factor model is clearly over-parameterized, having redundant parameters in both the covariance matrix Ψ_1 and mean μ_1^*.

In order to transform the working model parameters back to inferential model parameters, we let:

$$\begin{aligned} &\lambda_{jl} = \mathcal{S}(\lambda_{ll}^*)\lambda_{jl}^*\psi_l^{-1/2}, \quad \eta_t = S_d\left[\Psi_1^{1/2}(\eta_t^* - \mu_1^*)\right], \quad \alpha_i = \alpha_i^* + \Lambda_i^*\mu_1^*, \\ &A_\eta = A_\eta^*, \quad \mu_\eta = S_d\left[\Psi_1^{1/2}\left\{\mu_\eta^* + (A_\eta^* - I)\mu_1^*\right\}\right], \end{aligned} \tag{2.16}$$

where $\mathcal{S}(x) = -1$ for $x < 0$, $\mathcal{S}(x) = 1$ for $x \geqslant 0$, and $S_d = \text{diag}\{\mathcal{S}(\lambda_{ll}^*)\}$. Then, instead of placing priors directly on α_i and Λ_i, with $\eta_t \sim N(0, I)$, we specify priors

on α^*, Λ_i^*, μ_1^* and Ψ_1^*. In particular, we let

$$\begin{aligned}
&\lambda_{jl}^* \sim N(0,1) \ \text{ for all lower triangular elements,} \\
&\psi_l \sim \text{inv} - \Gamma(a_l, b_l), \\
&\alpha_i^* \sim N(0, \phi_\alpha), \quad A_i^* \sim \text{Unif}(-1,1), \quad \mu_1^* \sim N(0, \Phi_B).
\end{aligned} \tag{2.17}$$

Hence, in the one-factor case, we obtain

$$\begin{aligned}
&\Lambda_i = S(\Lambda_i^*)\Psi_1^{-1/2}\Lambda_i^*, \quad \eta_t = S(\Lambda_i^*)\Psi_1^{1/2}(\eta_t^* - \mu_1^*), \quad \alpha_i = \alpha_i^* + \Lambda_i^*\mu_1^*, \\
&A_\eta = A_\eta^*, \quad \mu_\eta = S(\Lambda_i^*)\Psi_1^{1/2}\left\{\mu_\eta^* + (A_\eta^* - 1)\mu_1^*\right\}
\end{aligned} \tag{2.18}$$

with priors

$$\begin{aligned}
&\Lambda_i^* \sim N(0,1) \ \text{ for all lower triangular elements,} \\
&\Psi_1 \sim \text{inv} - \Gamma(a, b), \\
&\alpha_i^* \sim N(0, \phi_\alpha), \quad A_i^* \sim \text{Unif}(-1,1), \quad \mu_1^* \sim N(0, \Phi_B)
\end{aligned} \tag{2.19}$$

To efficiently draw samples of parameters from the inferential model, we draw samples from the working model, followed by a post-processing step to transform working-model draws back to the parameters in the inferential model.

2.3.6 Sampling Model Parameters – Latent Factor Normalization

Parameter expansion can greatly reduce the autocorrelation of factor loadings, however, it has limited effect on improving the mixing of other parameters (e.g. the intercept term, see Figure (2.21)). To further improve computational efficiency, we combine latent factor normalization with parameter expansion in sampling the model parameters from the full conditional distributions. In detail, suppose $\eta_{1:T}^*$ is a realization of latent factors in model (2.15) with $\eta_t^* = (\eta_{1,t}^*, \ldots, \eta_{m,t}^*)$, let $\tilde{\eta}_t = (\tilde{\eta}_{1,t}, \ldots, \tilde{\eta}_{m,t})'$ denote the normalized latent factors, with

$$\tilde{\eta}_{i,t} = \frac{\eta_{i,t}^* - mean(\eta_{i,1:T}^*)}{std(\eta_{i,1:T}^*)}, \quad t = 1, \ldots, T; i = 1, \ldots, m.$$

Then in matrix form we have:

$$\tilde{\eta}_t = \nu + \Delta\eta_t^* \tag{2.20}$$

where ν is a $m \times 1$ vector with $\nu_i = -\frac{\text{mean}(\eta^*_{i,1:T})}{\text{std}(\eta^*_{i,1:T})}$, and Δ is a $m \times m$ diagonal matrix with $\Delta_{i,i} = \frac{1}{\text{std}(\eta^*_{i,1:T})}$.

PX with Latent Factor Normalization Algorithm:

Considering the working PX-factor model (2.15) and priors (2.17), given a realization of latent factors $\eta^*_{1:T}$, in order to obtain a sample of model parameter $\Theta = (\alpha_i, \Lambda_i)$ $(i = 1, \ldots, p)$ from the full conditional distribution $f(\Theta|\eta_{1:T})$, we can first sample $\tilde{\Theta}$ from the full conditional distribution using the *normalized PX* model (2.21) via normalized latent factors $\tilde{\eta}_{1:T}$ $(i = 1, \ldots, m)$:

$$h_i(\theta_{it}) = \tilde{\alpha}_i + z'_{it}\beta_i + \tilde{\Lambda}_i\tilde{\eta}_t, \tag{2.21}$$

with the induced priors (2.22) for the parameters of normalized PX-factor model derived from (2.17):

$$\begin{aligned} &\tilde{\lambda}_{jl} \sim N(0, \Delta_{ll}^2) \quad \text{for all lower triangular elements,} \\ &\tilde{\alpha}_i \sim N\left(0, (1/\phi_{\alpha_i} + \sum_{l=1}^{\min(i,m)} v_l^2/\Delta_{ll}^2)^{-1}\right). \end{aligned} \tag{2.22}$$

and transform $\tilde{\Theta}$ to obtain a sample of Θ via the following transformations via (2.23).

$$\begin{aligned} \Lambda_i^* &= \tilde{\Lambda}_i\Delta, \\ \alpha_i^* &= \tilde{\alpha}_i + \tilde{\Lambda}_i\nu. \end{aligned} \tag{2.23}$$

2.4 Simulation Examples

2.4.1 Mixed-Measurement Time Series

To access the efficiency of the algorithms, we first consider the simulation example jointly modeling a 3-dimensional mixed-measurement time series (D_t) consisting of

Poisson counts (Y_t), continuous (Z_t) and binary responses (W_t), with T=50, 100, 150. 200 and 500 respectively (as shown in (2.24)). In each case, 10% of the data in each time series are randomly missing, allowing us to evaluate the model in dealing with missing values. The time-varying latent factors are modeled to follow an O-U process.

$$\begin{aligned} & D_t = [Y_t, Z_t, W_t], \\ & Y_t \sim Poiss(\lambda_t), \quad Z_t \sim N(\mu_t, 1/\sigma_z^2), \quad W_t \sim Bernoulli(p_t), \\ \text{where } & \log(\lambda_t) = a + b\eta_t, \quad \mu_t = c + d\eta_t, \quad logit(p_t) = e + f\eta_t, \\ \text{and } & d\eta_t = -\beta(\eta_t - \alpha)dt + \sigma dW_t \end{aligned} \tag{2.24}$$

We run the sampler for 10,000 iterations, discarding the first 2,000 iterations as burn-in. The results are later compared with those obtained by simply using WinBUGS or extended Kalman filters. Taking T=200 case as an example, we set the values of a-f to be 2, 2, 3, 2, 0.5 and 1 respectively, $\sigma_z^2 = 1$ and for the time-varying latent factors $\alpha = 0$, $\beta = 1$ and $\sigma = 1$. The simulated time series are shown in Figure 2.1. We introduce priors on the parameters in the PX working model as shown in (2.17). Specifically, half-Cauchy priors are placed on the variance components of latent factor autoregressions as recommended by Gelman (2006), while Inv-Gamma(0.01,0.01) priors are given to all other variances. To visualize the mixing the MCMC chain, trace-plots of the posterior samples of model parameters (intercepts and factor loadings shown in Figure 2.2) and latent factors (Figure 2.3) are shown.

A summary of effective sample size and sample autocorrelation of posterior samples are further listed in Table 2.1. A summary of effective sample size and sample autocorrelation of posterior samples are further listed in Table 2.1.We compare the mixing of our MCMC sampler with that of WinBUGS (Traceplots shown in Figure 2.4 and 2.11, and the summary of effective sample size and autocorrelation shown in Table 2.5), which is based on a Gibbs sampler and use of adaptive-rejection sam-

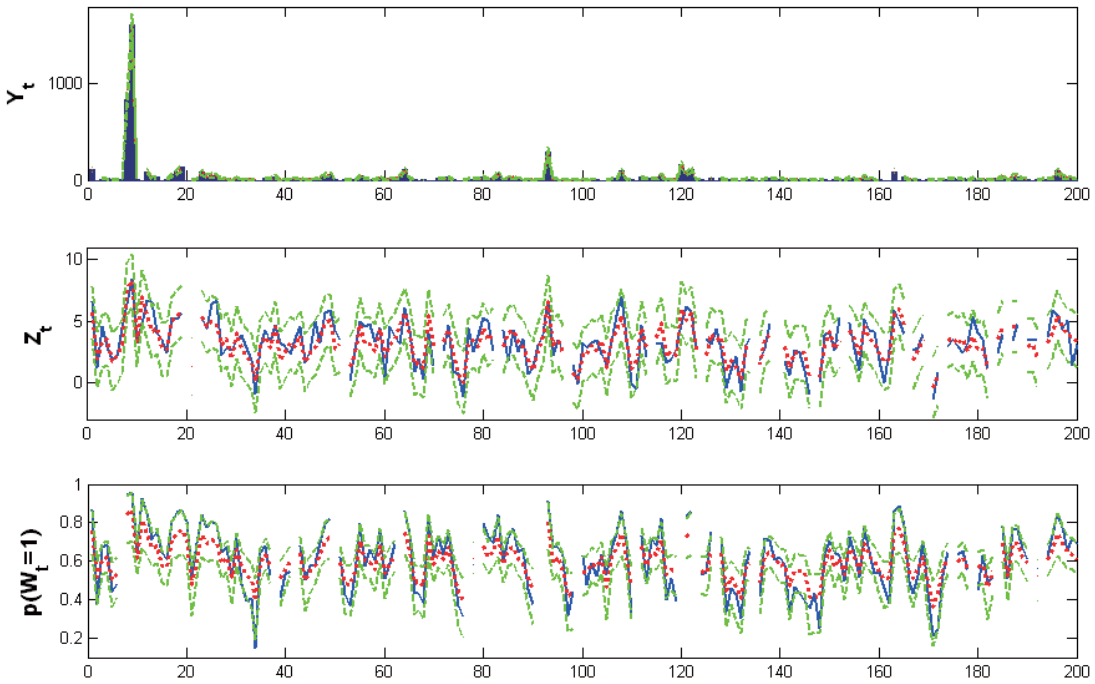

FIGURE 2.1: Illustration of simulated dataset and interval validation to show the model fitting. Solid lines (and the histogram for the first figure): simulated data. A 3-dimensional mixed-measurement time series $D_t = [Y_t, Z_t, W_t]$ up to T=200 is simulated, where Y_t is count time series, Z_t is continuous and W_t is binary. 10% data in each time series are randomly missing.; dotted lines: posterior mean; dashed lines: 95% posterior C.I.

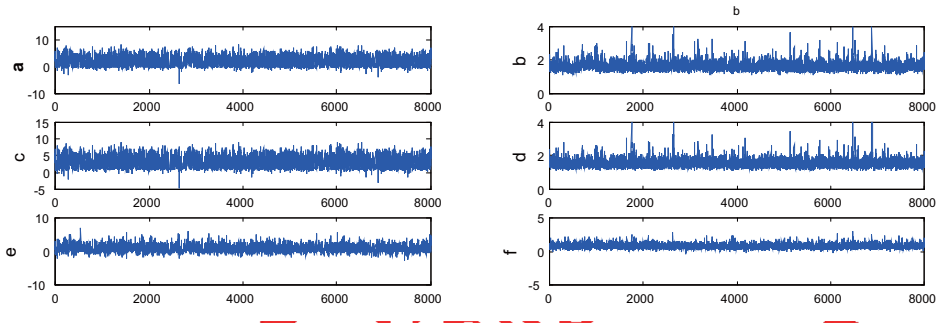

FIGURE 2.2: Traceplot of Posterior Samples of Model Parameters: Intercepts (a,c,e) and factor loadings (b,d,f) shown. The parameters are defined in (2.24).

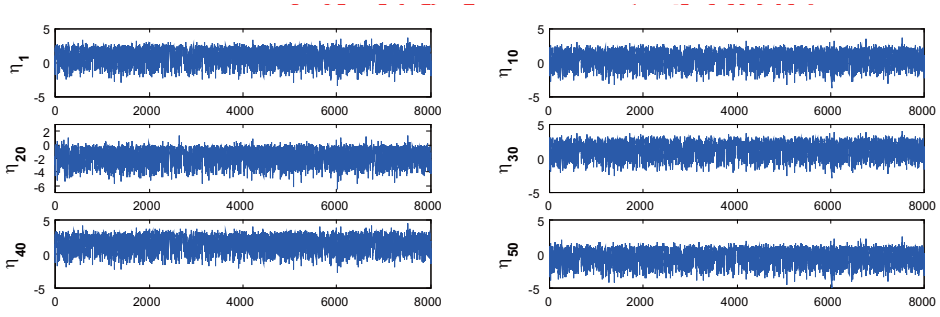

FIGURE 2.3: Traceplot of Posterior Samples of Latent Factors: Six time points shown. The latent factors are defined in (2.24).

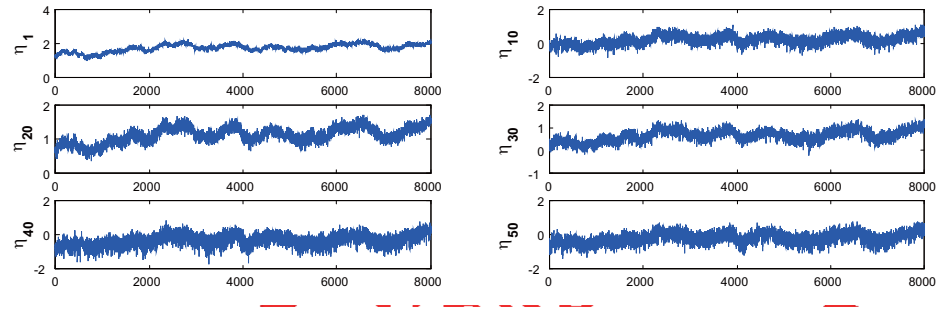

FIGURE 2.4: Traceplot Produced by WinBugs of Posterior Samples of Latent Factors. The traceplots are compared with those in Figure 2.3, also showing that our algorithm improve the mixing of the Markov Chain.

pler or random-walk MCMC for full conditional distributions without closed forms. Our sampler relying on GDKA, parameter expansion and latent factor normalization leads to tremendous gains in the mixing of MCMC chains for both model parameters and latent factors. Furthermore, it is important to note that all three techniques contribute to improving the efficiency of the MCMC sampler, as not implementing one of them tends to give poor mixing of the chains for at least some model parameters (see e.g. Figure 2.6 and 2.21). Especially, we find that while Metropolis Hastings via GDKA is critical for latent factor filtering and sampling, parameter expansion substantially improves computational performance for factor loadings, and

latent factor normalization helps the mixing of chains for other model parameters (e.g. intercepts).

On the other hand, since extended Kalman filters have been used to approximate non-Gaussian filtering and sampling distributions using Laplace normal approximations, especially when the length of time series is not long, we compared the results of our method with those of Kalman filters with all simulated data sets. A comparison of acceptance ratio clearly shows that our sampler gave much better acceptance ratio, especially with increased length of the time series. Low acceptance rates of extended Kalman filter inevitably lead to MCMC inefficiency, illustrated by the increased autocorrelation and decreased effective sample size (a comparison between Table 2.1 and Table 2.5).

To show the model fitting, internal validations are performed, with posterior means and 95% confidence intervals (C.I.s) of the simulated time series shown in Figure 2.1.

The use of our MCMC sampler seems to provide a reasonable inference to the unknown quantities. Furthermore, the framework and computational algorithms allows missing values on a subsect of the time series. Figure 2.12 shows that the model gives good interpolations of missing data based on the posterior marginal distributions.

Table 2.1: MCMC Summary for T=100 and T=150 respectively shows improved mixing, compared to Table 2.5, which is obtained using WinBUGS. Using T=100 and T=150 as examples, in each case, 10000 posterior samples are drawn for each unknown with the first 2000 as burn-in. Effective sample size (ESS) and sample autocorrelation with lag 10 and 30 of the 8000 posterior samples are shown for selected parameters as representatives (c, d, η_1 and η_{50} are defined as in (2.24)).

# of Time Points	T=100				T=150			
Parameters	c	d	η_1	η_{50}	c	d	η_1	η_{50}
ESS	1815	6500	1875	1740	1905	6025	1982	1887
Autocorr-Lag 10	0.045	0.008	0.043	0.039	0.019	0.025	0.019	0.018
Autocorr-Lag 30	0.007	-0.005	0.012	0.009	-0.009	-0.005	-0.011	-0.013

Efficient adaptive Metropolis Hastings via GDKA to sample time-varying latent factors is one of the key steps in our sampling scheme, and it is well known that the acceptance rates of Metropolis Hastings steps tend to decrease rapidly as the number of time points increases (Niemi and West (2010)). To evaluate this, we generated 100 simulated datasets for each T and look at the overall performance. Table 2.2 lists the acceptance rates of the Metropolis Hastings steps to sample latent factors using our sampling scheme, showing that the algorithm is efficient even for big time series with T=500, which has much higher acceptance ratio than extend Kalman filters. The accuracy of inference to unknown quantities is also further confirmed based on the 100 simulated datasets by checking whether the posterior estimates would recover the true values of model parameters under such a complex model, with Table 2.7 showing that in almost all cases the 95% posterior C.I.s cover the true values with all 100 simulated datasets.

Table 2.2: Acceptance Ratio for the Metropolis-Hastings update with GDKA for latent factors compared with extended Kalman filters (E-KF). Mean, minimum and maximum acceptance rates based on 100 simulated datasets are shown. Note that our algorithm both has higher acceptance ratio and more stable performance with mixed-measurement observations with possible missing values.

T (# of Time Points)	50	100	150	200	500
GDKA (mean)	73.1%	59.4%	43.3%	30.4%	12.4%
(min)	70.3%	57.3%	40.6%	27.4%	8.1%
(max)	76.1%	61.3%	46.8%	33.1%	13.3%
E-KF(mean)	41.1%	25.3%	1.7%	∼0	0
(min)	30.1%	14.1%	0%	0	0
(max)	43.3%	27.1%	7.7%	0.3%	0

Furthermore, it is worth mentioning that although we show that our algorithm updates latent factors in a single block and infers model parameters efficiently with relatively big time series (T=500), the performance of the proposed algorithm is expected to decrease with bigger time series. In those cases, a reasonable strategy

could be combining our algorithm with the idea of block sampling scheme, where the big time series can be divided to a few blocks with T smaller than 500 and updated sequentially in each block, which will definitely help in problems with extremely big time series.

2.4.2 Mixed Discrete Time Series

Time series with discrete outcomes having a variety of measurement scales (count, binary, categorical) are also commonly collected in various fields of study. Discrete variables contain less information than continuous variables (Chin and Lee (2008)), so it is of great help to share information between time series with mixed discrete outcomes whenever it is possible.

In this simulation study, 4-dimensional time series D_t consisting of a mixture of binomial counts (Y_t and Z_t) and binary responses (W_t and V_t) up to time T are jointly modeled, as shown in (2.25). Similarly, 10% of the simulated data in each time series are randomly missing to evaluate the model in dealing with missing data.

$$\begin{aligned} & D_t = [Y_t, Z_t, W_t, V_t], \\ & Y_t \sim Bino(N_y, p_t^{(y)}),\ Z_t \sim Bino(N_z, p_t^{(z)}), \\ & W_t \sim Bernoulli(p_t^{(w)}),\ V_t \sim Bernoulli(p_t^{(v)}), \\ \text{where } & logit(p_t^{(y)}) = a + b\eta_t,\ logit(p_t^{(z)}) = c + d\eta_t, \\ & logit(p_t^{(w)}) = e + f\eta_t,\ logit(p_t^{(v)}) = g + h\eta_t, \\ \text{and } & \eta_t = \gamma\eta_{t-1} + \epsilon_t,\ \epsilon_t \sim N(0, \phi_\epsilon). \end{aligned} \tag{2.25}$$

Taking T=100 case as an example, we set the values of a-h to be 0.5, 1, 1, -1.5, 1, -2, 0.5 and -3 respectively. An AR(1) autoregressive structure is applied to the time-varying latent factors, with $\gamma = 0.368$ and $\sigma_\epsilon = 0.658$. The simulated time series are shown in Figure 2.13. Similarly, to obtain posterior samples, we run our MCMC sampler for 10,000 iterations, discarding the first 2,000 samples as burn-in. Again,

the effective sample size and autocorrelation shown in Table 2.3 show good mixing of the Markov Chain, which is further illustrated by the trace-plots of intercepts, factor loadings (Figure 2.14) and latent factors (Figure 2.15). In addition, with T=100, the Metropolis Hastings step for latent factors update gives a good acceptance ratio of 56.1%.

Table 2.3: MCMC Summary for T=100 (selected parameters shown) shows good mixing of the Markov Chain. 10000 posterior samples are drawn for each unknown with the first 2000 as burn-in. Effective sample size (ESS) and sample autocorrelation with lag 10 and 30 of the 8000 posterior samples are shown for selected parameters as representatives ($a - h$, η_1 and η_{50} are defined as in (2.25)).

Parameters	a	b	c	d	e
ESS	5726	6449	5742	6488	5030
Autocorr-Lag 10	0.007	0.008	0.009	0.011	0.012
Autocorr-Lag 30	0.001	-0.006	0.003	-0.002	0.007
Parameters	f	g	h	η_1	η_{50}
ESS	2659	4057	2043	5998	5710
Autocorr-Lag 10	0.029	0.017	0.029	-0.004	-0.005
Autocorr-Lag 30	0.013	0.011	0.018	0.013	0.014

Internal validations with posterior mean and 95% C.I. of the simulated time series are shown in Figure 2.16. Again, the framework provides good fit to the simulated data, which is reinforced by the missing data interpolation results shown in Figure 2.17.

2.4.3 Mixed-Measurement-Mixed-Frequency Time Series

Another challenge with mixed-measurement time series is that time series are often sampled at different frequencies. For example, some time series are observed every month or quarter while others are recorded every year. In empirical financial analysis, daily and intra-daily time series are also often readily available for analysis. To model the time series simultaneously, the literature and methodology have been focused on either aggregating the higher-frequency data to the lower frequency, or

interpolating the lower-frequency data to the higher-frequency (Ghysels et al. (2005), Andreou et al. (2010) and Ghysels et al. (2007)). Neither of these is generally satisfactory. First of all, temporal aggregation loses information and may lead to poor prediction. Secondly, the approaches via data interpolation transform the mixed-frequency problem to a missing data problem, assuming that the model operates at the highest frequency of the time series, while data in the lower-frequency time series are periodically missing. However, commonly used interpolation methods generally do not efficiently and fully exploit the sample information, especially when massive missing data exist, which is very likely to occur when performing data interpolation in mixed-measurement-mixed-frequency time series analysis.

We apply our generalized dynamic factor model and computational algorithms to mixed-measurement-mixed-frequency time series analysis, in which the observed time series are simultaneously modeled via dynamic latent factors at specifically defined frequencies discussed later. Under this scenario, neither data aggregation nor interpolation is required in the inference or prediction, which improves the computational efficiency with mixed-frequency observations.

In detail, we treat the low-frequency time series as high-frequency time series with "missing data"and low "observed"frequency. For the p-dimensional mixed-measurement-mixed-frequency time series, let t be the time index of the highest frequency time series, and $\mathbf{y}_t = \{y_{it}\}$ $(i = 1, \ldots, p$ and $t = 1, \ldots, T)$, with observed frequencies $\omega = (\omega_1 \ldots, \omega_p)'$. Let $\mathbf{y}_\mathbf{t}^* = (y_{1t}^*, \ldots, y_{pt}^*)'$ be the corresponding ordered p-dimensional time series by descending frequency. Let $\omega^* = (\omega_1^*, \ldots, \omega_p^*)'$, with ω_i^* $(i = 1, \ldots, p)$ denoting the highest common frequency from y_{it}^* to y_{pt}^* . Then $\mathbf{y}_t^*$ can be modeled following (2.1), (2.2) and (2.3) via m common factors $\eta_t = (\eta_{1t}, \ldots, \eta_{mt})'$, with each latent factor η_{it} $(i = 1, \ldots, m)$ constructed to follow an AR(1) process with frequency ω_i^* $(i = 1, \ldots, m \leqslant p)$. In particular, in the one-factor model case, ω_1^* is the highest common frequency of all observed time series.

In the following simulation study, we considered jointly modeling a 3-dimensional mixed-measurement-mixed-frequency time series $\mathbf{y}_t$ with observed frequency $\omega = (3\omega_0, \omega_0, \omega_0)'$ respectively (shown in (2.26)). The simulated data are generated through two common latent factors ($\eta_t = (\eta_{1t}, \eta_{2t})'$) at the frequency of $3\omega_0$, and chosen in accordance with empirical applications to form multivariate mixed-measurement-mixed-frequency time series, where the first time series (Y_{1t}) can be considered as monthly recorded data, while the other two as quarterly recorded data. Specifically,

$$
\begin{aligned}
& Y_t = [Y_{1t}', Y_{2t}', Y_{3t}']', \\
& Y_{1t} \sim Bernoulli(p_{1t}), \quad Y_{2t} \sim N(\mu_{2t}, 1/\sigma^2), \quad Y_{3t} \sim Bino(N_3, p_{3t}), \\
\text{where } & logit(p_{1t}) = a_1 + b_1\eta_{1,t} + c_1\eta_{2,t} \\
& \mu_{2t} = a_2 + b_2\eta_{1,t} + c_2\eta_{2,t} \\
& logit(p_{3t}) = a_3 + b_3\eta_{1,t} + c_3\eta_{2,t} \\
\text{and } & \eta_{i,t} = \gamma_i\eta_{i,t-1} + \epsilon_{it}; \epsilon_{it} \sim N(0, \phi_{\eta_i})
\end{aligned}
\tag{2.26}
$$

The simulated data and latent factors up to T=90 are shown in Figure 2.18 (90 time points for Y_{1t}, and 30 time points observed for Y_{2t} and Y_{3t}). Again, analysis of posterior samples indicates good mixing the the Markov Chain (Table 2.6), and internal validation shows good fitting of the model to the simulated mixed-measurement-mixed-frequency data (Figure 2.19).

2.5 Intertwined Corporate Default, Recovery Rate and Business Cycle

It has been noted that corporate default rates and bond recovery rates on defaults are negatively correlated (Altman et al. (2005)). Since recovery rates play a critical role in pricing and risk models, treating recovery rates as either constant or a stochastic variable independent of default rates while neglecting inverse relationship leads to inaccurate estimation of the loss function and suboptimal capital allocation. Further-

more, according to the current Basel proposal, banks can opt to provide their own recovery rate forecasts for the calculation of regulatory capital (Creal et al. (2011)). Thus there is an immediate need for statistical models explaining the relationship between the corporate default risk and bond recovery rates (and probably some other credit risk indicators), which can be used in default prediction and credit risk modeling. Macroeconomic conditions, on the other hand, may affect both default and recovery rates (see e.g. Nickell et al. (2000)), where the number of defaults tends to increase with lower recovery rates in economic recessions (Altman et al. (2005)).

In this study, we apply the generalized Bayesian dynamic factor models to jointly modeling corporate default counts, bond recovery rates and business cycle indicators over time. Default counts are model by binomial distribution with the default probability modeled via logistic link function. Recovery rates, which are between 0 and 1, are modeled via logistic-Normal model. Unemployment rate is used as an indicator for business cycle. The relationship between them are modeled via shared common unobserved systematic risk factor(s).

2.5.1 Data

The 5 time series (shown in Figure 2.5) we have are annual recovery rates, corporate default counts and unemployment rate data from 1982 to 2008 (Latest public data obtained from *Moody's Default and Recovery Database*), denoted as:

$$\mathbf{y}_t = \{y_{it}\}\,; \quad (i = 1, \ldots, 5 \text{ and } t = 1, \ldots, 27)$$

where y_{1t}: annual default counts within investment-grade cooperates. y_{2t}: annual default counts within speculative-grade cooperates. y_{3t}: all-bonds annual recovery rates. y_{4t}: senior-secured bonds annual recovery rates. y_{5t}: annual unemployment rate of the United States.

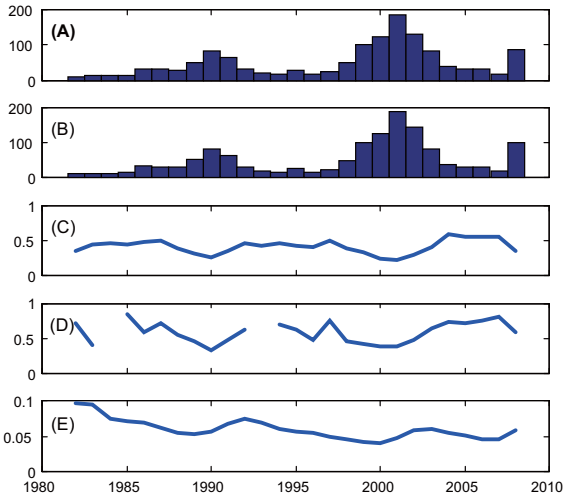

FIGURE 2.5: Data consisting of 5 Mixed-Measurement Time Series Shown. (A): Annual Default Counts: Speculative-Grade Cooperates. (B): Annual Default Counts: Investment-Grade Cooperates. (C): Annual Recovery Rates: All Bonds. (D): Annual Recovery Rates: Senior-Secured Bonds. (E): Annual Unemployment Rate.

2.5.2 Model Specification

Default counts are modeled as coming from binomial distributions. Recovery rates and unemployment rates are modeled as coming from logistic-Normal distributions. The inferential model is shown in (2.27). Default priors are placed on the corresponding parameter expanded working model.

$$
\begin{aligned}
& y_{1t} \sim Bino(N_{1t}, p_{1t}), \quad y_{2t} \sim Bino(N_{2t}, p_{2t}), \\
& \text{logit}(y_{3t}) \sim N(\mu_{3t}, \phi_3), \quad \text{logit}(y_{4t}) \sim N(\mu_{4t}, \phi_4), \quad \text{logit}(y_{5t}) \sim N(\mu_{5t}, \phi_5), \\
\text{where } & \text{logit}(p_{1t}) = \alpha_1 + \Lambda_1 \eta_t, \quad \text{logit}(p_{2t}) = \alpha_2 + \Lambda_2 \eta_t, \\
& \mu_{3t} = \alpha_3 + \Lambda_3 \eta_t, \quad \mu_{4t} = \alpha_4 + \Lambda_4 \eta_t, \quad \mu_{5t} = \alpha_5 + \Lambda_5 \eta_t, \\
\text{and } & \eta_1 \sim N(0, I_m), \quad \eta_t \sim N(a_\eta \eta_t + b_\eta, I_m), \quad |a_\eta| < 1
\end{aligned}
\tag{2.27}
$$

The 5 time series are jointly modeled via one or two factor models, based on

the previous studies showing that default counts and recovery rates are negatively correlated, and they may be further associated with but do not match business cycle. Inference of number of factors is based on comparing fit for the one and two factor models.

2.5.3 Results

We are interested in the inference of both parameters and latent factors, as well as the relationship between corporate default counts, bond recovery rates and business cycle.

In the **one-factor** model, the latent factor can be interpreted as credit risk within the period. We explore the posterior distribution of the unknowns using the algorithms discussed earlier. Two results show the efficiency of the MCMC sampler: Firstly, the acceptance ratio of latent factors in the Metropolis Hastings update is 81%. Secondly, parameter expansion and latent factor normalization promote the mixing of the Markov chain, and greatly improve the sampling efficiency (as shown in Figure 2.6), compared to those obtained via only parameter expansion (Figure 2.21) or WinBUGS (results not shown).

We constructed point and interval estimates of the credit risk factor within the period. This is shown in Figure 2.7. Clearly, the figure shows that credit pressure peaks around 1990, 2001 and 2008, which are all well-known financial crises.

Furthermore, Figure 2.8 shows the fitting of the model to the data, with posterior retrospective mean and 95% intervals shown with dotted lines and dashed lines respectively. Note that the fitting of annual recovery rates and default counts are good under the one-factor setting (The fitting of recovery rate for senior-secured bonds seems not as good as others, indicating the possibility of existence of additional factors. Two-factor model gives a better fit, as shown later), indicating that a common credit risk factor may be used to model the inverse relationship between

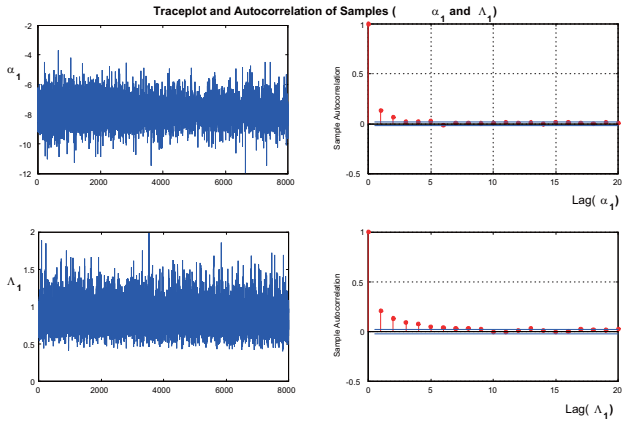

FIGURE 2.6: Traceplots and Autocorrelation Plots of Posterior Samples. Note that our modification can effectively reduce both the autocorrelation of factor loading matrix (Λ_1 as a representative here) and the intercept (α_1 as a representative) very well, as compared to those in Figure 2.21. α_1 and Λ_1 are defined in (2.27).

them. Fitting of unemployment rate data, however, is not as good, indicating the idea that business cycle, though may be related to credit cycle, have effects from additional latent factors.

In comparison, we also fit the data with a **two-factor** model, with the two factors chosen to be corresponding to the speculative-grade-cooperate default counts and unemployment rates time series, which can be considered as representing the credit risk factor and the business cycle factor. Again, we have high acceptance rates for the two latent factors (83% and 78% respectively) and the model fitting is shown in Figure 2.9. Clearly, the two-factor model gives improved fitting for unemployment data and senior-secured bonds recovery rate data, but has trivial improvement on others. This indicates that default and recovery risk, though related to, is not solely determined by the business cycle.

Furthermore, we specifically looked at the posterior distributions of factor loadings to obtain a better understanding of the relationship between the time series (Histograms of posterior samples of factor loadings shown in Figure 2.20 and the

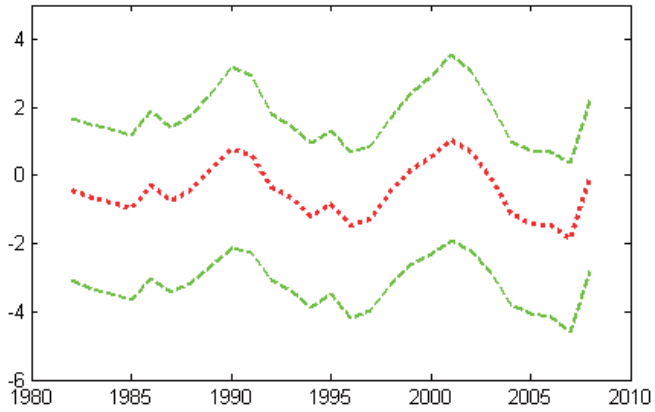

FIGURE 2.7: 1982-2008 Credit Pressure Reconstruction. Dotted line: posterior mean of the credit risk factor, and dashed lines: 95% posterior C.I.

posterior summaries are shown in Table 2.4). Table 2.4 clearly shows that default

Table 2.4: Posterior Summary of Factor Loadings Λ_{ij} ($i = 1, \ldots, 5$ and $j = 1, 2$) in the Two-Factor Model, where Λ_{ij} is the factor loading of the jth factor for the ith time series ($\Lambda_{12} = 0$).

Factor Loadings	Mean	Medium	SD	95% HPD C.I.
Λ_{11}	1.312	0.927	1.957	[0.478,3.127]
Λ_{21}	1.322	0.936	1.986	[0.503,3.195]
Λ_{22}	-0.0828	-0.0717	0.0808	[-0.225,0.035]
Λ_{31}	-0.592	-0.422	0.842	[-1.424,-0.172]
Λ_{32}	0.0642	0.0547	0.0735	[-0.038,0.196]
Λ_{41}	-0.683	-0.497	1.076	[-1.706,-0.128]
Λ_{42}	0.119	0.104	0.131	[-0.033,0.383]
Λ_{51}	0.0586	0.0440	0.115	[-0.077,0.213]
Λ_{52}	0.219	0.186	0.159	[0.099,0.454]

counts (the first and second times series) and bond recovery rates (the third and fourth time series) have strong negative correlation in terms of the credit risk factor (factor one), with the factor loadings of the first factor being strictly positive for default counts and strictly negative for bond recovery rates, at the 95% confidence level. In addition, we are not sure if unemployment rate is strongly correlated with default counts or bond recovery rates, although there is some weak evidence showing

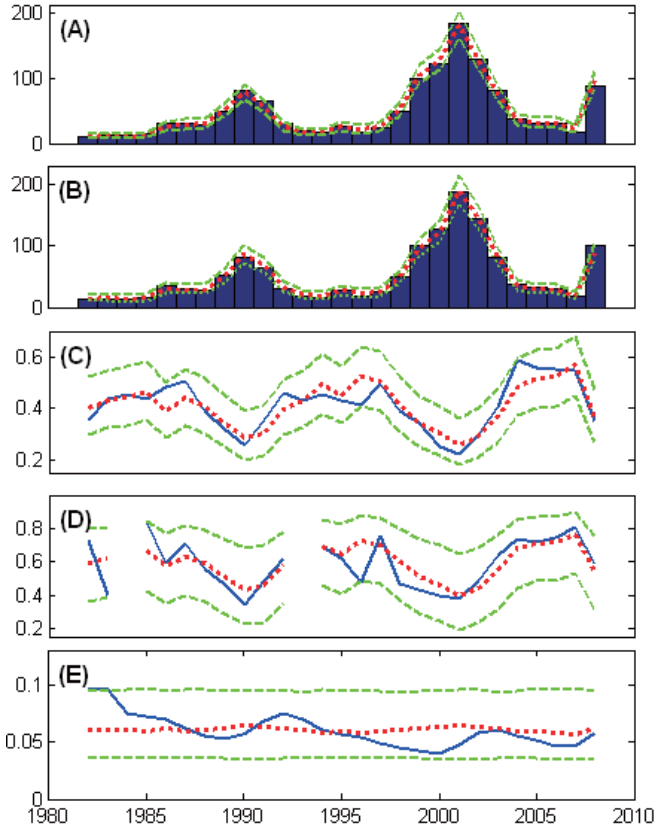

FIGURE 2.8: Model Fitting using One-Factor Generalized Dynamic Factor Model. The solid lines (and the histograms for the first two panels) show real historical data, the dotted lines indicate posterior means and the dashed lines show posterior 95% C.I. (A): Annual Default Counts: Speculative-Grade Cooperates. (B): Annual Default Counts: Investment-Grade Cooperates. (C): Annual Recovery Rates: All Bonds. (D): Annual Recovery Rates: Senior-Secured Bonds. (E): Annual Unemployment Rate.

that default counts may be negatively associated and bond recovery rates may be positively associated with unemployment rates. This further demonstrates that systematic default and recovery risk does not coincide with business cycle risk, which is consistent with the related findings in Das et al. (2007), Bruche and Gonzlez-Aguado (2010) and Creal et al. (2011).

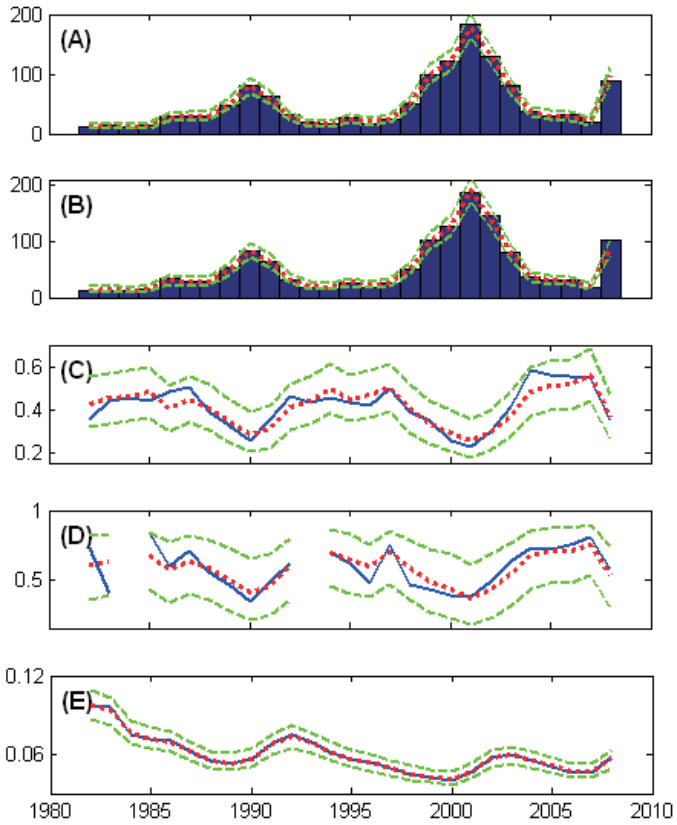

FIGURE 2.9: Model Fitting using Two-Factor Generalized Dynamic Factor Model. The solid lines (and the histograms for the first two panels) show real historical data, the dotted lines indicate posterior means and the dashed lines show posterior 95% C.I. (A): Annual Default Counts: Speculative-Grade Cooperates. (B): Annual Default Counts: Investment-Grade Cooperates. (C): Annual Recovery Rates: All Bonds. (D): Annual Recovery Rates: Senior-Secured Bonds. (E): Annual Unemployment Rate.

2.6 Discussion

This chapter proposes the use of Bayesian generalized dynamic factor models for jointly modeling mixed-measurement time series, and efficient computational algorithms are developed. The proposed framework allows mixed-scale measurements associated with each time series, with different measurements having different distributions in the exponential family, and thus provides a more flexible framework

for multivariate time series analysis. Although our illustration examples are based on specific k-factor models, the broader question of inference on the number of factors and model selection in (generalized) factor analysis, especially in the mixed-measurement observation framework, is interesting and challenging. Reversible jump MCMC algorithms can be used to allow for model uncertainly (Lopes and West (2004) and Lopes et al. (2011)). Alternatively, sparse spike-and-slab priors were proposed to induce sparsity in the factor loading matrix (see e.g. West (2003) and Carvalho et al. (2008b)). More recently, non-parametric infinite factor models were introduced using either a multiplicative gamma process shrinkage prior on factor loadings (Bhattacharya and Dunson (2011b)) or Indian-Buffet Process (IBP) prior on the factor loading matrix (Knowles and Ghahramani (2010)). However, translating these non-parametric approaches to the generalized dynamic factor model framework with mixed measurement is challenging and as yet unexplored.

2.7 Supplementary Materials

2.7.1 Greedy Density Kernel Approximation

Suppose that we have N training pairs $\{\Gamma_t^{(i)}, \eta_t^{(i)}\}$ $(i = 1, \ldots, N)$.

1. Start with σ^2 given, find the optimum density kernel approximation by:

   (a) Find the $\eta_t^{(i)}$ giving the maximum $\Gamma_t^{(i)}$, place one Gaussian kernel at this $\eta_t^{(i)}$.

   (b) Calculate the optimum weight(s) satisfying (2.8).

   (c) Use forward variable selection algorithm to add an additional Gaussian center. This basically sets the mean of an additional Gaussian kernel at the $\eta_t^{(i)}$ which gives the largest error according to approximation obtained in step (b).

(d) Go to step (b). Calculate the optimum weights satisfying (2.8), with one more Gaussian kernels.

(e) Continue the iteration until number of kernels reach the maximum.

2. Set $\sigma^2 = \sigma^2 \times c$ ($c > 1$, say c=2), find optimum approximation using step 1.

3. Compare the approximation errors in the two cases:

   (a) If error is reducing, go to step 2 and keep trying approximating using smaller variance

   (b) If error is increasing, set c=1/c, go to step 2, and trying to approximating using bigger variance

4. Continue exploring σ^2 until it gives the smallest error among σ^2, σ^2/c and $c\sigma^2$.

2.7.2 Supplementary Tables and Figures

Table 2.5: MCMC Summary using WinBUGS for T=100 and T=150 respectively, as compared to Table 2.1. Using T=100 and T=150 as examples, in each case, effective sample size (ESS) and sample autocorrelation with lag 10 and 30 of 8000 posterior samples are shown for selected parameters as representatives (c, d, η_1 and η_{50} are defined as in (2.24)).

# of Time Points	T=100				T=150			
Parameters	c	d	η_1	η_{50}	c	d	η_1	η_{50}
ESS	4.36	6.21	5.44	5.38	5.06	4.87	5.05	5.16
Autocorr-Lag 10	0.935	0.920	0.988	0.931	0.913	0.933	0.988	0.926
Autocorr-Lag 30	0.919	0.899	0.971	0.916	0.899	0.917	0.972	0.911

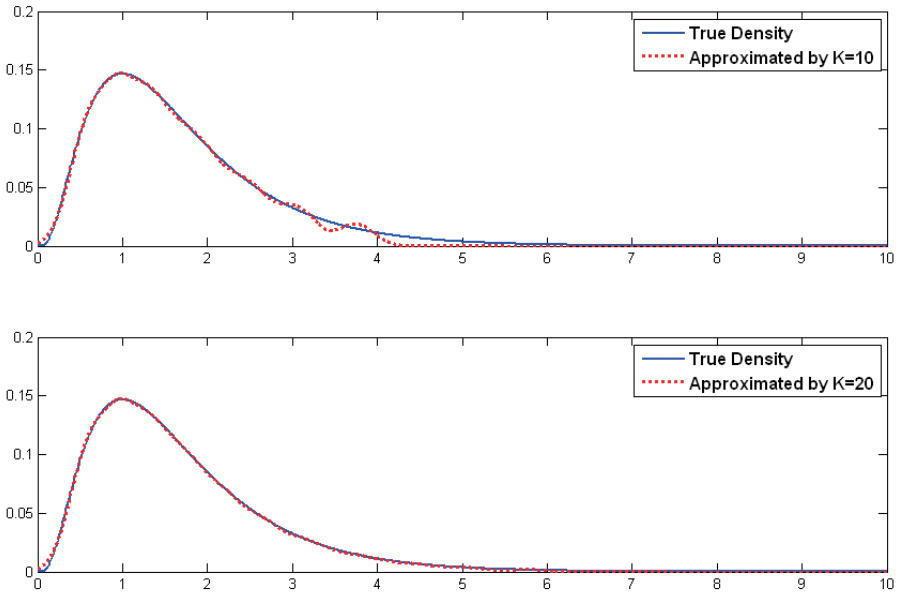

FIGURE 2.10: Arbitrary Density Kernel Approximation via a Mixture of Normal Kernels. An example of approximating an arbitrary density kernel $f(x) = e^{-x}x^2 \times$ lognormpdf$(x, 0, 1)$ on the compact domain [0,10], with a mixture of K=10 or 20 Gaussian kernels, are shown.

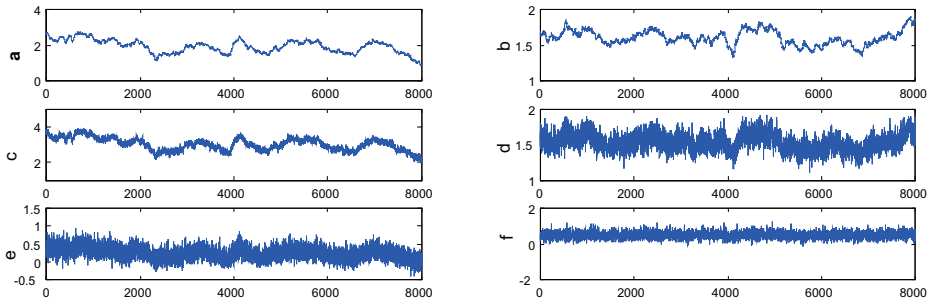

FIGURE 2.11: Traceplot Produced by WinBugs of Posterior Samples of Model Parameters. The traceplots are compared with those in Figure 2.2, showing that our algorithm improve the mixing of the Markov Chain.

Table 2.6: MCMC Summary for the mixed-measurement-mixed-frequency simulation study shows good mixing of the Markov Chain, with T=100 (selected parameters shown). 10000 posterior samples are drawn for each unknown with the first 2000 as burn-in. Effective sample size (ESS) and sample autocorrelation with lag 10 and 30 of the 8000 posterior samples are shown for selected parameters as representatives (defined as in (2.26)).

Parameters	a_2	b_2	c_2	$\eta_{1,1}$	$\eta_{1,90}$	$\eta_{2,1}$	$\eta_{2,90}$
ESS	3398	1818	7789	1605	5030	2659	4057
Autocorr-Lag 10	0.039	0.061	0.013	0.022	0.012	0.029	0.017
Autocorr-Lag 30	0.015	-0.017	0.013	-0.020	0.007	0.013	0.011

Table 2.7: Accuracy of inference to unknown quantities. Selected model parameters are shown. 100 simulated datasets are genenerated for different T, and the numbers of cases when posterior C.I.s do not cover the true values are shown.

T (# of Time Points)		50	100	150	200	500
True values not covered for	a	0/100	0/100	1/100	1/100	2/100
	b	0/100	0/100	0/100	0/100	0/100
	η_1	0/100	0/100	0/100	0/100	1/100
	η_{50}	0/100	0/100	0/100	0/100	0/100

Table 2.8: Acceptance Rate for the Metropolis Hastings update with GDKA for latent factors.

T (# of Time Points)	30	50	100	150	200
Acceptance Rate	79.5%	73.1%	55.4%	34.3%	25.4%

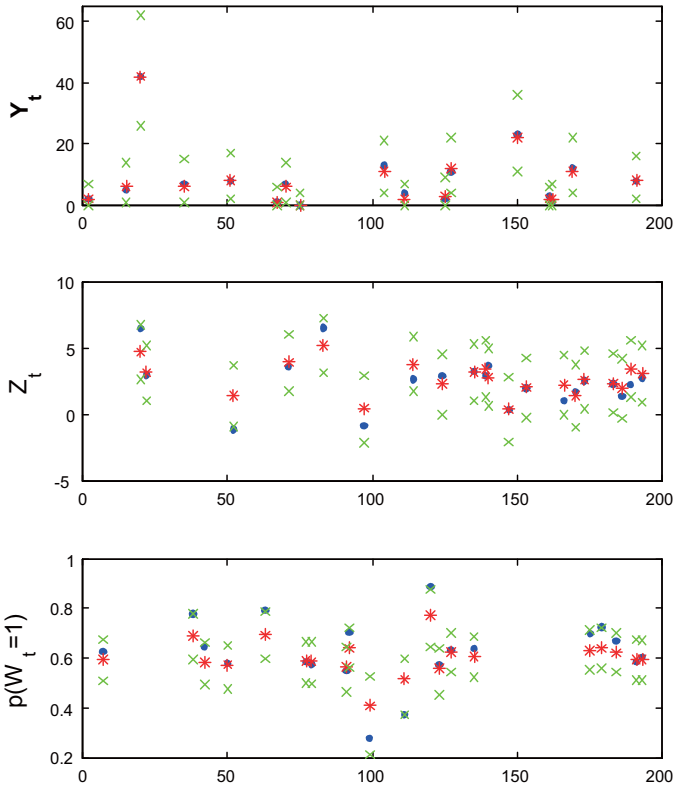

FIGURE 2.12: Missing Data Interpolation. Missing data in the 3 time series are interpolated from the marginal posterior distributions of the missing values. The dots are real simulated values, the stars are means of the interpolated data, and the crosses show the corresponding 95% intervals.

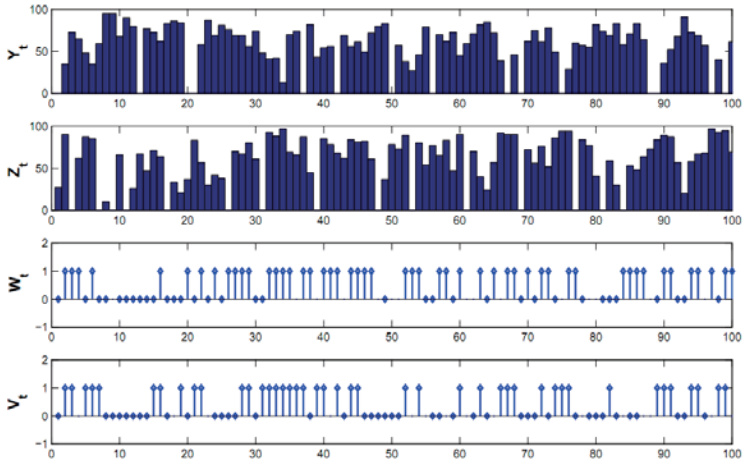

FIGURE 2.13: Simulated Data: a 4-dimensional mixed discrete time series $D_t = [Y_t, Z_t, W_t, V_t]$ up to T=100 is simulated, where Y_t and Z_t are time series with counts from binomial distributions, and W_t and V_t are binary time series.

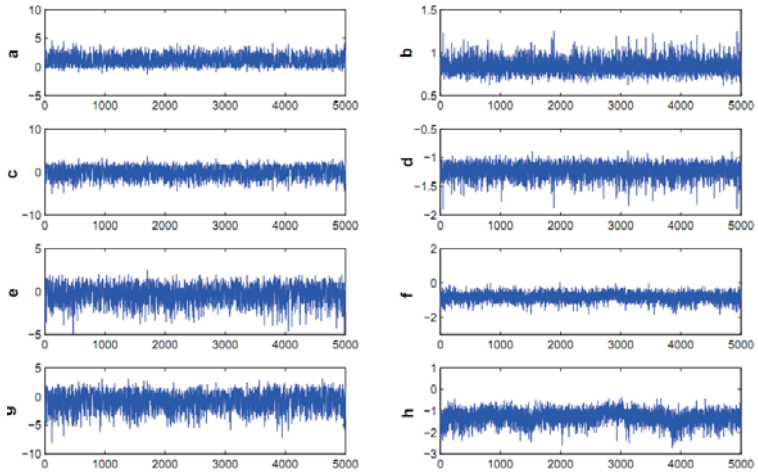

FIGURE 2.14: Traceplot of Posterior Samples of Model Parameters: Intercepts (a,c,e,g) and factor loadings (b,d,f,h) shown. The parameters are defined in (2.25).

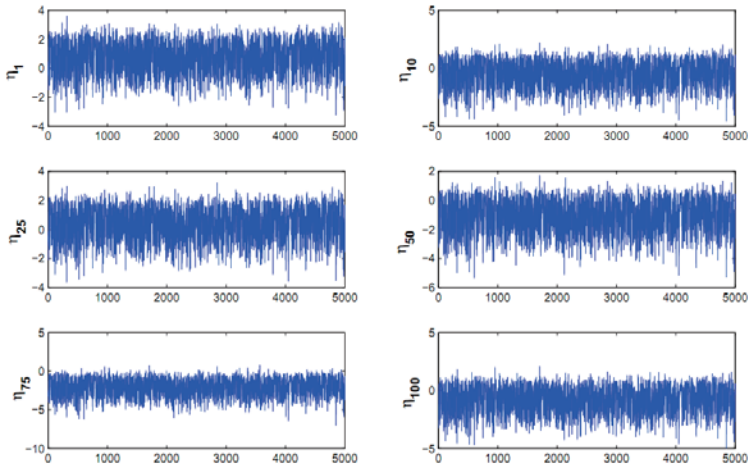

FIGURE 2.15: Traceplot of Posterior Samples of Latent Factors: Six time points shown. The latent factors are defined in (2.25).

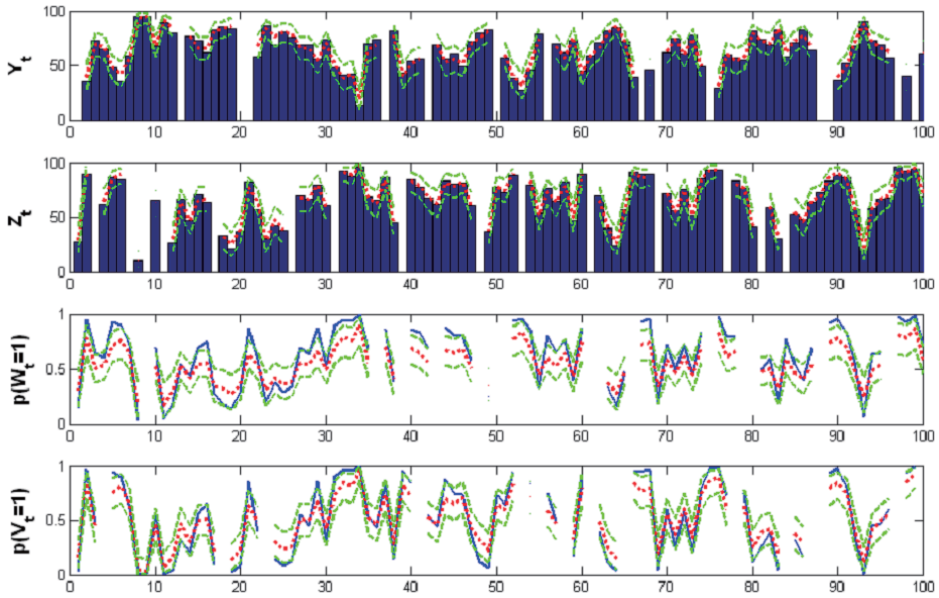

FIGURE 2.16: Interval Validation to show the model fitting. Solid lines (and the histogram for the first two panels): simulated data; dotted lines: posterior mean; dashed lines: 95% posterior C.I.

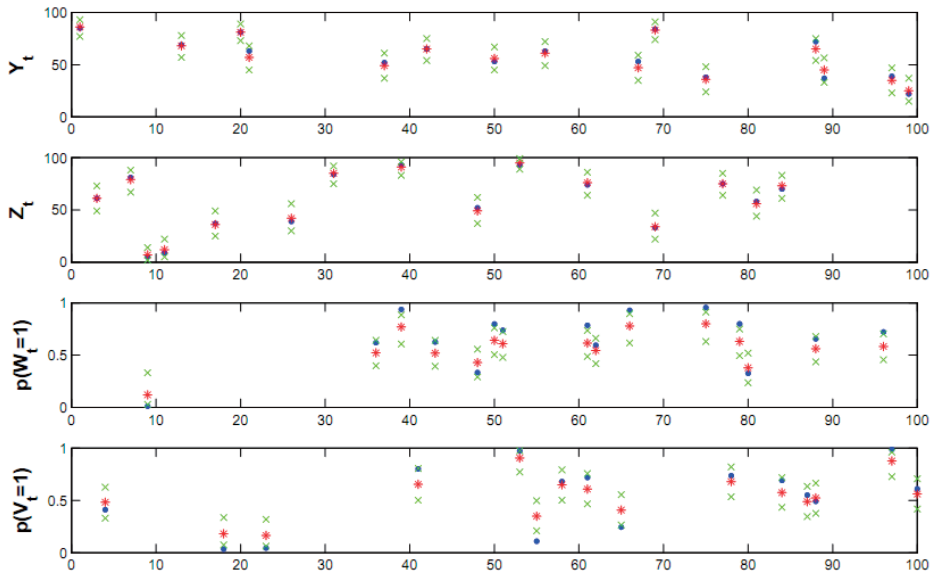

FIGURE 2.17: Missing Data Interpolation. Missing data in the 4 time series are interpolated from the marginal posterior distributions of the missing values. The dots are real simulated values, the stars are means of the interpolated data, and the crosses show the corresponding 95% intervals.

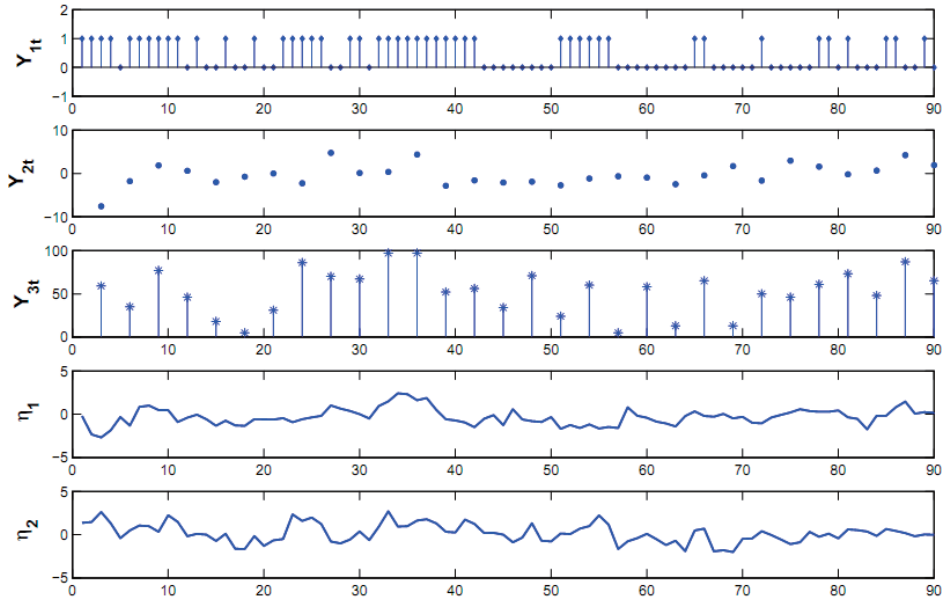

FIGURE 2.18: Simulated Data: a 3-dimensional mixed-measurement-mixed-frequency time series $Y_t = [Y_{1t}, Y_{2t}, Y_{3t}]$ is simulated, where Y_{1t} is binary with frequency $3\omega_0$, Y_{2t} is continuous with frequency ω_0, and Y_{3t} are counts from binomial distributions with frequency ω_0. The two time-varying latent factors η_1 and η_2 are also shown at the frequency of $3\omega_0$.

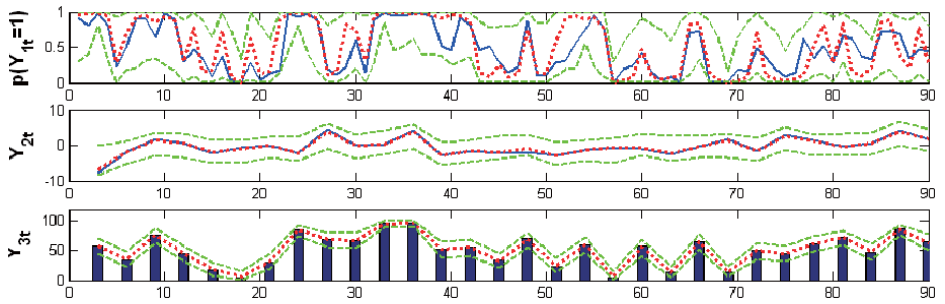

FIGURE 2.19: Interval Validation to show the model fitting for the mixed-measurement-mixed-frequency simulation study. Solid lines (and the histogram Y_{3t}): simulated mixed-measurement-mixed-frequency data (where Y_{1t} is binary with frequency $3\omega_0$, Y_{2t} is continuous with frequency ω_0, and Y_{3t} are counts from binomial distributions with frequency ω_0.); dotted lines: posterior mean; dashed lines: 95% posterior C.I.

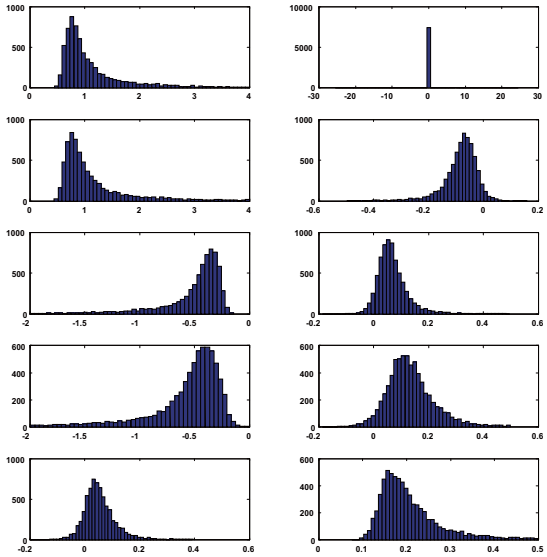

FIGURE 2.20: Histogram of Posterior Samples of Factor Loadings Λ_is ($i = 1, \ldots, 5$) in the Two-Factor Model. The factor loadings are defined in (2.27).

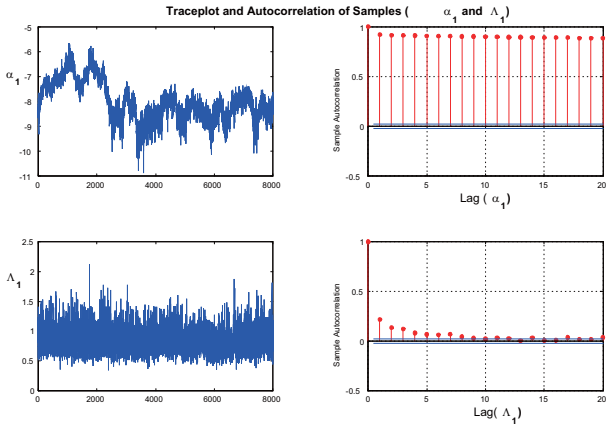

FIGURE 2.21: Traceplots and Autocorrelation Plots of Posterior Samples with parameter expansion but without the Latent Factor Normalization step. Compared to Figure 2.6, this indicates that Metropolis Hastings with GDKA, parameter expansion and latent factor normalization all contribute to improving the efficiency of the Markov Chain. α_1 and Λ_1 are defined in (2.27).

3

Bayesian Heavy-tailed and Skewed Factor Models

3.1 Introduction

In recent years a renewed attention has been devoted to joint modeling of multivariate and high-dimensional data with possible non-normality, due to the increased collection of large volumes of heavy-tailed and/or skewed data in a variety of application areas such as signal processing, network traffic, gene expression and finance. One relevant direction has been the construction of families of flexible multivariate distributions that can directly accommodate heavy tails and skewness. Several proposals have been put forward in the literature, including skew-symmetric distributions (Wang et al. (2004)), Gaussian variance-mean mixtures (Barndorff-Nielsen and Sorensen (1982)) and mixtures of multivariate normal or heavy-tailed distributions. Related models have been successfully applied to numerous applications from a wide range of fields, but essentially entirely in univariate or low-dimensional cases.

Extension to high-dimensional problems with potential heavy-tailed and skewed observations is very challenging. On the one hand, it is challenging to directly define multivariate distributions that can flexibly characterize heavy tails and skewness

in each dimension while also appropriately characterizing the types of dependence that arise in applications. For example, Gaussian copula models can be unrealistic as there is commonly tail-dependence. Even if a distributional family with sufficient flexibility could be defined, with limited numbers of observations, estimation of high-dimensional model parameters is problematic without dimensionality reduction.

Alternatively, factor models have been widely used to explain the dependence in p-dimensional continuous data via k latent factors with $k \ll p$. Gaussian factor models assume that the latent factor distribution is multivariate normal. This gives rise to a multivariate normal marginal distribution with a sparse decomposition of the $p \times p$ covariance matrix as $\Lambda\Lambda' + \Sigma$, where Λ is a $p \times k$ factor loadings matrix and Σ is a $p \times p$ diagonal matrix with nonnegative diagonal elements. The normal assumption was introduced mainly for convenience and computational efficiency, while Bartholomew (1988) suggested that the assumption is quite robust. However, this robustness was questioned by Hesketh et al. (2003) arguing that inappropriate specifications for latent factor distributions give more errors in the prediction of latent scores, in latent factor regression and lead to misinterpretation of the estimation results. Ma and Genton (2010) also described in the generalized linear latent variable model setting that imposing the normal assumption inappropriately biased the estimates.

Using alternatives to normal distributions for the latent factors to accommodate heavy tails and skewness is needed and has been of growing interest. In signal processing, independent component analysis uses a linear combination of non-Gaussian source signals to flexibly characterize a non-Gaussian signal (Comon (1994)). In low-dimensional problems with a small number of factors, Attias (1999) used univariate mixtures of normals for the latent factors. Yung (1997) and later Montanari and Viroli (2010a) also modeled latent factors using mixtures of normals, with Montanari and Viroli (2010b) instead using skew-normals. Motivated by non-Gaussian high-dimensional microarray gene expression data, Carvalho et al. (2008a) modeled the

latent factor distributions non-parametrically using Dirichlet process priors. Yang and Dunson (2010b) used a similar idea in structural equation models with latent variable distributions completely unknown and inferred via centered Dirichlet process (CDP) and CDP mixture priors.

Although it is conceptually appealing to use an extremely flexible nonparametric approach, allowing the distribution of the latent factors and data to follow essentially any form, there is a big price to be paid for such flexibility in terms of efficiency. One aspect of efficiency relates to the well known curse of dimensionality in which even the optimal rate of estimation for a p-variate density degrades rapidly with p, necessitating an enormous sample size n for accurate estimation in the absence of constraints. Another aspect is computational efficiency, with computation in mixture models for high-dimensional data quite challenging due in large part to the multimodality that arises and difficulties in adequately exploring possible modes. Also it is often questionable whether such flexibility is needed in most applications in which there may be deviations from normality such as heavy tails and skewness without multimodality.

With this motivation, our focus is on introducing a new framework for Bayesian sparse factor modeling of high-dimensional heavy-tailed and skewed data, with an emphasis on developing an approach that is flexible enough to characterize the majority of non-Gaussian data encountered in practice while also being computationally tractable to implement and leading to good estimation and prediction performance. The proposed model is a semiparametric Gaussian variance-mean mixture factor model, with the usual Gaussian factor model arising as a limiting case. We generalize the approach of Bhattacharya and Dunson (2011a) for efficient inference under Gaussian factor models with unknown numbers of factors to the heavy-tailed and skewed case.

3.2 Model Specification

3.2.1 Basic Factor Model Structure

The basic factor model is defined as follows:

$$y_i = \mu + \Lambda f_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Sigma), \quad i = 1, \ldots, n, \tag{3.1}$$

where Λ is a $p \times k$ matrix of factor loadings, $f_i = (f_{i1}, \ldots, f_{ik})'$ is a vector of latent factors, and ϵ_i is a residual with diagonal covariance matrix $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. The introduction of latent factors f_is induces dependence, as the covariance structure Ω of the marginal distribution of y_i has the form: $\Omega = \Lambda \text{Cov}(f)\Lambda' + \Sigma$. In practice, the number of latent factors is assumed to be small relative to the number of outcomes ($k << p$), leading to sparse models for Ω.

The latent factors f_is are assumed to be independently drawn from $F(\cdot)$. Traditionally, $F(f_i) = N(f_i|0, I)$ is a multivariate normal distribution with zero mean and identity covariance, inducing $y_i \sim N_p(0, \Omega)$ with $\Omega = \Lambda\Lambda' + \Sigma$. We propose to flexibly model the latent factor distribution via semi-parametric Gaussian variance-mean mixtures. This family of distributions is quite flexible in accommodating a rich variety of symmetric or asymmetric distributions with different skewness, conditional structures and tail behavior, including tail dependence. Appropriately accommodating tail dependence is crucial in many applications; if tail observations corresponding to anomalies or extreme events occur in concert, assuming tail independence can lead to a dramatic underestimation of risk. Such overly-simplified modeling may have played a large role in the recent financial crisis, for example.

3.2.2 Semi-Parametric Gaussian Variance-Mean Mixture Factor Model

Without going into too much detail on Gaussian variance-mean mixtures (GVMM) (see e.g. Barndorff-Nielsen (1982) and references therein), GVMM is a family of heavy-tailed and/or skew distributions introduced by Barndorff-Nielsen (1977) as a

flexible class of multivariate distributions induced through the following hierarchical model for $x_i = (x_{i1}, \ldots, x_{ip})'$,

$$x_i = \mu + \gamma V_i + \sqrt{V_i}\Sigma_z^{1/2} z_i, \quad z_i \sim N(0, I_p), \;\; V_i \sim G, \tag{3.2}$$

where $\mu \in \Re^p$ is a location parameter, $\gamma \in \Re^p$ is a drift or skewness parameter, G is a mixing distribution on $[0, \infty)$, $z_i \perp V_i$, and Σ_z is a positive definite matrix. With different choices of the mixing distribution G, members of this family include a variety of widely used distributions with heavy-tails and/or skewness, such as Student t, Laplace, hyperbolic, normal inverse Gaussian and variance gamma distributions (see e.g. Barndorff-Nielsen (1977), Barndorff-Nielsen et al. (1990), Barndorff-Nielsen (1982), Barndorff-Nielsen (1997), Arslan and Genc (2009) and Arslan (2010)).

Our model for high-dimensional data with potential non-normality maintains the basic factor model framework (3.1), while modeling the latent factor distributions via semi-parametric Gaussian variance mean mixtures. In details:

$$\begin{aligned} & y_i = \mu + \Lambda f_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Sigma), \\ & \text{where } f_i = (f_{i1}, \ldots, f_{ik})', \quad i = 1, \ldots, n. \\ & f_{ij} = \alpha_j + \gamma_j V_{ij} + \sqrt{V_{ij}} z_{ij}, \quad j = 1, \ldots, k. \\ & V_{ij} \sim G_j, \quad z_{ij} \sim N(0, 1), \end{aligned} \tag{3.3}$$

where γ_j controls the skewness of the marginal density for the jth factor, with $\gamma_j = 0$ inducing a symmetric scale mixture of Gaussians, and V_{ij} is a mixing variable that is independent of z_{ij}. By varying the mixing distributions G_j, the latent factor densities can be allowed to have a flexible variety of heavy-tailed and skewed forms. This is achieved by specifying Dirichlet process mixture (DPM) priors for the distribution of V_{ij}^*, where $V_{ij}^* = \log(V_{ij})$:

$$\begin{aligned} & V_{ij}^* = \log(V_{ij}) \sim N(\mu_j, \psi_j^{-1}), \\ & (\mu_j, \psi_j) \sim H_j, \quad H_j \sim DP(\alpha_0 H_0). \end{aligned} \tag{3.4}$$

Here, we use DPMs only for the densities of $\log(V_{ij})$ instead of characterizing the latent factor densities directly as DPMs of multivariate normals in order to maintain stronger regularity constraints on the densities; as mentioned above we do not want to allow arbitrarily complex multimodal densities. Such densities are very hard to estimate based on limited data, and this degree of flexibility is often not needed in practice.

Importantly, the model can be represented in a vector form, which shall be used for block updating of latent factors and factor loadings to gain computational efficiency:

$$\begin{aligned} y &= \mu + \Lambda f + \epsilon, \quad \epsilon \sim N_p(0, \Sigma), \\ f &\sim N(\alpha + \Gamma V, \Psi = \text{diag}(V)), \\ \log(V_{\cdot j}) &\sim N(\mu_j, \psi_j^{-1}), \\ (\mu_j, \psi_j) &\sim H_j, \quad H_j \sim DP(\alpha_0 H_0), \end{aligned} \tag{3.5}$$

where $f = (f_{\cdot 1}, \ldots, f_{\cdot k})'$ is the latent factor vector, $\alpha = (\alpha_1, \ldots, \alpha_k)'$, $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_k)$, $V = (V_{\cdot 1}, \ldots, V_{\cdot k})'$ is a vector of independent mixing variables ($\forall m \neq n$, $V_{\cdot m} \perp V_{\cdot n}$) with diagonal covariance matrix Σ_V, and $\Psi = \text{diag}(V)$ is also a diagonal covariance matrix.

Gaussian variance-mean mixture factor models are similar to Gaussian latent factor models in that the y_is are still conditionally Gaussian distributed and linearly related with the latent factors. The marginal covariance matrix of y_i is now given by $\Omega = \Lambda \Omega_f \Lambda' + \Sigma$, with Ω_f being the marginal covariance matrix of latent factors with $\Omega_f = \Gamma \Sigma_V \Gamma + \Psi$, where Σ, Γ, Σ_V and Ψ are defined as in (3.5).

However, non-Gaussian latent factor distributions lead to non-Gaussian marginal distributions of y_i and non-Gaussian conditional distributions between elements of y_i. The following theorem illustrates the flexible non-Gaussian conditional dependence structures between pairs of variables that can be obtained, showing conditional

moments, skewness, and kurtosis.

Theorem 1. *Let $y_i = (y_{i1}, \ldots, y_{ip})'$ be a sample of a $p \times 1$ random variable as defined in (3.5), with $G = (G_1, \ldots, G_k)$ being the mixing distributions. Then, for any arbitrary pair j and l ($1 \leqslant j \neq l \leqslant p$), the conditional distribution $y_{ij}|y_{il}$ is non-Gaussian with*

1. $E(y_{ij}|y_{il}) = \alpha_{j|l}^{(0)} + \alpha_{j|l}^{(1)}(y_{il} - \mu_l)$

   where $\alpha_{j|l}^{(0)} = \mu_l + \Lambda_j E_G\{\phi_1(V)\}$ *and* $\alpha_{j|l}^{(1)} = \Lambda_j E_G\{\phi_2(V)\}$

2. $Var(y_{ij}|y_{il}) = \beta_{j|l}^{(0)} + \beta_{j|l}^{(1)}(y_{il} - \mu_l) + \beta_{j|l}^{(2)}(y_{il} - \mu_l)^2$

   where $\beta_{j|l}^{(0)} = \Sigma_{jj} + \Lambda_j \left[E_G\{\phi_0(V)\} + Var_G\{\phi_1(V)\}\right] \Lambda_j',$

   $\beta_{j|l}^{(1)} = 2\Lambda_j Cov_G\{\phi_1(V), \phi_2(V)\}\Lambda_j',$

   $\beta_{j|l}^{(2)} = \Lambda_j \left[Var_G\{\phi_2(V)\}\right] \Lambda_j'.$

3. *Conditional skewness* $\gamma_{y_{ij}|y_{il}}^{(1)} = \dfrac{\sum_{m=0}^{3} E_G\{\tilde{\phi}_m(V)\}(y_{il} - \mu_l)^m}{(\beta_{j|l}^{(0)} + \beta_{j|l}^{(1)}(y_{il} - \mu_l) + \beta_{j|l}^{(2)}(y_{il} - \mu_l)^2)^{3/2}}$

4. *Conditional kurtosis* $\gamma_{y_{ij}|y_{il}}^{(2)} = \dfrac{\sum_{q=0}^{4} E_G\{\hat{\phi}_q(V)\}(y_{il} - \mu_l)^q}{(\beta_{j|l}^{(0)} + \beta_{j|l}^{(1)}(y_{il} - \mu_l) + \beta_{j|l}^{(2)}(y_{il} - \mu_l)^2)^2}$

where Λ_j denotes the j^{th} row of Λ; μ_l denotes the l^{th} element of μ. $\phi_0(\cdot)$, $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are $\mathcal{R}^k \to \mathcal{R}^k$ projection functions of the $k \times 1$ random variable V, and $\tilde{\phi}_m(\cdot)s$ and $\hat{\phi}_q(\cdot)s$ are $\mathcal{R}^k \to \mathcal{R}$ projection functions of V. Given the projection functions, $E_G(\phi_1(V))$, $Var_G\{\phi_1(V)\}$ and $Cov_G\{\phi_1(V), \phi_2(V)\}$ are the corresponding expectations, variances and covariances with respect to the mixing distributions $G = (G_1, \ldots, G_k)$.

The proof and derivations of the projection functions are shown in the appendix. Observe that although the conditional distributions are no longer Gaussian, the conditional expectation $E(y_{ij}|y_{il})$ is still linear with respect to y_{il} aiding interpretability.

However, the variance and higher moments of the conditional distribution of y_{ij} given y_{il} are polynomial functions of y_{il}, with these functions depending on model parameters.

These properties also illustrate that the model accommodates flexible modeling of the conditional distributions, which is crucial in many applications. For example, within a specific range of y_{il}, given different values of $\beta_{j|l}^{(m)}$s ($m = 0, 1, 2.$) and μ_l, the conditional variance $Var(y_{ij}|y_{il})$ can be either increasing, decreasing or hyperbolic with increased y_{il}, while the conditional variance of a Gaussian factor model is a special case with $\beta_{j|l}^{(1)} = \beta_{j|l}^{(2)} = 0$, leading to a constant conditional variance. To illustrate, simulated samples from GVMM factor models are plotted in Figure 3.1. Observe that (as shown in the first-column of the figure) with increased y_{i1}, it seems that $Var(y_{i2}|y_{i1})$ is decreasing, $Var(y_{i3}|y_{i1})$ is increasing while the conditional distribution of $y_{i4}|y_{i1}$ displays an interesting bipolar star shape that may result from a combination of local variations of conditional variance, skewness and kurtosis. The figure also shows a variety of conditional structures the framework can flexibly provide with additional skewness and kurtosis.

3.2.3 Sparse Gaussian Variance-Mean Mixture Infinite Factor Model

In practice the number of factors is typically unknown *a priori*, but can be challenging to infer from the data. Standard Bayesian methods for factor selection rely on choosing the model having the highest marginal likelihood of the data, but it can be highly challenging to accurately estimate the marginal likelihoods for models containing different numbers of factors. Lee and Song (2002) proposed to use path sampling in this context, with Ghosh and Dunson (2008) proposing path sampling with parameter expansion. Dutta and Ghosh (2010) argued that existing path sampling algorithms have flaws and suggested an improvement. Lopes and West (2004) instead proposed to use reversible jump MCMC. None of these approaches

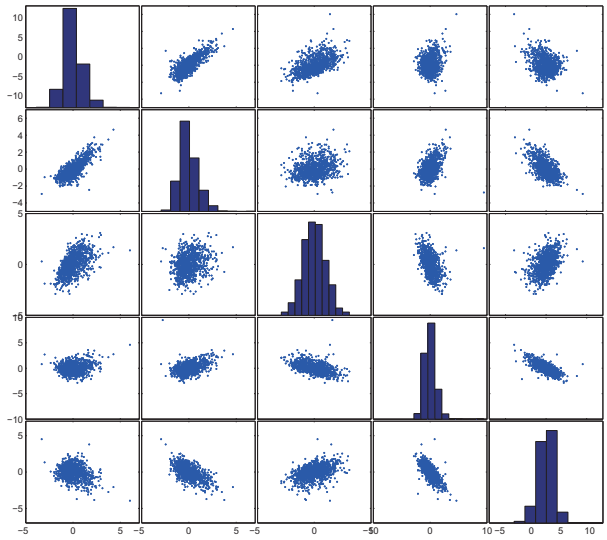

FIGURE 3.1: Scatter plots of samples from GVMM factor models.

scale very well to problems involving more than a few factors and modest numbers of observed variables, motivating Carvalho et al. (2008a) to use evolutionary stochastic search algorithms. Recently, Bhattacharya and Dunson (2011a) proposed a more efficient alternative based on embedding in a model with infinitely many factors; this approach has been successfully applied in about a dozen applications including with high-dimensional data; refer, for example to Runcie and Mukherjee (2012) and D. Durante (2012)

There is a parallel machine learning literature relying on latent feature models corresponding to hierarchical Beta and Indian buffet processes. For recent references, refer to Chen et al. (2010a) and Knowles and Ghahramani (2011). Such models allow each individual to select a finite number of latent features from an infinite "buffet", with the features typically drawn from Gaussian priors. This is a type of sparse Gaussian linear latent factor model, which accomplishes factor selection in a very different manner. By allowing different numbers of latent factors for each individual,

one induces a non-Gaussian marginal distribution, though it is not clear whether such marginals are flexible enough to approximate a rich variety of heavy-tailed and skewed distributions. In addition, posterior computation is daunting due to the vast number of different possibilities in terms of feature selection; it is an NP-hard problem to explore the space of possible allocations.

The Bhattacharya and Dunson (2011a) approach uses scale mixtures of Gaussian shrinkage priors on the loadings to facilitate efficient computation via blocked updating from conjugate distributions, while also allowing the number of factors to be unknown. This approach is dramatically more efficient computationally than alternatives in the Gaussian linear factor model setting, so we are motivated to generalize the approach to the heavy-tailed and skewed case. In particular, choosing an upper bound k^* for the number of factors needed, we let

$$\Lambda_{hj}|\phi_{hj},\tau_j \sim N(0,\phi_{hj}^{-1}\tau_j^{-1}), \quad \phi_{hj} \sim \text{Gamma}(v/2,v/2), \quad \tau_j = \prod_{l=1}^{j}\delta_l,$$

$$\delta_1 \sim \text{Gamma}(a_1,1), \quad \delta_l \sim \text{Gamma}(a_2,1),\ (l \geqslant 2).(j=1,\ldots,k^*; h=1,\ldots,p), \tag{3.6}$$

where τ_j is a global shrinkage prior for the j^{th} column of Λ and ϕ_{hj}s are local shrinkage parameters for the elements in the j^{th} column. Global-local shrinkage priors have been shown in simpler settings to have appealing properties (Polson and Scott (2011)).

The shrinkage priors in (3.6) favor increasing numbers of loadings that are very close to zero in the later columns of the loadings matrix through inducing a hierarchy on the variances. In particular, many loadings will effectively have a normal prior with variance very close to zero, with the approach allowing uncertainty in the specific loadings that are close to zero. The criteria for determining the effective number of factors, prediction and covariance matrix estimation based on the framework are discussed in the following sections.

3.3 Bayesian Computation and Posterior Analysis

3.3.1 MCMC schemes and computational efficiency

In addition to being able to flexibly model the latent factor distributions, another advantage of our framework is the computational tractability based on MCMC algorithms, which can be scaled up to high dimensions. Due to the conditional multivariate normal structure of our model (as described in (3.5)), the latent factors and factor loadings can be easily updated in blocks simply via conditional multivariate normal distributions. On the other hand, sampling parameters and mixing variables (V) in the semiparametric specification of latent factors are easily parallelizable, which leads to further computational speed-up via parallel computing toolboxes or GPU parallel computing (see e.g. Keckler et al. (2011) and Suchard et al. (2010), not covered in our study). Most of other parameters can also be updated efficiently given the conditionally conjugate priors. In details, the sequence of conditional posteriors to sample is as follows, and the mixing performances of MCMC are evaluated in data analysis sections.

1. Sample latent factors $f_{1:n}$, where $f_i = (f_{i1}, \ldots, f_{ik})'$, $i = 1, \ldots, n$.

   Letting $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_k)$, $V_i = (V_{i1}, \ldots, V_{ik})'$, and $\Psi_i = \text{diag}(V_i)$, the full conditional distribution of f_i is

   $$f_i | \ldots \sim N(\mu_{f_i}, \Sigma_{f_i})$$

   where $\Sigma_{f_i}^{-1} = \Lambda' \Sigma^{-1} \Lambda + \Psi_i^{-1}$, and $\mu_{f_i} = \Sigma_{f_i} \left(\Lambda' \Sigma^{-1} (y_i - \mu) + \Psi_i^{-1} (\alpha + \Gamma V_i) \right)$

2. Sample factor loading matrix Λ and covariance matrix $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ (or the corresponding precision matrix $\Phi = \text{diag}(\phi_1^2, \ldots, \phi_p^2)$).

   Given the multivariate normal-gamma conjugate priors for the factor loadings $\Lambda_{i\cdot}$ (where $\Lambda_{i\cdot}$ is the non-zeros elements of the i^{th} row of the factor loading

matrix Λ) and the precision ϕ_i, for any given i, the full conditional distribution of $\Lambda_{i\cdot}$ and ϕ_i is still multivariate normal-gamma distribution.

3. Sampling parameters in the Gaussian variance mean mixtures are trivial. This includes updating α, Γ and V in model (3.5). It is worth noting that sampling these parameters in a given MCMC iteration are parallelizable.

4. Sampling global and local shrinkage parameters ϕ_{jh} and τ_h in the semi-parametric Gaussian variance-mean infinite factor model, as described in Bhattacharya and Dunson (2011a).

To evaluate the computational efficiency of the approaches and mixing performance of MCMC up to high dimensions, we monitored the traceplot and autocorrelation of several norms of the marginal covariance matrix (Ω) samples, which are obtained as described later in Section 3.3.2. Matrix norms used in the studies include the matrix 1-norm ($||\Omega||_1$), which corresponds to the maximum of the column sums; the spectral norm ($||\Omega||_2$), which is the square root of the largest eigenvalue of $\Omega'\Omega$; and the Frobenius norm ($||\Omega||_F$). Mathematically, given the marginal covariance matrix $\Omega = \{\omega_{ij}\}$ $(i, j = 1, \ldots, p)$, the norms $||\Omega||_1$, $||\Omega||_2$ and $||\Omega||_F$ are defined as follows:

$$
\begin{aligned}
||\Omega||_1 &= \max_{1 \leqslant j \leqslant p} \sum_{i=1}^{p} |\omega_{ij}|, \\
||\Omega||_2 &= \sqrt{\lambda_{\max}(\Omega'\Omega)}, \\
||\Omega||_F &= \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{p} |\omega_{ij}|^2} = \sqrt{\mathrm{trace}(\Omega'\Omega)}.
\end{aligned}
\tag{3.7}
$$

3.3.2 Effective number of factors and covariance matrix estimation

In practice, the number of latent factors is often unknown *a priori* and expected to be significantly less than the dimension p. In our proposed Gaussian variance-mean

mixture infinite factor models framework, given a pre-specified truncated number k^* for the maximum latent factor number. The effective number of latent factors k can be obtained from the posterior samples of factor loadings Λ and latent factors f.

At the t^{th} iteration of MCMC, let $\hat{s}_j^{(t)}$ denote the sample standard deviation of the j^{th} latent factor $f_{\cdot j}^{(t)}$ $(j = 1, \ldots, k)$, and $\Lambda_{\cdot j}^{(t)}$ denote the j^{th} column of the factor loading matrix $\Lambda^{(t)}$. We define the normalized factor loading matrix $\tilde{\Lambda}^{(t)}$ at iteration t by letting

$$\tilde{\Lambda}_{\cdot j}^{(t)} = \Lambda_{\cdot j}^{(t)} / \hat{s}_j^{(t)}.$$

Further let $m^{(t)}$ denote the number of columns of the normalized factor loading matrix $\tilde{\Lambda}^{(t)}$ with all elements within a pre-specified threshold distance from zero. Then $k = k^{*(t)} - m^{(t)}$ is a sample of the effective number of latent factors at iteration t. This approach produces an accurate estimate of the true number of latent factors k in the simulation studies shown in later sections.

An approximated marginal covariance matrix $\Omega^{(t)}$ of y_i can also be obtained by letting $\Omega^{(t)} = \tilde{\Lambda}^{(t)}\tilde{\Lambda}^{(t)\prime} + \Sigma^{(t)}$ at the t^{th} iteration, and then averaging across the iterations after discarding a burn-in. Interval estimates for the elements of the covariance matrix and for any functional can be obtained in the usual manner from these samples.

3.3.3 Prediction and classification

In statistics and machine learning, it is common to be interested in prediction based on high-dimensional features using limited training data. Latent factor regression has been proposed in such "large p, small n"problems. Specifically, let $y_i = (z_i, x_i')'$, $i = 1, \ldots, N$, where x_is are p dimensional predictors and z_is are the response. Most previous studies assume that the predictors, response and latent factors are all continuous and normally distributed, and thus y_is can can be jointly modeled via a

typical Gaussian factor model. The normality assumption can badly restrict the predictive accuracy for non-Gaussian data.

If we use posterior predicative means as point estimates, prediction and classification with Gaussian variance-mean mixture factor models are similar to those with Gaussian latent factor regression models, because Theorem 3 clearly implies that $E[z_i|x_i] = \alpha^{(0)}_{z|x} + \alpha^{(1)}_{z|x}(x_i - \mu_x)$, where μ_x is a sub-vector of μ; $\alpha^{(0)}_{z|x}$ and $\alpha^{(1)}_{z|x}$ can be estimated using posterior samples of model parameters and G as shown in the theorem and appendix. However, since the specifications of non-Gaussian latent factor distributions lead to non-Gaussian marginal relationship between responses and predictors, to do prediction, we prefer combining the predictors in the testing set with all data in the training set and apply Gaussian variance-mean mixture factor models on the joint data set. Posterior predictive values of the responses in the testing set are sampled in each iteration of MCMC to obtain samples from the predictive distribution, which can be used to obtain point and interval estimates.

3.4 Simulation Studies

3.4.1 Study I: Number of factors known.

We consider a number of simulation examples to illustrate that our approach can effectively capture the underlying structure (e.g. the covariance matrix, marginal and latent factor distributions) with multivariate or high dimensional data sets. We simulated p-dimensional y_i given k-dimensional latent factors f_i, $i = 1, \ldots, n = 100$, from a p-dimensional conditional normal distribution $N(\mu + \Lambda f_i, \Sigma)$, where Λ is a $p \times k$ factor loading matrix and Σ is a diagonal matrix. We set Λ to be lower-triangular when generating simulated data. In order to illustrate the flexibility of our approach in modeling $F_j(\cdot)$, we assign a variety of either symmetric or asymmetric distributions with or without heavy tails and skewness to each $F_j(\cdot)$. Examples of such latent factor distributions used in the simulation studies include normal distribution, t-

distribution, log-normal distribution, and skewed distributions from the Gaussian variance-mean mixtures family with different mixing distributions.

We start with three different (p, k) combinations, including (200, 6), (500, 10) and (1500,10). For now, we first assume the number of latent factors is known, and the following prior distributions shown in (3.8) are placed on unknown quantities in the model specified in (3.1) and (3.5), as described previously. We run the Gibbs sampler for 10,000 iterations with a burn-in of 1000, and carefully monitor the convergence of the chain.

$$\begin{aligned}
&\mu|\Sigma \sim N(0, 100\Sigma), \quad \Lambda_{ij}|\Sigma \sim N(0, \sigma_i^2)\ (\forall j \leqslant i),\\
&\sigma_i^2 \sim \text{Inv-}\Gamma(\alpha_\sigma, \beta_\sigma)\ (i = 1, \ldots, p).\\
&\alpha_l \sim N(0, 100), \quad \gamma_l \sim N(\mu_{\gamma_l}, \sigma_{\gamma_l}^2), \quad (l = 1, \ldots, k).\\
&V_{.l}^* = \log(V_{.l}) \sim N(m_l, \psi_l), \ \ (m_l, \psi_l) \sim H_l, \ \ H_l \sim DP(\alpha_0 H_0),\\
&\text{where } H_0 : \psi_l \sim \text{Inv-}\Gamma\left(2, 2(log2/2)^2\right), \ \ m_l|\psi_l \sim N(0, 2\psi_l),\\
&\mu_{\gamma_l} \sim \Gamma(2, 2), \sigma_{\gamma_l}^2 \sim \Gamma(0.1, 0.1).
\end{aligned} \tag{3.8}$$

We will use the $(p, k) = (500, 10)$ case as an illustration. In this case, 10 latent factors are coming from 10 different distributions, including t distributions with different degrees of freedom, normal distribution, log-normal distribution and distributions from the Gaussian variance-mean mixtures, leading to a variety of non-Gaussian marginal distributions of the observations y_i (as shown in Figure 3.10). While we expect it to be hard for the traditional Gaussian factor models to capture the covariance structures accurately in this case, our proposed normal variance-mean mixture factor model estimates the covariance structure very well, illustrated by the comparison between the true covariance matrix (Figure 3.2) and the posterior mean of covariance matrices (Figure 3.3) obtained via MCMC. Furthermore, in all three cases we tested with different (p, k) combinations, the estimated covariance matrices are close to the true values, with small mean square error, average and maximum

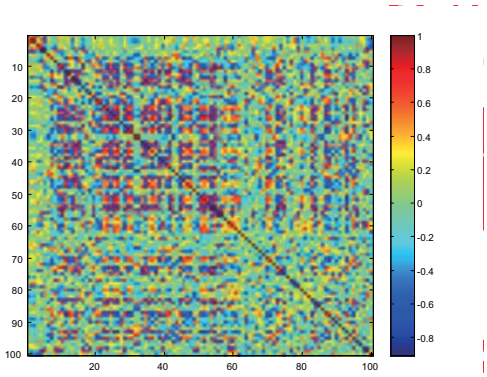

FIGURE 3.2: True covariance matrix of standardized y_i calculated based on model (3.1) and (3.5). y_i is 500 dimensional and 100 of the dimensions are shown as an illustration. This is compared with Figure 3.3.

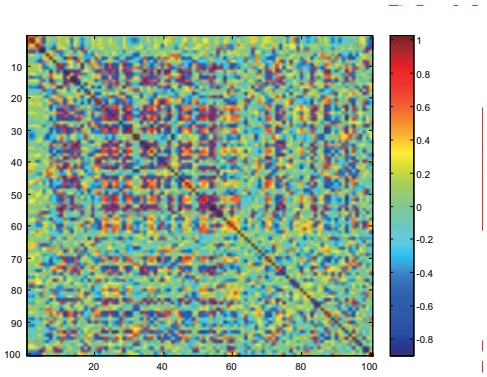

FIGURE 3.3: Posterior mean of the covariance matrix of standardized y_i using our proposed normal variance-mean factor model with number of factor fixed to be 10. y_i is 500 dimensional and 100 of the dimensions are shown as an illustration.

absolute bias. We also compared the results to those obtained by Gaussian factor models and the recent banding method by Rothman and Zhu (2010). A summary of the estimation errors and comparison is shown in Table 3.4.1. Based on Table 3.4.1, the proposed Gaussian variance-mean mixture factor model does significantly better than the other two.

We also checked whether the marginal distributions of y_i are correctly captured by looking at the inference of latent factor distributions. A comparison between Figure 3.4 (the true distributions) and Figure 3.5 (the inferred distributions) confirms that our framework captures the latent factor distributions well. Therefore, we conclude that the Gaussian variance-mean mixture factor model can effectively capture the underlying data structure up to high dimensions, when the number of factors is known.

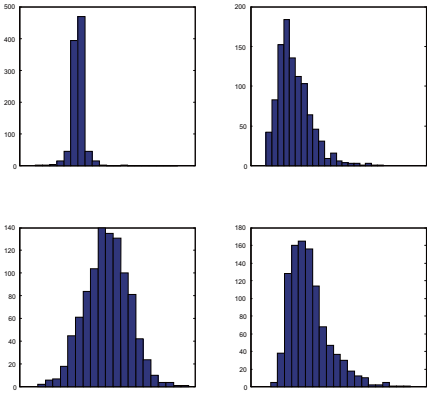

FIGURE 3.4: True latent factors distributions. The first 4 out of 10 latent factor distributions in the $(p, k) = (500, 10)$ case are shown.

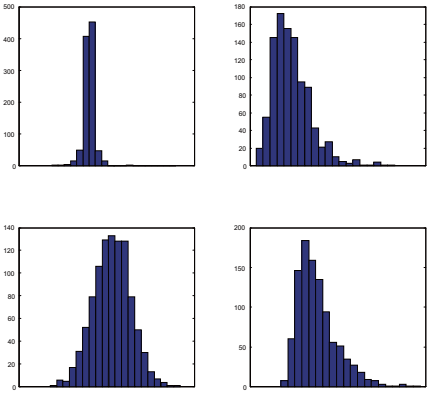

FIGURE 3.5: Latent factors distributions obtained via MCMC. The distributions are obtained in one MCMC iteration, where the posterior samples of latent factor of the first 4 dimensions are plotted.

3.4.2 Study II: with unknown number of factors

We further considered similar simulation examples while letting the number of factors be unknown, and tested whether both the covariance matrix and number of factors can be effectively inferred.

We generated $n = 100$ observations with four (p, k) combinations, where (p, k) takes (200,6), (500,10), (1500,10) and (5000, 15). The observed data were first fitted with the sparse Gaussian variance-mean mixtures infinite factor models of Bhattacharya and Dunson (2011a) truncated at a fixed level k^*. For all the three (p, k) combinations we tried, we used the truncation level $k^* = 5\log(p)$ as the maximum number of factors.

The priors we used are similar to those shown in (3.8) except that the multiplicative gamma process priors shown in (3.6) are placed on the factor loadings. The hyperparameter v in (3.6) is set to 3, while Gamma(2,1) hyperpriors are placed on

Table 3.1: Performance of covariance matrix estimation in the simulation studies when the numbers of factors are assumed to be known. Results using normal variance-mean mixture factor model (N-FM) are compared to those of the Gaussian factor model (G-FM) and the Banding method (Ba) proposed by Rothman and Zhu (2010). Mean square error (mse), average absolute bias (aab) and maximum absolute bias (mab) between estimated and true covariance matrices are compared.

(p,k)	(200,6)			(500,10)			(1500,10)		
method	N-FM	G-FM	Ba	N-FM	G-FM	Ba	N-FM	G-FM	Ba
rmse	0.0418	0.113	0.525	0.0442	0.163	0.453	0.0426	0.149	0.472
aab	0.0377	0.0854	0.452	0.0346	0.126	0.383	0.0340	0.116	0.405
mab	0.197	0.641	1.008	0.206	0.805	1.015	0.257	0.812	1.056

a_1 and a_2. We run the MCMC for 10,000 iterations with a burn-in of 5000. We monitored the columns in the normalized factor loading matrix $\tilde{\Lambda}$ having all normalized elements less than a threshold (10^{-4}) and determine the number of factors as described in Section 3.3.2. Covariance matrix estimation is also achieved as described in Section 3.3.2.

First of all, to evaluate the mixing performance of the MCMC up to high dimensions, we use the $(p,k) = (5000, 15)$ as an illustration and monitored the trace plots and autocorrelations of different norms of the marginal covariance matrix (Ω) samples, including the matrix 1-norm ($||\Omega||_1$), the spectral norm ($||\Omega||_2$) and the Frobenius norm ($||\Omega||_F$), as described in Section 3.3.1. Trace plots (Figure 3.11) suggest great mixing performance. Autocorrelations calculated (as shown in Table 3.4.2) also confirm comparable efficiency to the Gaussian infinite factor model, which has been noted to be quite computationally efficient relative to competitors (Bhattacharya and Dunson (2011a)).

The posterior means of the estimated number of factors in the four (p,k) combinations are 6.01, 10.3, 10.1 and 16.5 corresponding to k=6, 10, 10, and 15, with the empirical 95% C.I.s being (6,7), (10,11) and (10,11), (15,18) respectively. With the number of factors unknown, the estimated covariance matrix in each case is still close to the truth (illustrated in Figure 3.6 and 3.7), with significantly smaller

Table 3.2: Mixing Performance of Gaussian variance-mean mixture infinite factor model (N-IFM) compared to that of Gaussian infinite factor model (G-IFM). Autocorrelations (AC) of different norms of the MCMC samples of the marginal covariance matrix (Ω), including the matrix 1-norm ($||\Omega||_1$); the spectral norm ($||\Omega||_2$) and the Frobenius norm ($||\Omega||_F$), are monitored here, with $(p, k) = (5000, 15)$

	$\|\|\Omega\|\|_1$		$\|\|\Omega\|\|_2$		$\|\|\Omega\|\|_F$	
.	N-IFM	G-IFM	N-IFM	G-IFM	N-IFM	G-IFM
AC-Lag10	-0.0036	0.0023	0.015	-0.0089	0.0037	-0.0052

mean square error, average and maximum absolute bias than the other two methods compared (as shown in Table 3.4.2). Another quick comparison between Table 3.4.2 and Table 3.4.1 shows that the errors are slightly bigger than those when the numbers of factors are known, because we incorporate the uncertainty about the number of factors in the models, while still significantly smaller than those of the banding method. Furthermore, if we simply ignore the fact that y_is and latent factors are non-Gaussian and fit the sparse Gaussian infinite factor model proposed by Bhattacharya and Dunson (2011a), in addition to larger errors, we also mistakenly inferred too many factors (Table 3.4.2), which further increase as the values of (n, p) increase. The results indicate that it is crucial to be able to flexibly capture the underlying latent factor distributions, and that Gaussian factor models are less favored with high dimensional data with potential non-normality both in terms of dimension reduction and estimation accuracy.

We have shown that when the distributions of latent factors are potentially non-Gaussian, the semi-parametric Gaussian variance-mean mixture (infinite) factor model we proposed can correctly pick the number of latent factors, capture the underlying sparse structures and give good estimation of the covariance matrix, while fitting Gaussian (infinite) factor models with the normal specification leads to inference with bigger errors. Given that Gaussian distributions are also special cases of the Gaussian variance-mean mixtures family, in the following studies, we are testing

Table 3.3: Performance of covariance matrix estimation in the simulation studies when the numbers of factors are assumed unknown. Results of normal variance-mean mixture infinite factor model (N-IFM) are compared to those of Gaussian infinite factor model (G-IFM) developed by Bhattacharya and Dunson (2011a) and the Banding method proposed by Rothman and Zhu (2010). Mean square error (mse), average absolute bias (aab) and maximum absolute bias (mab) between estimated and true covariance matrices are tested.

(p,k)	(200,6)			(500,10)		
method	N-IFM	G-IFM	Ba	N-IFM	G-IFM	Ba
rmse	0.0474	0.0952	0.434	0.0523	0.149	0.425
aab	0.0374	0.0772	0.317	0.0417	0.122	0.344
mab	0.258	0.428	0.990	0.397	0.588	0.994
(p,k)	(1500,10)			(5000,15)		
method	N-IFM	G-IFM	Ba	N-IFM	G-IFM	Ba
rmse	0.0489	0.142	0.463	0.0801	0.169	0.446
aab	0.0387	0.114	0.393	0.0633	0.132	0.379
mab	0.316	0.706	1.056	0.601	0.995	1.198

Table 3.4: Posterior C.I. of number of factors with different (p,k,n) values. Results of using normal variance-mean mixture infinite factor model (N-IFM) are compared to those of Gaussian infinite factor model (G-IFM) developed by Bhattacharya and Dunson (2011a). Note that more "factors"are inferred with G-IFM, the number of which further increases when sample size n and dimension p increase.

(p,k,n)	(200,6,100)	(500,10,100)	(1500,10,100)
N-IFM 95% C.I.	[6,7]	[9,11]	[10,11]
G-IFM 95%C.I.	[7,9]	[12,13]	[17,18]
(p,k,n)	(200,6,500)	(500,10,500)	(1500,10,500)
N-IFM 95% C.I.	[6,7]	[9,10]	[10,10]
G-IFM 95%C.I.	[14,16]	[18,20]	[28,30]

the robustness of the proposed models when the underlying factor distributions are actually Gaussian, which is assumed unknown during inference and simultaneously inferred from the data. If robust results are obtained, we can argue that unless we are 100% sure that the latent factor distributions are Gaussian (which is often not the case in practice), the semi-parametric Gaussian variance-mean mixture (infinite) factor models allow uncertainty about the latent factor distributions and are more

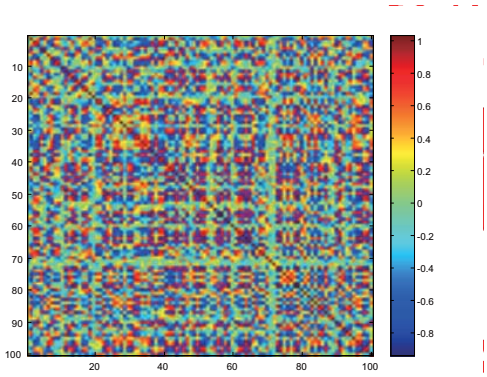

FIGURE 3.6: True covariance matrix of standardized y_i calculated based on model (3.1) and (3.5). y_i is 1500 dimensional and 100 of the dimensions are shown as an illustration. This is compared with Figure 3.3.

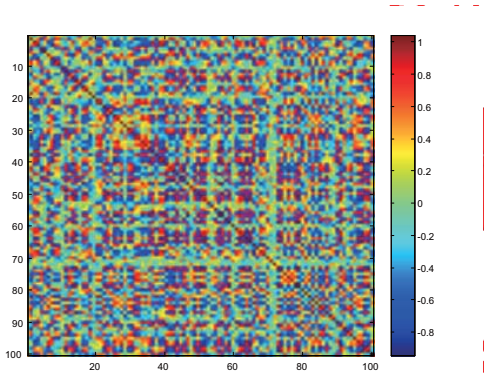

FIGURE 3.7: Mean of estimated covariance matrix of standardized y_i using our proposed normal variance-mean infinite factor model with number of factor unknown. y_i is 1500 dimensional and 100 of the dimensions are shown as an illustration.

robust approaches than Gaussian factor models.

To test this, we generated n=100 data using underlying $N(0, 1)$ latent factors with all four different (p, k) combinations, where (p, k) takes (200,6), (500,10), (1500,10) and (5000, 15). We fit the data using the semi-parametric Gaussian variance-mean mixture infinite factor model with unknown number of factors truncated at $k^* = 5\log(p)$. Posterior samples and estimates are obtained same as described before, which are all compared with those obtained from the Gaussian infinite factor model of Bhattacharya and Dunson (2011a).

The estimated covariance matrices are still close to the true values with small errors (as shown in Table 3.4.2). This is comparable to the results obtained by fitting a sparse Gaussian infinite factor model of Bhattacharya and Dunson (2011a), which assumes that the latent factor distributions are Gaussian. Overall, the simulation studies clearly highlight the merit of our proposed models in flexibly and robustly capturing the data structures.

Table 3.5: Performance of covariance matrix estimation in the simulation studies with true Gaussian factors when the numbers of factors are assumed unknown. Results of normal variance-mean mixture infinite factor model (N-IFM) are compared to those of Gaussian infinite factor model (G-IFM) developed by Bhattacharya and Dunson (2011a) and the Banding method proposed by Rothman and Zhu (2010). Mean square error (mse), average absolute bias (aab) and maximum absolute bias (mab) between estimated and true covariance matrices are tested. Note that N-IFM is robust when the latent factors are actually Gaussian.

(p,k)		(200,6)			(500,10)	
method	N-IFM	G-IFM	Ba	N-IFM	G-IFM	Ba
rmse	0.0472	0.0570	0.463	0.0840	0.0829	0.457
aab	0.0248	0.0186	0.305	0.0630	0.0611	0.333
mab	0.243	0.210	0.931	0.400	0.396	0.972
(p,k)		(1500,10)			(5000,15)	
method	N-IFM	G-IFM	Ba	N-IFM	G-IFM	Ba
rmse	0.0804	0.0788	0.502	0.103	0.0984	0.498
aab	0.0595	0.0591	0.368	0.0841	0.0827	0.373
mab	0.303	0.283	1.021	0.640	0.604	1.172

3.4.3 Study III: Prediction and Classification

In this simulation study, we tested the predictive performance of the Gaussian variance-mean mixture infinite factor model with high-dimensional non-normal predictors.

We simulated $n = 100$ observations with $(p, k) = (1500, 10)$ and considered the first dimension as the response with the others as predictors. We assume that all predictors and responses are continuous for the simplicity of exposition, and standard data augmentation procedures can be used otherwise. The 10 latent factors are coming from a variety of Gaussian or non-Gaussian distributions, leading to non-Gaussian distributions of predictors and responses. We divided the data into a training set of 80 and a validation set of 20 with responses missing to gauge the predictive performance. The accuracy of predictions is evaluated as described in Section 3.3.3.

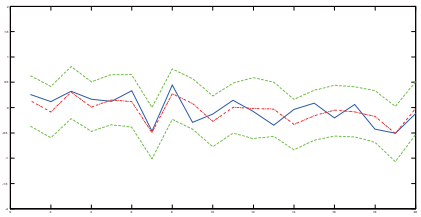

FIGURE 3.8: Predictive performance of GVMM factor model with non-Gaussian observations. Posterior predictive mean (in dotted red), 95% C.I. (in dashed green) and true values (in solid blue) are shown for the 20 continuous responses in the testing set. This is compared with Figure 3.9.

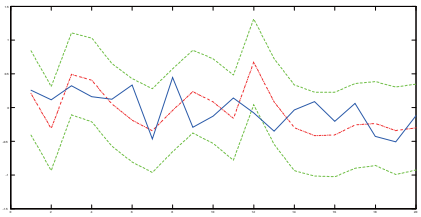

FIGURE 3.9: Predictive performance of Gaussian latent factor model with non-Gaussian observations. Posterior predictive mean (in dotted red), 95% C.I. (in dashed green) and true values (in solid blue) are shown for the 20 continuous responses in the testing set. This is compared to Figure 3.8.

As shown in Figure 3.8, the posterior predictive means based on our model are close to the truth in each case in the testing set, with 95% C.I.s covering the true values and root of mean squared prediction error (RMSEP) being 0.154. These are compared to the results obtained by applying a Gaussian sparse infinite factor model and latent factor regression (Figure 3.9) with RMSEP of 0.267. Clearly, ignoring the potential non-normality of the data leads to bad point estimates and wider interval estimates of the true values, due to the fact that the underlying structures are not correctly captured and also greater variance is needed to account for non-normality.

3.5 Data Analysis

In this section, we further highlight the merit of Gaussian variance-mean mixture infinite factor models for prediction in the $p \gg n$ setting with real data sets. In terms of prediction accuracy, we compare our Gaussian variance-mean mixture infinite factor models to five other methods used for prediction in $p \gg n$ settings: sparse Bayesian infinite factor models (G-IFM, Bhattacharya and Dunson (2011a)), principal component regression (PCR, Jolliffe (1982)), partial least squares regression (PLS, Mevik

and Wehrens (2007)), least-angle regression (LARS, Efron et al. (2004)) and elastic net (Zou and Hastie (2005)).

Four real data sets with continuous responses are analyzed using each method, which are all publicly available from the *R* packages pls (Mevik and Wehrens (2007)) and mixOmics (Le Cao et al. (2009)):

- **gasoline:** A data set consisting of octane number (response) and NIR spectra (predictors) of 60 gasoline samples. Dimension of NIR is 401, consisting of diffuse reflectance measurements from 900 to 1700nm.

- **yarn:** A data set consisting of the density (response) of 22 polyethylene terephthalate (PET) yarns and near-infrared spectra (NIR) measured at 268 different wavelengths as predictors.

- **multidrug:** The original data set contains the expression of 48 known human ABC transporters and the activity of 1429 drugs in 60 different cancer cell lines. We focus on 853 of the drug activity without missing values (predictors) and expression of one of the ABC transporters (ABCA3, response), which are also studied by Hahn et al. (2010).

- **nutrimouse:** A data set containing the expression measure of 120 genes (predictors) potentially relevant to nutritional problems and the concentration of C16 hepatic fatty acids (response). A total of 40 mice are studied.

To compare the models and methods in terms of prediction, we randomly split each data set into a training set of 80% observations and a validation set of 20% observations. For two of the Bayesian models (G-IFM and our N-IFM), the response and predictors are jointly modeled, and we use the posterior predictive means as predictive values in the validation set. For the remaining methods (PLS, PCR, LARS and ENET), each model is fit to the training data, with tuning parameters chosen by

ten-fold cross validation. After the tuning parameters have been chosen, predictions are performed on the validation set. Predictive performance is measured via root of mean squared prediction error (RMSEP).

Table 3.6: Root of mean squared prediction error (RMSEP) of each method on four data sets. Our Gaussian variance-mean mixtures infinite factors models (N-IFM) are compared with sparse Bayesian infinite factor models (G-IFM, Bhattacharya and Dunson (2011a)), principal component regression (PCR, Jolliffe (1982)), partial least squares regression (PLS, Mevik and Wehrens (2007)), least-angle regression (LARS, Efron et al. (2004)) and elastic-net (ENET, Zou and Hastie (2005)).

Data set	n	p	N-IFM	G-IFN	PLS	PCR	LARS	ENET
gasoline	60	401	**0.292**	0.378	0.335	0.338	0.441	0.359
yarn	22	268	**0.299**	0.382	0.431	0.583	0.313	0.311
nutrimouse	40	120	**2.094**	3.706	2.184	2.683	4.811	3.115
multidrug	60	853	0.887	0.935	0.942	**0.852**	0.964	0.897

As shown in Table 3.6, our Gaussian variance-mean mixtures infinite factor models (N-IFN) outperforms the other five methods on three of the four data sets and is only slightly worse than the best in the remaining multidrug example. The fact that our model always outperforms sparse Gaussian infinite factor models where the latent factor distributions are assumed Gaussian indicates the existence of potential non-normality in real data analysis and highlights the importance of more flexibility allowing non-normal latent factor distributions to obtain better prediction accuracy.

3.6 Discussion

In this paper, we proposed a semi-parametric specification for latent factor distributions in a model for multivariate data. Using alternatives to the normal distribution for latent factors is not new, but we advocate the use of our semi-parametric Gaussian variance-mean mixture factor models for several reasons. The first reason is that the extra flexibility of the framework is not at the cost of significant increase of computational burden. The conditional normality of the framework makes the

approach scalable up to moderately high-dimensional problems without the issues of isolated local modes that arise in unconstrained mixture models. Further scalability can potentially be obtained by replacing the MCMC algorithm with a fast deterministic approximation; in developing such approximations, the conditionally conjugate multivariate Gaussian structure should prove useful. The second reason lies in the results we showed that the model is robust even when the true latent factor distributions are Gaussian, taking advantage of the fact that the Gaussian distribution is a limiting case of the Gaussian variance-mean mixtures family. Therefore, by specifying semi-parametric GVMM distributions for the latent factors, we incorporate the uncertainty about potential (non-)normality in the models, allowing for either symmetric or skewed, heavier- or lighter-tailed latent factor distributions, and make the model flexible enough in practice to capture the structure of multivariate and high-dimensional observations with different marginal distributions.

3.7 Supplementary Materials

3.7.1 Proof

Let $y_i = (y_{i1}, \ldots, y_{ip})$ be a sample of a $p \times 1$ random variable as defined in (3.5), i.e.

$$\begin{aligned} y &= \mu + \Lambda f + \epsilon, \quad \epsilon \sim N_p(0, \Sigma), \\ f &\sim N(\alpha + \Gamma V, \Psi = \text{diag}(V)), \\ V &\sim G \end{aligned}$$

where $f = (f_{.1}, \ldots, f_{.k})'$ is the latent factor vector, and let $f_i = (f_{i1}, \ldots, f_{ik})$. $\alpha = (\alpha_1, \ldots, \alpha_k)'$, Γ is a diagonal matrix with $\text{diag}(\Gamma) = (\gamma_1, \ldots, \gamma_k)'$, $V = (V_{.1}, \ldots, V_{.k})'$ is a vector of independent mixing variables ($\forall m \neq n$, $V_{.m} \perp V_{.n}$) with mixing distribution $G = (G_1, \ldots, G_k)$ and diagonal covariance matrix Σ_V.

Proof. of Theorem 3. For any arbitrary pair j and l, $(1 \leqslant j \neq l \leqslant p)$, Let $\Lambda_{j.}$ denote

the j^{th} row of loadings matrix Λ, then

$$y_{ij}|f \sim N(\mu_j + \Lambda_j f_j, \Sigma_{jj})$$
$$f_i|y_{il}, V \sim N\left(\mu_{f_i|y_{.l},V}, \Sigma_{f_i|y_{.l},V}\right)$$
$$y_{ij}|y_{il}, V \sim N(\mu_j + \Lambda_j \mu_{f_i|y_{il},V}, \Lambda_j \Sigma_{f_i|y_{il},V} \Lambda_j' + \Sigma_{jj})$$

Mean and Variance

Marginalizing out f_i, we have:

$$E[y_{ij}|y_{il}] = \mu_j + \Lambda_j E[f_i|y_{il}]$$
$$Var[y_{ij}|y_{il}] = \Lambda_j Var[f_i|y_{il}]\Lambda_j' + \Sigma_{jj} \tag{3.9}$$

To derive $E[f_i|y_{il}]$ and $Var[f_i|y_{il}]$:

Given that

$$y_{il}|f_i \sim N(\mu_l + \Lambda_l f_i, \Sigma_{ll}), \quad f_i|V \sim N(\alpha + \Gamma V, \text{diag}(V))$$

the conditional posterior $f_i|y_{il}, V$ is also Gaussian with mean $\mu_{f_i|y_{il},V}$, precision matrix $\Phi_{f_i|y_{il},V}$ and covariance matrix $\Sigma_{f_i|y_{il},V} = \Phi^{-1}_{f_i|y_{il},V}$

$$\Phi_{f_i|y_{il},V} = \Gamma'\Psi^{-1}\Gamma + \Lambda_l'\Sigma_{ll}^{-1}\Lambda_l$$
$$\mu_{f_i|y_{il},V} = \Phi^{-1}_{f_i|y_{il},V}(\Psi^{-1}\alpha + \text{diag}(\Gamma) + \Lambda_l'\Sigma_{ll}^{-1}(y_{il} - \mu_l))$$

Use Sherman−Morrison−Woodbury formula to invert the precision matrix $\Phi_{f_i|y_{il},V}$, we have

$$\begin{aligned}\Sigma_{f_i|y_{il},V} &= \Psi\Gamma^{-2} + \Psi\Gamma^{-2}\Lambda_l'(\Sigma_{ll} + \Lambda\Psi\Gamma^{-2}\Lambda_l')\Lambda_l\Psi\Gamma^{-2} \\ &= \Psi\Gamma^{-2} + \Psi\Gamma^{-2}\Lambda_l'\left(\Sigma_{ll} + \sum_{m=1}^{k}\Lambda_{lm}^2 V_m/\gamma_m^2\right)\Lambda_l\Psi\Gamma^{-2}\end{aligned}$$
$$\mu_{f_i|y_{il},V} = \phi_1(V) + \phi_2(V)(y_{il} - \mu_l)$$

where $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are $\mathcal{R}^k \to \mathcal{R}^k$ projection functions defined by:

$$\phi_1(V) = \Sigma_{f_i|y_{il},V}\left(\Psi^{-1}\alpha + \text{diag}(\Gamma)\right)$$

$$\phi_2(V) = \Sigma_{f_i|y_{il},V}\Lambda_l'\Sigma_{ll}^{-1}$$

Let $\phi_0(V) \equiv \Sigma_{f_i|y_{il},V}$. Then, we have that :

$$\begin{aligned} E[f_i|y_{il}] =& E_G[\phi_1(V)] + E_G[\phi_2(V)](y_{il} - \mu_l) \\ Var[f_i|y_{il}] =& E_G[\phi_0(V)] + Var_G[\phi_1(V)] + 2Cov_G[\phi_1(V), \phi_2(V)](y_{il} - \mu_l) \\ &+ Var_G[\phi_2(V)](y_{il} - \mu_l)^2 \end{aligned} \tag{3.10}$$

Put (3.10) back to (3.9), we have that

$$E[y_{ij}|y_{il}] = \alpha_{j|l}^{(0)} + \alpha_{j|l}^{(1)}(y_{il} - \mu_l)$$

$$Var[y_{ij}|y_{il}] = \beta_{j|l}^{(0)} + \beta_{j|l}^{(1)}(y_{il} - \mu_l) + \beta_{j|l}^{(2)}(y_{il} - \mu_l)^2$$

where $\alpha_{j|l}^{(0)} = \mu_l + \Lambda_j E_G[\phi_1(V)]$

$$\alpha_{j|l}^{(1)} = \Lambda_j E_G[\phi_2(V)]$$

$$\beta_{j|l}^{(0)} = \Sigma_{jj} + \Lambda_j \left(E_G[\phi_0(V)] + Var_G[\phi_1(V)]\right)\Lambda_j'$$

$$\beta_{j|l}^{(1)} = 2\Lambda_j Cov_G[\phi_1(V), \phi_2(V)]\Lambda_j'$$

$$\beta_{j|l}^{(2)} = \Lambda_j \left(Var_G[\phi_2(V)]\right)\Lambda_j'$$

Skewness (γ_1) and Kurtosis(γ_2)

$$\gamma_1 = \frac{E[y_{ij}^3|y_{il}] - 3E[y_{ij}|y_{il}]Var[y_{ij}|y_{il}] - 3(E[y_{ij}|y_{il}]^3)}{(Var[y_{ij}|y_{il}])^{3/2}}$$

$$\gamma_2 = \frac{E[y_{ij}|y_{il} - E[y_{ij}|y_{il}]]^4}{(Var[y_{ij}|y_{il}])^2}$$

plug the previous results in, we obtain that γ_1 and γ_2 are of the forms

$$\gamma_1 = \frac{\sum_{m=0}^{3} E_G[\tilde{\phi}_m[V]](y_{il} - \mu_l)^m}{(\beta_{j|l}^{(0)} + \beta_{j|l}^{(1)}(y_{il} - \mu_l)^2)^{3/2}}$$

$$\gamma_2 = \frac{\sum_{q=0}^{4} E_G[\hat{\phi}_q[V]](y_{il} - \mu_l)^m}{(\beta_{j|l}^{(0)} + \beta_{j|l}^{(1)}(y_{il} - \mu_l)^2)^2}$$

where $\tilde{\phi}_m(\cdot)$s and $\hat{\phi}_q(\cdot)$s are all $\mathcal{R}^k \to \mathcal{R}$ projection functions of V defined by the model parameters. □

3.7.2 Supplementary Figures

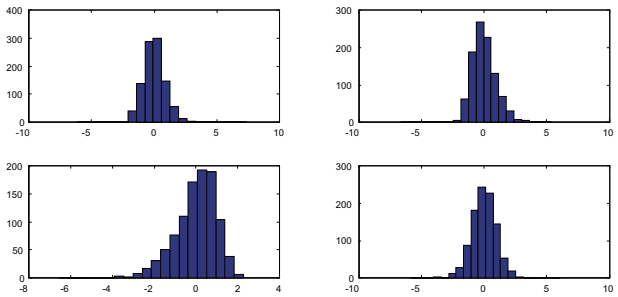

FIGURE 3.10: Samples of y_i from 4 out of 500 dimensions show different marginal distribution of observations, possibly with heavy tails, skewness or (seemly approximate) normality.

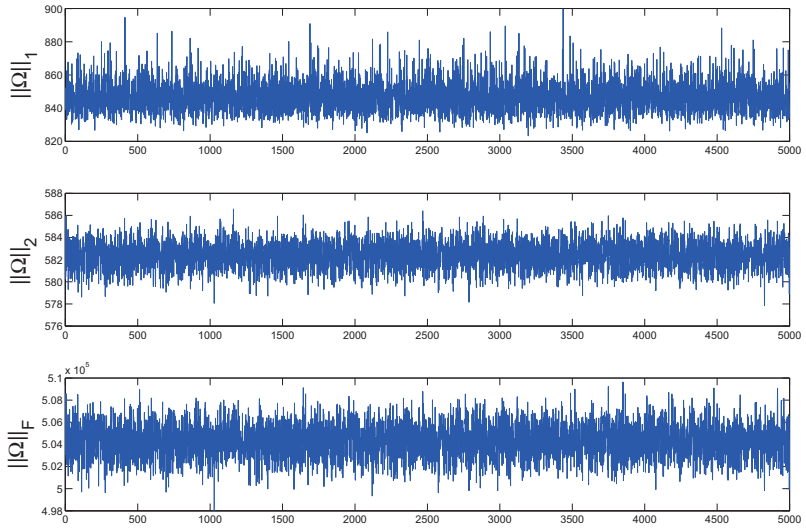

FIGURE 3.11: Trace plots of different norms of the MCMC samples of the marginal covariance matrix (Ω), including the matrix 1-norm ($||\Omega||_1$), the spectral norm ($||\Omega||_2$) and the Frobenius norm ($||\Omega||_F$), as described in Section 3.3.1. Results from $(p, k) = (5000, 15)$ are shown to illustrate the mixing performance with large p.

4

Semiparametric Gaussian Variance Mean Mixtures

4.1 Introduction

There is increasing awareness of the importance of developing new classes of multivariate distributions that flexibly characterize heavy tails and skewness, while accommodating tail dependence. Such tail dependence arises in many applications and is a natural consequence of dependence in outlying events. Such dependence is well known to occur in financial data, communication networks, weather and other settings, but is not adequately characterized by common approaches such as Gaussian copula models. Salmon (2012) provides a compelling commentary on how reliance on a single measure of correlation in two variables based on a Gaussian copula may have played a substantial role in the financial crisis. We need statistical methods based on new classes of distributions that do not rely on such unrealistic assumptions but that are still tractable to apply even in moderate to high-dimensional settings.

There is an existing literature relevant to this topic. Wang et al. (2004) proposed a class of skew-symmetric distributions having probability density functions (pdfs) of the form $2f(\mathbf{x})Q(\mathbf{x})$, where f is a continuous symmetric density and $Q : \Re^n \to [0, 1]$

is a skewing function. Choosing f as normal leads to the skew normal class (Azzalini (1985), Azzalini and Dalla Valle (1996)), with other special cases corresponding to skew-t (Sahu et al. (2003), Gupta (2003)), skew slash (Wang and Genton (2006)) and skew elliptical distributions (Genton and Loperfido (2005)). These parametric models are useful in providing computationally tractable distributions that have parameters regulating skewness and kurtosis in the data. However, choosing a specific parametric family for f and Q can be challenging in practice, with different choices yielding potentially different results. Although one can potentially conduct model selection or averaging, this adds to the computational burden.

Alternatively, nonparametric approaches have been explored to handle heavy-tailed and skewed observations with more flexibility. For instance, mixtures of normal distributions have been widely used to approximate arbitrary distributions. S. Venturini and Parmigiani (2008) use mixtures of gamma distributions over the shape parameter to model heavy-tailed medical expenditure data. Mixtures of other heavy-tailed distributions have also been proposed. Such nonparametric density estimation approaches face substantial challenges in multivariate cases due to the curse of dimensionality. Fully nonparametric density estimation is almost too flexible in allowing densities that have arbitrary numbers of modes and complex shapes, which are difficult to estimate accurately based on available data in many cases. There has been some attempt to reduce dimensionality in multivariate density estimation using mixtures of factor analyzers (Chen et al. (2010b)) and alternative approaches, but nonetheless the curse is only partly thwarted by such efforts.

We focus on Gaussian variance-mean mixtures (GVMM), introduced by Barndorff-Nielsen (1977) as a flexible class of multivariate distributions induced through the following hierarchical model for $y_i = (y_{i1}, \ldots, y_{ip})'$,

$$y_i = \mu + \gamma V_i + \sqrt{V_i}\Sigma^{1/2} Z_i, \quad Z_i \sim N(0, I_p), \;\; V_i \sim G, \tag{4.1}$$

where $\mu \in \Re^p$ is a location parameter, $\gamma \in \Re^p$ is a drift or skewness parameter, $V_i \sim G$, G is a mixture distribution on $[0, \infty)$, $Z_i \perp V_i$, and Σ is a positive definite matrix. Conditions such as $|\Sigma| = 1$ are typically imposed to avoid an unidentifiable scale factor. This model corresponds to a generalization of the multivariate normal distribution, which is obtained in the special case in which $\gamma = 0$.

Current literature on Gaussian variance-mean mixtures is mostly focused on univariate models in which the mixing distribution G belongs to a parametric family. Some special cases corresponding to different choices of G include Student t, Laplace, hyperbolic, normal inverse Gaussian and variance gamma distributions, and such models have been applied broadly (see e.g. Barndorff-Nielsen (1977), Barndorff-Nielsen et al. (1990), Barndorff-Nielsen (1982), Barndorff-Nielsen (1997), Arslan and Genc (2009) and Arslan (2010)). However, as with skew-symmetric distributions, limiting to a particular parametric class is clearly restrictive. In addition, some appealing subfamilies of GVMM, such as generalized hyperbolic (GH) distributions, are not analytically tractable. This is partially due to the flatness of the likelihood function (Eberlein and Prause (1998)), which makes it hard to obtain reliable parameter estimates without prior information or some form of penalization, even with large sample sizes (Aas and Haff (2006)).

We propose Bayesian semiparametric Gaussian variance-mean mixture models, in which the mixing distribution G is modeled nonparametrically to flexibly accommodate heavy-tails and skewness, while letting the data inform about the appropriate distribution choice. Efficient Bayesian computational strategies are developed for reliable inference on parameters.

4.2 Proposed Modeling Framework

4.2.1 Semi-Parametric GVMM

Consider the p-dimensional multivariate continuous heavy-tailed and/or skewed observations $y_i = (y_{i1}, \ldots, y_{ip}) \sim f$, $(i = 1, \ldots, N)$. Our goal is to obtain a flexible GVMM model for the density f via modeling the mixing distribution G nonparametrically. To achieve this, we use Dirichlet process mixture (DPM) of generalized log-normal (logGN) prior with unknown order s for G. Mathematically, the model is represented as:

$$\begin{aligned} y_i &= \mu + \gamma V_i + \sqrt{V_i}\Sigma^{1/2} Z_i, \\ V_i &\sim \log GN(m_i, \psi_i^{-1}, s), \quad Z_i \sim N(0, I), \\ (m_i, \psi_i) &\sim H, \quad H \sim DP(\alpha_0 H_0), \end{aligned} \tag{4.2}$$

where α_0 is the DP precision and H_0 is the base measure for m_i and ψ_i, which we choose independent normal and inverse-gamma respectively.

The family of logGN distributions considered here was initially introduced by Vianelli (1983), which can be obtained through the exponential transformation of a random variable that follows the generalized normal distribution (Varanasi and Aazhang (1989) and Nadarajah (2005)). The pdf of a logGN (m, ψ, s) distributed random variable V is given by:

$$f(v) = \frac{s}{2v\psi^{1/2}\Gamma\left(\frac{1}{s}\right)} \exp\left(-\left|\frac{\log(v) - m}{\psi^{1/2}}\right|^s\right) / \tag{4.3}$$

where $v > 0$, $m \in \Re$, $\psi > 0$ and $s > 0$. logGN densities are more flexible than log-normal densities in including an additional parameter s controlling tail behavior, with log-normal corresponding to the special case in which $s = 2$ and double-exponential obtained by taking $s = 1$.

Although logGN distributions are appealing in providing a simple generalization of the log-normal which is more flexible in the tails, such distributions have been rarely implemented even in simpler settings due to the computational hurdles involved. Fortunately, for Bayesian posterior computation via MCMC we can rely on a data augmentation algorithm based on Fact 2, with the adaption to our setting using DPMs of generalized log-normals shown in Section 4.3.2.

Fact 2. *Let V and W be two random variables such that*

1. $f(v|w) = \frac{1}{2\psi^{1/2}w^{1/s}v} I\left(\left|\frac{\log v - m}{\psi^{1/2}}\right|^s < w\right)$

2. $W \sim Gamma(1 + 1/s, 1)$

then $V \sim \log GN(m, \psi, s)$

DPMs of Gaussians provide a highly flexible approximation to arbitrary densities. As a prior for f directly, DPMs of Gaussians have appealing asymptotic properties and lead to minimax optimal adaptive rates of convergence under some conditions on the true data-generating density, with these conditions unfortunately ruling out heavy-tailed densities (Shen and Ghosal (2012)). As motivated above, our focus is on obtaining a flexible prior for multivariate heavy tailed and skewed densities, which is not fully nonparametric in the sense of allowing multimodal and other very irregular density shapes but can flexibly approximate a broad class of unimodal densities without inducing restrictive tail constraints. Hence, it is not appropriate to choose a DPM of Gaussians directly for the density of the data f.

Expression (4.2) instead uses a DPM of logGNs for the mixture distribution G within the GVMM framework. By using a such DPM prior and setting s unknown, we obtain large flexibility of the mixing distribution and its tail decay, which can adopt diverse degrees of heavy-tailness for the marginal density f of data, while allowing

the data to fully infer about the unknown mixing distribution. After considering a broad variety of alternatives, we have found this specification to be excellent at capturing a rich variety of multivariate heavy tailed and skewed distributions, while also allowing symmetric data. Some basic properties are detailed below.

4.2.2 Tail Behavior

It is important to understand the relationship between the tail behavior of the mixture distribution G and the induced tail behavior of f, the marginal distribution of y_i. We start by considering the univariate case. Theorem 3 of Barndorff-Nielsen and Sorensen (1982) describes the relationship in the special case in which $\gamma = 0$ and hence the marginal distribution is symmetric.

Theorem 3. (Barndorff-Nielsen 1982) *Suppose f(y) is the pdf of a Gaussian variance mixture as described in (4.1) with $\gamma = 0$, and that the tail of the mixing distribution G with pdf g satisfies:*

$$g(V) \sim e^{-\psi_+ V} V^{\lambda-1} L(V), \quad as\ V \to +\infty,$$

where $\psi_+ \geqslant 0$ and L is a function of slow variation with $\lim_{V\to\infty} \frac{L(aV)}{L(V)} = 1, \quad \forall a > 0$. Then

1. *If $\psi_+ = 0$, $f(y) \sim |y|^{2\lambda-1} L(y^2), \quad |y| \to \infty$.*

2. *If $\psi_+ > 0$, $f(y) \sim |y|^{\lambda-1} \exp(-(2\psi_+)^{1/2}|y|) L(|y|), \quad |y| \to \infty$*

Observe that the tail behavior of a Gaussian variance mixture (when $\gamma = 0$) mainly depends on the tail behavior of the mixing distribution. To generalize this to arbitrary Gaussian variance-mean mixtures, we first introduce the following Lemma.

Lemma 4. *Suppose $f_G(y)$ is the pdf of a Gaussian variance mixture with mixing distribution G, and let φ denote the moment generating function of G. Then $\forall \gamma \in \Re$*

s.t. $\varphi(\gamma^2/2) < \infty$*:*

$$f_{G^*}(y) = \frac{\exp(\gamma y)}{\varphi(\gamma^2/2)} f_G(y)$$

is the pdf of a Gaussian variance-mean mixture with skewness parameter γ *and mixing distribution* G^* *with pdf* g^* *satisfying:* $g^*(V) = \dfrac{\exp(\gamma^2 V/2)}{\varphi(\gamma^2/2)} g(V)$*, where* g *is the pdf of* G*. The converse is also true.*

Proofs can be found in the Appendix. This lemma provides a link between the tail behavior of a Gaussian variance-mean mixture and that of a Gaussian variance mixture, via the link between tails behaviors of the two mixing distributions. This relationship is used in the following Theorem.

Theorem 5. *Suppose* $f_G(y)$ *is the pdf of a Gaussian variance-mean mixture as described in (4.1), with skewness parameter* γ *and mixing distribution* G*. If the tail of the mixing density* g *satisfies:*

$$g(V) \sim e^{-\psi_+ V} V^{\lambda - 1} L(V) \quad \text{as } V \to +\infty$$

where $\psi_+ \geqslant 0$*, and* L *is a function of slow variation with* $\lim_{V\to\infty} \frac{L(aV)}{L(V)} = 1, \quad \forall a > 0$*, then:*

1. $f_G(y) \sim y^{\lambda-1} \exp\left(-(\sqrt{2\psi_+ + \gamma^2} - \gamma)y\right) L(y)$*, as* $y \to +\infty$*.*

2. $f_G(y) \sim |y|^{\lambda-1} \exp\left(-(\sqrt{2\psi_+ + \gamma^2} + \gamma)|y|\right) L(|y|)$*, as* $y \to -\infty$*.*

Theorem 5 shows that the tail behavior of a Gaussian variance-mean mixture also depends on that of the mixing distribution. Generally, heavier tails in the mixing distribution induce heavier tails of the Gaussian variance-mean mixture. Thus, by placing a DPM of generalized log-normals on the mixing distribution, we induce a prior on the density f with flexible degrees of tail decay. Another observation

is that if the mixing distributions have sub-exponential tails (such as log-concave distributions), then the Gaussian variance-mean mixture also has sub-exponential tails, which also illustrates the limitation in flexibility of using particular parametric cases of the Gaussian variance-mean mixtures to fit data.

4.2.3 Moments

To compute the moments of Gaussian variance-mean mixtures, we can directly apply the law of total cumulance. Let k_1, k_2, k_3 and k_4 denote the expectation, variance, skewness and kurtosis of the Gaussian variance-mean mixtures described in (4.1), and k_1^*, k_2^*, k_3^* and k_4^* denote those of the mixing distribution. Given they all exist, we have simply

$$k_1 = \mu + \gamma k_1^*, \quad k_2 = k_1^* + \gamma^2 k_2^*, \quad k_3 = \gamma k_2^* + \gamma^3 k_3^*. \tag{4.4}$$

More generally, we have $\varphi_f(s) = \exp(\mu s)\varphi(\gamma s + 1/2s^2)$, where φ_f and φ are moment generating functions of GVMM and the corresponding mixing distribution respectively. So clearly, the existence of moments of mixing distributions indicates the existence of moments of Gaussian variance-mean mixtures, and γ can control expectation, variance and kurtosis, in addition to being a skewness parameter.

4.3 Bayesian Computation

4.3.1 Priors

In the semi-parametric GVMM framework, the $|\Sigma| = 1$ constraint is typically imposed to guarantee model identifiability. To improve computational efficiency in our proposed Bayesian computational algorithm, we use a parameter-expansion approach in which priors are placed on parameters in an unidentifiable working model without the $|\Sigma| = 1$ constraint. We then include a post-processing step to transform the parameters back to an identifiable inferential model, which includes the $|\Sigma| = 1$

constraint. A related strategy was used in Gaussian factor models by Ghosh and Dunson (2009). We use fairly diffuse priors for unidentifiable parameters, as an aid to mixing and because it is difficult to elicit priors for these parameters. However, we avoid completely non-informative priors because flatness of the likelihood function in some GVMM models (Barndorff-Nielsen (1982)) can lead to unreliable inferences in the absence of some prior information or penalization. To address this problem, we propose an empirical Bayes approach to incorporate skewness information from the data in estimating hyperparameters for the skewness parameter γ.

In particular, we transform the original data to have positive sample skewness and unit sample variance. The data are first normalized and the sample skewness is calculated. If it is negative, we multiply the normalized data by negative one. In conducting inferences, we transform back to the scale and sign of the original data. As GVMMs are closed under linear transformations, this will induce a GVMM. Because the transformed data are more likely to be right skewed or symmetric, we can more easily elicit a default weakly informative prior for the skewness parameter γ, which we choose to be normal with positive mean μ_γ, with a gamma hyperprior placed on μ_γ to improve prior robustness. For the order parameter s, a truncated inverse-gamma prior is used. Diffuse priors are placed on the remaining unknowns. To summarize

$$\begin{aligned} &\gamma \sim N(\mu_\gamma, \phi_\gamma), \quad \mu_\gamma \sim \text{Gamma}(\alpha_\gamma, \beta_\gamma), \quad \phi_\gamma \sim \text{Gamma}(a_\phi, b_\phi), \\ &s \sim \text{Gamma}(\alpha_s, \beta_s), \\ &\mu \sim N(0, \phi_\mu), \quad \Sigma \sim \text{Inv-Wishart}\left(m_0, \Psi_0\right). \end{aligned} \tag{4.5}$$

These priors were used in an unconstrained working model for the transformed data and induce priors on the parameters in the identifiable inferential model having the $|\Sigma| = 1$ constraint.

As for the DPM of logGN prior for the mixing distribution G, for ease in com-

putation, we use a similar stick-breaking representation of the DPM as proposed by Sethuraman (1994) but with generalized log-normal instead of normal components, and truncated at M components following Ishwaran and James (2001). Furthermore, as pointed out before, since it is hard to directly update parameters of the generalized log-normal distributions, we further utilize Fact 2 to introduce augmented data $\boldsymbol{w} = (w_1, \ldots, w_N)$ to improve computational efficiency. To summarize, the data-augmented stick-breaking representation of the DPM of logGN prior is shown as follows:

$$f(v_i|\mathbf{m}, \boldsymbol{\psi}, \mathbf{K}, s, \boldsymbol{w}) = \frac{1}{2\psi_{K_i}^{1/2} w_i^{1/s} v_i} I\left(\left|\frac{\log v_i - m_{K_i}}{\psi_{K_i}^{1/2}}\right|^s < w_i\right)$$

$$w_i|s \sim \text{Gamma}(1 + 1/s, 1)$$

Note that these are eqivalent to: $V_i|\mathbf{m}, \boldsymbol{\psi}, \mathbf{K}, s \sim \log GN(V_i; m_{K_i}, \psi_{K_i}^{-1}, s)$

$$\begin{aligned} &K_i|\mathbf{p} \sim \sum_{k=1}^{M} p_k \delta_k(\cdot), \quad i = 1, \ldots, N \\ &(m_k, \psi_k) \sim H_0(m, \psi), \quad k = 1, \ldots, M \\ &p_1 = \nu_1, \quad p_k = \nu_k \Pi_{l=1}^{k-1}(1 - \nu_l), \quad k = 2, \ldots, M-1, \\ &\nu_k \sim \text{Beta}(1, \alpha_0), \; k = 1, \ldots, M-1, \quad \nu_M = 1, \end{aligned} \tag{4.6}$$

where $\mathbf{K} = (K_1, \ldots, K_N)$, $\mathbf{m} = (m_1, \ldots, m_M)$, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_M)$, $\mathbf{p} = (p_1, \ldots, p_M)$ and for H_0, we use independent normal and inverse gamma for m and ψ respectively. Here, K_i is augmented data representing the mixture component index for observation i.

4.3.2 Full Conditionals and Posterior Analysis

Given the model and priors as specified by (4.2), (4.5) and (4.6), we use a data-augmented Gibbs sampler to update the unknown quantities, including parameters in the GVMM framework (μ, γ and Σ), parameters in the DPM of logGN ($\boldsymbol{m}$, $\boldsymbol{\psi}$

and s), augmented data ($\boldsymbol{K}$ and $\boldsymbol{w}$), hyper-parameters and the mixing variables (V_i). The Gibbs sampler is computationally efficient and mixes rapidly as most of the full conditional distributions have closed forms, except those for μ_γ, V_i and s, which are all univariate and updated using Metropolis-Hastings steps within the Gibbs sampler. Key steps in each Gibbs sampler iteration are listed as follows:

I. Sample μ, γ and Σ. Given normal priors for (μ, γ) and inverse-Wishart prior for Σ and the model that $y_i|V_i \sim N(\mu + \gamma V_i, V_i\Sigma)$, (μ, γ) is sampled from conditionally normal distribution, and Σ from conditionally inverse-Wishart.

II. Sample $\boldsymbol{m}$ and $\boldsymbol{\psi}$. Given the following priors are used for m_k and ψ_k for the kth log-normal component, $k = 1, \ldots, M$:

$$m_k \sim N(\mu_m, \phi_m), \quad \psi_k \sim \text{Inv-Gamma}(\alpha_\psi, \beta_\psi)$$

Sample m_k, $k = 1, \ldots, M$, from conditionally truncated normal distribution:

$$m_k| \ldots$$

$$\sim N(\mu_m, \phi_m) I\left(\max_{l:K_l=k}\left\{\log V_l - \psi_k^{1/2} w_l^{1/s}\right\} < m_k < \min_{l:K_l=k}\left\{\log V_l + \psi_k^{1/2} w_l^{1/s}\right\}\right)$$

and sample ψ_k from conditionally truncated inverse-Gamma distribution:

$$\psi_k| \ldots \sim \text{Inv-Gamma}(\alpha_\psi + N_k/2, \beta_\psi) I\left(\psi^{1/2} > \max_{l:K_l=k}\left\{\frac{|\log V_l - m_k|}{w_l^{1/s}}\right\}\right)$$

where N_k is the total number of V_i in the k^{th} mixture component.

III. Sample $\boldsymbol{w}$. w_i, $i = 1, \ldots, N$, is sample from conditionally truncated exponential distribution:

$$w_i| \ldots \sim \text{Exp}(1) I\left(w_i > \left(\frac{|\log V_i - m_{K_i}|}{\psi_{K_i}^{1/2}}\right)^s\right)$$

IV. Sample s. Given the Gamma(α_s,β_s) for s, the full conditional distribution is sampled using Metropolis-Hastings algorithm, from the following full conditional truncated kernel:

$$f(s|\dots)\propto\frac{s^{\alpha_s+N-1}\exp(-\beta_s s)}{\Gamma^N(1/s)}I\left(\max_{i\in S^-}(0,a_i)<s<\min_{i\in S^+}(a_i)\right)$$

where $a_i=\dfrac{\log w_i}{\log\left(|\log V_i-m_{K_i}|/\psi_{K_i}^{1/2}\right)}$, $i=1,\dots,N$.

$S^-=\{i:\log\left(|\log V_i-m_{K_i}|/\psi_{K_i}^{1/2}\right)<0\}$,

and $S^+=\{i:\log\left(|\log V_i-m_{K_i}|/\psi_{K_i}^{1/2}\right)>0\}$.

V. Sampling $\boldsymbol{K}$. K_i, $i=1,\dots,N$ is sampled from the conditionally multinomial (MN) distribution, with

$$Pr(K_i=k)|\dots\propto\frac{p_k}{\psi_k^{1/2}}\exp\left\{-\left(\frac{|\log V_i-m_k|}{\psi_k^{1/2}}\right)^s\right\}$$

VI. ν_k, $k=1,\dots,M-1$ are updated from conditionally beta distribution as shown in Ishwaran and James (2001). Also update univariate V_i, $i=1,\dots,N$ using a Metropolis-Hastings step within the Gibbs sampler.

4.4 Simulation Study & Real Data Analysis

4.4.1 Univariate Semi-parametric GVMM

To test the semi-parametric framework, a simulated dataset from univariate GVMM is first modeled. Specifically, observations y_i ($i=1,\dots,N=1000$) are generated from model (4.2), with the true values of the parameters as shown in (4.7).

$$\mu=2,\ \ \gamma=2,\ \ \Sigma=1,\ \ V_i\sim Gamma(3,1),\forall i=1,\dots,N. \tag{4.7}$$

The histogram of simulated data shows significant heavy-tailness and skewness, with sample kurtosis and skewness being 4.60 and 1.01 respectively.

For Bayesian inference, we pre-process the original simulated data and place priors as described previously. We run the MCMC for 10000 iterations with the first 5000 as burn-in. Several aspects of the posterior distributions are analyzed to evaluate the model fitting. First of all, posterior samples of V_i and parameters allow us to reconstruct and visualize the unknown mixing distribution G. As shown in Figure 4.8, a comparison between the true mixing distribution (Gamma(3,1), panel (B)) and 100 reconstructed mixing distributions (panel (A)) show significant similarity, indicating that the model can effectively capture the underlying structure. This is further illustrated by the posterior distributions of μ and γ, which have 95% C.I.s being [0.8676,3.2390] and [1.2532,2.9660], and posterior means being 2.0804 and 2.1523 respectively, which also very well cover the true values. It is worth mentioning that later in the real data analysis where the true values are unknown, the posterior distribution of γ provides us with a useful tool to assess skewness, with significant positive values of γ indicating skewness.

Furthermore, we can reconstruct the dataset based on the posterior samples of model parameters and the mixing distribution G, and directly visualize whether the posterior predictive distribution resembles the observed one. We reconstructed 200 such datasets, each containing 5000 data points. While the reconstructed datasets resemble the observed one (Figure not shown), we specifically looked at the posterior quantile estimates of the fitted semi-parametric GVMM, which is done by getting the 95% posterior C.I.s for a series of quantiles based on the 200 reconstructed datasets. This is shown in Table 4.1, and compared to the quantile estimates obtained by fitting other models, such as normal, skewed normal and skewed t distributions (Table 4.2). Compared to the fact that simple Gaussian model fails to capture neither heavy-tailness and skewness, our GVMM model fits the data as well as skewed Gaussian

and skewed t distributions, while maintaining unique advantages of relatively easy forms, convenient sampling and interpretability.

Table 4.1: Posterior quantile estimation to show the model fitting. Posterior quantile C.I.s are obtained by simulating 200 reconstructed datasets based (each consisting of 5000 data points) on posterior samples of unknown quantities, each dataset giving one set of quantile point estimates. Observed quantiles are obtained from the 1000 observed simulated data.

Quantiles	Observed Quantile	Posterior Mean	Posterior 95% C.I.
2.5%	-1.474	-1.492	[-1.535, -1.450]
5%	-1.328	-1.322	[-1.363, -1.284]
25%	-0.738	-0.711	[-0.742, -0.676]
50%	-0.175	-0.162	[-0.190, -0.124]
75%	0.555	0.534	[0.492,0.584]
95%	1.783	1.871	[1.776, 1.985]
97.5%	2.394	2.414	[2.283,2.566]

Table 4.2: Quantile estimates are obtained from models fitted to the observed dataset using maximum likelihood estimation. To obtain the maximum likelihood estimators in skewed Gaussian and t distributions, the **sn** R package is used (Azzalini (2011)).

Quantiles	Observed	Gaussian	skewed-Gaussian	skewed-t
2.5%	-1.474	-1.960	-1.451	-1.446
5%	-1.328	-1.645	-1.306	-1.300
25%	-0.738	-0.674	-0.751	-0.749
50%	-0.175	0.000	-0.173	-0.185
75%	0.555	0.674	0.590	0.566
95%	1.783	1.645	1.887	1.895
97.5%	2.394	1.960	2.339	2.379

4.4.2 Modeling the S&P 500 returns

It is well known that stock returns do not always confirm well with a Gaussian distribution. Modeling both heavy-tailness and asymmetry of returns is becoming important in economics and finance. Here, we look at daily returns of Standard & Poor's 500 Composite (S&P 500) index from 01/02/1990 to 09/13/2011. Totally

5470 observations are shown in Figure 4.9A, with sample skewness being 0.189 (after pre-processing), which suggests that the return distribution may be slightly right skewed.

Similar univariate semi-parametric GVMM and prior setup are applied to the dataset to access the capability of the model in capturing the return distribution. To evaluate the model fitting, we also reconstructed 200 datasets based on the posterior samples of unknown quantities (each consisting of 5470 observations), and a quick comparison (Figure 4.9) between the observed and reconstructed datasets shows significant similarity, indicating that our model captures the return distribution well. We also look at posterior quantile estimates based on the 200 simulated datasets (Table 4.3), which further illustrate the goodness-of-fit of the model.

Furthermore, we specifically look at the posterior distribution of γ. Generally speaking, when the sample skewness is relatively small (as in this case), it is difficult to claim whether the true distribution is skewed or symmetric, because samples from symmetric heavy-tailed distributions can also exhibit significant sample skewness due to the presence of extreme values with a finite sample size. However, one feature of the Bayesian semi-parametric GVMM framework is that the skewness can be simultaneously inferred by looking at the sign of the γ parameter, which can test the actual existence of skewness directly against an "artificial sample skewness" due to heavy-tailness. To illustrate this, the histogram of 5000 posterior samples of γ is shown in Figure 4.1, which gives a 95% posterior C.I. [0.0033,0.0838] and thus claims that the distribution is slightly right skewed at the 95% confidence level. As a striking comparison, we generated 5000 samples from t-distribution (df=3, shown in Appendix). The sample skewness is 0.228, which appears to suggest a slight right skewness. However, when we fit the data using our Bayesian semi-parametric GVMM and look at the posterior distribution of γ (shown in Figure 4.2, as compared to Figure 4.1), the posterior 95% C.I. is [-0.0371, 0.0757], with posterior mean at 0.0213, which

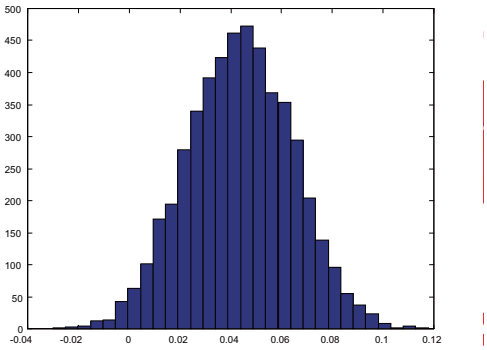

FIGURE 4.1: Posterior distribution of γ is presented. S&P500 daily returns are modeled via Bayesian semi-parametric GVMM, and γ confirms the existence of skewness with the sample skewness being 0.189.

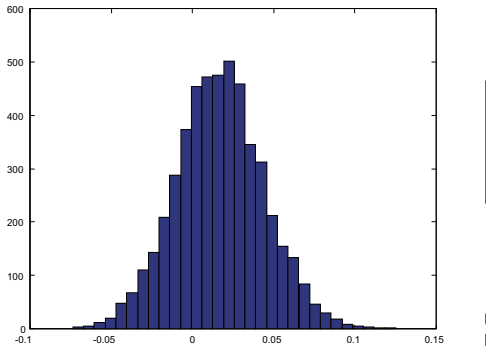

FIGURE 4.2: Posterior distribution of γ is presented. A set of random samples from t distribution with df=3 are modeled via Bayesian semi-parametric GVMM, and γ confirms no skewness although the sample skewness is 0.228.

argue against the existence of skewness and suggest that the sample skewness is due to the fact that t-distribution is heavy-tailed.

Table 4.3: Posterior quantile estimation to show the model fitting. Posterior quantile C.I.s are obtained by simulating 200 reconstructed datasets based (each consisting of 5470 data points) on posterior samples of unknown quantities, each dataset giving one set of quantile point estimates. Real data quantiles are obtained from the 5470 observed S&P 500 returns.

Quantiles	Real Data Quantile	Posterior Mean	Posterior 95% C.I.
2.5%	-1.965	-1.965	[-2.097,-1.867]
5%	-1.511	-1.513	[-1.589,-1.440]
25%	-0.493	-0.495	[-0.534, -0.469]
50%	-0.0315	-0.0175	[-0.0421,0.0087]
75%	0.467	0.476	[0.440,0.512]
95%	1.537	1.560	[1.482,1.649]
97.5%	2.056	2.064	[1.943,2,225]

4.4.3 Modeling Multivariate Monthly Precipitation

There has also been a growing interest for flexible families of non-Gaussian distributions allowing skewness and heavy tails in environmental science and climatology, as more heavy-tailed and skewed data are observed practically. Specifically, it is well known that monthly rainfall is strongly skewed to the right with high positive values of skewness coefficients (e.g.Wilson and Toumi (2005)). Various distributions have been suggested to model the precipitation data, among which there are the exponential, gamma(e.g.Wilks (2005) p.98), log-normal (e.g. Crow and Shimizu (1988)) and log-skewed-normal/t distributions (Marchenko and Genton (2010)). However, most of the studies are focusing on univariate modeling, and there has been little physical justification to why a specific distribution is used. Although skewed-Gaussian/t-type distributions have being extended to the multivariate setting (see e.g. Marchenko and Genton (2010) and Azzalini (2011)), they do not handle substantial skewness well (Aas and Haff (2006)). We are motivated to consider applying the Bayesian semi-parametric GVMM framework to model multivariate precipitation data. One appealing feature is that the extension to multivariate case is both straightforward and interpretable.

U.S. national and regional precipitation data are publicly available from the United States Historical Climatology Network (USHCN). For the purpose of exposition, we used monthly precipitation data measured in inches from four local stations (Albemarle, Chapel Hill, Edenton and Elizabeth City) in the state of North Carolina, for the period from 1895 through 2010 (116 data per station for each month). Figure 4.3 presents the histogram of log monthly precipitation data for July every year. Data from all four stations exhibit right skewness, with sample skewness coefficients being 0.675, 1.440, 0.632 and 0.971 respectively.

We fit the semi-parametric multivariate GVMM (4.2) to the data (dimension

p=4).

We run the Markov Chain for 10000 iterations, which shows good mixing and convergence, and discarded the first 5000 as burn-in. To illustrate the model fitting, we reconstruct a precipitation dataset with 5000 observations based on posterior samples of all unknown quantities, and compare the reconstructed posterior predictive distribution with the observed. Specifically, we test both whether the marginal univariate distributions of each dimension and the covariance structure are captured correctly. As shown in Figure 4.4, the curves represent the corresponding univariate kernel smoothing density estimates of the posterior predictive distributions based on 5000 samples from the fitted model, which show good fitting to the observed dataset shown with the histograms. The goodness of fit is further illustrated by the PP-plot between the observed and fitted distributions. Covariance structure is also modeled well (a comparison between Figure 4.6 and 4.7). All of these suggest that our framework effectively captures the underlying structure of the precipitation data.

As a comparison, we also fitted the log precipitation data with multivariate skewed-t distributions (Hansen (1994)), which have also been used to model skewed and heavy-tailed data, with two tails behaving as polynomials. The fitted model is obtained via maximum likelihood estimation using the **sn** R package provided by Azzalini (2011), and both the marginal distributions and covariance structure of the fitted multivariate skewed-t model are compared to those of the observed dataset. We can find that although the multivariate skewed-t distribution captures most of the univariate marginal distributions with a slightly lighter right tails and heavier left tails (Figure 4.10), it failed to find the correct covariance structure (Figure 4.11 compared to Figure 4.6 and 4.7). This could result from getting stuck in a local maximum and obtaining a suboptimal solution due to the relatively complicated likelihood function of multivariate skewed-t distributions. In any case, our Bayesian semi-parametric multivariate GVMM seems to provide a computationally and struc-

turally simpler yet more effective family of distributions for multivariate skewed and heavy-tailed data.

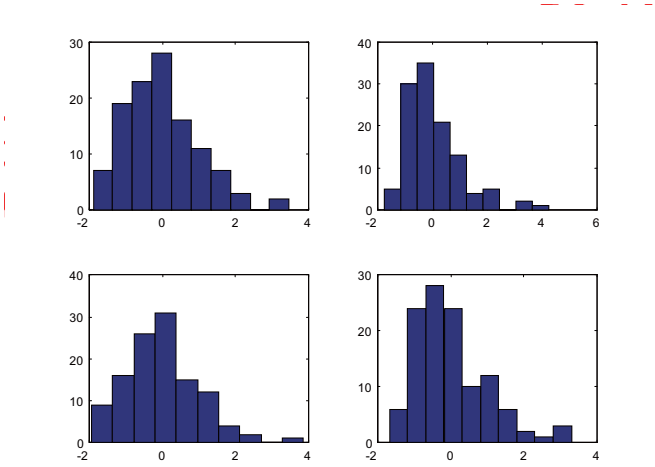

FIGURE 4.3: Monthly log precipitation data for July from 1895 to 2010 (116 observations) obtained from four stations in North Carolina show heavy right skewness.

4.5 Discussion

This chapter proposes the use of Bayesian semi-parametric Gaussian variance-mean mixtures as a flexible, interpretable and computationally tractable model for heavy-tailed and skewed observations. The model assumes the mixing distribution G in the general Gaussian variance-mean mixtures to be unknown, so the data inform about the appropriate distribution choice, the degree of skewness, heaviness of tails and shape of the distributions, and thus provide a more flexible framework for heavy-tailed and skewed data analysis. Although we test the model with univariate and multivariate simulated and real data, the broader question of scaling the framework to higher dimensions, possibly combined with factor models for sparse modeling, is interesting and challenging. Under the same scenario, the assumption that a single mixing variable is controlling the tails and skewness of all dimensions seems restric-

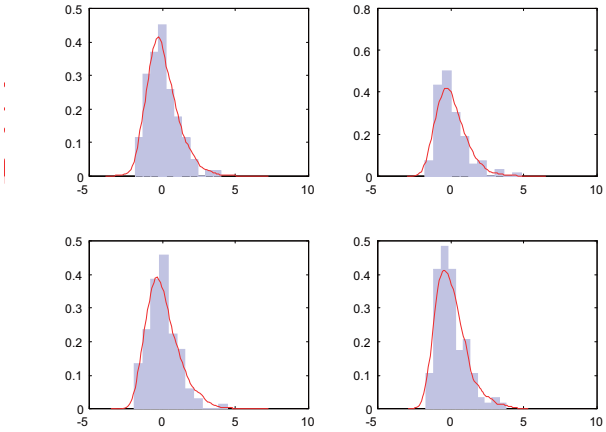

FIGURE 4.4: Monthly log precipitation data for July from 1895 to 2010 (116 observations) obtained from four stations in North Carolina (shown in histogram) are fitted using Bayesian semi-parametric GVMM. Red line shows the kernel density of fitted distributions for the stations, estimated from 5000 posterior predictive samples of the fitted GVMM.

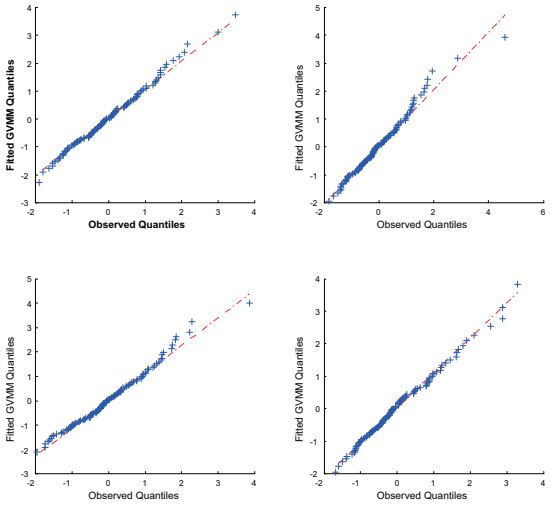

FIGURE 4.5: PP-plots for the Bayesian semi-parametric GVMM model fitted to log-precipitation for July in all four stations.

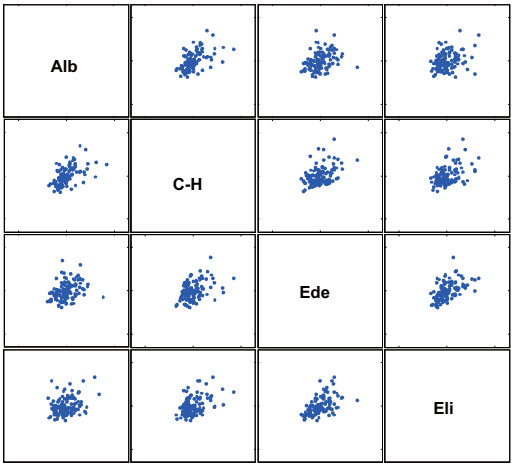

FIGURE 4.6: Sample covariance structure of monthly log precipitation data from four local stations in the state of North Carolina. Alb: Albemarle, C-H: Chapel Hill, Ede: Edenton and Eli: Elizabeth City

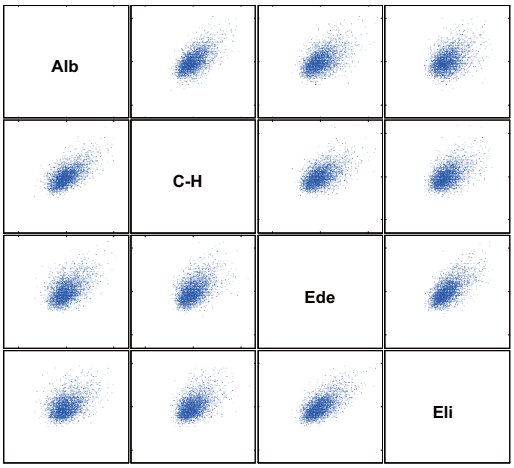

FIGURE 4.7: Covariance structure of monthly log precipitation when fitted with Bayesian semi-parametric GVMM.Alb: Albemarle, C-H: Chapel Hill, Ede: Edenton and Eli: Elizabeth City are four stations in the state of North Carolina.

tive, although this may be an acceptable assumption with low dimensions. Some hierarchical non-parametric models that allow multiple mixing variable/distributions may help in this case, but how to define an identifiable parametrization will definitely be an issue whenever multiple mixing variables are assumed, which is also worth more thoughts and theory to support. On the other hand, we consider time-varying semi-parametric GVMM a natural extension to the GVMM framework, taking advantage of the interpretability of model parameters. However, the more general class of spatial-temporal models, where additional structures and sources of information are included, is still challenging and yet unexplored.

4.6 Supplementary Materials

4.6.1 Proof

Proof. of Fact 2. Let V and W be two random variables such that

1. $f(v|w) = \frac{1}{2\psi^{1/2}w^{1/s}v} I\left(\left|\frac{\log v - m}{\psi^{1/2}}\right|^s < w\right)$
2. $W \sim \text{Gamma}(1 + 1/s, 1)$

then

$$\begin{aligned} f(v) &= \int f(v|w)f(w)dw \\ &\propto \int \frac{1}{w^{1/s}v} I\left(\left|\frac{\log v - m}{\psi^{1/2}}\right|^s < w\right) \cdot w^{1/s} exp(-w) dW \\ &\propto \int \frac{1}{v} I\left(\exp(-w) < \exp(-\left|\frac{\log v - m}{\psi^{1/2}}\right|^s)\right) \exp(-w)dw \\ &\propto \frac{1}{v} \exp\left(-\left|\frac{\log v - m}{\psi^{1/2}}\right|^s\right) \end{aligned}$$

which is the kernal of a $logGN(m, \psi, s)$ as defined in 4.3. □

Proof. of Lemma 4. Given p-dimensional observation y from a Gaussian variance mean mixtures described in (4.1). Let f_G^* denote the pdf of the Gaussian variance mean mixture with mixing distribution G^*, and let f_G denote the pdf of the Gaussian variance mixture ($\gamma = 0$) with mixing distribution G . Without loss of generality, we considered the case with $\mu = 0$ (otherwise, a linear transformation $\mu + y$ will complete the proof).

$$
\begin{aligned}
f(y) &\propto \int_0^\infty V^{-p/2} \exp\left(-\frac{1}{2} y' \frac{1}{V} \Sigma^{-1} y g(V) dV\right) \\
&\propto \int_0^\infty V^{-p/2} \exp\left(-\frac{1}{2} y' \frac{1}{V} \Sigma^{-1} y\right) \exp\left(-\frac{1}{2} \gamma' \Sigma^{-1} \gamma V\right) \exp\left(\frac{1}{2} \gamma' \Sigma^{-1} \gamma V\right) g(V) dV
\end{aligned}
$$

Let $\delta = \frac{1}{2}\gamma'\Sigma^{-1}\gamma$, with $\delta = \frac{1}{2}\gamma^2 \geqslant 0$ in the univariate case, and define

$$
g^*(V) = \frac{g(V)\exp(\delta V)}{\varphi(\delta)}
$$

where g and g^* are densities for G and G^* respectively, and φ is the moment generating function for g, with $\varphi(\delta) < \infty$. Plug g^* in, and we can see that

$$
f_{g^*}(y) = \frac{\exp(\gamma y)}{\varphi(\gamma^2/2)} f_g(y)
$$

. □

4.6.2 Supplementary Tables and Figures

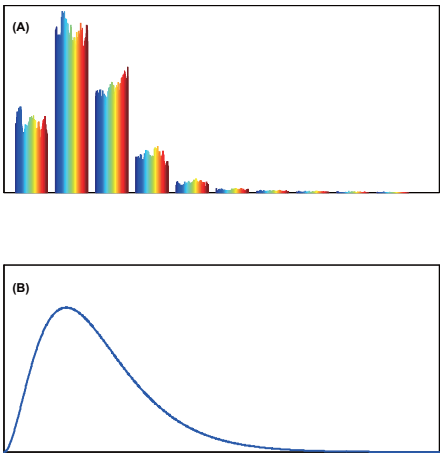

FIGURE 4.8: A comparison between the true mixing distribution (Gamma(3,1), panel (B)) and histograms of 100 reconstructed mixing distributions (panel (A)) show significant similarity.

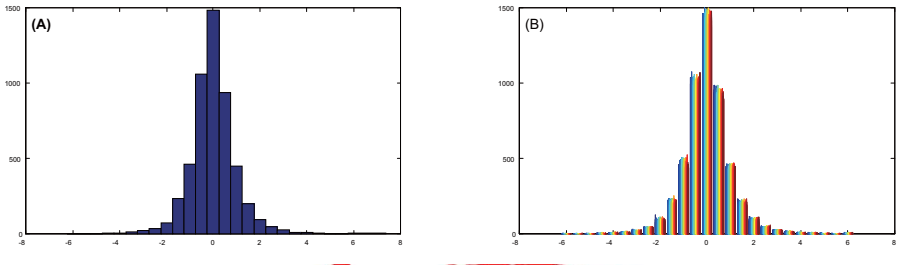

FIGURE 4.9: Fitting S&P 500 index return via univariate NVMM. Panel (A) shows the histogram of 200 reconstructed dataset (each consisting of 5470 observations) based on posterior samples. Panel (B) shows the real S&P 500 return from 01/02/1990 to 09/13/2011, totally 5470 observations. Significant similarity is observed.

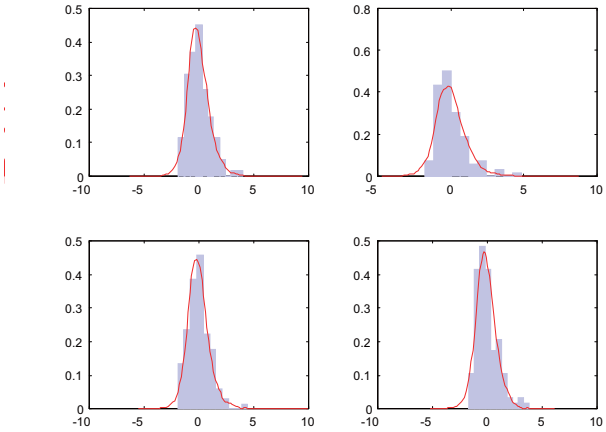

FIGURE 4.10: Monthly log precipitation data for July from 1895 to 2010 (116 observations) obtained from four stations in North Carolina (shown in histogram) are fitted using multivariate skewed-t distribution. Curves show the fitted distributions for the stations via maximum likelihood, using the **sn** R package.

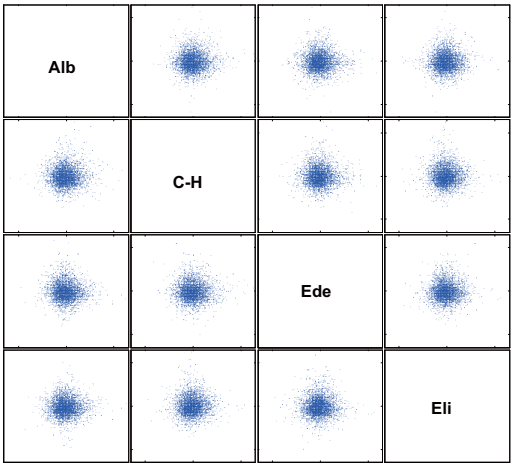

FIGURE 4.11: Covariance structure of monthly log precipitation when fitted with multivariate skewed-t distribution. Maximum likelihood estimation is used to obtain the fitted model using the **sn** R package.

Bibliography

Aas, K. and Haff, I. H. (Spring 2006), "The Generalized Hyperbolic Skew Students t-Distribution," *Journal of Financial Econometrics*, 4, 275–309.

Aguilar, O. and West, M. (2000), "Bayesian Dynamic Factor Models and Portfolio Allocation," *Journal of Business and Economic Statistics*, 18, 338–357.

Alspach, D. L. (1974), "Gaussian Sum Approximations in Nonlinear Filtering and Control," *Information Sciences*, 7, 271 – 290.

Altman, E. I., Brady, B., Resti, A., and Sironi, A. (2005), "The Link between Default and Recovery Rates: Theory, Empirical Evidence, and Implications," *Journal of Business*, 78, 2203–2228.

Anderson, T. (1963), "The use of factor analysis in the statistical analysis of multiple time series," *Psychometrika*, 28, 1–25.

Andreou, E., Ghysels, E., and Kourtellos, A. (2010), "Regression models with mixed sampling frequencies," *Journal of Econometrics*, 158, 246–261.

Arslan, O. (2010), "An alternative multivariate skew Laplace distribution: properties and estimation," *Statistical Papers*, 51, 865–887.

Arslan, O. and Genc, A. I. (2009), "The skew generalized t distribution as the scale mixture of a skew exponential power distribution and its applications in robust estimation," *Statistics*, 43, 481–498.

Attias, H. (1999), "Independent Factor Analysis," *Neural Computation*, 11, 803–851.

Azzalini, A. (1985), "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics*, 12, 171–178.

Azzalini, A. (2011), *R package* **sn***: The skew-normal and skew-t distributions (version 0.4-17)*, Università di Padova, Italia.

Azzalini, A. and Dalla Valle, A. (1996), "The multivariate skew-normal distribution," *Biometrika*, 83, 715–726.

Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171.

Bai, J. and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

Barndorff-Nielsen, O. E., K. J. L. and Sorensen, M. (1982), "Normal Variance-Mean Mixtures and z Distributions," *International Statistical Review*, 50, 145–159.

Barndorff-Nielsen, O. E. (1977), "Exponentially Decreasing Distributions for the Logarithm of Particle Size," *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 353, 401–419.

Barndorff-Nielsen, O. E. (1982), "The Hyperbolic Distribution in Statistical Physics," *Scandinavian Journal of Statistics*, 9, 43–46.

Barndorff-Nielsen, O. E. (1997), "Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling," *Scandinavian Journal of Statistics*, 24, 1–13.

Barndorff-Nielsen, O. E. and Shephard, N. (2001), "Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics," *Journal Of The Royal Statistical Society Series B*, 63, 167–241.

Barndorff-Nielsen, O. E., Jensen, J. L., and Sorensen, M. (1990), "Parametric Modelling of Turbulence," *Philosophical Transactions: Physical Sciences and Engineering*, 332, 439–455.

Bartholomew, D. J. (1988), "The Sensitivity of Latent Trait Analysis to Choice of Prior Distribution," *British Journal of Mathematical and Statistical Psychology*, 11, 101–107.

Bhattacharya, A. and Dunson, D. B. (2011a), "Sparse Bayesian infinite factor models," *Biometrika*, 98, 291–306.

Bhattacharya, A. and Dunson, D. B. (2011b), "Sparse Bayesian infinite factor models," *Biometrika*, 98, 291–306.

Boashash, B. (2006), *Time Frequency Signal Analysis and Processing: A Comprehensive Reference*, Oxford University Press.

Bruche, M. and Gonzlez-Aguado, C. (2010), "Recovery rates, default probabilities, and the credit cycle," *Journal of Banking & Finance*, 34, 754–764.

Carvalho, C., Chang, J., Lucas, J., Nevins, J. R., Wang, Q., and West, M. (2008a), "High-dimensional sparse factor modelling: applications in gene expression genomics," *Journal of the American Statistical Association*, 103, 1438–1456.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008b), "High-Dimensional Sparse Factor Modelling: Applications in Gene Expression Genomics," *Journal of the American Statistical Association*, 103, 1438–1456.

Chen, B., Chen, M., Paisley, J. W., Zaas, A. K., Woods, C. W., Ginsburg, G. S., Hero, A. O., Lucas, J. E., Dunson, D. B., and Carin, L. (2010a), "Bayesian Inference of the Number of Factors in Gene-Expression Analysis: Application to Human Virus Challenge Studies." *BMC Bioinformatics*, 11, 552.

Chen, M., Silva, J., Paisley, J. W., Wang, C., Dunson, D. B., and Carin, L. (2010b), "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: algorithm and performance bounds," *Trans. Sig. Proc.*, 58, 6140–6155.

Chin, R. and Lee, B. Y. (2008), *Principles and Practice of Clinical Trial Medicine*, Academic Press, 1 edn.

Comon, P. (1994), "Independent component analysis, A new concept?" *Signal Process.*, 36, 287–314.

Creal, D. D., Schwaab, B., Koopman, S. J., and Lucas, A. (2011), "Observation Driven Mixed-Measurement Dynamic Factor Models with an Application to Credit Risk," Tinbergen Institute Discussion Papers 11-042/2/DSF16, Tinbergen Institute.

Crow, E. L. and Shimizu, K. (1988), *Lognormal Distributions: Theory and Applications*, Marcel Dekker.

D. Durante, B. Scarpa, D. B. D. (2012), "Locally adaptive Bayesian covariance regression," *ArXiv e-prints*.

Das, S. R., Duffie, D., Kapadia, N., and Saita, L. (2007), "Common Failings: How Corporate Defaults Are Correlated," *Journal of Finance*, 62, 93–117.

Dunson, D. B. (2000), "Bayesian Latent Variable Models for Clustered Mixed Outcomes," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62, pp. 355–366.

Dunson, D. B. (2003), "Dynamic Latent Trait Models for Multidimensional Longitudinal Data," *Journal of the American Statistical Association*, 98, 555–563.

Dutta, R. and Ghosh, J. K. (2010), "Bayes Model Selection with Path Sampling: Factor Models and Other Examples," *ArXiv e-prints*.

Eberlein, E. and Prause, K. (1998), "The Generalized Hyperbolic Model: Financial Derivatives and Risk Measures," in *Mathematical Finance C Bachelier Congress 2000, Geman*, pp. 245–267, Springer.

Efron, B., Hastie, T., Johnstone, L., and Tibshirani, R. (2004), "Least angle regression," *Annals of Statistics*, 32, 407–499.

Gelman, A. (2004), "Parameterization and Bayesian Modeling," *Journal of the American Statistical Association*, 99, 537–545.

Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, 3, 514–534.

Genton, M. and Loperfido, N. (2005), "Generalized skew-elliptical distributions," *Annals of the Institute of Statistical Mathematics*, 57, 389–401.

Ghosh, J. and Dunson, D. B. (2008), *Bayesian model selection in factor analytic models, Random Effect and Latent Variable Model Selection, ed. D.B. Dunson*, John Wiley & Sons.

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, 306–320.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2005), "There is a risk-return trade-off after all," *Journal of Financial Economics*, 76, 509–548.

Ghysels, E., Sinko, A., and Valkanov, R. (2007), "MIDAS Regressions: Further Results and New Directions," *Econometric Reviews*, 26, 53–90.

Gupta, A. K. (2003), "Multivariate skewed t-distributions," *Statistics*, 37, 1.

Hahn, P. R., Mukherjee, S., and Carvalho, C. (2010), "Predictor-dependent shrinkage for linear regression via partial factor modeling," *ArXiv e-prints*.

Hamann, E., Deistler, M., and Scherrer, W. (2005), "Factor Models for Multivariate Time Series," in *Adaptive Information Systems and Modelling in Economics and Management Science*, eds. and A. Taudes, vol. 5 of *Interdisciplinary Studies in Economics and Management*, pp. 243–251, Springer Vienna.

Hansen, B. E. (1994), "Autoregressive Conditional Density Estimation," *International Economic Review*, 35.

Harrison, P. J. and Stevens, C. (1971), "A Bayesian Approach to Short-Term Forecasting," *Operational Research Quarterly (1970-1977)*, 22, pp. 341–362.

Hesketh, S., Pickles, A., and A., S. (2003), "Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation," *Statistical Modelling*, 3, 215–232.

Ishwaran, H. and James (2001), "Gibbs Sampling Methods for Stick Breaking Priors," *Journal of the American Statistical Association*, pp. 161–173.

Jolliffe, I. T. (1982), "A note on the use of principal components in regression," *Applied Statistics*, 31, 300.

Jungbacker, B. and Koopman, S. J. (2008), "Likelihood-based Analysis for Dynamic Factor Models," Tinbergen Institute Discussion Papers 08-007/4, Tinbergen Institute.

Keckler, S. W., Dally, W. J., Khailany, B., Garland, M., and Glasco, D. (2011), "GPUs and the future of parallel computing," *Micro, IEEE*, 31, 7–17.

Kinney, S. K. and Dunson, D. B. (2007), "Fixed and Random Effects Selection in Linear and Logistic Models," *Biometrics*, 63, 690–698.

Knowles, D. and Ghahramani, Z. (2010), "Nonparametric Bayesian Sparse Factor Models with application to Gene Expression modelling," *Annals of Applied Statistics*.

Knowles, D. and Ghahramani, Z. (2011), "Nonparametric Bayesian sparse factor models with application to gene expression modeling," *Annals of Applied Statistics*, 5, 1534–1552.

Le Cao, K. A., Gonzalez, I., and Dejean, S. (2009), "integrOmics: an R package to unravel relationships between two omics datasets," *Bioinformatics*, 25, 2855–2856.

Lee, S. Y. and Song, X. Y. (2002), "Bayesian Selection on the Number of Factors in a Factor Analysis Model," *Behaviormetrika*, 29, 23–40.

Liu, J. and West, M. (2001), "Combined parameter and state estimation in simulation-based filtering," in *Sequential Monte Carlo Methods in Practice*, eds. J. F. G. D. F. A. Doucet and N. J. Gordon, Springer-Verlag, New York.

Liu, J. S. and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

Lopes, H. F. and West, M. (2004), "Bayesian Model Assessment In Factor Analysis," *Statistica Sinica*, 14, 41–67.

Lopes, H. F., Gamerman, D., and Salazar, E. (2011), "Generalized spatial dynamic factor models," *Computational Statistics & Data Analysis*, 55, 1319–1330.

Ma, Y. and Genton, M. G. (2010), "Explicit estimating equations for semiparametric generalized linear latent variable models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 475–495.

Marchenko, Y. V. and Genton, M. G. (2010), "Multivariate log-skew-elliptical distributions with applications to precipitation data," *Environmetrics*, 21, 318–340.

Mevik, B. and Wehrens, R. (2007), "The pls Package: Principal Component and Partial Least Squares Regression in R," *Journal of Statistical Software*, 18, 1–24.

Mohamed, S., Heller, K., and Ghahramani, Z. (2008), "Bayesian Exponential Family PCA," *In Proceedings of the 21st Conference on Advances in Neural Information Processing Systems (NIPS21)*.

Montanari, A. and Viroli, C. (2010a), "Heteroscedastic factor mixture analysis," *Statistical Modelling*, 10, 1–24.

Montanari, A. and Viroli, C. (2010b), "A skew-normal factor model for the analysis of student satisfaction towards university courses," *Journal of Applied Statistics*, 37, 473–487.

Moustaki, I. and Knott, M. (2000), "Generalized latent trait models," *Psychometrika*, 65, 391–411.

Muthen, B. (1984), "A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators," *Psychometrika*, 49, 115–132.

Nadarajah, S. (2005), "A generalized normal distribution," *Journal of Applied Statistics*, 32, 685–694.

Nickell, P., Perraudin, W., and Varotto, S. (2000), "Stability of rating transitions," *Journal of Banking & Finance*, 24, 203–227.

Niemi, J. and West, M. (2010), "Adaptive mixture modelling Metropolis methods for Bayesian analysis of non-linear state-space models," *Journal of Computational and Graphical Statistics*, 19, 260–280.

Polson, N. G. and Scott, J. G. (2011), *Shrink globally, act locally: sparse Bayesian regularization and prediction. In Bayesian Statistics 9 (eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West)*, Oxford: Oxford University Press.

Ravines, H. S. M. and Schmidt, A. (2007), "An Efficient Sampling Scheme for Generalized Dynamic Models," Working papers, Universidade Federal do Rio de Janeiro.

Rothman, A. J., L. E. and Zhu, J. (2010), "A new approach to Cholesky-based covariance regularization in high dimensions." *Biometrika*, 97, 539–550.

Runcie, D. E. and Mukherjee, S. (2012), "Bayesian Sparse Factor Analysis of Genetic Covariance Matrices," *ArXiv e-prints*.

S. Venturini, F. D. and Parmigiani, G. (2008), "Gamma shape mixtures for heavy-tailed distributions," *The Annals of Applied Statistics*, 2, 756–776.

Sahu, S. K., Dey, D. K., and Branco, M. D. (2003), "A New Class of Multivariate Skew Distributions with Applications to Bayesian Regression Models," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 31, 129–150.

Salmon, F. (2012), "The formula that killed Wall Street," *Significance*, 9, 16–20.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Shen, W. and Ghosal, S. (2012), "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures." *arXiv:1109.6406v2*, 9, 16–20.

Sorenson, H. and Alspach, D. (1971), "Recursive Bayesian estimation using Gaussian sums," *Automatica*, 7, 465 – 479.

Stock, J. H. and Watson, M. W. (2002), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business & Economic Statistics*, 20, 147–62.

Stroud, J. R., Mller, P., and Polson, N. G. (2003), "Nonlinear State-Space Models with State-Dependent Variances," *Journal of the American Statistical Association*, 98, pp. 377–386.

Suchard, M. A., Wang, Q., Chan, C., Frelinger, J., Cron, A., and West, M. (2010), "Understanding GPU Programming for Statistical Computation: Studies in Massively Parallel Massive Mixtures." *Journal of computational and graphical statistics*, 19, 419–438.

Varanasi, M. K. and Aazhang, B. (1989), "Parametric generalized Gaussian density estimation," *Journal of the Acoustical Society of America*, 86, 1404–1415.

Vianelli, S. (1983), "The family of normal and lognormal distributions of order r," *Metron*, 41, 3–10.

Wang, J. and Genton, M. (2006), "The multivariate skew-slash distribution," *Journal of Statistical Planning and Inference*, 136, 209 – 220.

Wang, J., Boyer, J., and Genton, M. G. (2004), "A Skew-Symmetric Representation of Multivariate Distributions," *Statistica Sinica*, 14, 1259–1270.

West, M. (2003), "Bayesian Factor Regression Models in the "Large p, Small n"Paradigm," in *Bayesian Statistics*, pp. 723–732, Oxford University Press.

Wilks, D. S. (2005), *Statistical methods in the atmospheric sciences, Volume 91, Second Edition (International Geophysics)*, Academic Press.

Wilson, P. S. and Toumi, R. (2005), "Gibbs Sampling Methods for Stick Breaking Priors," *Geophisical Research Letters*, 32.

Yang, M. and Dunson, D. (2010a), "Bayesian Semiparametric Structural Equation Models with Latent Variables," *Psychometrika*, 75, 675–693.

Yang, M. and Dunson, D. B. (2010b), "Bayesian semiparametric structural equation models with latent variables," *Psychometrika*, 75, 675–693.

Yung, Y. F. (1997), "Finite mixtures in confirmatory factor-analysis models," *Psychometrika*, 62, 297–330.

Zhang, Z. and Nesselroade, J. R. (2007), "Bayesian Estimation of Categorical Dynamic Factor Models," *Multivariate Behavioral Research*, 42, 729.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320.

Biography

Kai Cui was born in Laiyang, Shandong Province, China on July. 12, 1983. In July 2006, he received his Bachelor's degree in Science from Tsinghua University, Beijing, China. He joined the Department of Statistical Science at Duke University in 2009 to pursue his PhD degree in statistics. At Duke, he worked on Bayesian statistics, statistical modeling and computational approaches. In 2011, he earned his *rn route* M.S. degree in Statistical Science from Duke University.