Entropy **2015**, *17*, 6481-6502; doi:10.3390/e17096481

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

A Bayesian Predictive Discriminant Analysis with Screened Data

Hea-Jung Kim

Department of Statistics, Dongguk University-Seoul, Pil-Dong 3Ga, Chung-Gu, Seoul 100-715, Korea; E-Mail: kim3hj@dongguk.edu; Tel.: +82-2-2260-3221

Academic Editors: Carlos De Bragança Pereira and Adriano Polpo

Received: 3 April 2015 / Accepted: 17 September 2015 / Published: 21 September 2015

Abstract: In the application of discriminant analysis, a situation sometimes arises where individual measurements are screened by a multidimensional screening scheme. For this situation, a discriminant analysis with screened populations is considered from a Bayesian viewpoint, and an optimal predictive rule for the analysis is proposed. In order to establish a flexible method to incorporate the prior information of the screening mechanism, we propose a hierarchical screened scale mixture of normal (HSSMN) model, which makes provision for flexible modeling of the screened observations. An Markov chain Monte Carlo (MCMC) method using the Gibbs sampler and the Metropolis–Hastings algorithm within the Gibbs sampler is used to perform a Bayesian inference on the HSSMN models and to approximate the optimal predictive rule. A simulation study is given to demonstrate the performance of the proposed predictive discrimination procedure.

Keywords: Bayesian predictive discriminant analysis; hierarchical model; MCMC method; optimal rule; scale mixture; screened observation

MSC classifications: 62H30; 62F15

1. Introduction

The topic of analyzing multivariate screened data has received a great deal of attention over the last few decades. In the standard multivariate problem, an analysis of data generated from a p-dimensional screened random vector $\boldsymbol{x} \stackrel{d}{=} [\boldsymbol{v}|\boldsymbol{v}_0 \in \mathbf{C}_q)]$ is our issue of interest, where the $p \times 1$ random vector $\boldsymbol{v}$ and the $q \times 1$ random vector $\boldsymbol{v}_0$ (called the screening vector) are jointly distributed with the correlation matrix $Corr(\boldsymbol{v}, \boldsymbol{v}_0^\top) \neq \mathbf{0}$. Thus, we observe $\boldsymbol{x}$ only when the unobservable screening vector $\boldsymbol{v}_0$ belongs to a known subset $\mathbf{C}_q$ of its space $\mathbb{R}^q$, such that $0 \leq P(\boldsymbol{v}_0 \in \mathbf{C}_q) \leq 1$. That is, $\boldsymbol{x}$ is subject to the

screening scheme or hidden truncation (or simply truncation if $\boldsymbol{v} = \boldsymbol{v}_0$). Model parameters underlying the joint distribution of $\boldsymbol{v}$ and $\boldsymbol{v}_0$ are then estimated from the screened data (*i.e.*, observations of $\boldsymbol{x}$) using the conditional density $f(\boldsymbol{x}|\ \boldsymbol{v}_0 \in \mathbf{C}_q)$.

The screening of sample (or sample selection) arises as open in practice as a result of controlling observability of the outcome of interest in the study. For example, the dataset consists of the Otis IQ test scores (the values of $\boldsymbol{x}$) of freshmen of a college. These students had been screened in the college admission process, which examines whether their prior school grade point average (GPA) and the Scholastic Aptitude Test (SAT) scores (*i.e.*, the screening values denoted by $\boldsymbol{v}_0$) are satisfactory. What the true vale of screening vector variable $\boldsymbol{v}_0$ of each student is may not be available due to a college regulation. The observations available are the IQ values of $\boldsymbol{x}$, the screened data. For the application with real screened data, one can refer to that with the student aid grants data given by [1] and that with the U.S. labor market data given by [2], as well. A variety of methods have been suggested for analyzing such screened data. See, e.g., [3–6], for various distributions for modeling and analyzing screened data; see, [7,8] for the estimative classification analysis with screened data; and see, e.g., [1,2,9,10], for the regression analysis with screened response data.

The majority of existing methods rely on the fact that $\boldsymbol{v}$ and $\boldsymbol{v}_0$ are jointly multivariate normal, and the screened observation vector $\boldsymbol{x}$ is subject to a univariate screening scheme defined by an open interval $\mathbf{C}_q$ with $q = 1$. In many practical situations, however, the screened data are generated from a non-normal joint distribution of $\boldsymbol{v}$ and $\boldsymbol{v}_0$, having a multivariate screening scheme defined by a q-dimensional ($q > 1$) rectangle region $\mathbf{C}_q$ of $\boldsymbol{v}_0$. In this case, a difficulty in applications with the screened data is that the empirical distribution of the screened data is skewed; its parametric model involves a complex density; and hence, standard methods of analysis cannot be used. See [4,6] for the conditional densities, $f(\boldsymbol{x}|\ \boldsymbol{v}_0 \in \mathbf{C}_q)$, useful for fitting the rectangle screened data generated from a non-normal joint distribution of $\boldsymbol{v}$ and $\boldsymbol{v}_0$. In this article, we develop yet another multivariate technique applicable for analyzing the rectangle screened data: we are interested in constructing a Bayesian predictive discrimination procedure for the data. More precisely, we consider a Bayesian multivariate technique for sorting, grouping and prediction of multivariate data generated from K rectangle screened populations. In the standard problem, a training sample $\mathcal{D} = \{(z_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\}$ is available, where, for each $i = 1, \ldots, n$, $\boldsymbol{x}_i$ is a $p \times 1$ rectangle screened observation vector coming from one of K populations and taking values in $\mathbb{R}^p$, and z_i is a categorical response variable representing the population membership, so that $z_i = k$ implies that the predictor $\boldsymbol{x}_i$ belongs to the k-th rectangle screened population (denoted by π_k), $k = 1, \ldots, K$. Using the training sample $\mathcal{D}$, the goal of the predictive discriminant analysis is to predict population membership of a new screened observation $\boldsymbol{x}$ based on the posterior probability of $\boldsymbol{x}$ belonging to π_k. The posterior probability is given by:

$$p(z = k\ |\mathcal{D}, \boldsymbol{x}) \propto p(\boldsymbol{x}\ |\mathcal{D}, z = k)p(z = k\ |\mathcal{D}),\ k = 1, \ldots, K, \tag{1}$$

where z is the the population membership of $\boldsymbol{x}$, $p(z = k\ |\mathcal{D})$ is the prior probability of π_k updated by the training sample $\mathcal{D}$ and:

$$p(\boldsymbol{x}\ |\mathcal{D}, z = k) = \int p(\boldsymbol{x}\ |\Theta_k)p(\Theta_k\ |\mathcal{D}, z = k)d\Theta_k, \tag{2}$$

$p(\boldsymbol{x}\ |\Theta_k) = p(\boldsymbol{x}\ |\boldsymbol{v}_0 \in \mathbf{C}_q, z = k)$ and $p(\Theta_k\ |\mathcal{D}, z = k)$, respectively, denote the predictive density, the probability density of $\boldsymbol{x}$ and the posterior density of parameters Θ_k associated with π_k. One of the

first and most applied predictive approaches by [11] is the case of unscreened and normally-distributed populations π_k with unknown parameters $\Theta_k = \{\boldsymbol{\mu}_k, \Sigma_k\}$, namely $\pi_k : N_p(\boldsymbol{\mu}_k, \Sigma_k)$ for $k = 1, \ldots, K$. This is called a Bayesian predictive discriminant analysis with normal populations ($BPDA_N$) in which a multivariate Student t distribution is obtained for Equation (2).

A practical example where the predictive discriminant analysis with the rectangle screened populations (π_k's) is applicable is in the discrimination between passed and failed pairs of applicants in a college admission process (the second screening process). Consider the case where college admission officers wish to set up an objective criterion (with a predictor vector $\boldsymbol{x}$) for admitting students for matriculation; however, the admission officers must first ensure that a student with observation $\boldsymbol{x}$ has passed the first screening process. The first screening scheme may be defined by the q-dimensional region $\mathbf{C}_q$ of the random vector $\boldsymbol{v}_0$ (consisting of SAT scores, high-school GPA, and so on), so that only the students who satisfy $\boldsymbol{v}_0 \in \mathbf{C}_q$ can proceed to the admission process. In this case, we encounter a crucial problem for applying the normal classification by [11]; given the screening scheme $\boldsymbol{v}_0 \in \mathbf{C}_q$, the assumption of the multivariate normal population distribution for $[\boldsymbol{x}|\ z = k] \overset{d}{=} [\boldsymbol{x}|\ \pi_k]$, $k = 1, 2, \ldots, K$ is not valid. The work in [7,12] found that the normal classification shows a lack of robustness to the departure from the normality of the population distribution, and hence, the performance of the normal classification can be very misleading, if used with the continuous, but non-normal or screened normal input vector $\boldsymbol{x}$.

Thus, the predictive density in Equation (2) has two specific features to be considered for Bayesian predictive discrimination with the rectangle screened populations, one about the prior distribution of the parameters Θ_k and the other about the distributional assumption of the population model with density $p(\boldsymbol{x}|\Theta_k)$. For the unscreened populations case, there have been a variety of studies that are concerned with the two considerations. See, for example, [11,13,14] for the choice of the prior distributions of Θ_k, and see [15,16] for copious references to the literature on the predictive discriminant analysis with non-normal population models. Meanwhile, for deriving Equation (2) of the rectangle screened observation $\boldsymbol{x}$, we need to develop a population model with density $p(\boldsymbol{x}|\Theta_k)$ that uses the screened sample information in order to maintain consistency with the underlying theory associated with the populations π_k generating the screened sample. Then, we propose a Bayesian hierarchical approach to flexibly incorporate the prior knowledge about Θ_k with the non-normal sample information, which is the main contribution of this paper to the literature on Bayesian predictive discriminant analysis.

The rest of this paper is organized as follows. Section 2 considers a class of screened scale mixture of normal (SSMN) population models, which well accounts for the screening scheme conducted through a q-dimensional rectangle region $\mathbf{C}_q$ of an external scale mixture of normal vector, $\boldsymbol{v}_0$. Section 3 proposes a hierarchical screened scale mixture of normal (HSSMN) model to derive the predictive density Equation (2) and proposes an optimal rule for Bayesian predictive discriminant analysis (BPDA) with the SSMN populations (abbreviated as $BPDA_{SSMN}$). Approximation of the rule is studied in Section 4 by using an MCMC method applied to the HSSMN model. In Section 5, a simulation study is done to check the convergence of the MCMC method and the performance of the $BPDA_{SSMN}$ by making a comparison between the $BPDA_{SSMN}$ and the $BPDA_N$. Finally, concluding remarks are given in Section 6.

2. The SSMN Population Distributions

Assume that the joint distribution of respective $q \times 1$ and $p \times 1$ vector variables $\boldsymbol{v}_0$ and $\boldsymbol{v}$, associated with π_k, is $F \in \mathcal{F}$, where:

$$\mathcal{F} = \left\{ F : N_s\left(\boldsymbol{\mu}_k^*, \kappa(\eta)\Sigma_k^*\right),\ \eta \sim g(\eta) \text{ with } \kappa(\eta) > 0, \text{ and } \eta > 0 \right\}, \tag{3}$$

$k = 1, \ldots, K$, $s = (q + p)$, η is a mixing variable with the pdf $g(\eta)$, $\kappa(\eta)$ is a suitably-chosen weight function and $\boldsymbol{\mu}_k^*$ and Σ_k^* are partitioned corresponding to the orders of $\boldsymbol{v}_0$ and $\boldsymbol{v}$:

$$\boldsymbol{v}^* = \begin{pmatrix} \boldsymbol{v}_0 \\ \boldsymbol{v} \end{pmatrix},\ \boldsymbol{\mu}_k^* = \begin{pmatrix} \boldsymbol{\mu}_{0k} \\ \boldsymbol{\mu}_k \end{pmatrix},\ \Sigma_k^* = \begin{pmatrix} \Sigma_{0k} & \Delta_k^\top \\ \Delta_k & \Sigma_k \end{pmatrix}. \tag{4}$$

Notice that $\mathcal{F}$ defined by Equation (3) denotes a class of scale mixture of multivariate normal (SMN) distributions (see, e.g., [17,18] for details), equivalently denoted as $SMN_s(\boldsymbol{\mu}_k^*, \Sigma_k^*, \kappa(\eta), G)$ in the remainder of the paper, where $G = G(\eta)$ denote the cdf of η.

Given the joint distribution $[\boldsymbol{v}^* |\ \pi_k] \sim SMN_s(\boldsymbol{\mu}_k^*, \Sigma_k^*, \kappa(\eta), G)$, the SSMN distribution is defined by the following screening scheme:

$$[\boldsymbol{x}|\pi_k] \overset{d}{=} [\boldsymbol{v}|\boldsymbol{v}_0 \in \mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}), \pi_k] \sim SSMN_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*, \kappa(\eta), G), \tag{5}$$

where $\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \{\boldsymbol{v}_0 \in \mathbb{R}^q |\ \boldsymbol{\alpha} \leq \boldsymbol{v}_0 \leq \boldsymbol{\beta}\}$ is a q-dimensional rectangle screening region in the space of $\boldsymbol{v}_0 \in \mathbb{R}^q$. Here, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_q)^T$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_q)^T$, and $\alpha_j < \beta_j$ for $j = 1, \cdots, q$. This region contains the cases of $\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\infty})$ and $\mathbf{C}_q(-\boldsymbol{\infty}, \boldsymbol{\beta})$ as special cases.

The pdf of $\boldsymbol{x}$ is given by:

$$f(\boldsymbol{x}|\boldsymbol{\mu}_k^*, \Sigma_k^*, \pi_k) = \frac{\int_0^\infty h_p(\boldsymbol{x}|\boldsymbol{\mu}_k^*, \Sigma_k^*, \kappa(\eta), G) dG(\eta)}{\int_0^\infty \bar{\Phi}_q(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_{0k}, \kappa(\eta)\Sigma_{0k}) dG(\eta)},\ \boldsymbol{x} \in \mathbb{R}^p, \tag{6}$$

where:

$$h_p(\boldsymbol{x}|\boldsymbol{\mu}_k^*, \Sigma_k^*, \kappa(\eta)) = \phi_p(\boldsymbol{x}\ ; \boldsymbol{\mu}_k, \kappa(\eta)\Sigma_k)\bar{\Phi}_q\big(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_{\boldsymbol{v}_{0k}|\boldsymbol{x}}, \kappa(\eta)\Sigma_{\boldsymbol{v}_{0k}|\boldsymbol{x}}\big),$$

$\boldsymbol{\mu}_{\boldsymbol{v}_{0k}|\boldsymbol{x}} = \boldsymbol{\mu}_{0k} + \Delta_k^\top \Sigma_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)$ and $\Sigma_{\boldsymbol{v}_{0k}|\boldsymbol{x}} = \Sigma_{0k} - \Delta_k^\top \Sigma_k^{-1} \Delta_k$. Here, $\phi_q(\cdot\ \ \boldsymbol{\mu}, \Sigma)$ and $\bar{\Phi}_q(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}, \Sigma)$, respectively, denote the pdf and the probability of the rectangle region of a random vector $w \sim N_q(\boldsymbol{\mu}, \Sigma)$. The latter is equivalent to $Pr(\boldsymbol{w} \in \mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}))$.

One particular member of the class of SSMN distributions is the rectangle-screened normal (RSN) distribution defined by Equations (5) and (6), for which $G(\eta)$ is degenerate with $\kappa(\eta) = 1$. The work in [4,8] studied properties of the distribution and denoted it as the $RSN_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*)$ distribution. Another member of the class is the rectangle-screened p-variate Student t distributions (RSt_p) considered by [8]. Its pdf is given by:

$$f(\boldsymbol{x}\ |\boldsymbol{\mu}_k^*, \Sigma_k^*, \pi_k) = t_p(\boldsymbol{x}\ |\boldsymbol{\mu}_k, \Sigma_k, \nu) \frac{\bar{T}_q\big(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_{\boldsymbol{v}_{0k}|\boldsymbol{x}}, \Gamma_{\boldsymbol{v}_{0k}|\boldsymbol{x}}, \nu + p\big)}{\bar{T}_q(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_{0k}, \Sigma_{0k}, \nu)},\ \boldsymbol{x} \in \mathbb{R}^p, \tag{7}$$

where $t_p(\cdot\ |\mathbf{a}, B, c)$ and $\bar{T}_p(\mathbf{C}_p\ ; \mathbf{a}, B, c)$ are the respective pdf and probability of a rectangle region $\mathbf{C}_p$ of the p-variate Student t distribution with the location vector $\mathbf{a}$, the scale matrix B, the degrees of freedom c and:

$$\Gamma_{\boldsymbol{v}_{0k}|\boldsymbol{x}} = (\nu + p)^{-1} \left\{\nu + (\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right\} \Sigma_{\boldsymbol{v}_{0k}|\boldsymbol{x}}.$$

Similar to the RSN distributions, the density Equation (7) of $[\boldsymbol{x}|\pi_k] \sim RSt_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*, \nu)$ is obtained by taking $\kappa(\eta) = 1/\eta$ and $\eta \sim Gamma(\nu/2, \nu/2)$, *i.e.*,

$$g(\eta) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \eta^{\nu/2-1} \exp\left\{-\frac{\nu}{2}\eta\right\}, \ \eta > 0.$$

The stochastic representations of the RSN and RSt_p distributions are immediately obtained by applying the following lemma, for which detailed proof can be found in [4].

Lemma 1. *Suppose* $[\boldsymbol{x}|\pi_k] \sim SSMN_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*, \kappa(\eta), G)$. *Then, it has the following stochastic representation in a hierarchical fashion,*

$$[\boldsymbol{x}|\eta, \pi_k] \overset{d}{=} \boldsymbol{\mu}_k + \Delta_k \Sigma_{0k}^{-1} \mathbf{Z}_{\mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k)} + (\Sigma_k - \Delta_k \Sigma_{0k}^{-1} \Delta_k^\top)^{1/2} \mathbf{Z}_p, \tag{8}$$

$$\eta \sim G(\eta) \ \textit{with} \ \kappa(\eta) > 0, \eta > 0, \tag{9}$$

where $\mathbf{Z}_p \sim N_p(\mathbf{0}, \kappa(\eta) I_p)$ *and* $\mathbf{Z}_{\mathbf{C}_q - \boldsymbol{\mu}_{0k}} \overset{d}{=} [\mathbf{Z}_q \ | \mathbf{Z}_q \in \mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k)]$ *are conditionally independent and* $\mathbf{Z}_q \sim N_q(\mathbf{0}, \kappa(\eta)\Sigma_{0k})$. *Here,* $\boldsymbol{a}_k = \boldsymbol{\alpha} - \boldsymbol{\mu}_{0k}$ *and* $\boldsymbol{b}_k = \boldsymbol{\beta} - \boldsymbol{\mu}_{0k}$.

Lemma 1 provides the following: (i) an intrinsic structure of the SSMN population distributions, which reveals a type of departure from the SMN law because the distribution of $[\boldsymbol{x}|\pi_k]$ reduces to the SMN distribution if $\Delta_k = \mathbf{0}$ (*i.e.*, $Cov(\boldsymbol{v}_0, \boldsymbol{v}|\pi_k) = \mathbf{0}$); (ii) the representation provides a convenient device for random number generation; (iii) it leads to a simple and direct construction of a HSSMN model for the BPDA with the SSMN populations, *i.e.*, $[\boldsymbol{x}|\pi_k] \sim SSMN_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*, \kappa(\eta), G)$.

3. The HSSMN Model

3.1. The Hierarchical Model

For a Bayesian predictive discriminant analysis, suppose we have K rectangle screened populations $\pi_k (k = 1, \ldots, K)$, each specified by the $SSMN_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*, \kappa(\eta), G)$ distribution. Let $\mathcal{D}_k = \{\boldsymbol{x}_{k1}, \ldots, \boldsymbol{x}_{kn_k}\}$ be a training sample obtained from the rectangle screened population π_k with the $SSMN_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*, \kappa(\eta), G)$ distribution, where the parameters $(\boldsymbol{\mu}_k^*, \Sigma_k^*)$ are unknown. The predictive discrimination analysis is to assess the relative predictive odds ratio or posterior probability that a screened multivariate observation $\boldsymbol{x}$ belongs to one of K populations, π_k. As noted by Equation (6), however, a complex likelihood function of $\mathcal{D}_k$ prevents us from choosing reasonable priors of the model parameters and obtaining the predictive density of $\boldsymbol{x}$ given by Equation (2). These problems are solved if we use the following hierarchical representation of the population models.

According to Lemma 1, we may rewrite the SSMN model for Equations (8) and (9) by a three-level hierarchy given by:

$$\begin{aligned}
[\boldsymbol{x}_{ki}|\eta_{ki}, \boldsymbol{f}_{ki}, \pi_k] &\overset{d}{=} \boldsymbol{\mu}_k + \Lambda_k \boldsymbol{f}_{ki} + \boldsymbol{\varepsilon}_{ki}, \ \ \boldsymbol{\varepsilon}_{ki} \overset{ind}{\sim} N_p(\mathbf{0}, \kappa(\eta_{ki})\Psi_k), \ i = 1, \ldots, n_k, \\
\boldsymbol{f}_{ki} &\overset{ind}{\sim} N_q(\mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k}) \mathrm{I}\big(\boldsymbol{f}_{ki} \in \mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k)\big), \ \ \kappa(\eta_{ki}) > 0, \\
\eta_{ki} &\overset{i.i.d.}{\sim} G(\eta) \ \text{ with } \ \eta_{ki} > 0,
\end{aligned} \tag{10}$$

where $\Lambda_k = \Delta_k \Sigma_{0k}^{-1}$, $\Psi_k = \Sigma_k - \Delta_k \Sigma_{0k}^{-1} \Delta_k^\top$, G is the scale mixing distribution of the independent η_{ki}'s, $\boldsymbol{f}_{ki}$ and $\boldsymbol{\varepsilon}_{ki}$ are independent conditional on η_{ki} and $N_q(\mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k})\mathrm{I}\big(\boldsymbol{f}_{ki} \in \mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k)\big)$ denotes a truncated $N_q(\mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k})$ distribution having the truncated space $\boldsymbol{f}_{ki} \in \mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k)$.

The first stage model in Equation (10) may be written in a compact form by defining the following vector and matrix notations,

$$\mathbf{X}_k = (\boldsymbol{x}_{k1} - \boldsymbol{\mu}_k, \ldots, \boldsymbol{x}_{kn_k} - \boldsymbol{\mu}_k),\ \mathbf{F}_k = (\boldsymbol{f}_{k1}, \ldots, \boldsymbol{f}_{kn_k}),$$
$$\mathbf{E}_k = (\boldsymbol{\varepsilon}_{k1}, \ldots, \boldsymbol{\varepsilon}_{kn_k}),\ \boldsymbol{\eta}_k = (\eta_{k1}, \ldots, \eta_{kn_k})^\top.$$

Then, the three-level hierarchy of the model Equation (10) can be expressed as:

$$\begin{array}{rcl} \mathbf{X}_k & = & \Lambda_k \mathbf{F}_k + \mathbf{E}_k,\ vec(\mathbf{E}_k) \sim N_{pn_k}(\mathbf{0}, D(\kappa(\boldsymbol{\eta}_k)) \otimes \Psi), \\ vec(\mathbf{F}_k) & \sim & N_{qn_k}(\mathbf{0}, D(\kappa(\boldsymbol{\eta}_k)) \otimes \Sigma_{0k})\mathrm{I}(\boldsymbol{f}_{ki} \in \mathbf{C}_q\big(\boldsymbol{a}_k, \boldsymbol{b}_k)\big),\ \ \mathrm{Cov}(vec(\mathbf{F}_k), vec(\mathbf{E}_k)^\top | \boldsymbol{\eta}_k) = \mathbf{O}, \\ \eta_{ki} & \overset{i.i.d.}{\sim} & g(\eta),\ \ i = 1, \ldots, n_k, \end{array} \tag{11}$$

where $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$, $vec(\mathbf{F}_k) = (\boldsymbol{f}_{k1}^\top, \ldots, \boldsymbol{f}_{kn_k}^\top)^\top$, $vec(\mathbf{E}_k) = (\boldsymbol{\varepsilon}_{k1}^\top, \ldots, \boldsymbol{\varepsilon}_{kn_k}^\top)^\top$, and $D(\kappa(\boldsymbol{\eta}_k)) = diag\{\kappa(\eta_{k1}), \ldots, \kappa(\eta_{kn_k})\}$ is an $n_k \times n_k$ diagonal matrix of the scale mixing functions. Note that the hierarchical population model Equation (11) adopts a robust discriminant modeling by the use of the scale mixture of normal, such as the SMN and the truncated SMN, and thus, it enables us to avoid the anomaly generated from the non-normal sample information.

The Bayesian analysis of the model in Equation (11) begins with the specification of the prior distributions of the unknown parameters. When the prior information is not available, a convenient strategy of avoiding improper posterior distribution is to use proper priors with their hyperparameters being fixed as appropriate quantities to reflect the flatness (or diffuseness) of priors (*i.e.*, limiting non-informative priors). For convenience, but not always optimal, we suppose that $\boldsymbol{\mu}_k$, $\boldsymbol{\mu}_{0k}$, (Λ_k, Ψ_k) and Σ_{0k} of the model in Equation (11) are independent *a priori*; prior distributions for $\boldsymbol{\mu}_k$ and $\boldsymbol{\mu}_{0k}$ are normal; an inverse Wishart prior distribution for Σ_{0k}; and a generalized natural conjugate family (see [19]) of prior distributions for (Λ_k, Ψ_k), so that we adopt the normal prior density for the Λ_k conditional on the matrix Ψ_k,

$$\begin{array}{rcl} P(\Lambda_k | \Psi_k) & \sim & |\Psi_k|^{-q/2} \exp\Big\{ -\dfrac{1}{2} \mathrm{tr}\,[\Psi_k^{-1}(\Lambda_k - \Lambda_{0k}) H_k (\Lambda_k - \Lambda_{0k})^\top] \Big\}, \\ \Psi_k & \sim & IW_p(R_k, \tau_k), \end{array}$$

where $W \sim IW_m(V, \nu)$ denotes the inverse Wishart distribution whose pdf $IW_m(W; V, \nu)$ is:

$$IW_m(W; V, \nu) \propto |W|^{-\nu/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}(W^{-1} V) \right\},\ \ V > 0.$$

Note that if $\Lambda_k = (\boldsymbol{\lambda}_{k1}, \ldots, \boldsymbol{\lambda}_{kq})$, $\boldsymbol{\lambda}_k \equiv vec(\Lambda_k) = (\boldsymbol{\lambda}_{k1}^\top, \ldots, \boldsymbol{\lambda}_{kq}^\top)^\top$ and $\boldsymbol{\lambda}_{0k} \equiv vec(\Lambda_{0k})$, then:

$$\mathrm{tr}\,[\Psi_k^{-1}(\Lambda_k - \Lambda_{0k}) H_k (\Lambda_k - \Lambda_{0k})^\top] = (\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{0k})^\top (H_k^{-1} \otimes \Psi_k)^{-1} (\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{0k}).$$

This prior elicitation of the parameters, along with the three-level hierarchical model Equation (11), produces a hierarchical screened scale mixture of normal population model, which is referred

to as *HSSMN*($\boldsymbol{\Theta}(k)$) in the rest of this paper, where $\boldsymbol{\Theta}(k) = \{\boldsymbol{\mu}_k, \boldsymbol{\mu}_{0k}, \Lambda_k, \Psi_k, \mathbf{F}_k, \Sigma_{0k}, \boldsymbol{\eta}_k\}$. The *HSSMN*($\boldsymbol{\Theta}(k)$) model is defined as follows.

$$\begin{aligned}
\left[\boldsymbol{x}_{ki}|\mathbf{F}_k, \boldsymbol{\mu}_k, \Psi_k, \boldsymbol{\eta}_k\right] &\overset{ind}{\sim} N_p(\boldsymbol{\mu}_k + \Lambda_k \boldsymbol{f}_{ki},\ \kappa(\eta_{ki})\Psi_k),\ i = 1, \ldots, n_k, \\
\left[\boldsymbol{f}_{ki}|\Psi_k, \Sigma_{0k}, \boldsymbol{\mu}_{0k}, \boldsymbol{\eta}_k\right] &\overset{ind}{\sim} N_q\Big(\mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k}\Big)\mathrm{I}\big(\boldsymbol{f}_{ki} \in \mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k)\big),\ i = 1, \ldots, n_k, \\
\boldsymbol{\mu}_k &\sim N_p(\boldsymbol{\theta}_k, \Omega_k), \\
\boldsymbol{\mu}_{0k} &\sim N_q(\boldsymbol{\theta}_{0k}, \Omega_{0k}), \\
\left[\boldsymbol{\lambda}_k|\ \Psi_k\right] &\sim N_{pq}(\boldsymbol{\lambda}_{0k}, H_k^{-1} \otimes \Psi_k), \\
\Psi_k &\sim IW_p(R_k, \tau_k),\ \ \tau_k > 2p, \\
\Sigma_{0k} &\sim IW_q(Q_k, \gamma_k),\ \ \gamma_k > 2q, \\
\eta_{ki} &\overset{ind}{\sim} g(\eta),\ i = 1, \ldots, n_k,
\end{aligned} \tag{12}$$

where $\boldsymbol{\lambda}_k \equiv vec(\Lambda_k)$, $\boldsymbol{\lambda}_{0k} \equiv vec(\Lambda_{0k})$ and hyperparameters $(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{0k}, \Omega_k, \Omega_{0k}, \Lambda_{0k}, H_k, R_k, \tau_k, Q_k, \gamma_k)$ are fixed as appropriate quantities to reflect the flatness of priors.

The last distributional specification is omitted in the RSN distribution case. For the *HSSMN*($\boldsymbol{\Theta}(k)$) model for the RSt_ν distribution, we may set $\eta_{ki} \overset{ind}{\sim} Gamma(\nu/2, \nu/2)$, $\nu \sim Gamma(1, 0.1)\mathrm{I}(\nu > 2)$, a truncated Gamma distribution (see, e.g., [20]). See, for example, [21,22] and the references therein for other choices of the prior distribution of ν.

3.2. Posterior Distributions

Based on the *HSSMN*($\Theta(k)$) model structure with the likelihood and the prior distributions in Equation (12), the joint posterior distribution of $\Theta(k)$ is given by:

$$\begin{aligned}
p(\Theta(k) \mid \mathcal{D}_k) \ \propto\ & \Big(\prod_{i=1}^{n_k} |\kappa(\eta_{ki})\Psi_k|^{-1/2}\Big) \exp\Big\{-\frac{1}{2}\mathrm{tr}\left[\Psi_k^{-1}(\mathbf{X}_k - \Lambda_k\mathbf{F}_k)\mathbf{D}(\kappa(\boldsymbol{\eta}_k))^{-1}(\mathbf{X}_k - \Lambda_k\mathbf{F}_k)^\top\right]\Big\} \\
& \times\ |\Psi_k|^{-(q+\tau_k)/2} \exp\Big\{-\frac{1}{2}\mathrm{tr}[\Psi_k^{-1}G_k]\Big\} \prod_{i=1}^{n_k} \frac{\phi_q(\boldsymbol{f}_{ki}; \mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k})}{\bar{\Phi}_q\left(\mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k); \mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k}\right)} \Big(\prod_{i=1}^{n_k} g(\eta_{ki})\Big) \\
& \times\ IW_q(\Sigma_{0k}; Q_k, \gamma_k)\phi_p(\boldsymbol{\mu}_k; \boldsymbol{\theta}_k, \Omega_k)\phi_q(\boldsymbol{\mu}_{0k}; \boldsymbol{\theta}_{0k}, \Omega_{0k}),
\end{aligned} \tag{13}$$

where:

$$\begin{aligned}
& \Big(\prod_{i=1}^{n_k} |\kappa(\eta_{ki})\Psi_k|^{-1/2}\Big) \exp\Big\{-\frac{1}{2}\mathrm{tr}\left[\Psi_k^{-1}(\mathbf{X}_k - \Lambda_k\mathbf{F}_k)\mathbf{D}(\kappa(\boldsymbol{\eta}_k))^{-1}(\mathbf{X}_k - \Lambda_k\mathbf{F}_k)^\top\right]\Big\} \\
\propto\ & \prod_{i=1}^{n_k} \phi_p(\boldsymbol{x}_{ki}; \boldsymbol{\mu}_k + \Lambda_k\boldsymbol{f}_{ki}, \kappa(\eta_{ki})\Psi_k),
\end{aligned}$$

$G_k = (\Lambda_k - \Lambda_{0k})H_k(\Lambda_k - \Lambda_{0k})^\top + R_k$ and $g(\eta_{ki})$'s denote the densities of the mixing variables η_{ki}'s. Note that the joint posterior of Equation (13) is not simplified in an analytic form of the known density and, thus, intractable for the posterior inference. Instead, we derived each of conditional posterior distribution of $\boldsymbol{\mu}_k$, $\boldsymbol{\mu}_{0k}$, $\boldsymbol{\lambda}_k \equiv vec(\Lambda_k)$, Σ_{0k}, $\mathbf{F}_k$, Ψ_k and η_{ki}'s, which will be useful for posterior inference based on Markov chain Monte Carlo methods (MCMC). All of the full conditional posterior distributions are as follows (see the Appendix for their derivations):

(1) The full conditional distribution of $\boldsymbol{\mu}_k$ is a p-variate normal given by:

$$\Big[\boldsymbol{\mu}_k \mid \ \Theta(k)_{\setminus \boldsymbol{\mu}_k}, \mathcal{D}_k\Big] \sim N_p\Big(\boldsymbol{\mu}_{\boldsymbol{\mu}_k}, \Sigma_{\boldsymbol{\mu}_k}\Big), \tag{14}$$

where $\boldsymbol{\mu}_{\boldsymbol{\mu}_k} = \Sigma_{\boldsymbol{\mu}_k}\Big(\Omega_k^{-1}\boldsymbol{\theta}_k + \sum_{i=1}^{n_k} \Psi_k^{-1}(\boldsymbol{x}_{ki} - \Lambda_k \boldsymbol{f}_{ki})/\kappa(\eta_{ki})\Big)$ and $\Sigma_{\boldsymbol{\mu}_k} = \Big(\sum_{i=1}^{n_k} \frac{1}{\kappa(\eta_{ki})}\Psi_k^{-1} + \Omega_k^{-1}\Big)^{-1}$.

(2) The full conditional density of $\boldsymbol{\mu}_{0k}$ is given by:

$$p(\boldsymbol{\mu}_{0k}|\ \Theta(k)_{\setminus \boldsymbol{\mu}_{0k}}, \mathcal{D}_k) \propto \frac{\phi_q(\boldsymbol{\mu}_{0k}; \boldsymbol{\theta}_{0k}, \Omega_{0k})}{\prod_{i=1}^{n_k} \bar{\Phi}_q\left(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_{0k}, \kappa(\eta_{ki})\Sigma_{0k}\right)}. \tag{15}$$

(3) The full conditional posterior distribution of $\boldsymbol{\lambda}_k$ is given by:

$$\Big[\boldsymbol{\lambda}_k \mid \Theta(k)_{\setminus \boldsymbol{\lambda}_k}, \mathcal{D}_k\Big] \sim N_{pq}\Big(\boldsymbol{\mu}_{\boldsymbol{\lambda}_k}, \Sigma_{\boldsymbol{\lambda}_k}\Big), \tag{16}$$

where:

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\lambda}_k} &= vec(\Lambda_k^*), \ \ \Lambda_k^* = (\mathbf{X}_k \mathbf{D}(\kappa(\boldsymbol{\eta}_k))^{-1}\mathbf{F}_k^\top + \Lambda_{0k} H_k) Q_k^{-1} \\ \Sigma_{\boldsymbol{\lambda}_k} &= Q_k^{-1} \otimes \Psi_k, \text{ and } \ Q_k = \mathbf{F}_k \mathbf{D}(\kappa(\boldsymbol{\eta}_k))^{-1}\mathbf{F}_k^\top + H_k. \end{aligned}$$

(4) The full conditional posterior distribution of Ψ_k is an inverse-Wishart distribution:

$$\Big[\Psi_k \mid \ \Theta(k)_{\setminus \Psi_k}, \mathcal{D}_k\Big] \sim IW_p\Big(V_k, \boldsymbol{\nu}_k\Big) \ \ \boldsymbol{\nu}_k > 2p, \tag{17}$$

where $V_k = (\mathbf{X}_k - \Lambda_k \mathbf{F}_k)\mathbf{D}(\kappa(\boldsymbol{\eta}_k))^{-1}(\mathbf{X}_k - \Lambda_k \mathbf{F}_k)^\top + (\Lambda_k - \Lambda_{0k}) H_k (\Lambda_k - \Lambda_{0k})^\top + R_k$ and $\boldsymbol{\nu}_k = n_k + q + \tau_k$.

(5) The full conditional posterior distribution of $\boldsymbol{f}_{ki}$ is the q-variate truncated normal given by:

$$\Big[\boldsymbol{f}_{ki} \mid \ \Theta(k)_{\setminus \boldsymbol{f}_{ki}}, \mathcal{D}_k\Big] \overset{ind}{\sim} N_q\Big(\boldsymbol{\mu}_{\boldsymbol{f}_{ki}}, \kappa(\eta_{ki})\Sigma_{\boldsymbol{f}_{ki}}\Big)\mathrm{I}\big(\boldsymbol{f}_{ki} \in \mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k)\big), \ i = 1, \ldots, n_k, \tag{18}$$

where $\boldsymbol{\mu}_{\boldsymbol{f}_{ki}} = \Sigma_{\boldsymbol{f}_{ki}} \Lambda_k^\top \Psi_k^{-1}(\boldsymbol{x}_{ki} - \boldsymbol{\mu}_k)$ and $\Sigma_{\boldsymbol{f}_{ki}} = \Big(\Sigma_{0k}^{-1} + \Lambda_k^\top \Psi_k^{-1} \Lambda_k\Big)^{-1}$.

(6) The full conditional posterior density of Σ_{0k} is given by:

$$p(\Sigma_{0k} \mid \ \Theta(k)_{\setminus \Sigma_{0k}}, \mathbf{y}_k) \propto IW_q(\Sigma_{0k}; Q_k, \gamma_k) \prod_{i=1}^{n_k} \frac{\phi_q(\boldsymbol{f}_{ki}; \mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k})}{\bar{\Phi}_q\left(\mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k); \mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k}\right)}. \tag{19}$$

(7) The full conditional posterior densities of η_{ki}'s are given by:

$$\begin{aligned} p(\eta_{ki}|\ \ \Theta(k)_{\setminus \eta_{ki}}, \mathbf{y}_k) \ \ &\propto \ \ \kappa(\eta_{ki})^{-\frac{p}{2}} \exp\left\{-\frac{\boldsymbol{z}_{ki}^\top \Psi_k^{-1} \boldsymbol{z}_{ki}}{2\kappa(\eta_{ki})}\right\} \\ &\times \ \ \frac{\phi_q(\boldsymbol{f}_{ki}; \mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k})}{\bar{\Phi}_q\left(\mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k); \mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k}\right)} g(\eta_{ki}), \ i = 1, \ldots, n_k, \end{aligned} \tag{20}$$

where $\boldsymbol{z}_{ki} = \boldsymbol{x}_{ki} - \boldsymbol{\mu}_k - \Lambda_k \boldsymbol{f}_{ki}$ and η_{ki}'s are independent.

Based on the above full conditional posterior distributions and the stochastic representations of the SSMN in Lemma 1, one can easily obtain Bayes estimates of the k-th SSMN population mean $\boldsymbol{\mu}_{\pi_k} = E[\boldsymbol{x}|\pi_k]$ and covariance matrix $\Sigma_{\pi_k} = Cov(\boldsymbol{x}|\pi_k)$, $k = 1, \ldots, K$. Specifically, the mean and

covariance matrix of an observation $\boldsymbol{x}$ belonging to $\pi_k : SSMN_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*, \kappa(\eta), G)$, which are used for calculating their Bayes estimates via Rao–Blackwellization, are given by:

$$\begin{aligned} \boldsymbol{\mu}_{\pi_k} &= \boldsymbol{\mu}_k + \Omega_{21k}\Omega_{22k}^{-1}\boldsymbol{\xi}_k \\ \Sigma_{\pi_k} &= \Omega_{22k} - \Omega_{21k}(\Omega_{11k}^{-1} - \Omega_{11k}^{-1}\mathbf{T}_k\Omega_{11k}^{-1})\Omega_{21k}^{\top}, \end{aligned} \tag{21}$$

where $\Omega_{21k} = \kappa(\eta)\Delta_k$, $\Omega_{11k} = \kappa(\eta)\Sigma_{0k}$, $\Omega_{22k} = \kappa(\eta)\Sigma_k$,

$$\boldsymbol{\xi}_k = \int_{\mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k)} \frac{\mathbf{z}}{\zeta_k(2\pi)^{q/2}|\Omega_{11k}|^{1/2}} \exp\Big\{ -\frac{\mathbf{z}^{\top}\Omega_{11k}^{-1}\mathbf{z}}{2} \Big\} d\mathbf{z},$$

$\zeta_k = \bar{\Phi}_q(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_{0k}, \Omega_{11k})$, $\mathbf{T}_k = \mathbf{P}_k - \boldsymbol{\xi}_k\boldsymbol{\xi}_k^{\top}$ and:

$$\mathbf{P}_k = \int_{\mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k)} \frac{\mathbf{z}\mathbf{z}^{\top}}{\zeta_k(2\pi)^{q/2}|\Omega_{11k}|^{1/2}} \exp\Big\{ -\frac{\mathbf{z}^{\top}\Omega_{11k}^{-1}\mathbf{z}}{2} \Big\} d\mathbf{z}.$$

We see that these moments of Equation (21) agree with the formula for the mean and covariance matrix of the untruncated marginal distribution of a general multivariate truncated distribution given by [23]. Readers are referred to [24] with the R package `tmvtnorm` and [25] with the R package `mvtnorm` for implementing calculations of $\boldsymbol{\xi}_k$ and $\mathbf{P}_k$ involved in the first and second moments.

When the sampling information, *i.e.*, the observed training samples, is augmented by the proper information of prior knowledge, the anomalies of the maximum likelihood estimate of the SSMN model, investigated by [16], would disappear in the *HSSMN* $(\Theta(k))$ model. Furthermore, note that the conditional distribution of $\boldsymbol{\lambda}_k$ in Equation (16) is a pq-dimensional one; and hence, its Gibbs sampling needs to be performed by using the inverse of the matrix of order pq, which may cause computational costs in implementing the MCMC method. For large q, a more computationally-convenient Gibbs sampler can be considered based on the full conditional posterior distributions of $\boldsymbol{\lambda}_{kj}$, $j = 1, \ldots, p$, than the Gibbs sampler with $\boldsymbol{\lambda}_k$ in Equation (16), where $\boldsymbol{\lambda}_k \equiv Vec(\Lambda_k)$ and $\Lambda_k \equiv (\boldsymbol{\lambda}_{k1}, \ldots, \boldsymbol{\lambda}_{kq})$.

For this purpose, we defined the following notations: for $j = 1, \ldots, p$,

$$\tilde{\boldsymbol{\lambda}}_k(j) = (E_j \otimes I_p)\boldsymbol{\lambda}_k, \quad \tilde{\boldsymbol{\theta}}_k(j) = (E_j \otimes I_p)\boldsymbol{\mu}_{\boldsymbol{\lambda}_k},$$

$$\tilde{\Omega}_k(j) = (E_j \otimes I_p)\Sigma_{\boldsymbol{\lambda}_k}(E_j \otimes I_p)^{\top}, \quad E_i = (\mathbf{e}_j, \mathbf{e}_1, \ldots, \mathbf{e}_{j-1}, \mathbf{e}_{j+1}, \ldots, \mathbf{e}_q)^{\top},$$

where $\mathbf{e}_j$ denotes the j-th column of I_q, namely an elementary vector with unity for its j-th element and zeros elsewhere. Furthermore, we consider the following partitions:

$$\tilde{\boldsymbol{\lambda}}_k(j) = \begin{pmatrix} \boldsymbol{\lambda}_{kj} \\ \boldsymbol{\lambda}_{kj}^* \end{pmatrix}, \quad \tilde{\boldsymbol{\theta}}_k(j) = \begin{pmatrix} \tilde{\boldsymbol{\theta}}_k(1j) \\ \tilde{\boldsymbol{\theta}}_k(2j) \end{pmatrix}, \quad \text{and} \quad \tilde{\Omega}_k(j) = \begin{pmatrix} \tilde{\Omega}_{k11}(j) & \tilde{\Omega}_{k12}(j) \\ \tilde{\Omega}_{k21}(j) & \tilde{\Omega}_{k22}(j) \end{pmatrix},$$

where the orders of $\boldsymbol{\lambda}_{kj}^*$, $\tilde{\boldsymbol{\theta}}_k(1j)$, $\tilde{\Omega}_{k11}(j)$ and $\tilde{\Omega}_{k21}(j)$ are $(p-1)q \times 1$, $q \times 1$, $q \times q$ and $(p-1)q \times q$, respectively. Under these partitions, the conditional property of a multivariate normal distribution leads to the full conditional posterior distributions of $\boldsymbol{\lambda}_{kj}$ given by:

$$\left[\boldsymbol{\lambda}_{kj}|\Theta(k)_{\backslash\boldsymbol{\lambda}_{kj}}, \mathbf{y}_k\right] \sim N_q\left(\boldsymbol{\mu}_{\boldsymbol{\lambda}_{kj}}, \Sigma_{\boldsymbol{\lambda}_{kj}}\right), \tag{22}$$

for $j = 1, \ldots, q$, where:

$$\boldsymbol{\mu}_{\boldsymbol{\lambda}_{kj}} = \tilde{\boldsymbol{\theta}}_k(1j) + \tilde{\Omega}_{k12}\tilde{\Omega}_{k22}^{-1}(\boldsymbol{\lambda}_{kj}^* - \tilde{\boldsymbol{\theta}}_k(2j)) \text{ and } \Sigma_{\boldsymbol{\lambda}_{kj}} = \tilde{\Omega}_{k11}(j) - \tilde{\Omega}_{k12}(j)\tilde{\Omega}_{k22}(j)^{-1}\tilde{\Omega}_{k21}(j).$$

When p is large, we may partition $\boldsymbol{\lambda}_{kj}$ into two vectors with smaller dimensions, say $\boldsymbol{\lambda}_k = (\boldsymbol{\lambda}_{kj}(1)^\top, \boldsymbol{\lambda}_{kj}(2),^\top)^\top$, then use their full conditional normal distributions for the Gibbs sampler.

Now, the posterior sampling can be implemented by using all of the conditional posterior Equations (14)–(20). The Gibbs sampler and Metropolis–Hastings algorithm within the Gibbs sampler may be used to obtain posterior samples of all of the unknown parameters $\Theta(k)$. Note that in the case where the pq-dimensional matrix is too large to manipulate for computation, the Gibbs sampler can be modified by replacing the full conditional posterior Equation (16) with Equation (22). That is, as indicated by Equation (22), the modified Gibbs sampler based on Equation (22) would be more convenient for numerical computation than the first one using Equation (16). The detailed Markov chain Monte Carlo algorithm with Gibbs sampling is discussed in the next subsection.

3.3. Markov Chain Monte Carlo Sampling Scheme

It is not complicated to construct an MCMC sampling scheme working with $\Theta(k) = \{\boldsymbol{\mu}_k, \boldsymbol{\mu}_{0k}, \Lambda_k, \Psi_k, \mathbf{F}_k, \Sigma_{0k}, \boldsymbol{\eta}_k\}$, since a routine Gibbs sampler would work to generate posterior samples of $(\boldsymbol{\mu}_k, \Lambda_k, \Psi_k, \mathbf{F}_k)$ based on each of their full conditional posterior distributions obtained in Section 3.2. In the posterior sampling of $\boldsymbol{\mu}_{0k}$, Σ_{0k} and $\boldsymbol{\eta}_k$, Metropolis–Hastings within the Gibbs algorithm would be used, since their conditional posterior densities do not have explicit forms of known distributions as in Equations (15), (19) and (20).

Here, for simplicity, we considered the MCMC algorithm based on the *HSSMN*$(\Theta(k))$ model with a known screening scheme, in which $\boldsymbol{\mu}_{0k}$ and Σ_{0k} are assumed to be known. The extension to the general *HSSMN*$(\Theta(k))$ model with unknown $\boldsymbol{\mu}_{0k}$ and Σ_{0k} can be made without difficulty.

The MCMC algorithm starts with some initial values $\boldsymbol{\mu}_k^{[0]}$, $\boldsymbol{\lambda}_k^{[0]}$, $\Psi_k^{[0]}$, $\mathbf{F}_k^{[0]}$ and $\boldsymbol{\eta}_k^{[0]}$. The detailed posterior sampling steps are as follows:

Step 1: generate $\boldsymbol{\mu}_k$ by using the full conditional posterior distribution in Equation (14).
Step 2: generate $\boldsymbol{\lambda}_k$ by using the full conditional posterior distribution in Equation (16).
Step 3: generate inverse-Wishart random matrix ψ_k by using the full conditional posterior distribution in Equation (17).
Step 4: generate independent q-variate truncated normal random variables $\boldsymbol{f}_{ki}$ by using the full conditional posterior distribution in Equation (18).
Step 5: given the current values $\{\boldsymbol{\mu}_k, \Lambda_k, \Psi_k, \mathbf{F}_k\}$, we independently generate a candidate η_{ki} from a proposal density $q(\eta_{ki}^*|\eta_{ki}) = g(\eta_{ki}^*)$, as suggested by [26], which is used for a Metropolis–Hastings algorithm. Then, accept the candidate value with the acceptance rate:

$$\alpha(\eta_{ki}, \eta_{ki}^*) = \min\left\{\frac{p(\Theta(k)|\eta_{ki}^*)}{p(\Theta(k)|\eta_{ki}))}, 1\right\}$$

$i = 1, \ldots, n_k$. Because the target density is proportional to $p(\Theta(k)|\eta_{ki})g(\eta_{ki})$ and $p(\Theta(k)|\eta_{ki})$ is uniformly bounded for $\eta_{ik} > 0$ where:

$$p(\Theta(k)|\eta_{ki}) = \phi_p(\boldsymbol{x}_{ki}; \boldsymbol{\mu}_k + \Lambda_k \boldsymbol{f}_{ki}, \kappa(\eta_{ki})\Psi_k) \frac{\phi_q(\boldsymbol{f}_{ki}; \mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k})}{\bar{\Phi}_q(\mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k); \mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k})}$$

and $g(\cdot)$ is the density of mixing variable η_{ik}. Note that $\boldsymbol{\eta}_k = (\eta_{k1}, \ldots, \eta_{kn_k})^\top$.

When one conducts a posterior inference of the *HSSMN* ($\Theta(k)$) model using the samples obtained from the MCMC sampling algorithm, the following points should be noted.

(i) See, e.g., [18], for the sampling method for η_{ki} from various mixing distributions $g(\eta_{ki})$ of the SMN distributions, such as the multivariate t, multivariate $logit$, multivariate $stable$ and multivariate $exponential\ power$ models.

(ii) Suppose the *HSSMN*($\Theta(k)$) model involves unknown $\boldsymbol{\mu}_{0k}$. Then, as indicated by the full conditional posterior of $\boldsymbol{\mu}_{0k}$ in Equation (15), the complexity of the conditional distribution prevents us from using straightforward Gibbs sampling. Instead, we may use a simple random walk Metropolis algorithm that uses a normal proposal density $q(\boldsymbol{\mu}^*_{0k}|\boldsymbol{\mu}_{0k}) = q(|\boldsymbol{\mu}^*_{0k} - \boldsymbol{\mu}_{0k}|)$ to sample from the conditional distribution of $\boldsymbol{\mu}^*_{0k}$; that is, given the current point is $\boldsymbol{\mu}_{0k}$, the candidate point is $\boldsymbol{\mu}^*_{0k} \sim N_q(\boldsymbol{\mu}_{0k}, D)$, where a diagonal matrix D should be turned, so that the acceptance rate of the candidate point is around 0.25 (see, e.g., [26]).

(iii) When the *HSSMN*($\Theta(k)$) model involves unknown Σ_{0k}: The MCMC sampling algorithm, using the full conditional posterior Equation (19) is not straightforward, because the conditional posterior density is unknown and complex. Instead, we may apply a Metropolized hit-and-run algorithm, described by [27], to sample from the conditional posterior of Σ_{0k}.

(iv) One can easily calculate the posterior estimate of $\Theta_k = (\boldsymbol{\mu}^*_k, \Sigma^*_k)$ by using that of $\Theta(k)$, because the re-parameterizing relations are $\Psi = \Sigma_k - \Delta_k \Sigma^{-1}_{0k} \Delta^\top_k$ and $\Lambda_k = \Delta_k \Sigma^{-1}_{0k}$.

4. The Predictive Classification Rule

Suppose we have K populations π_k, $k = 1, \ldots, K$, each specified by the *HSSMN*($\Theta(k)$) model. For each of the populations, we have the screened training sample $\mathcal{D}_k$ comprised of a set of independent observations $\{\boldsymbol{x}_{ki},\ i = 1, \ldots, n_k\}$ whose population level is $z_{ki} = k$. Let $\boldsymbol{x}$ be assigned to one of the K populations, with prior probability p_k of belonging to π_k, $\sum_{k=1}^K p_k = 1$. Then, the predictive density of $\boldsymbol{x}$ given $\mathcal{D}$ under the *HSSMN*($\Theta(k)$) model with the space $\Theta(k) \in \boldsymbol{\Theta}(k)$ is:

$$p(\boldsymbol{x}\ |\mathcal{D}, z = k) = \int_{\boldsymbol{\Theta}(k)} p(\boldsymbol{x}\ |\Theta(k)) p(\Theta(k)\ |\mathcal{D}) d\Theta(k),\ k = 1, \ldots, K, \tag{23}$$

and the posterior probability that $\boldsymbol{x}$ belongs to π_k, *i.e.*, $p(z = k\ |\mathcal{D}, \boldsymbol{x}) = p(\boldsymbol{x} \in \pi_k\ |\mathcal{D}, \boldsymbol{x})$, is:

$$p(\boldsymbol{x} \in \pi_k\ |\mathcal{D}, \boldsymbol{x}) = \frac{p(\boldsymbol{x}\ |\mathcal{D}, z = k) p(z = k\ |\mathcal{D})}{\sum_{j=1}^K p(\boldsymbol{x}\ |\mathcal{D}, z = j) p(z = j\ |\mathcal{D})},\ k = 1, \ldots, K, \tag{24}$$

where $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k$, $p(\boldsymbol{x} \mid \Theta(k))$ is equal to Equation (6) and $p(\Theta(k) \mid \mathcal{D})$ is the joint posterior density given in Equation (13). We see from Equation (24) that the total posterior probability of misclassifying $\boldsymbol{x}$ from π_i to π_j, $i \neq j$ is defined by:

$$TPM(j) = \sum_{i \neq j; i=1}^{K} \frac{p(\boldsymbol{x} \mid \mathcal{D}, z = i) p(z = i \mid \mathcal{D})}{\sum_{\ell=1}^{K} p(\boldsymbol{x} \mid \mathcal{D}, z = \ell) p(z = \ell \mid \mathcal{D})}. \tag{25}$$

We minimize the misclassification error at this point if we choose j, so as to minimize Equation (25); that is, we select k that gives the maximum posterior probability $p(\boldsymbol{x} \in \pi_k \mid \mathcal{D}, \boldsymbol{x})$ (see, e.g., Theorem 6.7.1 of [28] (p. 234). Thus, an optimal Bayesian predictive discrimination rule that minimizes the classification error is to classify $\boldsymbol{x}$ into π_k, if $\boldsymbol{x} \in R_k$, where the optimal classification region is given by:

$$R_k : \; p(\boldsymbol{x} \mid \mathcal{D}, z = k) p(z = k \mid \mathcal{D}) > p(\boldsymbol{x} \mid \mathcal{D}, z = j) p(z = j \mid \mathcal{D}), \quad \text{for all} \quad j \neq k; k = 1, \ldots, K, \tag{26}$$

$p(z = k \mid \mathcal{D})$ is the posterior probability of population π_k given the dataset $\mathcal{D}$. If we assume the values of p_k's are *a priori* known, then $p(z = k \mid \mathcal{D}) = p_k$.

Since we are unable to obtain an analytic solution of Equation (26), a numerical approach is required. Thus, we used the MCMC method of the previous section to draw samples from the posterior density of the parameters, $p(\Theta(k) \mid \mathcal{D})$, to approximate the predictive density, Equation (23), by:

$$p(\boldsymbol{x} \mid \mathcal{D}, z = k) \approx \frac{1}{N_k - M} \sum_{t=M+1}^{N_k} p(\boldsymbol{x} \mid \Theta_k^t), \; k = 1, \ldots, K, \tag{27}$$

where Θ_k^t's are posterior samples generated from the MCMC process under the *HSSMN*$(\Theta(k))$ model and M and N_k are the burn-in period and run length, respectively.

If we assume Dirichlet priors for p_k, that is:

$$[p_1, \ldots, p_{K-1}] \sim Dirichlet(d_1, \ldots, d_{K-1}; d_K)$$

(see, e.g., [19] (p. 143) for the distributional properties), then:

$$p(z = k \mid \mathcal{D}) = E[p_k \mid \mathcal{D}] = \frac{d_k + n_k}{\sum_j^K d_j + n_j}, \; k = 1, \ldots, K - 1 \tag{28}$$

and $p(z = K \mid \mathcal{D}) = 1 - \sum_{j=k}^{K-1} p(z = k \mid \mathcal{D})$.

Thus, the posterior probabilities in Equation (24) and the minimum error classification region R_k in Equation (26) can be generated within the MCMC scheme, which uses Equation (27) to approximate the predictive densities involved in Equations (24) and (26).

5. Simulation Study

This section presents results of a simulation study to show the convergence of the MCMC algorithm and the performance of the $BPDA_{SSMN}$. Simulation of the training sample observations, model estimation by the MCMC algorithm and a comparison of classification results among three BPDA methods were implemented by coding the R package program. The three methods consist of two proposed $BPDA_{SSMN}$ methods (*i.e.*, $BPDA_{RSN}$ and $BPDA_{RSt}$ for classifying RSN and RSt populations) and $BPDA_N$ by [11] (for classifying unscreened normal populations).

5.1. A Simulation Study: Convergence of the MCMC Algorithm

This simulation study considers inference of the *HSSMN*$(\Theta(k))$ model with a two-dimensional case by generating a training sample of one thousand observations, $n_k = 1000$, from each population π_k, $k = 1, 2, 3$. We considered the following specific choice of parameters, *i.e.*, $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2})^\top$, Σ_{0k}, $\boldsymbol{\mu}_{0k}$, $\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta})$, $\boldsymbol{\lambda}_k = Vec(\Lambda_k^\top) = (\lambda_{k1}, \ldots, \lambda_{k4})^\top$ and $\Psi_k = \Sigma_k - \Delta_k \Sigma_{0k}^{-1} \Delta_k = \{\psi_{kij}\}$ matrices, for generating a synthetic data from π_k,

$$\begin{aligned} \boldsymbol{\mu}_k &= \begin{pmatrix} 1+k \\ -2+k \end{pmatrix}, \ \Sigma_{0k} = \begin{pmatrix} 7+\varepsilon_k & -2 \\ -2 & 4+\varepsilon_k \end{pmatrix}, \ \boldsymbol{\alpha} = \begin{pmatrix} -0.5 \\ -0.5 \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \\ \boldsymbol{\mu}_{0k} &= (0,0)^\top, \ \Delta_k = \begin{pmatrix} 2 & 1 \\ 2 & 0 \end{pmatrix}, \ \Sigma_k = \begin{pmatrix} 3+\varepsilon_k & 0 \\ 0 & 1+\varepsilon_k \end{pmatrix}, \ \text{and } \varepsilon_k = 0.1 \times k. \end{aligned}$$

Based on the above parameter values with $p = q = 2$, we simulated 200 sets of three training samples of each size $n_k = 1000$ from three populations π_k, $k = 1, 2, 3$. Two cases of screened populations were assumed, that is $\pi_k : RSN_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*)$ and $\pi_k : RSt_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*, \nu = 5)$. The respective datasets were generated by using the stochastic representation of each population (see Lemma 1 for the representation). Given a generated training sample, corresponding population parameters were estimated by using the MCMC algorithm based on the *HSSMN*$(\Theta(k))$ model, Equation (12), for each screened population, π_k, distribution. We used $\boldsymbol{\mu}_k = \mathbf{0}$, $\Delta_k = I_2$ and $\Psi_k = I_2$ as the initial values of the MCMC algorithm. To satisfy an objective Bayesian perspective considered by [29], we need to specify the hyper-parameters $(\boldsymbol{\theta}_k, \boldsymbol{\delta}_k, \Omega_k, \Omega_{0k}, H_k, R_k, Q_k, \tau_k, \gamma_k)$ of the *HSSMN*$(\Theta(k))$ model, so as to be insensitive to changes of the priors. Thus, we assumed that we have no information about the parameters. To specify this, we adopted $\boldsymbol{\theta}_k = \mathbf{0}$, $\boldsymbol{\delta}_k = \mathbf{0}$, $\Omega_k = 10^3 I_p$, $\Omega_{0k} = 10^3 I_q$, $H_k = 10^{-3} I_q$, $R_k = 10^{-3} I_p$, $Q_k = 10^{-3} I_q$, $\tau_k = 10^{-3} + p + 1$ and $\gamma_k = 10^{-3} + q + 1$ (see, e.g., [18]).

The MCMC samplers were based on 20,000 iterations as burn-in, followed by a further 20,000 iterations with a thinning size of 10. Thus, the final MCMC samples with a size of 2000 were obtained for each *HSSMN*$(\Theta(k))$ model. Table 1 only provides posterior summaries for the parameters of the π_1 distribution for the sake of saving space. From Column 4–Column 9 of the table list, the mean and three quantiles of 200 sets of posterior samples, which were obtained from the MCMC method, were repeatedly applied to the 200 sets of training sample of size $n_1 = 1000$. Then, the remaining two columns of the table list formal convergence test results of the MCMC algorithm. In estimating the Monte Carlo error (MC error) in Column 5, we used the batch mean method with 50 batches, see e.g., [30] (pp. 39–40). The low values of the MC errors indicate that the variability of each estimate due to the simulation is well controlled. The table also compares the MCMC results with the true parameter values (listed in Column 3): (i) each parameter value in Column 3 is located in the credible interval (2.5% quantile, 97.5% quantile); (ii) for each parameter, we see that the difference between its true value and corresponding posterior mean is less than 2 $\times$ the standard error (s.e.). Thus, the posterior summaries, obtained by using the weakly informative priors, indicate that the MCMC method based on the *HSSMN*$(\Theta(1))$ model performs well in estimating the population parameters, regardless of the SSMN models (RSN and RSt) considered.

Table 1. Posterior summaries of 200 Markov chain Monte Carlo (MCMC) results for the π_1 models.

Model (π_1)	Parameter	True	Mean	MC Error	s.e.	2.5%	Median	97.5%	R_c	p-Value
RSN	μ_{11}	2.000	1.966	0.003	0.064	1.882	1.964	2.149	1.014	0.492
	μ_{12}	−1.000	−0.974	0.002	0.033	−1.023	−0.974	−0.903	1.011	0.164
	λ_{11}	0.312	0.320	0.008	0.159	0.046	0.322	0.819	1.021	0.944
	λ_{12}	0.406	0.407	0.007	0.164	0.030	0.417	0.872	1.018	0.107
	λ_{13}	0.250	0.253	0.004	0.083	0.082	0.256	0.439	1.019	0.629
	λ_{14}	0.125	0.133	0.004	0.067	0.003	0.133	0.408	1.017	0.761
	ψ_{111}	1.968	2.032	0.005	0.130	1.743	2.008	2.265	1.034	0.634
	ψ_{112}	−0.625	−0.627	0.002	0.098	−0.821	−0.617	−0.405	1.022	0.778
	ψ_{122}	0.500	0.566	0.001	0.039	0.465	0.557	0.638	1.018	0.445
RSt	μ_{11}	2.000	2.036	0.004	0.069	1.867	2.050	2.166	1.015	0.251
	μ_{12}	−1.000	−1.042	0.003	0.036	−1.137	−1.054	−0.974	1.012	0.365
	λ_{11}	0.312	0.318	0.008	0.072	0.186	0.320	0.601	1.017	0.654
	λ_{12}	0.406	0.405	0.006	0.074	0.262	0.414	0.562	1.019	0.712
	λ_{13}	0.250	0.255	0.005	0.051	0.113	0.257	0.387	1.023	0.661
	λ_{14}	0.125	0.136	0.005	0.055	0.027	0.133	0.301	1.019	0.598
	ψ_{111}	1.968	1.906	0.006	0.108	1.781	1.996	2.211	1.023	0.481
	ψ_{112}	−0.625	−0.620	0.003	0.101	−0.818	−0.615	−0.422	1.021	0.541
	ψ_{122}	0.500	0.459	0.002	0.044	0.366	0.457	0.578	1.016	0.412

Some of the trace plots from an MCMC run are provided in Figure 1. Each plot demonstrates a parallel zone centered near the true parameter value of interest with no obvious tendency or periodicity. These plots and the small MC error values listed in Table 1 convince us of the convergence of the MCMC algorithm. For a formal diagnostic check, we calculated the Brooks and Gelman diagnostic statistic R_c (adjusted shrinkage factor introduced by [31]) using a MCMC runs with three chains in parallel, each one starting from different initial values. The calculated R_c value for each parameter is listed in the 10th column of Table 1. Table 1 shows that all of the R_c values are close to one, indicating the convergence of the MCMC algorithm. For another formal diagnostic check, we applied the Heidelberger–Welch diagnostic tests of [32] to single-chain MCMC runs, which were used to plot Figure 1. They consist of the stationarity test and the half-width test for the MCMC runs of each parameter. The 11th column of Table 1 lists the p-value of the test for the stationarity of the single Markov chain, where all of the p-values are larger than 0.1. Furthermore, all of the the half-width tests, testing the convergence of the Markov chain of a single parameter, were passed. Thus, all of the diagnostic checking methods (formal and informal methods) advocate the convergence of the proposed MCMC algorithm, and hence, we can say that it generates an MCMC sample that comes from the marginal posterior distributions of interest (*i.e.*, the SSMN population parameters). It is seen that the similar estimation results in Table 1 apply to the posterior summaries of the other parameters in π_2 and π_3 distributions. According to these simulation results, we can say that the MCMC algorithm constructed in Section 3.3 provides an efficient method for estimating the SSMN distributions. To achieve this quality of MCMC algorithm for the higher dimensional case (with large p and/or q values), the diagnostic tests,

considered in this section, should be used to monitor the convergence of the algorithm; for more details, see [30].

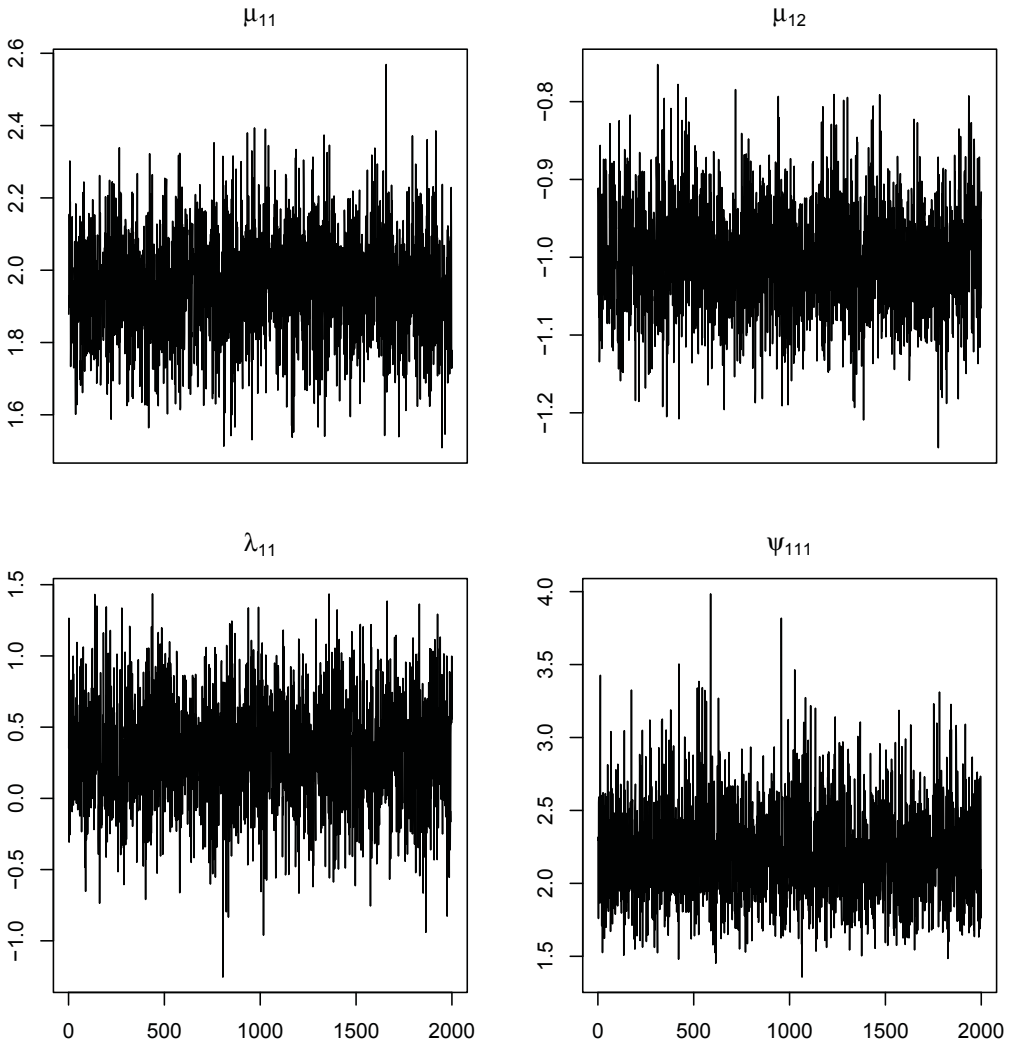

Figure 1. Trace plots of μ_{11}, μ_{12}, λ_{11} and ψ_{111} generated from *HSSMN*$(\Theta(1))$ of the RSt with $\nu = 5$ model.

5.2. A Simulation Study: Performance of the Predictive Methods

This simulation study compares the performance of three BPDA methods using training samples generated from three rectangle screened populations, π_k $(k = 1, 2, 3)$. The three methods compared are $BPDA_{RSN}$, $BPDA_{RSt}$ with degrees of freedom $\nu = 5$, and $BPDA_N$ (a standard predictive method with no screening). Two different cases of rectangle screened population distributions were used to generate the training samples. One case is the rectangle screened population π_k with the $RSN_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*)$ distribution. The other case is π_k with the $RSt_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*, \nu = 5)$ distribution in order to examine the robustness of $BPDA_{RSt}$ in discriminating observations from heavily-tailed empirical distributions. For each case, we obtained 200 sets of training and validation (or testing) samples of each size $n_k = 20, 50, 100$ generated from the rectangle screened distribution of π_k. They are denoted by $\mathcal{D}_k(i)$ and $\mathcal{V}_k(i)(i = 1, \ldots, 200)$. The i-th validation sample $\mathcal{V}_k(i)$ that corresponds

to the training $\mathcal{D}_k(i)$ sample was simply obtained by setting $\mathcal{V}_k(i) = \mathcal{D}_k(i-1)(i = 1, \ldots, 200)$, where $\mathcal{D}_k(0) = \mathcal{D}_k(200)$.

The parameter values of the screened population distributions of the three populations π_k were given by:

$$\boldsymbol{\mu}_k^* = \begin{pmatrix} \mathbf{0}_q \\ \varepsilon(-2+k) \ * \mathbf{1}_p \end{pmatrix}, \ \Sigma_k^* = \begin{pmatrix} I_q & \Delta^\top \\ \Delta & \sqrt{k} * I_p \end{pmatrix}, \ \Delta = \rho J'_{p\times q}, \ \boldsymbol{\alpha} = a\mathbf{1}_q, \ \boldsymbol{\beta} = \mathbf{1}_q$$

for $p = 2, 5$, $q = 2$ and $k = 1, 2, 3$. Further, we assumed that the parameters $\boldsymbol{\mu}_{0k}$ and Σ_{0k} of the underlying q-dimensional screening vector $\boldsymbol{v}_0$ and the rectangle screening region $C_q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ were known as given above. Thus, we may investigate the performance of the BPDA methods by varying the values of correlation ρ, dimension p of the predictor vector, rectangle screened region and differences among the three population means and covariance matrices whose expressions can be found in [4]. Here, $\mathbf{1}_r$ is a $r \times 1$ summing vector whose every element is unity, and $J'_{p\times q}$ denote a $p \times 2$ matrix whose every odd row is equal to (1, 0) and every even row is (0, 1).

Using the training samples, we calculated the approximate predictive densities Equation (27) by the MCMC algorithm proposed in Section 3.3. In this calculation, we assumed that $p_k = 1/3$, because $n_1 = n_2 = n_3 = n$. Thus, the posterior probabilities in Equation (24) and the minimum error classification region R_k in Equation (26) can be estimated within the MCMC scheme, which uses Equation (27) to approximate the predictive densities involved in both Equations (24) and (26). Then, we estimated the classification error rates of the three BPDA methods by using the validation samples, $\mathcal{V}_k(i)(i = 1, \ldots, 200)$. To apply the $BPDA_{RSN}$ and $BPDA_{RSt}$ methods for classifying the simulated training samples, we used the optimal classification rule, which uses Equation (26), while we used the posterior odds ratio given in [11] to implement the $BPDA_N$ method. Then, we compare the classification results in terms of error rates. The error rate of each population (ER_{π_k}) and the total error rate (TotalER) were estimated by:

$$\text{Total}ER = \sum_{k=1}^{3} p_k ER_{\pi_k} \ \text{ and } \ ER_{\pi_k} = \frac{n_k^*}{n_k}, \ k = 1, 2, 3,$$

where n_k^* is the number of misclassified observations out of n_k validation sample observations from π_k.

For each case of π_k distributions, the above procedure was implemented on each set of 200 validation samples to evaluate the error rates of the BPDA methods. Here, [Case 1]denotes that the training (and validation) samples were generated from $\pi_k : RSN_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*)$, and [Case 2] indicates that they were generated from $\pi_k : RSt_p(\mathbf{C}_q(\boldsymbol{\alpha}, \boldsymbol{\beta}); \boldsymbol{\mu}_k^*, \Sigma_k^*, \nu = 5)$, $k = 1, 2, 3$. For each case, Table 2 compares the mean of classification error rates obtained from the 200 replicated classifications by using the BPDA methods. The error rates and their standard errors in Table 2 are indicated as follows. (i) Both the $BPDA_{RSN}$ and $BPDA_{RSt}$ methods work reasonably well in classifying screened observations, compared to the $BPDA_N$ method. This implies that, in BPDA, they provide better classification results than the $BPDA_N$, provided that π_k's are screened by a rectangle screening scheme. (ii) The performance of the $BPDA_{SSMN}$ methods becomes better as the correlation (ρ) between the screening variables and predictor variables becomes larger. (iii) For a comparison of the error rates with respect to the values of a, we see that the $BPDA_{SSMN}$ methods tends to yield better performance in the discrimination of a screened by a small rectangle screened region. (iv) The performance of the three BPDA methods

improves when the differences of the mean increases. (v) An increase in the sizes of dimension p and training sample n also tends to yield a better performance of the BPDA methods. (vi) As expected, the performance of the $BPDA_{RSN}$ in [Case 1] is better than the other two methods, because the estimates of error rates are not covered by the corresponding two standard errors. Further, a considerable gain in the error rates over the $BPDA_N$ manifests the utility of the $BPDA_{RSN}$ in the discriminant analysis. (vii) As for [Case 2], the table indicates that the performance of the $BPDA_{RSt}$ is better than the two other methods. This demonstrates the robustness of the $BPDA_{RSt}$ method in the discrimination with screened and heavy tailed data.

Table 2. Classification error rates: the respective standard errors are in parenthesis.

p	n	a	Method	$\rho = 0.5$		$\rho = 0.9$	
				$\varepsilon = 0.1$	$\varepsilon = 0.4$	$\varepsilon = 0.1$	$\varepsilon = 0.4$
				[Case 1]			
2	20	0.5	$BPDA_{RSN}$	0.322(0.0025)	0.174(0.0022)	0.281(0.0024)	0.106(0.0020)
			$BPDA_{RSt}$	0.335(0.0025)	0.185(0.0023)	0.306(0.0025)	0.115(0.0021)
			$BPDA_N$	0.350(0.0025)	0.206(0.0023)	0.356(0.0025)	0.205(0.0021)
		-0.5	$BPDA_{RSN}$	0.329(0.0027)	0.182(0.0023)	0.301(0.0025)	0.134(0.0021)
			$BPDA_{RSt}$	0.348(0.0024)	0.193(0.0022)	0.319(0.0024)	0.142(0.0021)
			$BPDA_N$	0.349(0.0025)	0.201(0.0023)	.349(0.0025)	0.192(0.0020)
	100	0.5	$BPDA_{RSN}$	0.303(0.0016)	0.161(0.0014)	0.266(0.0015)	0.097(0.0013)
			$BPDA_{RSt}$	0.316(0.0017)	0.165(0.0013)	0.275(0.0015)	0.101(0.0013)
			$BPDA_N$	0.351(0.0025)	0.186(0.0023)	0.356(0.0025)	0.186(0.0021)
		-0.5	$BPDA_{RSN}$	0.306(0.0015)	0.163(0.0014)	0.282(0.0014)	0.116(0.0013)
			$BPDA_{RSt}$	0.318(0.0017)	0.168(0.0015)	0.291(0.0015)	0.121(0.0013)
			$BPDA_N$	0.338(0.0024)	0.172(0.0023)	0.337(0.0026)	0.170(0.0021)
5	20	0.5	$BPDA_{RSN}$	0.318(0.0025)	0.158(0.0022)	0.240(0.0024)	0.101(0.0020)
			$BPDA_{RSt}$	0.327(0.0026)	0.175(0.0023)	0.276(0.0025)	0.114(0.0021)
			$BPDA_N$	0.337(0.0026)	0.183(0.0023)	0.332(0.0025)	0.184(0.0020)
		-0.5	$BPDA_{RSN}$	0.321(0.0025)	0.165(0.0023)	0.231(0.0025)	0.109(0.0021)
			$BPDA_{RSt}$	0.330(0.0026)	0.207(0.0023)	0.318(0.0025)	0.141(0.0021)
			$BPDA_N$	0.345(0.0026)	0.216(0.0024)	0.346(0.0025)	0.218(0.0021)
	100	0.5	$BPDA_{RSN}$	0.280(0.0015)	0.150(0.0014)	0.233(0.0015)	0.084(0.0012)
			$BPDA_{RSt}$	0.291(0.0016)	0.153(0.0015)	0.249(0.0015)	0.092(0.0013)
			$BPDA_N$	0.307(0.0025)	0.186(0.0023)	0.308(0.0025)	0.189(0.0021)
		-0.5	$BPDA_{RSN}$	0.291(0.0016)	0.163(0.0014)	0.239(0.0015)	0.103(0.0013)
			$BPDA_{RSt}$	0.294(0.0016)	0.169(0.0015)	0.253(0.0015)	0.117(0.0013)
			$BPDA_N$	0.305(0.0024)	0.175(0.0022)	0.301(0.0025)	0.176(0.0021)

Table 2. *Cont.*

p	n	a	Method	$\rho = 0.5$		$\rho = 0.9$	
				$\varepsilon = 0.1$	$\varepsilon = 0.4$	$\varepsilon = 0.1$	$\varepsilon = 0.4$
				[Case 2]			
2	20	0.5	$BPDA_{RSN}$	0.351(0.0025)	0.189(0.0022)	0.310(0.0025)	0.114(0.0021)
			$BPDA_{RSt}$	0.320(0.0024)	0.175(0.0023)	0.293(0.0024)	0.105(0.0020)
			$BPDA_{N}$	0.367(0.0026)	0.185(0.0023)	0.365(0.0024)	0.191(0.0020)
		−0.5	$BPDA_{RSN}$	0.349(0.0026)	0.192(0.0022)	0.317(0.0024)	0.149(0.0022)
			$BPDA_{RSt}$	0.321(0.0023)	0.183(0.0021)	0.304(0.0023)	0.132(0.0021)
			$BPDA_{N}$	0.356(0.0025)	0.210(0.0023)	0.357(0.0025)	0.199(0.0020)
	100	0.5	$BPDA_{RSN}$	0.313(0.0016)	0.164(0.0015)	0.273(0.0015)	0.098(0.0014)
			$BPDA_{RSt}$	0.306(0.0015)	0.158(0.0013)	0.265(0.0014)	0.091(0.0012)
			$BPDA_{N}$	0.346(0.0023)	0.179(0.0022)	0.341(0.0024)	0.175(0.0022)
		−0.5	$BPDA_{RSN}$	0.321(0.0015)	0.170(0.0014)	0.287(0.0015)	0.119(0.0015)
			$BPDA_{RSt}$	0.310(0.0014)	0.164(0.0013)	0.281(0.0013)	0.112(0.0013)
			$BPDA_{N}$	0.329(0.0025)	0.181(0.0025)	0.327(0.0027)	0.176(0.0022)
5	20	0.5	$BPDA_{RSN}$	0.329(0.0024)	0.181(0.0024)	0.281(0.0023)	0.119(0.0021)
			$BPDA_{RSt}$	0.317(0.0023)	0.164(0.0020)	0.265(0.0021)	0.094(0.0020)
			$BPDA_{N}$	0.340(0.0027)	0.196(0.0024)	0.314(0.0026)	0.152(0.0022)
		−0.5	$BPDA_{RSN}$	0.342(0.0025)	0.205(0.0024)	0.332(0.0024)	0.194(0.0024)
			$BPDA_{RSt}$	0.328(0.0022)	0.171(0.0022)	0.275(0.0022)	0.118(0.0021)
			$BPDA_{N}$	0.351(0.0026)	0.224(0.0025)	0.329(0.0025)	0.175(0.0025)
	100	0.5	$BPDA_{RSN}$	0.284(0.0016)	0.155(0.0018)	0.283(0.0016)	0.154(0.0013)
			$BPDA_{RSt}$	0.271(0.0014)	0.149(0.0014)	0.238(0.0014)	0.086(0.0011)
			$BPDA_{N}$	0.294(0.0026)	0.192(0.0024)	0.274(0.0026)	0.161(0.0024)
		−0.5	$BPDA_{RSN}$	0.289(0.0016)	0.177(0.0015)	0.288(0.0016)	0.175(0.0013)
			$BPDA_{RSt}$	0.278(0.0013)	0.162(0.0013)	0.231(0.0014)	0.107(0.0011)
			$BPDA_{N}$	0.312(0.0025)	0.178(0.0025)	0.270(0.0026)	0.141(0.0022)

6. Conclusions

In this paper, we proposed an optimal predictive method (BPDA) for the discriminant analysis of multidimensional screened data. In order to incorporate the prior information about a screening mechanism flexibly in the analysis, we introduced the SSMN models. Then, we provided the *HSSMN*$(\Theta(k))$ model for Bayesian inference of the SSMN populations, where the screened data were generated. Based on the *HSSMN*$(\Theta(k))$ model, posterior distributions of $\Theta(k)$ were derived, and the calculation of the optimal predictive classification rule was discussed by using an efficient MCMC method. Numerical studies with simulated screened observations were given to illustrate the convergence of the MCMC method and the usefulness of the BPDA.

The methodological results of the Bayesian estimation procedure proposed in the paper can be extended to other multivariate linear models that incorporate non-normal errors, a general covariance matrix and truncated random covariates. For example, the seemingly unrelated regression (SUR) model and the factor analysis model (see, e.g., [19]) can be explained in the same framework of the proposed *HSSMN*$(\Theta(k))$ in Equation (12). The former is a special case of the *HSSMN*$(\Theta(k))$ model in which

$\mathbf{Z}_k$'s are observable as predictors. Therefore, when the regression errors are non-normal, it would be plausible to apply the proposed approach by using the *HSSMN*$(\Theta(k))$ model to work with a robust SUR model, whereas the latter is a natural extension of the oblique factor analysis model to the case of that with non-normal measurement errors. The *HSSMN*$(\Theta(k))$ model can also be extended to accommodate missing, values as done in the other models by [33,34]. We are hopeful to address these issues, as well, in the near future.

Acknowledgments

The research of Hea-Jung Kim was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2013R1A2A2A01004790).

Author Contributions

The author developed a Bayesian predictive method for the discriminant analysis of screened data. For the multivariate technique, the author introduced a predictive discrimination method with the SSMN populations ($BPDA_{SSMN}$) and provided a Bayesian estimation methodology, which is suited to the $BPDA_{SSMN}$. The methodology consists of constructing a hierarchical model for the SSMN populations (*HSSMN*$(\Theta(k))$) and using an efficient MCMC algorithm to estimate the SSMN models, as well as an optimal rule for the $BPDA_{SSMN}$.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix

Derivations of the full conditional posterior distributions

(1) The full conditional posterior density of $\boldsymbol{\mu}_k$ given $\boldsymbol{\mu}_{0k}, \boldsymbol{\lambda}_k, \Psi_k, F_k, \Sigma_{0k}, \boldsymbol{\eta}_k$ and $\mathcal{D}_k$ is proportional to:

$$\begin{aligned} & \prod_{i=1}^{n_k} \phi_p(\boldsymbol{x}_{ki}; \boldsymbol{\mu}_k + \Lambda_k \boldsymbol{f}_{ki}, \kappa(\eta_{ki})\Psi_k)\phi_p(\boldsymbol{\mu}_k; \boldsymbol{\theta}_k, \Omega_k) \\ \propto \quad & \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{\boldsymbol{\mu}_k})^\top \Sigma_{\boldsymbol{\mu}_k}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{\boldsymbol{\mu}_k})\right\} \end{aligned}$$

which is a kernel of the $N_p(\boldsymbol{\mu}_{\boldsymbol{\mu}_k}, \Sigma_{\boldsymbol{\mu}_k})$ distribution.

(2) It is obvious from the joint posterior density in Equation (13).

(3) It is straightforward to see from Equation (13) that the full conditional posterior density of $\boldsymbol{\lambda}_k$ is given by:

$$\begin{aligned} p(\Lambda_k \mid \Theta(k)_{\setminus \Lambda_k}, \mathcal{D}_k) &\propto \exp\Big\{ -\frac{1}{2}\text{tr}\,[\Psi_k^{-1} V_k]\Big\} \\ &\propto \exp\Big\{ -\frac{1}{2}\text{tr}\,[\Psi_k^{-1}(\Lambda_k - \Lambda_k^*) Q_k (\Lambda_k - \Lambda_k^*)^\top]\Big\} \\ &\propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\lambda}_k - \boldsymbol{\mu}_{\boldsymbol{\lambda}_k})^\top \Sigma_{\boldsymbol{\lambda}_k}^{-1}(\boldsymbol{\lambda}_k - \boldsymbol{\mu}_{\boldsymbol{\lambda}_k})\right\}. \end{aligned}$$

This is a kernel of $N_{pq}(\boldsymbol{\mu}_{\boldsymbol{\lambda}_k}, \Sigma_{\boldsymbol{\lambda}_k})$, where $V_k = (\mathbf{X}_k - \Lambda_k \mathbf{F}_k)\mathbf{D}(\kappa(\boldsymbol{\eta}_k))^{-1}(\mathbf{X}_k - \Lambda_k \mathbf{F}_k)^\top + (\Lambda_k - \Lambda_{0k}) H_k (\Lambda_k - \Lambda_{0k})^\top + R_k$ and $\nu_k = n_k + q + \tau_k$.

(4) We see from Equation (13) that the full conditional posterior density of Ψ_k is given by:

$$\begin{aligned} & p(\Psi_k \mid \Theta(k)_{\setminus \boldsymbol{\Psi}_k}, \mathcal{D}_k) \\ \propto\ & |\Psi_k|^{-(n_k+q)/2} \exp\Big\{ -\frac{1}{2}\text{tr}\,[\Psi_k^{-1}(\mathbf{X}_k - \Lambda_k \mathbf{F}_k)\mathbf{D}(\kappa(\boldsymbol{\eta}_k))^{-1}(\mathbf{X}_k - \Lambda_k \mathbf{F}_k)^\top\Big\} \\ \times\ & \exp\Big\{ -\frac{1}{2}\text{tr}\,[\Psi_k^{-1}(\Lambda_k - \Lambda_{0k}) H_k (\Lambda_k - \Lambda_{0k})^\top]\Big\} \times IW_p(\Psi_k; R_k, \tau_k) \\ \propto\ & |\Psi_k|^{-(n_k+\tau_k+q)/2} \exp\left\{ -\frac{1}{2}\text{tr}\left[\Psi_k^{-1} V_k\right]\right\}. \end{aligned}$$

This is a kernel of $IW_p(V_k, \nu_k)$.

(5) We see, from Equation (13), that the full conditional posterior densities of $\boldsymbol{f}_{ki}$'s are independent, and each density is given by:

$$\begin{aligned} p(\boldsymbol{f}_{ki} \mid \Theta(k)_{\setminus \boldsymbol{f}_{ki}}, \mathcal{D}_k) &\propto \phi_q(\boldsymbol{f}_{ki}; \mathbf{0}, \kappa(\eta_{ki})\Sigma_{0k}) \phi_p(\boldsymbol{x}_{ki}; \boldsymbol{\mu}_k + \Lambda_k \boldsymbol{f}_{ki}, \kappa(\eta_{ki})\Psi_k) \mathrm{I}\big(\boldsymbol{f}_{ki} \in (\boldsymbol{a}_k, \boldsymbol{b}_k)\big) \\ &\propto \exp\Big\{ -\frac{1}{2\kappa(\eta_{ki})}\Big[\boldsymbol{f}_{ki}^\top \Big(\Sigma_{0k}^{-1} + \Lambda_k^\top \Psi_k^{-1} \Lambda_k\Big)\boldsymbol{f}_{ki} \\ &- 2\boldsymbol{f}_{ki}^\top \Lambda_k^\top \Psi_k^{-1}(\boldsymbol{x}_{ki} - \boldsymbol{\mu}_k)\Big]\Big\} \mathrm{I}\big(\boldsymbol{f}_{ki} \in \mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k)\big) \\ &\propto \exp\Big\{ -\frac{1}{2\kappa(\eta_{ki})}(\boldsymbol{f}_{ki} - \boldsymbol{\mu}_{\boldsymbol{f}_{ki}})^\top \Sigma_{\boldsymbol{f}_{ki}}^{-1}(\boldsymbol{f}_{ki} - \boldsymbol{\mu}_{\boldsymbol{f}_{ki}})\Big\} \mathrm{I}\big(\boldsymbol{f}_{ki} \in \mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k)\big) \end{aligned}$$

which is a kernel of the q-variate truncated normal $N_q\Big(\boldsymbol{\mu}_{\boldsymbol{f}_{ki}}, \kappa(\eta_{ki})\Sigma_{\boldsymbol{f}_{ki}}\Big)\mathrm{I}\big(\boldsymbol{f}_{ki} \in \mathbf{C}_q(\boldsymbol{a}_k, \boldsymbol{b}_k)\big)$.

(6) It is obvious from the joint posterior density in Equation (13).

(7) It is obvious from the joint posterior density in Equation (13).

References

1. Catsiapis, G.; Robinson, C. Sample selection bias with multiple selection rules: An application to student aid grants. *J. Econom.* **1982**, *18*, 351–368.
2. Mohanty, M.S. Determination of participation decision, hiring decision, and wages in a double selection framework: Male-female wage differentials in the U.S. labor market revisited. *Contemp. Econ. Policy* **2001**, *19*, 197–212.
3. Kim, H.J. A class of weighted multivariate normal distributions and its properties. *J. Multivar. Anal.* **2008**, *99*, 1758–1771.

4. Kim, H.J.; Kim, H.-M. A class of rectangle-screened multivariate normal distributions and its applications. *Statisitcs* **2015**, *49*, 878–899.
5. Lin, T.I.; Ho, H.J.; Chen, C.L. Analysis of multivariate skew normal models with incomplete data. *J. Multivar. Anal.* **2009**, *100*, 2337–2351.
6. Arellano-Valle, R.B.; Branco, M.D.; Genton, M.G. A unified view of skewed distributions arising from selections. *J. Can. Stat.* **2006**, *34*, 581–601.
7. Kim, H.J. Classification of a screened data into one of two normal populations perturbed by a screening scheme. *J. Multivar. Anal.* **2011**, *102*, 1361–1373.
8. Kim, H.J. A best linear threshold classification with scale mixture of skew normal populations. *Comput. Stat.* **2015**, *30*, 1–28.
9. Marchenko, Y.V.; Genton, M.G. A Heckman selection-t model. *J. Am. Stat. Assoc.* **2012**, *107*, 304–315.
10. Sahu, S.K.; Dey, D.K.; Branco, M.D. A new class of multivariate skew distributions with applications to Bayesian regession models. *Can. J. Stat.* **2003**, *31*, 129–150.
11. Geisser, S. Posterior odds for multivariate normal classifications. *J. R. Stat. Soc. B* **1964**, *26*, 69–76.
12. Lachenbruch, P.A.; Sneeringer, C.; Revo, L.T. Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Commun. Stat.* **1973**, *1*, 39–57.
13. Wang, Y.; Chen, H.; Zeng, D.; Mauro, C.; Duan, N.; Shear, M.K. Auxiliary marke-assited classification in the absence of class identifiers. *J. Am. Stat. Assoc.* **2013**, *108*, 553–565.
14. Webb, A. *Statistical Pattern Recognition*; Wiley: New York, NY, USA, 2002.
15. Aitchison, J.; Habbema, J.D.F.; Key, J.W. A critical comparison of two methods of statistical discrimination. *Appl. Stat.* **1977**, *26*, 15–25.
16. Azzalini, A.; Capitanio, A. Statistical application of the multivariate skew-normal distribution. *J. R. Stat. Soc. B* **1999**, *65*, 367–389.
17. Branco, M.D. A general class of multivariate skew-elliptical distributions. *J. Multivar. Anal.* **2001**, *79*, 99–113.
18. Chen, M-H.; Dey, D.K. Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions. *Indian J. Stat.* **1998**, *60*, 322–343.
19. Press, S.J. *Applied Multivariate Analysis*, 2nd ed.; Dover: New York, NY, USA, 2005.
20. Reza-Zadkarami, M.; Rowhani, M. Application of skew-normal in classification of satellite image. *J. Data Sci.* **2010**, *8*, 597–606.
21. Wang, W.L.; Fan, T.H. Bayesian analysis of multivariate t linear mixed models using a combination of IBF and Gibbs samplers. *J. Multivar. Anal.* **2012**, *105*, 300–310.
22. Wang, W.L.; Lin, T.I. Bayesian analysis of multivariate t linear mixed models with missing responses at random. *J. Stat. Comput. Simul.* **2015**, *85*, doi:10.1080/00949655.2014.989852.
23. Johnson, N.L.; Kotz, S. *Distribution in Statistics: Continuous Multivariate Distributions*; Wiley: New York, NY, USA, 1972.
24. Wilhelm, S.; Manjunath, B.G. tmvtnorm: Truncated multivariate normal distribution and student t distribution. *R J.* **2010**, *1*, 25–29.

25. Genz, A.; Bretz, F. *Computation of Multivariate Normal and t Probabilities*; Springer: New York, NY, USA, 2009.
26. Chib, S.; Greenberg, E. Understanding the Metropolis-Hastings algorithm. *Am. Stat.* **1995**, *49*, 327–335.
27. Chen, H.-M.; Schmeiser, R.W. Performance of the Gibbs, hit-and-run, and metropolis samplers. *J. Comput. Gr. Stat.* **1993**, *2*, 251–272.
28. Anderson, T.W. *Introduction to Multivariate Statistical Analysis*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2003.
29. Adwards, W.H.; Lindman, H.; Savage, L.J. Bayesian statistical inference for psychological research. *Psycol. Rev.* **1963**, *70*, 192–242.
30. Ntzoufras, I. *Bayesian Modeling Using WinBUGS*; Wiley: New York, NY, USA, 2009.
31. Brooks, S.; Gelman, A. Alternative methods for monitoring convergence of iterative simulations. *J. Comput. Gr. Stat.* **1998**, *7*, 434–455.
32. Heidelberger, P.; Welch, P. Simulation run length control in the presence of an initial transient. *Oper. Res.* **1992**, *31*, 1109–1144.
33. Lin, T.I. Learning from incomplete data via parameterized t mixture models through eigenvalue decomposition. *Comput. Stat. Data Anal.* **2014**, *71*, 183–195.
34. Lin, T.I.; Ho, H.J.; Chen, C.L. Analysis of multivariate skew normal models with incomplete data. *J. Multivar. Anal.* **2009**, *100*, 2337–2351.