M.S. THESIS

Robust Sparse Bayesian Infinite Factor Models

강건 희소 베이즈 무한 인자 모형

BY

LEE JAE-JOON

FEBRUARY 2021

DEPARTMENT OF STATISTICS
SEOUL NATIONAL UNIVERSITY

M.S. THESIS

Robust Sparse Bayesian Infinite Factor Models

강건 희소 베이즈 무한 인자 모형

BY

LEE JAE-JOON

FEBRUARY 2021

DEPARTMENT OF STATISTICS
SEOUL NATIONAL UNIVERSITY

Robust Sparse Bayesian Infinite Factor Models

강건 희소 베이즈 무한 인자 모형

지도교수 이 재 용

이 논문을 이학석사 학위논문으로 제출함

2020년 10월

서울대학교 대학원

통계학과

이 재 준

이재준의 이학석사 학위 논문을 인준함

2020년 12월

위 원 장: 임 요 한 (인)

부위원장: 이 재 용 (인)

위 원: 정 성 규 (인)

Abstract

Most of previous works and applications of Bayesian factor model have assumed the normal likelihood regardless of its validity. We propose a Bayesian factor model for heavy-tailed high-dimensional data based on multivariate Student-t likelihood to obtain better covariance estimation. We use multiplicative gamma process shrinkage prior and factor number adaptation scheme proposed in Bhattacharya and Dunson [*Biometrika* (2011) 291–306]. Since a naive Gibbs sampler for the proposed model suffers from slow mixing, we propose a Markov Chain Monte Carlo algorithm where fast mixing of Hamiltonian Monte Carlo is exploited for some parameters in proposed model. Simulation results illustrate the gain in performance of covariance estimation for heavy-tailed high-dimensional data. We also provide a theoretical result that the posterior of the proposed model is weakly consistent under reasonable conditions. We conclude the paper with the application of proposed factor model on breast cancer metastasis prediction given DNA signature data of cancer cell.

keywords: Bayesian modeling, Factor model, Multiplicative gamma process prior, Multivariate t-distribution, Hamiltonian Monte Carlo

student number: 2019-24162

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Factor model is a highly efficient tool to understand the covariance structure of high-dimensional data. The covariance structure is captured by representing the p-dimensional observation as the sum of linear transformation of latent factors ($k \ll p$) and an error term. In the factor model, the covariance matrix Ω has the form of $\Omega = \Lambda\Lambda^T + \Sigma$, where Λ is a $p \times k$ factor loading and Σ is a $p \times p$ diagonal error variance matrix. Due to the parsimony of representing $p \times p$ covariance with only $p(k+1)$-dimensional parameters, the factor model is widely used for covariance estimation in many applications with high-dimensional data, e.g. spatial analysis (Lopes et al., 2008) and genomics (Carvalho et al., 2008).

The number of latent factors k is a key element in the factor model. Variations of the factor model have been proposed for the estimation of the number of factors. Lopes & West (2004) updated the number of latent factors in the posterior sampling process using reversible jump Markov Chain Monte Carlo (Green, 1995). Ando (2009) determined the number of latent factors by maximizing the marginal likelihood, which is analytically derived with a chosen prior distribution. Bhattacharya & Dunson (2011) proposed *multiplicative gamma process shrinkage prior*, which is a prior for the infinite factor model and encourages factor loadings with large indices to be close to 0. In the posterior sampling, the number of factors is adapted by adding or deleting latent factors depending on the sparsity of the current factor loading estimate. Such adaptation in Bhattacharya & Dunson (2011) is desirable in that an additional calculation is not required. Moreover, it is guaranteed that the Markov Chain Monte Carlo (MCMC) algorithm using factor adaptation is ergodic.

For the last few decades, many approaches have been made to obtain sparse estimator under the high-dimensional setting. Variations of factor model have been proposed in a similar vein. West (2003) and Carvalho et al. (2008) used the spike-and-slab prior on factor loadings, which is a mixture of a point mass at 0 and a continuous density. Although the point mass mixture prior is intuitive and does induce sparse estimates, it has a critical disadvantage of slow mixing and convergence. Later, due to the advantage in posterior computation over point mass mixture prior, factor models using global-local shrinkage prior (Polson & Scott, 2010) have been suggested. For example, the aforementioned infinite factor model of Bhattacharya & Dunson (2011) assigned multiplicative gamma process shrinkage prior on factor loadings, and Ferrari & Dunson (2020) proposed a factor regression model using Dirichlet-Laplace shrinkage prior (Bhattacharya et al., 2015) on factor loadings.

Most of the factor models aforementioned are based on the normality assumption which, however, is ill-suited when outliers are present. Ando (2009) proposed a factor model with matrix-variate t distribution to obtain robust estimate. Zhang et al. (2014) proposed a robust version of the factor model utilizing the fact that a multivariate t distribution can be represented as a scale mixture of normal distributions. To the best of our knowledge, however, no approach has been proposed for both robustness against outliers and sparsity of the estimate.

This work proposes a *robust sparse Bayesian infinite factor model*, which estimates covariance robustly under heavy tail distribution. Specifically, it is an extension of the *sparse Bayesian infinite factor model* (Bhattacharya & Dunson, 2011), utilizing the multivariate t likelihood instead of normal likelihood. Under the heavy tail distribution, the proposed model has improved performance of covariance estimation over the normal-likelihood factor model of Bhattacharya & Dunson (2011). Also, we show that under the assumption of known degrees of freedom of t-distribution, the posterior is consistent under the weak topology. Despite the optimal value of t degrees of freedom is not given in the real data analysis, simulation results indicate that the proposed model outperforms the normal-likelihood factor model by choosing a sufficiently small number as the degrees of freedom of t distribution.

In Chapter 2 basic concept and previous approaches of Bayesian factor model are introduced. In Chapter 3 we propose robust factor model with Student's t-likelihood. The posterior computation algorithm is also presented. In Chapter 4 we show theoretical properties of the proposed model. In Chapter 5 performance of the proposed model is demonstrated through simulation studies. In Chapter 6 the proposed model is applied to prediction of breast carcinoma metastasis using microarray data of cancer tissue. The discussion is given in Chapter 7.

Chapter 2

Factor Models

2.1 Settings

Let $\mathbf{Y} \in \mathbb{R}^{n\times p}$ be an independent (centered) sample of n observations with p variables. Factor model is a latent variable model which assumes that the mean of p-dimensional observation $\mathbf{y}_i$ is determined by latent factor η_i with lower dimensions $k \ll p$. The formulation of factor model is as follows:

$$\begin{aligned}
&\mathbf{y}_i = \Lambda\eta_i + \varepsilon_i, \;\; i = 1, \ldots, n, \\
&\eta_i \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k), \;\; \varepsilon_i \sim \mathcal{N}_p(\mathbf{0}, \Sigma), \;\; \Sigma = \text{diag}(\sigma_1^2, \cdots, \sigma_p^2),
\end{aligned}$$

where $\Lambda \in \mathbb{R}^{p\times k}$ is a factor loading matrix and $\varepsilon_i \in \mathbb{R}^p$ is an error term for ith observation. In this paper, for simplicity, we only consider the case of independent sample, i.e., the case where error covariance matrix Σ is a diagonal matrix. The diagonality assumption on error covariance matrix can be relaxed according to dependence structure of observations, for example, a banded matrix.

By the property of normal distribution, the conditional distribution and marginal distribution of $\mathbf{y}_i$ can be easily derived as follows:

$$\begin{aligned}
&\mathbf{y}_i|\eta_i \sim \mathcal{N}_p(\Lambda\eta_i, \Sigma) \\
&\mathbf{y}_i \sim \mathcal{N}_p(\mathbf{0}, \Omega), \;\; \Omega = \Lambda\Lambda^T + \Sigma.
\end{aligned}$$

Note that the covariance of observation $\mathbf{y}_i$ is $\Omega = \Lambda\Lambda^T + \Sigma$ in marginal likelihood. In practical applications with high-dimensional data of large p, this characterization of unknown covariance Ω is useful for capturing the low-dimensional($k \ll p$) covariance structure of the data.

For inference of parameters of interest, frequentist approach often uses expectation-maximization (EM) algorithm or its variation, because of the presence of unobserved latent variable η. The other example of frequentist approach is a principal component method. Principal component method calculates truncated singular value decomposition of covariance matrix of rank k and performs Cholesky factorization to obtain factor loading estimate. On the other hand, Bayesian approach assigns prior distribution on parameters of interest and performs inference on posterior distribution.

2.2 Bayesian Factor Models

Assuming number of latent factors k is given, one can consider the following simple prior distribution on factor loading Λ and error covariance Σ:

$$\lambda_{jh} \sim \mathcal{N}(0, \sigma_\Lambda^2), \quad j = 1, \dots, p, \;\; h = 1, \dots, k,$$
$$\sigma_j^{-2} \sim \text{Ga}(a_\sigma, b_\sigma), \quad j = 1, \dots, p,$$

where λ_{jh} is (j,h)-th entry of Λ and σ_j^2 is (j,j)-th entry of Σ. In this case posterior sampling is straightforward by Gibbs sampler. This simple formulation may be enough for some applications when data is small or the number of latent factors is given. There, however, are several issues need to be addressed when implementing factor model: unidentifiability of factor loading and unknown number of latent factors. The following sections focus on introducing preceding works of Bayesian factor models and how they dealt with each issue.

2.2.1 Unidentifiability of Factor Loading

For a statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, we say that $\mathcal{P}$ is identifiable if $\theta_1 = \theta_2$ implies $P_{\theta_1} = P_{\theta_2}$ for all $\theta_1, \theta_2 \in \Theta$. Under factor model formulation, unidentifiability of factor loading Λ is due to rotation invariance of Λ, i.e., for any orthogonal matrix $Q \in \mathbb{R}^{k \times k}$, the following holds:

$$\Omega = \Lambda\Lambda^T + \Sigma = \Lambda Q Q^T \Lambda^T + \Sigma = \Lambda Q (\Lambda Q)^T + \Sigma = \tilde{\Lambda}\tilde{\Lambda}^T + \Sigma.$$

Since a rotated factor loading $\tilde{\Lambda} = \Lambda Q$ induces the same covariance for any orthogonal matrix Q, the factor model is not identifiable by nature.

A well-known solution is to impose lower-triangular constraint and positive diagonal constraint (Geweke & Zhou, 1996). This approach has a limitation that choosing the order of first k variables whose factor loadings are under constraint becomes an important modeling decision. Carvalho et al. (2008) imposed lower-triangular and positive diagonal constraint on factor loading and suggested an model search algorithm to determine the set of variables to be included in the model and first k variables' order. It repeatedly adds variable and latent factor to initial choice of variable and latent factor dimension. However, the method is inefficient in that each iteration of the model search consists of refitting factor model under new set of variables and number of latent factors. Refer to Carvalho et al. (2008) for detailed procedure of the model search algorithm.

The other solution is to post-process the posterior samples of factor loading Λ. This approach is computationally more efficient than imposing constraints on factor loading because additional calculation or modification in prior specification is not needed. Ghosh & Dunson (2009) assigns prior on factor loading satisfying lower-triangular constraint and post-process the obtained posterior samples to satisfy positive diagonal constraint. McParland et al. (2014) used Procrustean method where the posterior samples of factor loading are rotated or reflected to be as close as possible to a reference factor loading matrix.

It is critical to obtain identifiablity of factor loading Λ in some applications where interpretation of factor loading is necessary, say factor analysis of spatial data analysis (Lopes et al., 2008). However, when performing covariance matrix estimation or latent factor regression using factor model, this identifiability issue of factor loading is not a problem under Bayesian framework, as it is in frequentist settings, as long as posterior distribution is proper (Bhattacharya & Dunson, 2011; Ferrari & Dunson, 2020).

2.2.2 Unknown Latent Factor Dimension

Factor model is a dimension reduction technique where high-dimensional data is represented as a linear transformation of low-dimensional latent factors. The number of latent factors k represents the dimension of low-dimensional subspace in which data lie. Thus the number of latent factors is a key component of factor model. In Bayesian framework, it is harder than it is in frequentist setting to compare estimates obtained from factor models of different numbers of latent factors based on certain information criteria, as Bayesian inference involves multiple iterations of posterior sampling. Thus most of recent Bayesian works on factor model choose either to determine the number of factors before main posterior computation or to update the number of factors within posterior computation.

The model search algorithm of Carvalho et al. (2008) can be an example of determining the number of latent factors in advance. The algorithm chooses the number of factors along with the set of variables to include in the model. Ando (2009) suggested a factor model with matrix-variate t prior on factor loadings and derived marginal likelihood of the number of latent factors analytically. The number of factors is then determined by maximizing the marginal likelihood. The two methods have limitation that the methods of determining the number of latent factors is not available in different prior settings.

On the other hand, there have been approaches that update the number of factors within parameter updates. Lopes & West (2004) updated the number of factors using reversible jump MCMC, which has birth/death move step at the end of every iteration. In adaptation step, the number of factors is updated with modified Metropolis-Hastings ratio. Bhattacharya & Dunson (2011) considered factor loading of all possible number of factors and dynamically truncated the latent factors with loadings close to 0, along with the prior which imposes stronger shrinkage to 0 for the factor with large index. This is efficient in that additional computation is not needed. Also the ergodicity is guaranteed for MCMC algorithm using the adaptation method of Bhattacharya & Dunson (2011), by condition of diminishing adaptation in Roberts & Rosenthal (2007). Detailed procedure of the method is illustrated in Section 3.1.

Chapter 3

Robust Sparse Bayesian Infinite Factor Models

3.1 Sparse Bayesian Infinite Factor Models

The *sparse Bayesian infinite factor model* (Bhattacharya & Dunson, 2011) is a Bayesian factor model specialized for high-dimensional covariance estimation. The joint distribution of observation $\mathbf{y}_i \in \mathbb{R}^p$ and latent factor $\eta_i \in \mathbb{R}^k$ is as follows:

$$\begin{bmatrix} \mathbf{y}_i \\ \eta_i \end{bmatrix} \Big| \Lambda, \Sigma \overset{iid}{\sim} \mathcal{N}_{p+k} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Lambda\Lambda^T + \Sigma & \Lambda \\ \Lambda^T & \mathbf{I}_d \end{bmatrix} \right), \quad i = 1, 2, \ldots, n.$$

The model is differentiated from the preceding Bayesian factor models in mainly two points: its expanded parameterization on factor loading Λ and the adaptation on the number of factors k.

Choosing the number of the latent factor k is an important issue. The model addresses this issue by first allowing the parameter space Θ_Λ to contain all possible numbers of latent factor and by dynamically truncating the insignificant latent factors in posterior computation. The parameter space of factor loading Λ and error covariance

Σ are as follows:

$$\Theta_\Lambda = \left\{ \Lambda = (\lambda_{jh}), j = 1, \ldots, p, h = 1, \ldots, \infty, \max_{1 \le j \le p} \sum_{h=1}^{\infty} \lambda_{jh}^2 < \infty \right\},$$
$$\Theta_\Sigma = \left\{ \Sigma \in \mathbb{R}^{p \times p} : \Sigma_{jj} > 0 \ \ \forall j = 1, \ldots, p, \ \ \Sigma_{ij} = 0 \ \ \forall 1 \le i \ne j \le p \right\},$$

where Σ_{ij} is the (i, j)th element of matrix Σ. Note that the condition

$$\max_{1 \le j \le p} \sum_{h=1}^{\infty} \lambda_{jh}^2 < \infty$$

is a necessary and sufficient condition for all the entries of $\Lambda\Lambda^T$ to be finite so that the resulting covariance matrix $\Omega = \Lambda\Lambda^T + \Sigma$ is defined.

For prior Π_Λ on factor loadings with infinitely many latent factors, *the multiplicative gamma process prior* is proposed. It is a global-local shrinkage prior (Polson & Scott, 2010) having entry-wise and column-wise variance components as local and global variance components, respectively. Also, choosing $a_2 \ge 1$, it is designed so that the strong shrinkage is imposed for the factors with large column index. The full prior specification of sparse Bayesian infinite factor models is as follows:

$$\begin{aligned} &\lambda_{jh} | \phi_{jh}, \tau_h \sim \mathcal{N}(0, \phi_{jh}^{-1} \tau_h^{-1}), \ \ \phi_{jh} \sim \text{Ga}(\kappa/2, \kappa/2), \ \ \tau_h = \prod_{l=1}^{h} \delta_l, \\ &\delta_1 \sim \text{Ga}(a_1, 1), \ \ \delta_l \sim \text{Ga}(a_2, 1), \ \ l \ge 2, \ \ a_1 \sim \text{Ga}(2, 1), \ \ a_2 \sim \text{Ga}(2, 1) \\ &\Sigma = \text{diag}(\sigma_1^{-2}, \ldots, \sigma_p^{-2}), \ \ \sigma_j^{-2} \sim \text{Ga}(a_\sigma, b_\sigma), \ \ j = 1, \ldots, p. \end{aligned} \tag{1}$$

The number of factors k is determined adaptively by adding or removing latent factor within MCMC iterations, inspecting current factor loading estimate $\hat{\Lambda}^{(t)}$. At the tth iteration of MCMC, the chain goes through adaptation step with decreasing probability $p(t)$, say $p(t) = 1/\exp(1 + 0.0005t)$. In adaptation step, if there are columns of the current value $\Lambda^{(t)}$ whose entries are all close to zero under prespecified threshold, the columns are removed, otherwise new columns are generated from the prior distribution and are added to the current factor loadings. Also corresponding columns of latent factor matrix η, variance components ϕ_{jh}, δ_h for deleted (added) column of factor loadings are also deleted (added) accordingly. The adaptation procedure is to keep only the effective latent factors whose factor loadings take up a large part of current posterior sample of the covariance.

This adaptive method has a significant advantage of computation, compared to other methods which needed additional MCMC step (Lopes & West, 2004) or comparison of other model selection criteria (Ando, 2009). As justification for their adaptation scheme, Bhattacharya & Dunson (2011) showed that, with the prior specified as equation 1, the prior probability of approximated covariance $\Omega_H = \Lambda_H^T \Lambda_H + \Sigma$ being arbitrarily close to $\Omega = \Lambda \Lambda^T + \Sigma$ converges to 1 at exponential rate as H goes to ∞, where Λ_H is a truncated factor loading of Λ with first H columns. Furthermore, the adaptation procedure satisfies the diminishing adaptation condition in Roberts & Rosenthal (2007). Thus the convergence of the MCMC algorithm is guaranteed.

3.2 Robust Sparse Bayesian Infinite Factor Models

The sparse Bayesian infinite factor model is a factor model based on the normal likelihood. Even though the model has proven its success in high-dimensional covariance estimation, the model may not be the best option when there are outliers in the data or the error distribution has heavy tail. We extend the model by replacing the normal distribution with t-distribution which has heavier tail and propose *robust sparse Bayesian infinite factor model*.

A multivariate t distribution has a polynomial tail instead of exponential one. The probability density function of multivariate t distribution is as follows:

$$f(\mathbf{y}|\nu, \mu, \Omega) = \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{p/2} \det(\Omega)^{1/2}} \left[1 + \frac{(\mathbf{y}-\mu)^T \Omega^{-1} (\mathbf{y}-\mu)}{\nu}\right]^{-\frac{\nu+p}{2}},$$

where $\Gamma(x)$ is a gamma function and $\det(A)$ is the determinant of a square matrix A. When extending normal likelihood to the t likelihood, we use an equivalent represen-

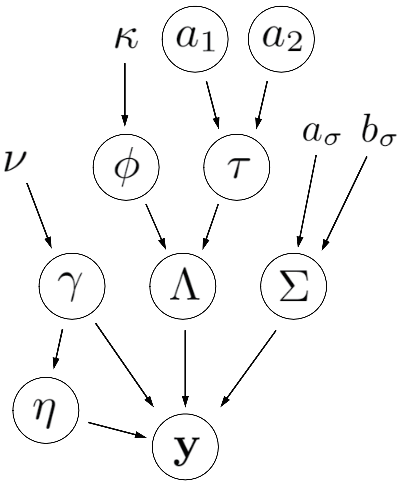

Figure 3.1: The directed acyclic graph representation for the proposed models

tation of multivariate t distribution as a scale mixture of normal distributions.

$$\begin{bmatrix}\mathbf{y}_i\\ \eta_i\end{bmatrix}\Big|\Lambda,\Sigma,\nu \overset{ind}{\sim} t_{p+k}\left(\nu,\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\Lambda\Lambda^T+\Sigma & \Lambda\\ \Lambda^T & \mathbf{I}_d\end{bmatrix}\right),\quad i=1,2,\dots,n.$$

$$\iff \begin{bmatrix}\mathbf{y}_i\\ \eta_i\end{bmatrix}\Big|\gamma_i,\Lambda,\Sigma \overset{ind}{\sim} \mathcal{N}_{p+k}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\frac{1}{\gamma_i}\begin{bmatrix}\Lambda\Lambda^T+\Sigma & \Lambda\\ \Lambda^T & \mathbf{I}_d\end{bmatrix}\right),\quad i=1,2,\dots,n. \tag{2}$$

$$\gamma_i|\nu \overset{iid}{\sim} \text{Ga}\left(\frac{\nu}{2},\frac{\nu}{2}\right),\quad i=1,2,\dots,n.$$

It is desirable to use the representation in equation 2 because posterior computation is a straightforward Gibbs update, exploiting conjugacy of the normal model with the normal prior distribution. The directed acyclic graph representation for the proposed model is illustrated in Fig. 3.1. The details of the posterior computation of the proposed model are explained in Section 3.3.

For the prior distribution of factor loading Λ and error variance Σ, we follow the prior Π_Λ and Π_Σ as defined in equation 1. As we are dealing with multivariate t-distribution, we have ν, the degrees of freedom, as an additional parameter. We fix ν

at sufficiently small value in all analyses. From extensive simulation studies, we found that datasets with moderate size have only dim information for ν and unspecified ν render slow mixing in the posterior sampling. Thus, the improved performance can be attained by choosing sufficiently small value of ν in the presence of outliers. The simulation results under different choices of ν are demonstrated in Chapter 5.

3.3 Inference

While most of the posterior computation steps of the proposed model are similar to those in Bhattacharya & Dunson (2011), a Gibbs update step can be modified to incorporate the auxiliary variable γ_i which extends the normal likelihood to the multivariate Student-t likelihood. For the number of latent factors k, we use the same factor adaptation strategy to adaptively determine k as illustrated in Section 3.1.

Since all conditionals are tractable distributions, a straightforward Gibbs sampler can be implemented for posterior computation of the proposed factor model with t likelihood, as in the normal-likelihood factor model of Bhattacharya & Dunson (2011). However, in high-dimensional setting ($n \ll p$), we have observed slow mixing of Markov chain when naive Gibbs sampler is implemented on the proposed model. To cope with the computational issue arising from more complicated model structure, we made two additional modifications upon Gibbs sampler; collapsing and Hamiltonian Monte Carlo.

The collapsed Gibbs sampler (Liu, 1994) is a variation of Gibbs sampler which utilizes the conditional of *collapsed* version of joint distribution with some parameters are marginalized out of the condition term. Decoupling some dependencies between conditionals, it is known that the collapsed Gibbs sampler leads to faster mixing than that of the Gibbs sampler. We apply this collapsing idea on η and γ. This is equivalent to regarding η and γ as a block of single parameter and updating them at a single step of a Gibbs sampler.

$$
\begin{aligned}
p(\gamma_i, \eta_i | \mathbf{y}_i, \cdots) &= p(\eta_i | \mathbf{y}_i, \cdots) p(\gamma_i | \eta_i, \mathbf{y}_i, \cdots) \\
p(\eta_i | \mathbf{y}_i, \cdots) &\sim t_k \left(\eta_i : \nu + p, (I + \Lambda\Sigma\Lambda)^{-1}\Lambda\Sigma\mathbf{y}_i, \frac{\nu + \mathbf{y}_i^T \mathbf{y}_i}{\nu + p} (I + \Lambda\Sigma\Lambda)^{-1} \right) \\
p(\gamma_i | \eta_i, \mathbf{y}_i, \cdots) &\sim \text{Ga} \left(\gamma_i : \frac{\nu + p + k}{2}, \frac{\nu + (\mathbf{y}_i - \Lambda\eta_i)^T \Sigma^{-1} (\mathbf{y}_i - \Lambda\eta_i) + \eta_i^T \eta_i}{2} \right)
\end{aligned}
$$

Fundamentally, the Gibbs sampler is a random-walk Metropolis algorithm with full conditional as a proposal distribution. Both methods explore parameter space via random walk which is highly inefficient for high-dimensional parameter space. Nowadays, in such a case with high-dimensional parameters, the Hamiltonian Monte Carlo is considered to be a gold-standard for posterior computation and has proven empirical success in many applications. The Hamiltonian Monte Carlo uses an auxiliary variable (momentum) and the information from gradient of the log-posterior to perform better search.

To deal with the complicated model structure of γ which affects both latent variable η and observation $\mathbf{y}$, we apply No-U-Turn sampler (Hoffman & Gelman, 2014) for updating η. The No-U-Turn sampler is a variation of the Hamiltonian Monte Carlo which automatically tunes the path length of Hamiltonian approximation. Though the No-U-Turn sampler is often used to update all parameters in the model, we applied single No-U-Turn sampler update per iteration. This is comparable to commonly used Metropolis-within-Gibbs scheme, which updates some parameter with Metropolis update while updating the others with Gibbs sampler. Applying No-U-Turn sampler on $n \times k$ dimensional η, we aim to keep both simplicity of overall posterior computation and better mixing of Hamiltonian Monte Carlo in posterior inference.

For the Metropolis-Hastings updates of a_1 and a_2, we used Gaussian proposal with lower bound constraint of $a_1 > 2$ and $a_2 > 3$, respectively. It is a sufficient condition that induced prior on each entry of covariance Ω has finite second moment. Refer to Section 2.2 of Bhattacharya & Dunson (2011) for the detailed explanation. Also Durante (2017) suggests that choosing a_2 moderately higher than a_1 facilitates better shrinkage of factor loadings, which motivates higher lower bound for a_2 than a_1. The MCMC algorithm for robust sparse Bayesian infinite factor models given the number of factors k is as follows:

1. Sample λ_j, the jth row of factor loading Λ, for $j = 1, \dots, p$ from normal distribution:
$$p(\lambda_j | \cdots) \sim \mathcal{N}_k \left(\lambda_j : \Psi_\Lambda^j \left(\sigma_j^{-2} \sum_{i=1}^n \gamma_i y_{ij} \eta_i \right), \Psi_\Lambda^j \right),$$
$$\text{where } \Psi_\Lambda^j = \left(\sigma_j^{-2} \sum_{i=1}^n \gamma_i \eta_i \eta_i^T + \text{diag}(\phi_{jh} \tau_h) \right)^{-1}.$$

2. Sample σ_j^{-2}, for $j = 1, \dots, p$ from gamma distributions:
$$p(\sigma_j^{-2} | \cdots) \sim \text{Ga} \left(\sigma_j^{-2} : a_\sigma + \frac{n}{2}, b_\sigma + \frac{\sum_{i=1}^n \gamma_i (y_{ij} - \lambda_j^T \eta_i)^2}{2} \right).$$

3. Sample η_i, for $i = 1, \dots, n$ with a single iteration of No-U-Turns-Sampler with step size ϵ from t distribution:
$$p(\eta_i | \mathbf{y}_i, \cdots) \sim t_k \left(\eta_i : \nu + p, (I + \Lambda \Sigma \Lambda)^{-1} \Lambda \Sigma \mathbf{y}_i, \frac{\nu + \mathbf{y}_i^T \mathbf{y}_i}{\nu + p} (I + \Lambda \Sigma \Lambda)^{-1} \right).$$

4. Sample γ_i, for $i = 1, \dots, n$ from gamma distributions:
$$p(\gamma_i | \eta_i, \mathbf{y}_i, \cdots) \sim \text{Ga} \left(\gamma_i : \frac{\nu + p + k}{2}, \frac{\nu + (\mathbf{y}_i - \Lambda \eta_i)^T \Sigma^{-1} (\mathbf{y}_i - \Lambda \eta_i) + \eta_i^T \eta_i}{2} \right).$$

5. Sample ϕ_{jh}, for $j = 1, \dots, p$, $h = 1, \cdots, k$ from gamma distributions:
$$p(\phi_{jh} | \cdots) \sim \text{Ga} \left(\phi_{jh} : \frac{\kappa + 1}{2}, \frac{\kappa + \tau_h \lambda_{jh}^2}{2} \right).$$

6. Sample δ_h, for $h = 1, \dots, k$ from gamma distributions:
$$p(\delta_1 | \cdots) \sim \text{Ga} \left(\delta_1 : a_1 + \frac{pk}{2}, 1 + \frac{\sum_{\ell=1}^k \sum_{j=1}^p \tau_\ell \phi_{j\ell} \lambda_{j\ell}^2}{2} \right),$$
$$p(\delta_h | \cdots) \sim \text{Ga} \left(\delta_h : a_2 + \frac{p(k - h + 1)}{2}, 1 + \frac{\sum_{\ell=h}^k \sum_{j=1}^p \tau_\ell \phi_{j\ell} \lambda_{j\ell}^2}{2} \right), \quad h \geq 2.$$

7. Sample a_1, a_2 by Metropolis-Hastings update, using Gaussian proposal with lower bound constraint of $a_1 > 2$ and $a_2 > 3$.

Chapter 4

Theoretical Properties

Bhattacharya & Dunson (2011) showed the weak consistency of the posterior density of their model. In this chapter, we show that the posterior density of the proposed model is weakly consistent, given that the degrees of freedom ν of the t-distribution is well-specified. All proofs for theorems can be found in Chapter 8.

For the sake of coherence, we follow the notation of Bhattacharya & Dunson (2011). Π_Λ and Π_Σ are prior distribution on Θ_Λ and Θ_Σ, respectively. Θ_Ω is a space of $p \times p$ positive semi-definite matrices, and an open ray $\Theta_\nu = (2, \infty)$ is a parameter space for the degrees of freedom ν. Let $g : \Theta_\Lambda \times \Theta_\Sigma \to \Theta_\Omega$ be a mapping which maps (Λ, Σ) to covariance matrix as follows:

$$g(\Lambda, \Sigma) = \Lambda\Lambda^T + \Sigma.$$

Let $\tilde{g} : \Theta_\nu \times \Theta_\Lambda \times \Theta_\Sigma \to \Theta_\nu \times \Theta_\Omega$ be a mapping such that:

$$\tilde{g}((\nu, \Lambda, \Sigma)) = (\nu, g(\Lambda, \Sigma)) = (\nu, \Lambda\Lambda^T + \Sigma).$$

The parameters of multivariate t likelihood are (ν, Ω). Then full prior distribuion Π on $\Theta_\nu \times \Theta_\Omega$ is $\Pi = (\Pi_\nu \otimes \Pi_\Lambda \otimes \Pi_\Sigma) \circ \tilde{g}^{-1}$ which is induced by Π_ν, Π_Λ, Π_Σ. If we prespecify the degrees of freedom ν, say $\nu = \tilde{\nu}$, then it is equivalent to choosing Π_ν as a Dirac probability measure at some point $\tilde{\nu}$.

Theorem 1 *Let*

$$B_{\varepsilon}^{\infty}((\nu_0, \Omega_0)) = \Big\{(\nu, \Omega) \in \Theta_{\nu} \times \Theta_{\Omega} \ \ : \ \ |\nu - \nu_0| < \varepsilon, \ \ d_{\infty}(\Omega, \Omega_0) < \varepsilon\Big\},$$

where $d_{\infty}(A, B) = \max_{1 \leq i,j \leq p} |a_{ij} - b_{ij}|$ *denotes a max-norm distance for two* $p \times p$ *matrices. If* $\nu_0 > 2$ *and* Ω_0 *is any* $p \times p$ *covariance matrix, then* $\Pi\{B_{\varepsilon}^{\infty}((\nu_0, \Omega_0))\} > 0$ *for any* $\varepsilon > 0$*.*

Theorem 2 *For fixed* ν_0 *and* Ω_0*, and for any* $\varepsilon > 0$*, there exists* $\varepsilon^* > 0$*, such that*

$$B_{\varepsilon}^{\infty}((\nu_0, \Omega_0)) \subset \Big\{(\nu, \Omega) \in \Theta_{\nu} \times \Theta_{\Omega} : \mathit{KL}\big((\nu_0, \Omega_0), (\nu, \Omega)\big) < \varepsilon\Big\},$$

where $\mathit{KL}((\nu_0, \Omega_0), (\nu, \Omega))$ *denotes the Kullback-Leibler divergence between two multivariate Student-t distribution,* $t(\nu_0, \mathbf{0}, \Omega_0)$ *and* $t(\nu, \mathbf{0}, \Omega)$*.*

Theorem 1 states that the support of prior Π is large enough so that arbitrarily small neighborhood of any $(\nu_0, \Omega_0) \in \Theta_{\nu} \times \Theta$ has strictly positive prior probability. Along with Theorem 1, Theorem 2 ensures that, Kullback-Leibler support condition is satisfied for any (ν, Ω) for the proposed prior Π. Thus if we prespecify t degrees of freedom correctly, i.e., if we choose $\Pi_{\nu} = \delta_{\nu_0}$ for true t degrees of freedom ν_0, the weak posterior consistency holds by Schwartz (1965).

Chapter 5

Simulation Study

In this chapter, we illustrate a simulation study of covariance estimation under high-dimensional data and compare its performance with the normal-likelihood factor model of Bhattacharya & Dunson (2011). We generated $\mathbf{y}_i, i = 1, \ldots, n$ from heavy-tailed multivariate t distribution with parameter ν_0 and $\Omega_0 = \Lambda_0 \Lambda_0^T + \Sigma_0$. The true covariance of synthetic data is then $\frac{\nu_0}{\nu_0 - 2}\Omega_0$. We let factor loading Λ_0 be sparse so that 70–80$\%$ of entries of Ω_0 are zero. The diagonal terms of error variance matrix Σ_0 is generated by the inverse gamma distribution of shape 1 and rate $1/4$. Code for estimating covariance using the proposed model is available on https://github.com/lee-jaejoon/robust-sparse-bayesian-infinite-factor-models.

The covariance estimation is conducted in two cases: when ν is well-specified and misspecified. In the well-specified case, the true degrees of freedom ν_0 and the prespecified degrees of freedom ν in the model were set as $\nu_0 = \nu = 3$. In misspecified case, the degrees of freedom was $\nu = 3$, while the true degrees of freedom was $\nu_0 = 7$. For each settings of (p, k), 10 repeated simulations were conducted. We ran 20,000 iterations of Markov Chain Monte Carlo as described in Section 3.3 with 5,000 burn-in steps. Learning rate ϵ for updating η is set at $\epsilon = 0.025, 0.015, 0.01$ for $(p, k) = (200, 10)$, $(500, 15)$, $(1000, 20)$, respectively. The adaptation probability in t th iteration $p(t)$ is chosen $p(t) = \exp(-1.2 - 0.0004t)$. In the adaptation step, we

Model		**Normal likelihood**					**Multivariate t likelihood**				
(p, k)		1-norm	2-norm	MSE	AAB	MAB	1-norm	2-norm	MSE	AAB	MAB
	mean	33.3858	9.1841	0.0098	0.0561	1.0011	30.0996	8.1040	0.0083	0.0530	0.8983
$p = 200$	min	29.3328	8.1465	0.0077	0.0496	0.9280	28.7672	8.0434	0.0071	0.0486	0.8428
$k = 10$	median	33.4496	9.0973	0.0099	0.0566	0.9881	29.8370	8.1069	0.0081	0.0526	0.9042
	max	36.4696	11.3066	0.0111	0.0593	1.0978	32.1881	8.1538	0.0093	0.0562	0.9511
	mean	89.6723	25.3974	0.0116	0.0677	1.0639	78.2845	23.2956	0.0100	0.0656	0.9032
$p = 500$	min	82.4762	23.7311	0.0101	0.0638	0.9714	74.6951	23.1118	0.0088	0.0609	0.8452
$k = 15$	median	85.1119	24.2223	0.0118	0.0684	1.0143	78.1479	23.2990	0.0098	0.0649	0.8963
	max	107.7094	31.6338	0.0127	0.0715	1.2807	82.4753	23.5229	0.0117	0.0714	0.9890
	mean	217.1357	46.8709	0.0142	0.0752	1.6856	205.3249	37.6516	0.0131	0.0755	1.4546
$p = 1000$	min	198.5997	38.9138	0.0129	0.0715	1.5203	200.2763	37.5020	0.0116	0.0716	1.3352
$k = 20$	median	217.1729	45.1398	0.0132	0.0735	1.6032	205.5342	37.6245	0.0134	0.0763	1.4816
	max	234.4568	57.5753	0.0164	0.0809	1.9884	211.1231	37.8733	0.0141	0.0786	1.5669

Table 5.1: The simulation result of the covariance estimation when the true degrees of freedom is $\nu_0 = 3$ and the model degrees of freedom is $\nu = 3$

deleted the factors 70% of whose loading entries are closer to 0 than 0.01. The proposal variances of Metropolis-Hastings update for a_1 and a_2 are tuned so that the acceptance rates be 50–70%. After sampling from the posterior distribution is done, the covariance estimate is obtained by averaging the posterior samples of covariance. The estimated covariance is then evaluated with the matrix 1-norm (maximum absolute column sum), the matrix 2-norm (maximum singular value), the mean squared error (MSE), the average absolute bias (AAB), and the maximum absolute bias (MAB). The simulation result for well-specified case and misspecified case are displayed in Table 5.1 and 5.2, respectively.

Table 5.1 shows the simulation results of the well-specified case where both true

Model		**Normal likelihood**					**Multivariate t likelihood**				
(p, k)		1-norm	2-norm	MSE	AAB	MAB	1-norm	2-norm	MSE	AAB	MAB
	mean	36.7584	10.9573	0.0105	0.0573	1.4114	30.9032	8.5313	0.0088	0.0544	1.3421
$p = 200$	min	30.5850	8.5565	0.0085	0.0511	1.2321	25.9585	7.1489	0.0078	0.0517	1.1556
$k = 10$	median	35.1613	10.6633	0.0100	0.0561	1.3626	31.4622	8.5725	0.0087	0.0546	1.3700
	max	57.0237	17.8385	0.0178	0.0766	1.6724	36.5888	10.1715	0.0104	0.0587	1.4878
	mean	91.8864	24.8418	0.0123	0.0709	1.1925	81.9975	24.4893	0.0090	0.0617	1.0973
$p = 500$	min	81.3002	24.5005	0.0102	0.0647	1.0277	79.2123	24.4243	0.0083	0.0587	0.9954
$k = 15$	median	90.7399	24.5695	0.0117	0.0699	1.1620	81.3976	24.4942	0.0088	0.0608	1.0718
	max	104.1514	25.6546	0.0150	0.0791	1.5117	86.9735	24.5355	0.0100	0.0663	1.2767
	mean	196.8423	44.7915	0.0113	0.0655	1.6796	193.8782	39.5985	0.0137	0.0769	1.3940
$p = 1000$	min	177.2834	39.5988	0.0108	0.0647	1.5423	188.0602	39.3843	0.0130	0.0746	1.3285
$k = 20$	median	200.2282	43.2991	0.0114	0.0652	1.6653	193.3312	39.4151	0.0133	0.0764	1.3789
	max	213.9246	55.1276	0.0118	0.0667	1.9318	200.3391	40.3553	0.0151	0.0809	1.4485

Table 5.2: The simulation result of the covariance estimation when the true degrees of freedom is $\nu_0 = 7$ and the model degrees of freedom is $\nu = 3$

and model degrees of freedoms are $\nu_0 = \nu = 3$ for covariance estimation. The proposed model performs better than the normal likelihood model in all cases. In $(p, k) = (1000, 20)$, MSE and AAB of normal likelihood model shows smaller value than that of the proposed t likelihood model. However, observing that maximum absolute bias of normal likelihood model is larger, we can presume that the scale of covariance entries is underestimated in normal likelihood model's case, which leads to biased estimation. Though the estimation performance was slightly poor, the normal-likelihood factor model estimated the number of factors in a stable manner, even with the data from heavy-tailed distribution. Mean elapsed times for the proposed model are 4.18, 15.51, 46.19 minutes, which are about 1.52, 1.47, 1.50 times longer than those of the normal model in $(p, k) = (200, 10)$, $(500, 15)$, $(1000, 20)$, respectively. As we set up sparse true covariance matrix 70 to 80% of whose entries are zero, we can monitor and compare the covariance estimates of the two model for those strictly zero covariance entries. For covariance entries whose true values are zero, 10th and 90th percentile of estimated covariance entries from the proposed model are $(-0.0608, 0.0816)$, $(-0.0873, 0.1010)$, $(-0.0992, 0.1097)$ on average, while the normal model showed $(-0.0651, 0.0813)$, $(-0.0835, 0.0937)$, $(-0.0950, 0.1003)$ in $(p, k) = (200, 10)$, $(500, 15)$, $(1000, 20)$, respectively. This demonstrates that the proposed model and the normal model have similar shrinkage for the true zero entries.

Table 5.2 shows the simulation results of the misspecified case where model degrees of freedom is $\nu = 3$ while true degrees of freedom is $\nu_0 = 7$. Even when the degrees of freedom is misspecified, we can see that using the proposed model with small enough degrees of freedom yields better covariance estimation performance than the normal model. Likewise, we can observe the same possible bias in the estimate of normal likelihood model when $(p, k) = (1000, 20)$ as in Table 5.1. Also the proposed model does not lose the capability of estimating the number of latent factors under misspecification of the degrees of freedom. Mean elapsed times for the proposed model are 4.43, 17.19, 48.86 minutes, which are about 1.60, 1.59, 1.58 times longer than those of the normal model in $(p,k) = (200,10)$, $(500,15)$, $(1000,20)$, respectively. From the proposed model, the 10th and 90th percentile of estimated covariance entries whose true values are zero are $(-0.0612, 0.0892)$, $(-0.0744, 0.0890)$, $(-0.1023, 0.1133)$ on average, while the normal model showed $(-0.0640, 0.0858)$, $(-0.0882, 0.1023)$, $(-0.0791, 0.0860)$ in $(p,k) = (200,10)$, $(500,15)$, $(1000,20)$, respectively. This implies that, even under misspecified degrees of freedom, the proposed model still shows similar shrinkage for true zero entries compared to normal model.

Chapter 6

Real Data Analysis : T1T2 Node-Negative Breast Cancer Application

6.1 Background and Previous Researches

Carcinoma is a type of cancer that develops from epithelial cells. Invasive ductal carcinoma is a type of breast carcinoma which begins growing in a milk duct and invades adjacent tissue of the breast. It is the most common type of breast cancer, accounting for 80% of all breast cancer diagnoses. Cancer cells are developed by accumulations of multiple DNA mutations that are not repaired by their own repair mechanisms. Gravier et al. (2010) analyzed the DNA signature of tumor cells from 168 patients with small invasive ductal carcinomas without axillary lymph node involvement (T1T2N0) to predict metastasic progression in 5 years after diagnosis.

Gene expression of each patient's tumor cell was obtained by array comparative genomic hybridization(aCGH). aCGH is a technique to detect the change in chromosomal copy number. DNAs of tumor cell and normal cell are labelled with green and red fluorescent protein, respectively. The DNAs are then mixed and undergone hybridization: the process of single stranded DNA binding to its complementary DNA strand. Next, green-to-red ratio is measured by fluorescent microscopy, which repre-

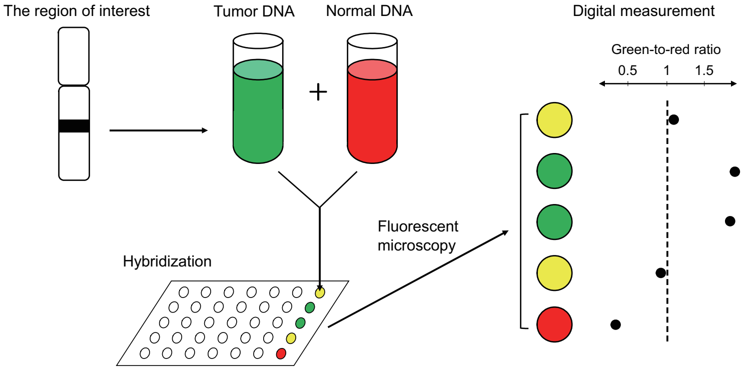

Figure 6.1: The overview of array comparative genomic hybridization (aCGH) procedure

sents the chromosomal gain or loss of tumor DNA in the region of interest. The overall procedure of data acquisition through aCGH is illustrated in Fig. 6.1.

The training set contained 2,905 predictor variables (log2 transformed) representing genomic signatures of chromosome 2p22.2, 3p23, and 8q21-24. Among 168 patients, 111 patients did not have any metastatic event in 5 years after initial diagnosis, while early metastasis of breast carcinoma was reported in other 57 patients. The dataset analysed during the current study is available in the Gene Expression Omnibus (GEO) repository database with accession number GSE19159, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19159. In order to predict the progression of metastasis, Gravier et al. (2010) combined the outcome of multiple classifiers each of which are based on logistic regression.

The latent factor regression is an efficient method under high-dimensional setting of $p \gg n$, where joint covariance structure of continuous dependent variable z_i and predictor variable $\mathbf{x}_i$ is estimated by performing factor model on $\mathbf{y}_i = (z_i, \mathbf{x}_i^T)^T$. The predictive distribution for z_{new} can be obtained as follows:

$$p(z_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{y}_1, \ldots, \mathbf{y}_n) = \int p(z_{\text{new}}|\mathbf{x}_{\text{new}}, \Omega)p(\Omega|y_1, \ldots, \mathbf{y}_n)d\Omega,$$

Under joint normality assumption, the conditional distribution of z given $\mathbf{x}, \Omega$ is as follows:

$$z|\mathbf{x}, \Omega \sim \mathcal{N}(\mathbf{x}^T \Omega_{xx}^{-1} \Omega_{xz}, \Omega_{zz} - \Omega_{zx} \Omega_{xx}^{-1} \Omega_{xz}).$$

Here $\beta = \Omega_{xx}^{-1} \Omega_{xz}$ can be considered as a regression coefficient in latent factor regression. There have been a few approaches to analyze high-dimensional microarray data with latent factor regression (Carvalho et al., 2008; Bhattacharya & Dunson, 2011). Bhattacharya & Dunson (2011) implemented their shrinkage prior to induce shrinkage in regression coefficient estimate. Then the feature selection is performed by sorting predictor variables by absolute value of estimated regression coefficient.

6.2 Model and Results

Our goal is to predict the progression of metastasis given DNA signature data of cancer cells (0: no metastasis, 1: metastasis). By investigating Q-Q plot of each variable, we have observed a heavy-tailed structure of the data. Though the proposed factor model is an efficient tool to estimate the low-dimensional structure of heavy-tailed high-dimensional data, we cannot implement the latent factor regression method with the proposed model because the dependent variable is not continuous but binary.

Instead, we implemented discriminant analysis using the covariance estimate obtained by the proposed model. We divide the data set into the training set (118 of 168 patients) and the test set (50 of 168 patients). Covariance estimates for patients with metastasis (36 of 118 patients) and without metastasis (82 of 118 patients) are obtained separately from the training set. We ran Markov chain Monte Carlo algorithm as in Section 3.3 for posterior computation for $20,000$ iterations with $5,000$ burn-in steps. The degrees of freedom of t likelihood is set to $\nu = 5$. The step size ϵ for Hamiltonian Monte Carlo update for $\eta_i, i = 1, \dots, n$ is set to $\epsilon = 0.2$. The estimated number of factors for patients with metastasis and without metastasis are 50 and 70 with 95% credible interval $(48, 52)$ and $(65, 71)$, respectively.

After estimating covariance for patients with metastasis and without metastasis, we calculated the log likelihood ratio of observations in test set as follows:

$$\log\left(\frac{t(\mathbf{y};\nu,\hat{\mu}_1,\hat{\Sigma}_1)}{t(\mathbf{y};\nu,\hat{\mu}_0,\hat{\Sigma}_0)}\right)=\log\left(\frac{\left|\hat{\Sigma}_1\right|^{-\frac{1}{2}}\left\{\nu+(\mathbf{y}-\hat{\mu}_1)^T\hat{\Sigma}_1(\mathbf{y}-\hat{\mu}_1)\right\}^{-\frac{\nu+p}{2}}}{\left|\hat{\Sigma}_0\right|^{-\frac{1}{2}}\left\{\nu+(\mathbf{y}-\hat{\mu}_0)^T\hat{\Sigma}_0(\mathbf{y}-\hat{\mu}_0)\right\}^{-\frac{\nu+p}{2}}}\right),$$

where $\hat{\mu}_1, \hat{\mu}_0$ are training sample mean of patients with metastasis and without metastasis, respectively. The covariance estimates of patients with metastasis and without metastasis obtained by the proposed model are denoted as $\hat{\Sigma}_1$ and $\hat{\Sigma}_0$, respectively. If the log likelihood ratio is greater than a threshold ξ, we classified the observation as a patient with metastasis. We determined the value of threshold $\xi = 0$ in our case. Sensitivity is the proportion of true positives which are correctly identified by classifier, while specificity is the proportion of true negatives which are correctly identified by classifier. Both are measures of classification performance widely used in medicine. The test accuracy was 86% which outperforms the classfier suggested by Gravier et al. (2010). The classfier of Gravier et al. (2010) showed test accuracy of 78%. Test sensitivity of 66.7% and test specificity of 90.2% are observed, while Gravier et al. (2010) showed 84% and 66%, respectively.

Chapter 7

Discussion

In this paper, we have proposed a Bayesian infinite factor model with multiplicative gamma process shrinkage prior for robust covariance estimation under heavy-tailed high-dimensional data. Also we have shown the fact that, under well-specified degrees of freedom of t distribution, the posterior density from the proposed model is weakly consistent.

There are a few research directions which are worthy of further study. Kleijn et al. (2006) and Ramamoorthi et al. (2015) have studied posterior consistency under model misspecification. In the same spirit, theoretical properties of the proposed model under misspecification of the degrees of freedom can be potential avenues of exploration. Murphy et al. (2020) has introduced *the infinite mixture of infinite factor analysers* (IMIFA) model, which is a Pitman-Yor mixture of the model of Bhattacharya & Dunson (2011). The same extension of the proposed model from normal likelihood to Student's t-likelihood can also be made when some or all of the mixture components are suspected to follow heavy-tailed distribution. Finally, the proposed model is not completely choice-free, due to step size parameter ϵ used in No-U-Turn sampler update for η. Hoffman & Gelman (2014) suggested a method of adaptive setting for the value of ϵ. This, however, is not directly applicable in our settings, because we are using a single iteration of No-U-Turn sampler whose target function changes as estimates of the other parameters change. Devising a method of tuning ϵ would be an improvement on our work.

Chapter 8

Appendix

8.1 Proof of Theorem 1

Let $\varepsilon > 0$ be fixed, and let

$$B_\varepsilon\big((\nu_0, \Lambda_0, \Sigma_0)\big) = \Big\{(\nu, \Lambda, \Sigma) : |\nu - \nu_0| < \varepsilon, d_2(\Lambda, \Lambda_0) < \varepsilon, d_\infty(\Sigma, \Sigma_0) < \varepsilon\Big\}.$$

By Lemma 2 of Bhattacharya & Dunson (2011), there exists $\varepsilon_1 > 0$ such that

$$\begin{aligned}\tilde{g}\Big(B_{\varepsilon_1}\big((\nu_0, \Lambda_0, \Sigma_0)\big)\Big) &\subset B_\varepsilon^\infty\Big(\tilde{g}\big((\nu_0, \Lambda_0, \Sigma_0)\big)\Big)\\ &= B_\varepsilon^\infty\big((\nu_0, g(\Lambda_0, \Sigma_0))\big) = B_\varepsilon^\infty\big((\nu_0, \Omega_0)\big).\end{aligned}$$

Thus, we have

$$B_{\varepsilon_1}\big((\nu_0, \Lambda_0, \Sigma_0)\big) \subset \tilde{g}^{-1}\Big(B_\varepsilon^\infty\big((\nu_0, \Omega_0)\big)\Big).$$

Denoting the prior distribution as $\Pi = \Pi_\nu \otimes \Pi_\Omega = (\Pi_\nu \otimes \Pi_\Lambda \otimes \Pi_\Sigma) \circ \tilde{g}^{-1}$, we have

$$\begin{aligned}(\Pi_\nu \otimes \Pi_\Lambda \otimes \Pi_\Sigma)\Big\{B_{\varepsilon_1}\big((\nu_0, \Lambda_0, \Sigma_0)\big)\Big\} &\leq (\Pi_\nu \otimes \Pi_\Lambda \otimes \Pi_\Sigma)\Big\{\tilde{g}^{-1}\Big(B_\varepsilon^\infty\big((\nu_0, \Omega_0)\big)\Big)\Big\}\\ &= \Pi\Big(B_\varepsilon^\infty\big((\nu_0, \Omega_0)\big)\Big).\end{aligned}$$

Thus, if

$$\begin{aligned}
&\Pi_\nu\Big(\big\{\nu \in \Theta_\nu : |\nu - \nu_0| < \varepsilon_1\big\}\Big) > 0\\
&\Pi_\Lambda\Big(\big\{\Lambda \in \Theta_\Lambda : d_2(\Lambda, \Lambda_0) < \varepsilon_1\big\}\Big) > 0\\
&\Pi_\Sigma\Big(\big\{\Sigma \in \Theta_\Sigma : d_\infty(\Sigma, \Sigma_0) < \varepsilon_1\big\}\Big) > 0,
\end{aligned}$$

we obtain the conclusion.

Since the support of Π_ν and Π_Σ are Θ_ν and Θ_Σ, respectively, the inequalities for ν and Σ hold. For the inequality of the Λ, we can apply the proof of Proposition 2 of Bhattacharya & Dunson (2011).

8.2 Proof of Theorem 2

Let $\nu_0 > 2, \Omega_0 \in \Theta_\Omega$ be true parameter. We wish to show that, for any $\varepsilon > 0$, we can choose $\varepsilon^* > 0$ such that

$$\mathrm{KL}\Big((\nu_0, \Omega_0), (\nu, \Omega)\Big) < \varepsilon, \ \text{ for all } |\nu_0 - \nu| < \varepsilon^* \text{ and } d_\infty(\Omega_0, \Omega) < \varepsilon^*. \tag{3}$$

Let $\varepsilon > 0$ be given. By the definition of Kullback-Leibler divergence, we have

$$\begin{aligned}
&\mathrm{KL}\Big((\nu_0, \Omega_0), (\nu, \Omega)\Big)\\
&= \int \log \frac{t(\mathbf{y}; \nu_0, \Omega_0)}{t(\mathbf{y}; \nu, \Omega)} t(\mathbf{y}; \nu_0, \Omega_0) d\mathbf{y}\\
&= \mathbb{E}_{(\nu_0,\Omega_0)}\left[\log\left(\frac{\frac{\Gamma[(\nu_0+p)/2]}{\Gamma(\nu_0/2)(\nu_0\pi)^{p/2}\det(\Omega_0)^{1/2}}\left[1 + \frac{\mathbf{y}^T\Omega_0^{-1}\mathbf{y}}{\nu_0}\right]^{-(\nu_0+p)/2}}{\frac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)(\nu\pi)^{p/2}\det(\Omega)^{1/2}}\left[1 + \frac{\mathbf{y}^T\Omega^{-1}\mathbf{y}}{\nu}\right]^{-(\nu+p)/2}}\right)\right]\\
&= \log\left(\frac{\frac{\Gamma[(\nu_0+p)/2]}{\Gamma(\nu_0/2)}\nu_0^{\nu_0/2}}{\frac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)}\nu^{\nu/2}}\right) + \frac{1}{2}\log\left(\frac{\det(\Omega)}{\det(\Omega_0)}\right) + \mathbb{E}_{(\nu_0,\Omega_0)}\left[\log\frac{\left[\nu_0 + \mathbf{y}^T\Omega_0^{-1}\mathbf{y}\right]^{-(\nu_0+p)/2}}{\left[\nu + \mathbf{y}^T\Omega^{-1}\mathbf{y}\right]^{-(\nu+p)/2}}\right]\\
&\leq \left|\log\left(\frac{\Gamma[(\nu_0+p)/2]}{\Gamma(\nu_0/2)}\nu_0^{\nu_0/2}\right) - \log\left(\frac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)}\nu^{\nu/2}\right)\right|\\
&\quad + \left|\frac{1}{2}\log\left(\det(\Omega)\right) - \frac{1}{2}\log\left(\det(\Omega_0)\right)\right| + \left|\mathbb{E}_{(\nu_0,\Omega_0)}\left[\log\frac{\left[\nu_0 + \mathbf{y}^T\Omega_0^{-1}\mathbf{y}\right]^{-\frac{\nu_0+p}{2}}}{\left[\nu + \mathbf{y}^T\Omega^{-1}\mathbf{y}\right]^{-\frac{\nu+p}{2}}}\right]\right|.
\end{aligned} \tag{4}$$

By continuity of the functions in the equation 4, we can choose $\varepsilon_1^*, \varepsilon_2^* > 0$ that bounds the first and second terms of equation 4 with $\varepsilon/3$, respectively. By the triangle inequality, the third term of equation 4 is

$$\begin{aligned}
&\left|\mathbb{E}_{(\nu_0,\Omega_0)}\left[\log\frac{\left[\nu_0+\mathbf{y}^T\Omega_0^{-1}\mathbf{y}\right]^{-(\nu_0+p)/2}}{\left[\nu+\mathbf{y}^T\Omega^{-1}\mathbf{y}\right]^{-(\nu+p)/2}}\right]\right| \\
&\leq \left|\frac{\nu+p}{2}\mathbb{E}\log\left[\nu+\mathbf{y}^T\Omega^{-1}\mathbf{y}\right]-\frac{\nu_0+p}{2}\mathbb{E}\log\left[\nu+\mathbf{y}^T\Omega^{-1}\mathbf{y}\right]\right| \\
&\quad+\left|\frac{\nu_0+p}{2}\mathbb{E}\log\left[\nu+\mathbf{y}^T\Omega^{-1}\mathbf{y}\right]-\frac{\nu_0+p}{2}\mathbb{E}\log\left[\nu_0+\mathbf{y}^T\Omega_0^{-1}\mathbf{y}\right]\right| \qquad (5) \\
&= A+B.
\end{aligned}$$

Denote the first and second terms of equation 5 as A and B, respectively. For A, we have

$$\begin{aligned}
A &= \frac{|\nu-\nu_0|}{2}\left|\mathbb{E}\log\left[\nu+\mathbf{y}^T\Omega^{-1}\mathbf{y}\right]\right| \\
&\leq \frac{|\nu-\nu_0|}{2}\mathbb{E}\Big[\left|\log\left[\nu+\mathbf{y}^T\Omega^{-1}\mathbf{y}\right]\right|\Big] \\
&= \frac{|\nu-\nu_0|}{2}\mathbb{E}\Big[\log\left[\nu+\mathbf{y}^T\Omega^{-1}\mathbf{y}\right]\Big] \\
&\leq \frac{|\nu-\nu_0|}{2}\mathbb{E}\Big[\nu-1+\mathbf{y}^T\Omega^{-1}\mathbf{y}\Big].
\end{aligned}$$

Using the fact that the expectation of quadratic form of $\mathbf{y} \sim t(\nu_0, \mathbf{0}, \Omega_0)$ is $\mathbb{E}[\mathbf{y}^T\Omega^{-1}\mathbf{y}] = \frac{\nu_0}{\nu_0-2}\mathrm{tr}(\Omega^{-1}\Omega_0)$, we have

$$\begin{aligned}
A &= \frac{|\nu-\nu_0|}{2}\Big[\nu-1+\frac{\nu_0}{\nu_0-2}\mathrm{tr}(\Omega^{-1}\Omega_0)\Big] \\
&= \frac{|\nu-\nu_0|}{2}\left[\nu-1+\frac{\nu_0}{\nu_0-2}\sum_{j=1}^{p}\lambda_j(\Omega^{-1}\Omega_0)\right] \\
&\leq \frac{|\nu-\nu_0|}{2}\Big[|\nu-\nu_0|+\nu_0-1+\frac{\nu_0}{\nu_0-2}p\lambda_{\max}(\Omega^{-1}\Omega_0)\Big].
\end{aligned}$$

Let $\lambda_{\max}(\Omega^{-1}\Omega_0)$ be the largest eigenvalue of $\Omega^{-1}\Omega_0$. For an eigenvector $v \in \mathbb{R}^p$ corresponding to $\lambda_{\max}(\Omega^{-1}\Omega_0)$ and sufficiently large $M_1 > 0$, the following holds:

$$\begin{aligned}
\lambda_{\max}(\Omega^{-1}\Omega_0) &= \|\lambda_{\max}(\Omega^{-1}\Omega_0)v\|_2 \\
&\le p^{1/2}\|\lambda_{\max}(\Omega^{-1}\Omega_0)v\|_\infty \\
&= p^{1/2}\|\Omega^{-1}\Omega_0 v\|_\infty \\
&\le p^{1/2}\|\Omega^{-1}\|_\infty\|\Omega_0\|_\infty\|v\|_\infty \\
&\le p^{1/2}(\|\Omega^{-1}-\Omega_0^{-1}\|_\infty + \|\Omega_0^{-1}\|_\infty)\|\Omega_0\|_\infty\|v\|_\infty \\
&\le p^{1/2}(\|\Omega^{-1}-\Omega_0^{-1}\|_\infty + M_1)M_1\|v\|_\infty \\
&\le p^{1/2}(\|\Omega^{-1}-\Omega_0^{-1}\|_\infty + M_1)M_1 \\
&= p^{1/2}(\|\Omega^{-1}-\Omega_0^{-1}\|_\infty + M_1)M_1.
\end{aligned}$$

With this upper bound of $\lambda_{\max}(\Omega^{-1}\Omega_0)$, we have

$$\begin{aligned}
A &\le \frac{|\nu-\nu_0|}{2}\Big[|\nu-\nu_0| + \nu_0 - 1 + \frac{\nu_0}{\nu_0-2}p\lambda_{\max}(\Omega^{-1}\Omega_0)\Big] \\
&\le \frac{|\nu-\nu_0|}{2}\Big[|\nu-\nu_0| + \nu_0 - 1 + \frac{\nu_0}{\nu_0-2}p^{3/2}(\|\Omega^{-1}-\Omega_0^{-1}\|_\infty + M_1)M_1\Big].
\end{aligned}$$

By continuity of matrix inversion, we can choose $\tilde{\varepsilon} > 0$ such that $\|\Omega - \Omega_0\|_\infty < \tilde{\varepsilon}$ implies $\|\Omega^{-1} - \Omega_0^{-1}\|_\infty < 1$. Plus we can choose $\varepsilon_3^* \in (0, \tilde{\varepsilon})$ small enough so that A is bounded above by $\varepsilon/6$. So we have

$$\begin{aligned}
A &< \frac{|\nu-\nu_0|}{2}\Big[|\nu-\nu_0| + \nu_0 - 1 + \frac{\nu_0}{\nu_0-2}p^{3/2}(1 + M_1)M_1\Big] \\
&< \frac{\varepsilon_3^*}{2}\Big[\varepsilon_3^* + \nu_0 - 1 + \frac{\nu_0}{\nu_0-2}p^{3/2}(1 + M_1)M_1\Big] \\
&< \frac{\varepsilon}{6}.
\end{aligned}$$

For B, by Jensen's inequality, we have

$$\begin{aligned}
B &= \left|\frac{\nu_0+p}{2}\mathbb{E}\log\left[\nu + \mathbf{y}^T\Omega^{-1}\mathbf{y}\right] - \frac{\nu_0+p}{2}\mathbb{E}\log\left[\nu_0 + \mathbf{y}^T\Omega_0^{-1}\mathbf{y}\right]\right| \\
&= \frac{\nu_0+p}{2}\left|\mathbb{E}\log\left[\frac{\nu + \mathbf{y}^T\Omega^{-1}\mathbf{y}}{\nu_0 + \mathbf{y}^T\Omega_0^{-1}\mathbf{y}}\right]\right| \\
&\le \frac{\nu_0+p}{2}\mathbb{E}\left|\log\left[\frac{\nu + \mathbf{y}^T\Omega^{-1}\mathbf{y}}{\nu_0 + \mathbf{y}^T\Omega_0^{-1}\mathbf{y}}\right]\right|. \qquad (6)
\end{aligned}$$

For a fixed unit vector $\omega \in \mathbb{R}^p$, let $g_\omega(t)$ be a function defined on $t > 0$ as follows:

$$\begin{aligned} g_\omega(t) &= \left.\frac{\nu + \mathbf{y}^T\Omega^{-1}\mathbf{y}}{\nu_0 + \mathbf{y}^T\Omega_0^{-1}\mathbf{y}}\right|_{\mathbf{y}=t\omega} \\ &= \frac{\nu + t^2\omega^T\Omega^{-1}\omega}{\nu_0 + t^2\omega^T\Omega_0^{-1}\omega}. \end{aligned}$$

Investigating critical points and limits of $t > 0$, we have the following bound of $g_\omega(t)$,

$$\frac{\omega^T\Omega^{-1}\omega}{\omega^T\Omega_0^{-1}\omega} \wedge \frac{\nu}{\nu_0} \le g_\omega(t) \le \frac{\omega^T\Omega^{-1}\omega}{\omega^T\Omega_0^{-1}\omega} \vee \frac{\nu}{\nu_0}. \tag{7}$$

Equation 7 holds for any unit vector $\omega \in \mathbb{R}^p$. Thus by taking infimum and supremum on lower and upper bounds, respectively, we have

$$\left[\left(\inf_{\|\omega\|=1} \frac{\omega^T\Omega^{-1}\omega}{\omega^T\Omega_0^{-1}\omega}\right) \wedge \frac{\nu}{\nu_0}\right] \le g_\omega(t) \le \left[\left(\sup_{\|\omega\|=1} \frac{\omega^T\Omega^{-1}\omega}{\omega^T\Omega_0^{-1}\omega}\right) \vee \frac{\nu}{\nu_0}\right]. \tag{8}$$

For $\tilde{\omega} = \Omega_0^{-1/2}\omega/\|\Omega_0^{-1/2}\omega\|_2$, we yield the following inequality of $\frac{\omega^T\Omega^{-1}\omega}{\omega^T\Omega_0^{-1}\omega}$,

$$\lambda_{\min}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}) \le \frac{\omega^T\Omega^{-1}\omega}{\omega^T\Omega_0^{-1}\omega} \le \lambda_{\max}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}), \tag{9}$$

which is obtained by following result,

$$\begin{aligned} \frac{\omega^T\Omega^{-1}\omega}{\omega^T\Omega_0^{-1}\omega} &= \frac{\omega^T\Omega_0^{-1/2}\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}\Omega_0^{-1/2}\omega}{\omega^T\Omega_0^{-1/2}\Omega_0^{-1/2}\omega} \\ &= \frac{\tilde{\omega}^T\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}\tilde{\omega}}{\tilde{\omega}^T\tilde{\omega}}. \end{aligned}$$

Here $\lambda_{\min}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2})$ and $\lambda_{\max}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2})$ are the smallest and the largest eigenvalues of $\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}$, respectively. By equation 8 and equation 9, $\log g_\omega(t)$ is bounded as follows:

$$\left[\lambda_{\min}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}) \wedge \frac{\nu}{\nu_0}\right] \le g_\omega(t) \le \left[\lambda_{\max}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}) \vee \frac{\nu}{\nu_0},\right]$$

$$\log\left[\lambda_{\min}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}) \wedge \frac{\nu}{\nu_0}\right] \le \log g_\omega(t) \le \log\left[\lambda_{\max}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}) \vee \frac{\nu}{\nu_0}\right]. \tag{10}$$

Note that equation 10 holds for any $\omega \in \mathbb{R}^p, \|\omega\|_2 = 1$. For any $\mathbf{y} \in \mathbb{R}^p$, $\mathbf{y}$ can be written as $\mathbf{y} = \|\mathbf{y}\| \frac{\mathbf{y}}{\|\mathbf{y}\|} = t\omega,\ t \overset{let}{=} \|\mathbf{y}\|,\ \omega \overset{let}{=} \frac{\mathbf{y}}{\|\mathbf{y}\|}$. Thus we have the upper bound of the integrand of equation 6 as follows:

$$\begin{aligned}
&\left|\log\left[\frac{\nu + \mathbf{y}^T\Omega^{-1}\mathbf{y}}{\nu_0 + \mathbf{y}^T\Omega_0^{-1}\mathbf{y}}\right]\right| \\
&\leq \max\left\{\left|\log\left[\lambda_{\min}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}) \wedge \frac{\nu}{\nu_0}\right]\right|, \left|\log\left[\lambda_{\max}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}) \vee \frac{\nu}{\nu_0}\right]\right|\right\}.
\end{aligned}$$

Here we use the following limiting property of eigenvalue as $\Omega \to \Omega_0$ in max-norm sense:

$$\lambda_{\min}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}) \to 1,\ \ \lambda_{\max}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}) \to 1,\ \ \nu/\nu_0 \to 1.$$

So we can choose sufficiently small $\varepsilon_4^* > 0$ such that the following inequalities hold for all $d(\Omega, \Omega_0) < \varepsilon_4^*$,

$$\begin{aligned}
B &\leq \frac{\nu_0 + p}{2}\mathbb{E}\left|\log\left[\frac{\nu + \mathbf{y}^T\Omega^{-1}\mathbf{y}}{\nu_0 + \mathbf{y}^T\Omega_0^{-1}\mathbf{y}}\right]\right| \\
&\leq \frac{\nu_0 + p}{2}\max\left\{\mathbb{E}\left|\log\left[\lambda_{\min}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}) \wedge \frac{\nu}{\nu_0}\right]\right|, \mathbb{E}\left|\log\left[\lambda_{\max}(\Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2}) \vee \frac{\nu}{\nu_0}\right]\right|\right\} \\
&< \frac{\varepsilon}{6}.
\end{aligned}$$

Therefore, letting $\varepsilon^* = \min\{\varepsilon_1^*, \varepsilon_2^*, \varepsilon_3^*, \varepsilon_4^*\}$, $|\nu_0 - \nu| < \varepsilon^*$ and $d_\infty(\Omega_0, \Omega) < \varepsilon^*$ imply the following.

$$
\begin{aligned}
&\mathrm{KL}\Big((\nu_0, \Omega_0), (\nu, \Omega)\Big) \\
&\leq \left|\log\left(\frac{\Gamma[(\nu_0+p)/2]}{\Gamma(\nu_0/2)}\nu_0^{\nu_0/2}\right) - \log\left(\frac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)}\nu^{\nu/2}\right)\right| \\
&\quad + \left|\frac{1}{2}\log\left(\det(\Omega)\right) - \frac{1}{2}\log\left(\det(\Omega_0)\right)\right| \\
&\quad + \left|\mathbb{E}_{(\nu_0,\Omega_0)}\left[\log\frac{\left[\nu_0 + \mathbf{y}^T\Omega_0^{-1}\mathbf{y}\right]^{-(\nu_0+p)/2}}{\left[\nu + \mathbf{y}^T\Omega^{-1}\mathbf{y}\right]^{-(\nu+p)/2}}\right]\right| \\
&\leq \left|\log\left(\frac{\Gamma[(\nu_0+p)/2]}{\Gamma(\nu_0/2)}\nu_0^{\nu_0/2}\right) - \log\left(\frac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)}\nu^{\nu/2}\right)\right| \\
&\quad + \left|\frac{1}{2}\log\left(\det(\Omega)\right) - \frac{1}{2}\log\left(\det(\Omega_0)\right)\right| \\
&\quad + \left|\frac{\nu+p}{2}\mathbb{E}\log\left[\nu + \mathbf{y}^T\Omega^{-1}\mathbf{y}\right] - \frac{\nu_0+p}{2}\mathbb{E}\log\left[\nu + \mathbf{y}^T\Omega^{-1}\mathbf{y}\right]\right| \\
&\quad + \left|\frac{\nu_0+p}{2}\mathbb{E}\log\left[\nu + \mathbf{y}^T\Omega^{-1}\mathbf{y}\right] - \frac{\nu_0+p}{2}\mathbb{E}\log\left[\nu_0 + \mathbf{y}^T\Omega_0^{-1}\mathbf{y}\right]\right| \\
&< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{6} + \frac{\varepsilon}{6} \\
&= \varepsilon
\end{aligned}
$$

In other words, for any $\varepsilon > 0$, we can choose $\varepsilon^* > 0$ such that

$$\mathrm{KL}\Big((\nu_0, \Omega_0), (\nu, \Omega)\Big) < \varepsilon, \text{ for all } |\nu_0 - \nu| < \varepsilon^* \text{ and } d_\infty(\Omega_0, \Omega) < \varepsilon^*.$$

Thus equation 3 is proved and we have

$$\Big\{(\nu, \Omega) : |\nu_0 - \nu| < \varepsilon^* \text{ and } d_\infty(\Omega_0, \Omega) < \varepsilon^*\Big\} \subset \Big\{(\nu, \Omega) : \mathrm{KL}\Big((\nu_0, \Omega_0), (\nu, \Omega)\Big) < \varepsilon\Big\}.$$

The proof of Theorem 2 is done.

Bibliography

Ando, T. Bayesian factor analysis with fat-tailed factors and its exact marginal likelihood. *Journal of Multivariate Analysis*, 100(8):1717–1726, 2009.

Bhattacharya, A. and Dunson, D. B. Sparse bayesian infinite factor models. *Biometrika*, pp. 291–306, 2011.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512): 1479–1490, 2015.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

Durante, D. A note on the multiplicative gamma process. *Statistics & Probability Letters*, 122:198–204, 2017.

Ferrari, F. and Dunson, D. B. Bayesian factor analysis for inference on interactions. *Journal of the American Statistical Association*, pp. 1–12, 2020.

Geweke, J. and Zhou, G. Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2):557–587, 1996.

Ghosh, J. and Dunson, D. B. Default prior distributions and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320, 2009.

Gravier, E., Pierron, G., Vincent-Salomon, A., Gruel, N., Raynal, V., Savignoni, A., De Rycke, Y., Pierga, J.-Y., Lucchesi, C., Reyal, F., et al. A prognostic dna signature for t1t2 node-negative breast cancer patients. *Genes, chromosomes and cancer*, 49 (12):1125–1134, 2010.

Green, P. J. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Hoffman, M. D. and Gelman, A. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Kleijn, B. J., van der Vaart, A. W., et al. Misspecification in infinite-dimensional bayesian statistics. *Ann. Stat.*, 34(2):837–877, 2006.

Liu, J. S. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427): 958–966, 1994.

Lopes, H. F. and West, M. Bayesian model assessment in factor analysis. *Statistica Sinica*, pp. 41–67, 2004.

Lopes, H. F., Salazar, E., Gamerman, D., et al. Spatial dynamic factor analysis. *Bayesian Analysis*, 3(4):759–792, 2008.

McParland, D., Gormley, I. C., McCormick, T. H., Clark, S. J., Kabudula, C. W., and Collinson, M. A. Clustering south african households based on their asset status using latent variable models. *The annals of applied statistics*, 8(2):747, 2014.

Murphy, K., Viroli, C., Gormley, I. C., et al. Infinite mixtures of infinite factor analysers. *Bayesian Analysis*, 15(3):937–963, 2020.

Polson, N. G. and Scott, J. G. Shrink globally, act locally: Sparse bayesian regularization and prediction. In Bernardo, J. M., Bayarri, M., Berger, J. O., Dawid, A., Heckerman, D., Smith, A. F., and West, M. (eds.), *Bayesian Statistics 9*. New York: Oxford University Press, 2010.

Ramamoorthi, R., Sriram, K., Martin, R., et al. On posterior concentration in misspecified models. *Bayesian Analysis*, 10(4):759–789, 2015.

Roberts, G. O. and Rosenthal, J. S. Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of applied probability*, 44(2):458–475, 2007.

Schwartz, L. On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26, 1965.

West, M. Bayesian factor regression models in the “large p, small n” paradigm. In Bernardo, J. M., Dawid, A. P., Berger, J. O., West, M., Heckerman, D., Bayarri, M., and Smith, A. F. (eds.), *Bayesian Statistics 7*. New York: Oxford University Press, 2003.

Zhang, J., Li, J., and Liu, C. Robust factor analysis using the multivariate t-distribution. *Statistica Sinica*, 24(1):291–312, 2014.

초 록

베이즈 인자 모형에 대한 대부분의 선행 연구는 자료가 따르는 분포가 정규분포임을 가정한다. 이 연구는 다변수 t 가능도를 사용함으로써, 이상치가 존재하는 고차원 자료에 대해 더 개선된 공분산 추정 성능을 갖는 베이즈 인자 모형을 제시한다. 잠재인자의 수를 결정하기 위해서 본 모형은, 무한히 많은 잠재인자에 대해 수축사전분포를 부여하고 이를 동적으로 절단해나가는 Bhattacharya와 Dunson [*Biometrika* (2011) 291–306]의 방법을 적용했다. 일반적인 깁스 샘플러는 느린 믹싱으로 인해 본 연구에서 제안한 모형의 사후분포를 계산하는 데 한계가 있기 때문에, 본 연구는 모형 내 일부 모수에 대해 해밀토니안 몬테 카를로 방법을 사용한 사후분포 계산 알고리즘을 제시한다. 본 연구는 제안된 모형으로부터 유도된 사후분포가 특정 조건 하에서 사후일치성을 만족한다는 이론적 성질을 증명하였다. 모의실험을 통해 본 연구에서 제안된 모형이 이상치가 존재하는 고차원 자료 하에서 개선된 공분산 추정 성능을 보인다는 것을 확인할 수 있다. 또한 암 조직의 DNA 시그니처 자료에 본 연구에서 제안한 공분산 추정 모형을 적용하여 유방암 전이 여부를 예측하는 분석 사례를 소개한다.

주요어: 베이즈 모델링, 인자 모형, 곱 감마 과정 사전분포, 다변수 t 분포, 해밀토니안 몬테 카를로

학번: 2019-24162