Bayesian Classification of Multiclass Functional Data

Xiuqi Li
Department of Operations Research
North Carolina State University
Raleigh, NC 27695
U.S.A.
e-mail: xli35@ncsu.edu

and

Subhashis Ghosal
Department of Statistics
North Carolina State University
Raleigh, NC 27695
U.S.A.
e-mail: sghosal@stat.ncsu.edu

Abstract: We propose a Bayesian approach to estimating parameters in multiclass functional models. Unordered multinomial probit, ordered multinomial probit and multinomial logistic models are considered. We use finite random series priors based on a suitable basis such as B-splines in these three multinomial models, and classify the functional data using the Bayes rule. We average over models based on the marginal likelihood estimated from MCMC output. Posterior contraction rates for the three multinomial models are computed. We also consider linear and quadratic discriminant analysis, and apply functional principal component analysis to reduce the dimension. A simulation study is conducted to compare these methods on different types of data.

Keywords and phrases: Multiclass functional data, Multinomial probit models, B-splines, Posterior contraction rate, Discriminant analysis.

1. Introduction

Functional data analysis is increasingly drawing attentions in many areas, such as biomedicine, environment, and economics (Ullah and Finch, 2013). Classification of functional data, especially when the data units can come from more than two categories, is a fundamental problem of interest.

Generalized linear models are often used to classify the functional data (Müller and Stadtmüller, 2005; James, 2002). Zhu, Brown, and Morris (2012) proposed a functional mixed model to classify functional data. Linear discriminant analysis is also fitted to functional data classification (James and Hastie, 2001). Partial least squares regression on functional data is proposed for linear discriminant analysis (Preda, Saporta, and Lévéder, 2007). There are also nonparametric approaches for functional data classification (Biau, Bunea, and Wegkamp, 2005; Ferraty and Vieu, 2003). F. Rossi and Villa (2006) adapted support vector machine to functional data classification. Wavelets approaches are also applied to classify and cluster functional data (Ray and Mallick, 2006).
In this paper, we consider a polytomous response $Y$ taking values $k = 1, \ldots, K$, with functional covariate $\{X(t), t \in [0,1]\}$. The main problem is to estimate the probability $P(Y = k | X)$, which can be conveniently modeled by a linear function of $\int \beta(t)X(t)dt$

$$P(Y = k | X) = H \left( \int \beta(t)X(t)dt \right),$$

where $H$ is a cumulative distribution function, and $\beta(\cdot)$ is an unknown coefficient function. Unordered multinomial probit, ordered multinomial probit and multinomial logistic models are considered in this paper which correspond to different choices of $H$. For ordered multinomial probit model, there are additional order restrictions. Finite random series priors (Shen and Ghosal, 2015) are applied to the three multinomial models. As a comparison, we consider Bayesian linear and quadratic discriminant analysis. Functional principal components give a dimension reduction in the two discriminant analyses. Following a Bayesian approach, the posterior distribution of the parameters are obtained using the training data, and then classification rules are applied on the test data using posterior probability of class membership.

The paper is organized as follow. In Section 2, the three functional multinomial models are described. Section 3 gives description of applying the finite random series prior on these models. The model averaging methods are described in Section 4. In Section 5, the posterior contraction rates of the three functional multinomial models with finite random series prior are computed. As a comparison, Section 6 describes the Bayesian discriminant analysis. In Section 7, a simulation study is conducted on different types of data, and B-splines bases for finite random series priors are applied. In Section 8, the three multinomial models and Bayesian discriminant analysis are tested on the real data.

2. Model

2.1. Ordered Multinomial Probit Model

Let $X_i(t), i = 1, \ldots, N, t \in [0,1]$, be the observed functional data associated with a categorical variable $Y_i$ taking possible values $1, \ldots, K$. We assume that $(X_i, Y_i), i = 1, \ldots, N$ are independent and identically distributed (i.i.d) observations.

Following Albert and Chib (1993), we consider the model described implicitly as follow: there exists a latent variable $W_i$ distributed as $N(\int \beta(t)X_i(t)dt, 1)$, for $i = 1, \ldots, N$, and that $Y_i = k$ if $\gamma_{k-1} < W_i \leq \gamma_k$, where $k = 1, \ldots, K, \gamma_0 = -\infty, \gamma_K = \infty$, and $\beta(\cdot)$ is an unknown coefficient function. To ensure identifiability, we set $\gamma_1 = 0$. Under the assumed model, the probability of choosing a category $k$ is given by

$$P(Y_i = k | X_i) = \Phi \left( \gamma_k - \int \beta(t)X_i(t)dt \right) - \Phi \left( \gamma_{k-1} - \int \beta(t)X_i(t)dt \right),$$

where $\Phi$ stands for the distribution function of the standard normal distribution.
2.2. Unordered Multinomial Probit Model

The unordered multinomial probit model can be described by the following data augmentation method. As in Albert and Chib (1993), let $W'_{i} = (W'_{i1}, \ldots, W'_{iK})^T, i = 1, \ldots, N$, be latent varible, such that $W'_{il}$ follows a linear model

$$W'_{il} = \int \beta'_l(t)X_i(t)dt + \epsilon'_{il}, \quad (3)$$

where $\epsilon'_{il} \sim N(0, 1), i = 1, \ldots, N, l = 1, \ldots, K$, are i.i.d. standard normal variables.

Consider the latent variables $W_i = (W_{i1}, \ldots, W_{iK-1})^T, W_{il} = W'_{il} - W'_{iK}$,

$$W_{il} = \int \beta'_l(t)X_i(t)dt - \int \beta'_K(t)X_i(t)dt + \epsilon_{il}, \quad (4)$$

where $\epsilon_{il} = \epsilon'_{il} - \epsilon'_{iK}$, and $l = 1, \ldots, K - 1$. Let $\epsilon_{i} = (\epsilon_{i1}, \ldots, \epsilon_{iK-1})^T$. Then $\epsilon_{i}$ follows $N(0, \Sigma)$, where $\Sigma$ is a $(K-1) \times (K-1)$ matrix with 2 at diagonal entries and 1 at all off-diagonal entries.

The probability of choosing the $k$th ($k = 1, \ldots, K-1$) alternative is given by

$$P(Y_i = k|X_i) = P(W_{ik} > W_{il}, \text{ for all } l \neq k, \text{ and } W_{ik} > 0), \quad (5)$$

and the probability of choosing alternative $K$ is given by

$$P(Y_i = K|X_i) = P(W_{il} < 0 \text{ for all } l = 1, \ldots, K - 1). \quad (6)$$

2.3. Multinomial Logistic Model

In this model, the probability of choosing category $k$ is given by

$$P(Y_i = k|X_i) = \frac{\exp[\int \beta_k(t)X_i(t)dt]}{\sum_{l=1}^{K-1} \exp[\int \beta_l(t)X_i(t)dt]}, \quad (7)$$

To ensure model identification, set $\beta_K(t) = 0$. Then the probability of choosing category $k$ ($k = 1, \ldots, K - 1$) is given by

$$P(Y_i = k|X_i) = \frac{\exp[\int \beta_k(t)X_i(t)dt]}{1 + \sum_{l=1}^{K-1} \exp[\int \beta_l(t)X_i(t)dt]}, \quad (8)$$

and the probability of choosing category $K$ is given by

$$P(Y_i = k|X_i) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp[\int \beta_l(t)X_i(t)dt]}, \quad (9)$$
3. Finite Random Series Prior

The functional parameter $\beta(t)$ is given a prior which is a finite linear combination of a certain chosen basis functions: $\beta(t) = \sum_{j=1}^{J} \theta_j \psi_j(t)$, where $\{\psi_1(t), \ldots, \psi_J(t)\}$ is a basis, for example, B-splines, Fourier, or wavelets etc, and $(\theta_1, \ldots, \theta_J)$ is given a prior distribution.

The advantage of a using finite random series prior is that the inner product between the functional paramenter and the functional data $\int \beta(t)X_i(t)dt$ is reduced to a simple linear combination

$$\int \beta(t)X_i(t)dt = \int \sum_{j=1}^{J} \theta_j \psi_j(t)X_i(t)dt = \sum_{j=1}^{J} \theta_j Z_{ij},$$

where $Z_{ij} = \int \psi_j(t)X_i(t)dt$.

3.1. Ordered Multinomial Probit Model

Using a finite random series $\beta(t) = \sum_{j=1}^{J} \theta_j \psi_j(t)$, the model in (2) can be rewritten as

$$P(Y_i = k|X_i) = \Phi\left(\gamma_k - \sum_{j=1}^{J} \theta_j Z_{ij}\right) - \Phi\left(\gamma_{k-1} - \sum_{j=1}^{J} \theta_j Z_{ij}\right),$$

where $Z_{ij} = \int \psi_j(t)X_i(t)dt$.

Define $\theta = (\theta_1, \ldots, \theta_J)^T$, and $Z_i = (Z_{i1}, \ldots, Z_{iJ})^T$. Then (11) can be written compactly as

$$P(Y_i = k|X_i) = \Phi(\gamma_k - Z_i^T \theta) - \Phi(\gamma_{k-1} - Z_i^T \theta).$$

Clearly the unobserved latent variable $W_i$ follows $N(Z_i^T \theta, 1)$. Assign $\theta$ a conjugate prior $N(\theta_0, B_0)$. Then the posterior distribution of $\theta$ is given by

$$\theta|Y, W \sim N(\theta_n, B_n), \quad B_n = (B_0^{-1} + Z^T Z)^{-1}, \quad \theta_n = B_n(B_0^{-1} \theta_0 + Z^T W),$$

where $Z = (Z_1^T, \ldots, Z_N^T)^T$, and $W = (W_1, \ldots, W_N)^T$.

We follow the scheme introduced by Albert and Chib (1993). The posterior distribution of $W_i$ is given by

$$W_i|(\theta, \gamma, Y_i = k) \sim TN(Z_i^T \theta, 1, \gamma_{k-1}, \gamma_k),$$

where $TN(Z_i^T \theta, 1, \gamma_{k-1}, \gamma_k)$ is the truncation of the normal distribution with mean $Z_i^T \theta$ and variance 1 at the interval $(\gamma_{k-1}, \gamma_k)$.

Albert and Chib (1993) assigned a diffuse prior on the cut-points. However, model averaging needs a proper prior. A normal prior is not appropriate due to the order restriction on $\gamma_1, \ldots, \gamma_K$. Albert and Chib (1997) proposed a transformation of $\gamma = (\gamma_1, \ldots, \gamma_K)$ which avoids the order restriction.

$$\alpha_1 = \log(\gamma_2), \quad \alpha_j = \log(\gamma_{j+1} - \gamma_j), \quad 2 \leq j \leq K - 2.$$
Note that $\gamma_1 = 0$ and by the inverse map

$$\gamma_j = \sum_{i=1}^{j-1} \exp(\alpha_i), \quad 2 \leq j \leq K - 1. \quad (16)$$

Then $\gamma$ can be reparameterized by $\alpha = (\alpha_1, \ldots, \alpha_{K-2})$. Assign a multivariate normal prior with mean $\alpha_0$ and covariance $A_0$ on $\alpha$. To sample $\gamma$, apply the following steps of Metropolis-Hastings algorithm.

1. Sample $\alpha'$ from the proposal distribution $q(\alpha', \alpha|Y, \theta, W)$. Here we allow the proposal density to depend on the data and the two remaining blocks for the convenience of computing the marginal likelihood in the future.

2. Move to $\alpha'$ from the current $\alpha$ with probability

$$m(\alpha, \alpha'|Y, \theta, W) = \min \left\{ \frac{f(Y|\alpha', \theta, W) \pi(\alpha', \theta)}{f(Y|\alpha, \theta, W) \pi(\alpha, \theta)} \cdot \frac{q(\alpha', \alpha|Y, \theta, W)}{q(\alpha, \alpha'|Y, \theta, W)} , 1 \right\}. \quad (17)$$

3. Compute $\gamma$ by the inverse map (16).

To implement the MCMC sampling, first draw $\gamma$ by the above steps. Then sample from the posterior distributions (14) and (13).

### 3.2. Unordered Multinomial Probit Model

Let $\beta_l(t) = \sum_{j=1}^{J} \theta_{lj} \psi_j(t)$, where $l = 1, \ldots, K$. Then (4) can be rewritten as

$$W_{il} = \sum_{j=1}^{J} \theta_{lj} Z_{ij} - \sum_{j=1}^{J} \theta'_{Kj} Z_{ij} + \epsilon_{il} = \sum_{l=1}^{J} (\theta'_{il} - \theta'_{Kl}) Z_{ij} + \epsilon_{il}, \quad (18)$$

where $Z_{ij} = \int \psi_j(t) X_i(t) dt$.

Let $\theta_{lj} = \theta'_{lj} - \theta'_{Kj}$, where $j = 1, \ldots, J$. Define $\theta_l = (\theta_{l1}, \ldots, \theta_{lJ})^T$, and $Z_i = (Z_{i1}, \ldots, Z_{iJ})^T$. Then (18) is given by

$$W_{il} = Z_i^T \theta_l + \epsilon_{il}, \quad (19)$$

where $i = 1, \ldots, N, l = 1, \ldots, K - 1$.

Define a $J \times (K - 1)$ matrix $\Theta = (\theta_1, \ldots, \theta_{K-1})$. Then we have $W_i = Z_i^T \Theta + \epsilon_i$, where $W_i = (W_{i1}, \ldots, W_{iK-1})^T$, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iK-1})^T$, and $\epsilon_i \sim N(0, \Sigma)$.

In the model described in Section 2, $\Sigma$ is known with 2 on diagonal entries and 1 on all off-diagonal entries. The only parameter needs to be estimated is $\Theta$. In order to draw the matrix $\Theta$ using the Gibbs sampling, we can stack the data in a matrix form which is given by

$$W = Z \Theta + \epsilon, \quad (20)$$

where $W = (W_1^T, \ldots, W_N^T)^T$ is an $N \times (K - 1)$ matrix, $Z = (Z_1^T, \ldots, Z_N^T)^T$ is an $N \times J$ matrix, and $\epsilon = (\epsilon_1^T, \ldots, \epsilon_N^T)^T$ is an $N \times (K - 1)$ matrix.
This results in a matrix normal distribution. The density function of matrix normal distribution $\text{MN}_{n \times p}(M, U, V)$ is

$$(2\pi)^{-np/2}|V|^{-n/2}|U|^{-p/2}\exp\left(-\frac{1}{2}\text{tr}[V^{-1}(X - M)^TU^{-1}(X - M)]\right),$$

(21)

where $M$ is an $n \times p$ mean matrix, $U$ is an $n \times n$ row variance matrix, $V$ is a $p \times p$ column variance matrix, $\text{tr}$ represents the trace of a matrix, and $|U|$ and $|V|$ represent the determinants of $U$ and $V$ respectively.

Thus $W|\Theta \sim \text{MN}_{N \times (K-1)}(Z\Theta, I_N, \Sigma)$. Here the row variance-covariance matrix $I_N$ is an identity matrix of rank $N$, since $W_1, \ldots, W_N$ are independent. We consider the matrix normal prior $\Theta \sim \text{MN}_{J \times (K-1)}(U_0, V_0, \Sigma)$. Then the posterior is given by

$$\Theta|Y, W \sim \text{MN}_{J \times (K-1)}(U_n, V_n, \Sigma),$$

$$V_n = (Z^TZ + V_0)^{-1}, U_n = V_n(Z^TW + V_0^{-1}U_0).$$

(22)

To draw a sample of $W$, we use the method introduced by McCulloch and P. E. Rossi (1994). Let $W_{i, -l}$ denote $(W_{i, 1}, \ldots, W_{i, l-1}, W_{i, l+1}, \ldots, W_{i, K-1})^T$, $Z_{i, .}$ denote the $i$th row of $Z$, the vector $\Theta_{.-l}$ denote the $l$th column of $\Theta$, the matrix $\Theta_{.-l,-l}$ denote $\Theta$ without the $l$th column, the scalar $\Sigma_{l,l}$ denote the $(l,l)$th entry of $\Sigma$, $\Sigma_{l,-l}$ denote $\Sigma$ without the $l$th row and the $l$th column, $\Sigma_{.-l,-l}$ denote $\Sigma$ without the $l$th row and the $l$th column, and $\Sigma_{l,-l}$ denote $\Sigma$ without the $l$th entry. We draw $W_{il}$ from the conditional truncated normal distribution described below:

$$W_{il}|(W_{i, -l}, \Theta, Y_i) \sim \text{TN}(m_{il}, \tau_{il}^2, a, b),$$

$$m_{il} = Z_{i, l} + \Sigma_{l,-l}^{-1}(W_{i, -l} - Z_{i, -l}),$$

$$\tau_{il}^2 = \Sigma_{l,l} - \Sigma_{l,-l}\Sigma_{l,-l}^{-1}\Sigma_{l,l},$$

$$a, b = \begin{cases} 
(\max\{W_{i,-l}, 0\}, \infty), & \text{if } Y_i = l, l = 1, \ldots, K - 1 \\
(-\infty, \max\{W_{i,-l}\}), & \text{if } Y_i \neq l, l = 1, \ldots, K - 1, \\
(-\infty, 0), & \text{if } Y_i = K 
\end{cases}$$

(23)

$$i = 1, \ldots, N, l = 1, \ldots, K - 1.$$

To implement the Gibbs sampling, we draw samples from (22) and (23).

### 3.3. Multinomial Logistic Model

Let $\beta_k(t) = \sum_{j=1}^{J} \theta_{kj} \psi_j(t)$. Then (8) and (9) can be rewritten as

$$P(Y_i = k|X_i) = \frac{\exp[\sum_{j=1}^{J} \theta_{kj} Z_{ij}]}{1 + \sum_{l=1}^{K-1} \exp[\sum_{j=1}^{J} \theta_{lj} Z_{ij}]}, \quad k = 1, \ldots, K - 1,$$

(24)

$$P(Y_i = K|X_i) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp[\sum_{j=1}^{J} \theta_{lj} Z_{ij}]}.$$

(25)
where $Z_{ij} = \int \psi_j(t) X_i(t) dt$.

Define $\theta_k = (\theta_{k1}, \ldots, \theta_{kJ})^T, k = 1, \ldots, K - 1$, and $Z_i = (Z_{i1}, \ldots, Z_{ij})^T$. Then (24) and (25) are given by

$$P(Y_i = k|X_i) = \frac{\exp[Z_i^T \theta_k]}{1 + \sum_{l=1}^{K-1} \exp[Z_l^T \theta_l]}, k = 1, \ldots, K - 1,$$

(26)

and

$$P(Y_i = K|X_i) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp[Z_l^T \theta_l]},$$

(27)

For each $\theta_k, k = 1, \ldots, K - 1$, we assign a multivariate normal prior $N(\mu_k, \Sigma_k)$, and apply Metropolis-Hastings algorithm to sample $\theta_k$.

1. Sample $\theta'_k$ from the proposal distribution $q(\theta'_k, \theta_k|Y, \theta_{-k})$.
2. Move to $\theta'_k$ from the current $\theta_k$ with probability

$$m(\theta_k, \theta'_k|Y, \theta_{-k}) = \min \left\{ \frac{f(Y|\theta'_k, \theta_{-k})\pi(\theta'_k, \theta_{-k})}{f(Y|\theta_k, \theta_{-k})\pi(\theta_k, \theta_{-k})}, 1 \right\},$$

(28)

where $\theta_{-k}$ denotes all the blocks except the $k$th one.

4. Model Averaging

For the three multinomial models, we put the finite random series priors on the functional coefficients. The number of basis function $J$ in these priors is unknown. However, for computational purposes $J$ should take a finite number of values: $\{J_1, \ldots, J_M\}$. For each given $J_m, m = 1, \ldots, M$, we have the misclassification rate $r_m$. If we can get the marginal likelihood $m(Y|J_m)$, then we can compute the average misclassification rate $\bar{r}$ for each multinomial model.

$$\bar{r} = \frac{M}{\sum_{m=1}^{M} m(Y|J_m)} \cdot r_m,$$

(29)

The marginal likelihood can be written as the normalizing constant of the posterior density

$$m(Y|J_m) = \frac{f(Y|J_m, B)\pi(B|J_m)}{\pi(B|Y, J_m)},$$

(30)

where $B$ is a set of parameter blocks, and (30) holds for any $B$. The numerator is the product of the likelihood and the prior. The denominator is the posterior density of $B$. For a given $B^*$, the posterior density $\pi(B^*|Y, J_m)$ can be estimated from the Gibbs output (Chib, 1995), and the Metropolis-Hasting output (Chib and Jeliazkov, 2001). Then the estimated marginal likelihood in the logarithm scale is

$$\log \hat{m}(Y|J_m) = \log f(Y|J_m, B^*) + \log \pi(B^*|J_m) - \log \hat{\pi}(B^*|Y, J_m).$$

(31)
4.1. Ordered Multinomial Probit Model

There are two parameter blocks in this model, $\theta$ and $\alpha$, where $\alpha$ is the transformation of $\gamma$ as in (15). Given $\theta^* = G^{-1} \sum_{g=1}^G \theta(g)$, and $\alpha^* = G^{-1} \sum_{g=1}^G \alpha(g)$, where $\{\theta(g), \alpha(g)\}_{g=1}^G$ are from the MCMC output, the joint posterior density can be written as

$$\pi(\theta^*, \alpha^* | Y, J_m) = \pi(\alpha^* | Y, J_m) \pi(\theta^* | Y, J_m, \alpha^*),$$  

(32)

where

$$\pi(\theta^* | Y, J_m, \alpha^*) = \int \pi(\theta^* | Y, J_m, \alpha^*, W) \pi(W | Y, J_m, \alpha^*) dW.$$  

(33)

The Monte Carlo estimate of $\pi(\theta^* | Y, J_m, \alpha^*)$ is

$$\hat{\pi}(\theta^* | Y, J_m, \alpha^*) = M^{-1} \sum_{m=1}^M \pi(\theta^* | Y, J_m, \alpha^*, W^{(m)}),$$  

(34)

where $\{W^{(m)}\}_{m=1}^M$ are sampled from distribution $[W | Y, J_m, \alpha^*]$. The draws of $W$ from the Gibbs sampler are from the distribution $[W | Y, J_m]$, so $\pi(\theta^* | Y, J_m, \alpha^*, W)$ cannot be averaged directly by the Gibbs sampling output. Additional sampling for $W$ is needed. We sample $\{\theta^{(m)}\}$ from the density $\pi(\theta | Y, J_m, \alpha^*, W)$, and given that $\theta^{(m)}$, we sample $\{W^{(m)}\}$ from $\pi(W | Y, J_m, \theta, \alpha^*)$.

The explicit distribution of $\alpha^*$ given $(Y, J_m)$ is unknown, and hence the draws of $\alpha$ are obtained from a Metropolis-Hastings sampling. By the local reversibility condition (see Chib and Jeliazkov, 2001 for details), the posterior density of $\alpha$ can be written as

$$\pi(\alpha^* | Y, J_m) = \frac{E_1\{m(\alpha, \alpha^* | Y, J_m, \theta, W) q(\alpha^* | Y, J_m, \theta, W)\}}{E_2\{m(\alpha^*, \alpha | Y, J_m, \theta, W)\}},$$  

(35)

where $m(\alpha, \alpha^* | Y, J_m, \theta, W)$ is defined in (17), $q(\alpha^* | Y, J_m, \theta, W)$ is the proposal density, the expectation $E_1$ is with respect to the distribution $\pi(\theta, W | Y, J_m)$, and $E_2$ is with respect to the distribution $\pi(\theta, W | Y, J_m, \alpha^*) \times q(\alpha^*, \alpha | Y, J_m, \theta, W)$.

Then the estimate of $\pi(\alpha^* | Y, J_m)$ is given by

$$G^{-1} \sum_{g=1}^G m(\alpha(g), \alpha^* | Y, J_m, \theta(g), W(g)) q(\alpha(g), \alpha^* | Y, J_m, \theta(g), W(g)) \frac{1}{M^{-1} \sum_{m=1}^M m(\alpha^*, \alpha^* | Y, J_m, \theta^{(m)}, W^{(m)})},$$  

(36)

where $\{\theta(g), \alpha(g), W(g)\}_{g=1}^G$ are obtained from the MCMC output. $\{\theta^{(m)}, W^{(m)}\}$ are obtained from $\pi(\theta | Y, J_m, \alpha^*, W)$ and $\pi(W | Y, J_m, \theta, \alpha^*)$, and then given $\{\theta^{(m)}, W^{(m)}\}$, sample $\alpha^{(m)}$ from $q(\alpha^*, \alpha | Y, J_m, \theta^{(m)}, W^{(m)})$.

4.2. Unordered Multinomial Probit Model

The only unknown parameter is $\Theta$. Given $\Theta^* = G^{-1} \sum_{g=1}^G \Theta^{(g)}$, where $\{\Theta^{(g)}\}$ are from the Gibbs sampling output, the posterior density of $\Theta$ at $\Theta^*$ can be written as

$$\pi(\Theta^* | Y, J_m) = \int \pi(\Theta^* | Y, J_m, W) \pi(W | Y, J_m) dW.$$

(37)

---

Li and Ghosal/Functional Data Classification
Then the Monte Carlo estimate of \( \pi(\Theta^*|Y, J_m) \) is

\[
\hat{\pi}(\Theta^*|Y, J_m) = \sum_{g=1}^{G} \pi(\Theta^*|Y, J_m, W^{(g)}),
\]

(38)

where \( \{W^{(g)}\}^G_{g=1} \) are from the Gibbs sampling output.

From Section 3.2, \( \Theta = (\theta_1, \ldots, \theta_{K-1}) \), where \( \theta_l = \theta'_l - \theta'_{K}, \ l = 1, \ldots, K-1 \).

Then (6) can be rewritten as

\[
P(Y = K)
= \frac{1}{(2\pi)^{(K-1)/2} |\Sigma|^{1/2}} \int_{-\infty}^{-Z^T \Theta_{-1,1}} \cdots \int_{-\infty}^{-Z^T \Theta_{-1,K-1}} \exp\left(-\frac{1}{2} U^T \Sigma^{-1} U\right) dU,
\]

(39)

where \( \Theta_{l,1} \) denotes the \( l \)th column of \( \Theta \).

For \( l \neq K \), let \( \Theta^l = (\theta_l - \theta_1, \ldots, \theta_{l-1} - \theta_l, \theta_{l+1} - \theta_l, \ldots, \theta_{K-1} - \theta_l, -\theta_l) \), then

\[
P(Y = l)
= \frac{1}{(2\pi)^{(K-1)/2} |\Sigma|^{1/2}} \int_{-\infty}^{-Z^T \Theta^l_{1,1}} \cdots \int_{-\infty}^{-Z^T \Theta^l_{1,K-1}} \exp\left(-\frac{1}{2} U^T \Sigma^{-1} U\right) dU.
\]

(40)

Due to the exchangeable correlation structure of \( \Sigma \), (40) can be reduced to a one dimensional integral (Dunnett, 1989) given by

\[
P(Y = l)
= \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} \left\{ \prod_{k=1}^{K-1} \Phi(-u\sqrt{2} - Z^T \Theta^l_{1,k}) + \prod_{k=1}^{K-1} \Phi(u\sqrt{2} - Z^T \Theta^l_{1,k}) \right\} e^{-u^2} du.
\]

(41)

The expression in (39) can also be reduced to the same form as in (41). Then (41) can be approximated by Gaussian quadrature as follow:

\[
P(Y = l) \approx \frac{1}{w_q} \left\{ \prod_{k=1}^{K-1} \Phi(-\sqrt{2x_q} - Z^T \Theta^l_{1,k}) + \prod_{k=1}^{K-1} \Phi(\sqrt{2x_q} - Z^T \Theta^l_{1,k}) \right\},
\]

(42)

where \( w_q \) and \( x_q \) are the weights and roots of the Laguerre polynomial of order \( Q \).

The likelihood of this unordered multinomial probit model can be approximated using (42).

### 4.3. Multinomial Logistic Model

There are \( K - 1 \) unknown parameters: \( \theta_1, \ldots, \theta_{K-1} \). Given \( \theta_k^* = G^{-1} \sum_{g=1}^{G} \theta_k^{(g)} \), \( k = 1, \ldots, K-1 \), where \( \{\theta_k^{(g)}\}^G_{g=1} \) are from the Metropolis-Hastings sampling output, the joint posterior density can be written as

\[
\pi(\theta_1^*, \ldots, \theta_{K-1}^*|Y, J_m) = \prod_{i=1}^{K-1} \pi(\theta_i|Y, J_m, \theta_1^*, \ldots, \theta_{i-1}^*).
\]

(43)
By the local reversibility, each full conditional density can be written as
\[
\pi(\theta_i|Y, J_m, \theta_i^*, \ldots, \theta_{i-1}^*) = \frac{E_1\{m(\theta_i, \theta_i^*|Y, J_m, \Psi_{i-1}^*, \Psi_{i+1}^*)q(\theta_i^*, \theta_i^*|Y, J_m, \Psi_{i-1}^*, \Psi_{i+1}^*)\}}{E_2\{m(\theta_i^*, \theta_i^*|Y, J_m, \Psi_{i-1}^*, \Psi_{i+1}^*)\}},
\]
where \(\Psi_{i-1} = (\theta_1, \ldots, \theta_{i-1}), \Psi_{i+1} = (\theta_{i+1}, \ldots, \theta_K-1)\), \(E_1\) is the expectation with respect to the distribution \(\pi(\theta_i, \Psi_{i+1}|Y, J_m, \Psi_{i-1}^*)\), and \(E_2\) is that with respect to \(\pi(\Psi_{i+1}|Y, J_m, \Psi_{i-1}^*, \theta_i^*)\).

Then the estimate of \(\pi(\theta_i|Y, J_m, \theta_i^*, \ldots, \theta_{i-1}^*)\) is given by
\[
\hat{\pi}(\theta_i|Y, J_m, \theta_i^*, \ldots, \theta_{i-1}^*) = \frac{G^{-1} \sum_{g=1}^{G} m(\theta_i^{(g)}, \theta_i^*|Y, J_m, \Psi_{i-1}^*, \Psi_{i+1}^{(g)}|Y, J_m, \Psi_{i-1}^*, \Psi_{i+1}^{(g)})}{M^{-1} \sum_{m=1}^{M} m(\theta_i^{(m)}, \theta_i^*|Y, J_m, \Psi_{i-1}^*, \Psi_{i+1}^{(m)})},
\]
where \(\{\theta_i^{(g)}, \Psi_{i+1}^{(g)}\}_{g=1}^{G}\) are obtained from \(\pi(\theta_i, \Psi_{i+1}|Y, J_m, \Psi_{i-1}^*)\), \(\{\Psi_{i+1}^{(m)}\}\) are obtained from \(\pi(\Psi_{i+1}|Y, J_m, \Psi_{i-1}^*, \theta_i^*)\), and then for each \(\{\Psi_{i+1}^{(m)}\}\), sample \(\theta_i^{(m)}\) from \(q(\theta_i^*, \theta_i|Y, J_m, \Psi_{i-1}^*, \Psi_{i+1}^{(m)})\).

5. Posterior Contraction Rate

The posterior contraction rates of the three multinomial models with finite random series prior can be obtained using the results by Shen and Ghosal (2015) on posterior contraction rates for finite random series. We use \(\lesssim\) to denote an inequality up to a constant multiple. \(f \lesssim g\) for \(f \lesssim g \lesssim f\). For a vector \(\theta \in \mathbb{R}^d, ||\theta||_p = \{\sum_{i=1}^{d} |\theta_i|^p\}^{1/p}\), where \(1 \leq p < \infty\) and \(||\theta||_{\infty} = \max_{1 \leq i \leq d} |\theta_i|\). Similarly, for a function \(f\) with respect to a measure \(G\), we define \(||f||_{p,G} = \{\int |f(x)|^p dG\}^{1/p}\), where \(1 \leq p < \infty\), and \(||f||_{\infty,G} = \sup_{x} |f(x)|\). Let \(N(\epsilon, T, d)\) denote the \(\epsilon\)-covering number of a set \(T\) for a metric \(d\). Let \(h^2(p, q) = \int (\sqrt{p} - \sqrt{q})^2 d\mu\), be the squared Hellinger distance, \(K(p, q) = \int p \log(p/q) d\mu, V(p, q) = \int p \log^2(p/q) d\mu\), be the Kullback-Leibler (KL) divergences.

Suppose that \((X_1, Y_1), i = 1, \ldots, n\), are the independent observations. Let \(p\) denote the joint probability of \((X, Y)\), where \(Y\) takes values \(1, \ldots, K\) and \(p_0\) denote the true joint probability. Let \((X^{(n)}, Y^{(n)})\) be the vector of \(n\) observations following the probability \(p^{(n)}\). Let \(\pi_k = P(\cdot|X)\) be the probability of the \(k\)th category conditioned on \(X\), and \(\pi_{0k}\) be the true probability of the \(k\)th category conditioned on \(X\). Define the probability vector \(\pi = (\pi_1, \ldots, \pi_K)^T\), where \(\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k\), and \(\pi_0 = (\pi_0, \ldots, \pi_{0K})^T\), where \(\pi_{0K} = 1 - \sum_{k=1}^{K-1} \pi_{0k}\). For these multinomial models, the KL divergences \(K(p, p_0), V(p, p_0),\) and the squared Hellinger distance \(h^2(p_1, p_2)\) can be reduced to \(K(\pi, \pi_0), V(\pi, \pi_0),\) and \(h^2(\pi_1, \pi_2)\). We define a distance metric as
\[
d(\pi_0, \pi) = \sqrt{\sum_{k=1}^{K} E_X |\pi_k(X) - \pi_{0k}(X)|^2}.
\]
Let $\Pi$ be the generic notation for priors on the number of bases $J$, and the parameter blocks $B$, where $B = (\theta, \gamma)$ for ordered multinomial probit model, and $B = (\theta_1, \ldots, \theta_{K-1})$ for unordered multinomial probit and multinomial logistic models. According to Shen and Ghosal (2015), the priors on $J$ and the coefficients of the basis functions $\theta = (\theta_1, \ldots, \theta_J)^T$ need to satisfy the conditions (A1) and (A2). For ordered multinomial probit model, we add condition (A3).

(A1) For some $c_1, c_2 > 0$, $0 \leq t_2 \leq t_1 \leq 1$, $\exp\{-c_1 J \log t_1\} \leq \Pi(J = j) \leq \exp\{-c_2 J \log t_2\}$.

(A2) Given $J$, $\Pi(\|\theta - \theta_0\|_2 \leq \epsilon) \geq \exp\{-c_3 J \log(1/\epsilon)\}$ for every $\|\theta_0\|_\infty \leq H$, where $c_3$ is some positive constant, $H$ is chosen sufficiently large, and $\epsilon > 0$ is sufficiently small. Also, assume that $\Pi(\theta \notin [-M, M]^J) \leq J \exp\{-CM^3\}$ for some constant $C$, $t_3 > 0$.

(A3) Given $K$ categories, $\Pi(\|\gamma - \gamma_0\|_2 \leq \epsilon) \geq \exp\{-c_4 K \log(1/\epsilon)\}$, where $c_4$ is some positive constant.

**Theorem 1.** Let $\epsilon_n \geq \bar{\epsilon}_n$ be two sequences of positive numbers satisfying $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$. Suppose that the priors satisfy the conditions (A1), (A2) and (A3). Let $J_n$, $J$, and $M_n$ be sequences of positive numbers, and $W_{J_n, M_n} = \{\pi : J \leq J_n, \|B\|_\infty \leq M_n\}$ such that the following hold:

\[
\log \mathcal{N}(\epsilon_n, W_{J_n, M_n}, d) \lesssim n\epsilon_n^2, \tag{47}
\]

\[
\Pi(\pi \notin W_{J_n, M_n}) \lesssim \exp\{-n\epsilon_n^2\}, \tag{48}
\]

\[
-\log \Pi\left(d^2(\pi, \pi_0) \leq \epsilon_n^2\right) \lesssim n\epsilon_n^2. \tag{49}
\]

Then for every $M_n \to \infty$, we have $\Pi\left(d(\pi, \pi_0) \geq M_n\epsilon_n|X^{(n)}, Y^{(n)}\right) \to 0$ in probability.

**Proof.** This theorem is derived from Theorem 2 of Shen and Ghosal (2015), and Theorem 4 of Ghosal and van der Vaart (2007a). For more details, see Ghosal and van der Vaart (2017).

The squared Hellinger distance between $\pi_1$ and $\pi_2$ is

\[
h^2(\pi_1, \pi_2) = E_X \left\{ \sum_{k=1}^{K} \left( \sqrt{\pi_{1k}(X)} - \sqrt{\pi_{2k}(X)} \right)^2 \right\}, \tag{50}
\]

where $\pi_1 = (\pi_{11}, \ldots, \pi_{1K})^T$, and $\pi_2 = (\pi_{21}, \ldots, \pi_{2K})^T$. The squared Hellinger distance can be bounded by

\[
h^2(\pi_1, \pi_2) = E_X \sum_{k=1}^{K} \frac{|\pi_{1k}(X) - \pi_{2k}(X)|^2}{|\sqrt{\pi_{1k}(X)} + \sqrt{\pi_{2k}(X)}|^2} \lesssim E_X \sum_{k=1}^{K} |\pi_{1k}(X) - \pi_{2k}(X)|^2. \tag{51}
\]

The Kullback-Leibler divergences for multinomial models are

\[
K(\pi_0, \pi) = E_X \left\{ \sum_{k=1}^{K} \pi_{0k}(X) \log \frac{\pi_{0k}(X)}{\pi_k(X)} \right\} \tag{52}
\]
Then we can bound $K$ according to Lemma 8 of Ghosal and van der Vaart (2007b), we have

$$K(\pi_0, \pi) \lesssim h^2(\pi_0, \pi) \left( 1 + \log \left\| \frac{\pi_0}{\pi} \right\|_\infty \right),$$

(54)

$$V(\pi_0, \pi) \lesssim h^2(\pi_0, \pi) \left( 1 + \log \left\| \frac{\pi_0}{\pi} \right\|_\infty \right)^2.$$  

(55)

Then we can bound $K(\pi_0, \pi)$ and $V(\pi_0, \pi)$ by

$$\max\{K(\pi_0, \pi), V(\pi_0, \pi)\} \lesssim E_X \sum_{k=1}^{K} |\pi_{1k}(X) - \pi_{2k}(X)|^2. \quad (56)$$

Thus we have

$$\Pi(K(\pi_0, \pi) \leq c_1^2, V(\pi_0, \pi) \leq c_2^2) \geq \Pi(d^2(\pi_0, \pi) \leq c_2) \quad \Box$$

To obtain the posterior contraction rate, we need to verify the conditions (47)–(49), and we also need additional assumptions on the basis. We use $\theta^T \psi(t)$ to approximate $\beta(t)$, where $\theta = (\theta_1, \ldots, \theta_J)^T$, and $\psi(t) = (\psi_1(t), \ldots, \psi_J(t))^T$. Let $\beta_0(t)$ be the true value, and $r = 2$ or $\infty$. Assume that there exist a $\theta_0 \in \mathbb{R}^J$, $\|\theta_0\|_\infty \leq H$ and $K_0 \geq 0$ such that

$$\|\beta_0(\cdot) - \theta_0^T \psi(\cdot)\|_r \lesssim J^{-\alpha},$$

(58)

$$\|\theta_1^T \psi(\cdot) - \theta_2^T \psi(\cdot)\|_r \lesssim J^{K_0} \|\theta_1 - \theta_2\|_2, \quad \theta_1, \theta_2 \in \mathbb{R}^J. \quad (59)$$

Remark 2 of Shen and Ghosal (2015) gave examples of bases satisfying relations (58) and (59).

5.1. Ordered Multinomial Probit Model

Let $\gamma = (\gamma_0, \ldots, \gamma_K)^T$ be the vector of the threshold points, and $\gamma_0 = (\gamma_{00}, \ldots, \gamma_{0K})^T$ be the vector of the true values of the threshold points. Let $\beta(t)$ be the parameter function on $[0, 1]$, and $\beta_0(t)$ be the true parameter function on $[0, 1]$. Let $\pi_k(X) = \Phi(\gamma_k - \int \beta(t)X(t)dt - \Phi(\gamma_{k-1} - \int \beta(t)X(t)dt)$ and $\pi_{0k}(X) = \Phi(\gamma_{0k} - \int \beta_0(t)X(t)dt - \Phi(\gamma_{0k-1} - \int \beta_0(t)X(t)dt)$.

Theorem 2. Suppose that $EX(t)^2$ is uniformly bounded away from 0 and $\infty$ for every $t \in (0, 1)$, the priors satisfy the conditions (A1), (A2) and (A3), and that the basis $\psi(t)$ satisfy (58) and (59) with $r = 2$. The posterior contraction rate of the ordered multinomial probit model is $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)(\log n)^{\alpha/(2\alpha+1)+(1-\alpha)^2}}$ relative to $d(\pi_0, \pi)$. More explicitly, for every $M_n \to \infty$, $\Pi(\beta : \rho(\beta_0, \beta) \geq M_n \epsilon_n | X^{(n)}, Y^{(n)}) \to 0$ in probability, where $\rho(\beta_0, \beta) = \sqrt{EX| \int (\beta(t) - \beta_0(t))X(t)dt|_2^2}$, and $\Pi(\gamma : \max_j |\gamma_j - \gamma_{0j}| \geq M_n \epsilon_n | X^{(n)}, Y^{(n)}) \to 0$ in probability.
Proof. By the Lipschitz continuity of \( \Phi \), and uniformly boundness of \( E X(t)^2 \), we have

\[
\sum_{k=1}^{K} E_X|\pi_k(X) - \pi_0k(X)|^2 \lesssim \sum_{k=1}^{K} \left\{ \| \gamma_k - \gamma_0k \|^2 + E_X \left\| \int (\beta(t) - \beta_0(t)X(t))dt \right\|^2 \right\} \\
\lesssim \sum_{k=1}^{K} \left\{ \| \gamma_k - \gamma_0k \|^2 + \int \| \beta(t) - \beta_0(t) \|^2 dt \right\} \\
\lesssim \| \gamma - \gamma_0 \|^2 + \| \beta(\cdot) - \beta_0(\cdot) \|^2.
\]

(60)

Observe that with the finite random series prior, the \( L_2 \)-distance between \( \beta(\cdot) \) and \( \beta_0(\cdot) \) is bounded by

\[
\| \beta(\cdot) - \beta_0(\cdot) \|_2 = \| \theta^T \psi(\cdot) - \theta_0^T \psi(\cdot) + \theta_0^T \psi(\cdot) - \beta_0(\cdot) \|_2 \\
\leq \| \theta^T \psi(\cdot) - \theta_0^T \psi(\cdot) \|_2 + \| \theta_0^T \psi(\cdot) - \beta_0(\cdot) \|_2.
\]

(61)

Then we have

\[
\Pi \left( \sum_{k=1}^{K} E_X|\pi_k(X) - \pi_0k(X)|^2 \leq \varepsilon_n^2 \right) \\
\geq \Pi(\| \gamma - \gamma_0 \|_2 \leq \bar{\varepsilon}_n/\sqrt{2}) \Pi(\| \beta(\cdot) - \beta_0(\cdot) \|_2 \leq \varepsilon_n/\sqrt{2}) \\
\geq \Pi(\| \gamma - \gamma_0 \|_2 \leq \bar{\varepsilon}_n/\sqrt{2}) \Pi(\| \theta - \theta_0 \| \leq \varepsilon_n/(2\sqrt{2} J_n K_0)) \\
\gtrsim \exp \left\{ -K \log(\sqrt{2}/\bar{\varepsilon}_n) \right\} \exp \left\{ -J_n \log(2\sqrt{2} J_n K_0/\varepsilon_n) \right\}.
\]

To satisfy the relation (49), we need

\[
K \log(\sqrt{2}/\bar{\varepsilon}_n) + J_n \log(2\sqrt{2} J_n K_0/\varepsilon_n) \leq n \bar{\varepsilon}_n^2.
\]

(63)

Thus, (63) leads to the conditions that \( J_n \log n \leq n \bar{\varepsilon}_n^2 \), and \( J_n^{-\alpha} \leq \bar{\varepsilon}_n \). Then we obtain the preliminary contraction rate \( \varepsilon_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)} \), for \( J_n \asymp (n/\log n)^{1/(2\alpha+1)} \).

(Using 60), we obtain

\[
\log \mathcal{N}(\varepsilon_n, \mathcal{W}_{J_n, M_n}, d) \lesssim \log \mathcal{N}(\varepsilon_n, \mathcal{W}_{J_n, M_n}, \| \cdot \|_2) \lesssim n \varepsilon_n^2.
\]

(64)

According to Theorem 2 of Shen and Ghosal (2015), to satisfy (64), we need

\[
J_n \{ (K_0 + 1) \log J_n + \log M_n + C_0 \log n \} \leq n \varepsilon_n^2,
\]

(65)

for some positive constant \( C_0 \). To satisfy (48), we need

\[
b n \varepsilon_n^2 \leq J_n \log J_n, \log J_n + n \varepsilon_n^2 \leq M_n^\alpha,
\]

(66)

for some \( b > 0 \). For \( M_n = n^{1/t_3} \), (66) implies that \( J_n \log J_n \gtrsim n \varepsilon_n^2 \). Thus \( J_n \asymp n^{1/(2\alpha+1)}(\log n)^{2\alpha/(2\alpha+1)-t_2} \). (65) implies that \( J_n \log n \lesssim n \varepsilon_n^2 \).
As a result, the posterior contraction rate is \( \epsilon_n \approx n^{-\alpha/(2\alpha+1)} (\log n)^{\alpha/(2\alpha+1)} + (1-t_2)/2 \) relative to \( \sqrt{\sum_{k=1}^{K} \mathbb{E}_X |\pi_k(X) - \pi_{0k}(X)|^2} \).

Further, if \( k = 1 \) and \( |\int \beta(t)X(t)dt| < C \), then

\[
\mathbb{E}_X |\pi_k(X) - \pi_{01}(X)|^2 \geq \mathbb{E}_X \left| \phi(\int \beta(t)X(t)dt) - \phi(\int \beta_0(t)X(t)dt) \right|^2,
\]

(67)

Hence if \( \mathbb{E}_X |\pi_1(X) - \pi_{01}(X)|^2 \) is small, then \( \mathbb{E}_X \left| \int \beta(t)X(t)dt - \int \beta_0(t)X(t)dt \right|^2 \) is also small.

For any \( k \), we have

\[
\mathbb{E}_X |\pi_k(X) - \pi_{0k}(X)|^2 \geq |\gamma_k - \gamma_{0k}|^2 - \mathbb{E}_X \left| \int \beta(t)X(t)dt - \int \beta_0(t)X(t)dt \right|^2.
\]

(68)

Thus \( |\gamma_k - \gamma_{0k}| \) is also small.

\( \square \)

### 5.2. Unordered Multinomial Probit Model

The probability of the \( k \)th category is expressed by the latent variable. Let \( \beta_l(t) = \beta'_l(t) - \beta''_l(t) \), and \( \varepsilon_l = \varepsilon'_l - \varepsilon''_l \). Then \( W_l = \int \beta_l(t)X(t)dt + \varepsilon_l \), where \( l = j \) if \( j < k \), \( l = j - 1 \) if \( j > k \) so that \( l = 1, \ldots, K - 1 \).

\[
\pi_k(X) = \mathbb{P}(W_1 \leq 0, \ldots, W_{K-1} \leq 0) = \mathbb{P}(\varepsilon_1 \leq -\int \beta_1(t)X(t)dt, \ldots, \varepsilon_{K-1} \leq -\int \beta_{K-1}(t)X(t)dt)
\]

\[
= \frac{1}{(2\pi)^{(K-1)/2} |\Sigma|^{1/2}} \int_{-\infty}^{-\int \beta_1(t)X(t)dt} \cdots \int_{-\infty}^{-\int \beta_{K-1}(t)X(t)dt} \exp \left( -\frac{1}{2} Z^T \Sigma^{-1} Z \right) dZ,
\]

(69)

where \( \Sigma \) is a \((K - 1) \times (K - 1)\) matrix with 2 on diagonal entries and 1 on all off-diagonal entries.

Due to the exchangeable correlation structure of \( \Sigma \), \( \pi_k \) can be reduced to a one-dimensional integral (Dunnett, 1989)

\[
\pi_k(x) = \frac{1}{\sqrt{\pi}} \int_0^{\infty} \left\{ \prod_{l=1}^{K-1} \Phi(-z\sqrt{2} - \int \beta_l(t)X(t)dt) + \prod_{l=1}^{K-1} \Phi(z\sqrt{2} - \int \beta_l(t)X(t)dt) \right\} e^{-z^2} dz
\]

(70)
**Theorem 3.** Suppose that $EX(t)^2$ is uniformly bounded away from 0 and $\infty$ for every $t \in (0, 1)$, the priors satisfy the conditions (A1) and (A2), and that the basis $\psi(t)$ satisfy (58) and (59) with $r = 2$. The posterior contraction rate of the unordered multinomial probit model is $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_2)/2}$ relative to $d(\pi_0, \pi)$.

**Proof.** By the Lipschitz continuity of the function $\Phi$, and uniformly boundness of $EX(t)^2$, we have

$$\sum_{k=1}^{K} E_X |\pi_k(X) - \pi_{0k}(X)|^2 \lesssim \sum_{k=1}^{K-1} E_X \left\{ \left| \int \beta_k(t)X(t)dt - \int \beta_{0k}(t)X(t)dt \right|^2 \right\} \lesssim \sum_{k=1}^{K-1} \|\beta_k(\cdot) - \beta_{0k}(\cdot)\|_2^2. \tag{71}$$

The $L_2$-distance between $\beta_k(\cdot)$ and $\beta_{0k}(\cdot)$ is bounded by

$$\|\beta_k(\cdot) - \beta_{0k}(\cdot)\|_2 = \|\theta_k^T \psi(\cdot) - \theta_{0k}^T \psi(\cdot) + \theta_{0k}^T \psi(\cdot) - \beta_{0k}(\cdot)\|_2 \leq \|\theta_k^T \psi(\cdot) - \theta_{0k}^T \psi(\cdot)\|_2 + \|\theta_{0k}^T \psi(\cdot) - \beta_{0k}(\cdot)\|_2. \tag{72}$$

Then we have

$$\Pi(\sum_{k=1}^{K} E_X |\pi_k(X) - \pi_{0k}(X)|^2 \leq \bar{\epsilon}_n^2) \geq \Pi(\sum_{k=1}^{K-1} \|\beta_k(\cdot) - \beta_{0k}(\cdot)\| \leq \bar{\epsilon}_n) \geq \Pi(||\theta_k - \theta_{0k}|| \leq \bar{\epsilon}_n/\sqrt{2K - 1}\bar{J}_n K_0) \geq \exp\{-\bar{J}_n \log(2\sqrt{K - 1}\bar{J}_n K_0/\bar{\epsilon}_n)\}. \tag{73}$$

To satisfy the relation (49), we need

$$\bar{J}_n \log(2\sqrt{K - 1}\bar{J}_n K_0/\bar{\epsilon}_n) \leq n\bar{\epsilon}_n^2. \tag{74}$$

Thus, (74) leads to the conditions that $\bar{J}_n \log n \leq n\bar{\epsilon}_n^2$, and $\bar{J}_n^{-\alpha} \leq \bar{\epsilon}_n$. Then we obtain the preliminary contraction rate $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)}$, for $\bar{J}_n \asymp (n/\log n)^{1/(2\alpha+1)}$.

Following the same argument as (64)–(66), the posterior contraction rate is $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_2)/2}$ relative to $\sqrt{\sum_{k=1}^{K} E_X |\pi_k(X) - \pi_{0k}(X)|^2}$. □

**5.3. Multinomial Logistic Model**

Let $\beta_k(t)$, $k = 1, \ldots, K - 1$, be the coefficient functions on $[0, 1]$, and $\beta_{0k}(t)$, $k = 1, \ldots, K - 1$, be the true coefficient functions on $[0, 1]$. 

Theorem 4. Suppose that $EX(t)^2$ is uniformly bounded away from 0 and $\infty$ for every $t \in (0, 1)$, the priors satisfy the conditions (A1) and (A2), and that the basis $\psi(t)$ satisfy (57) and (58) with $r = 2$. The posterior contraction rate of the multinomial logistic model is $\epsilon_n \sim n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-\alpha)/2}$ relative to $d(\pi_0, \pi)$.

Proof. The proof is similar to that of Theorem 3. \qed

6. Discriminant Analysis

As a comparison to those multinomial models, we use Bayesian discriminant analysis to classify the functional data. Classical discriminant analysis applies only to multivariate data. For functional data, we can use certain orthogonal linear functions to determine the classification probabilities:

$$
(f_1, \ldots, f_m)^T = \left( \int \beta_1(t)X_i(t)dt, \ldots, \int \beta_m(t)X_i(t)dt \right)^T
$$

(75)

Ideally these $\beta_1(t), \ldots, \beta_m(t)$ are unknown, but putting a prior on these functions with identifiability restrictions is complicated. We instead take $\beta_1(t), \ldots, \beta_m(t)$ to be known as the first $m$ principal components (Ramsay and Silverman, 2005), but let the means and the covariance matrices be unknown. Then discriminant analysis can be applied to the $m$ principal components.

6.1. Linear Discriminant Analysis

Linear discriminant analysis assumes that for each of the $K$ category, the set of linear function $(f_1, \ldots, f_m)$ follows a normal distribution with the same covariance matrix: $(f_{il1}, \ldots, f_{ilm})^T \sim N(\mu_l, \Sigma)$, where $\mu_l$ is the population mean of category $l$, $l = 1, \ldots, K$, $i = 1, \ldots, N_l$, and $N_l$ is the number of data in category $l$. Then the probability of choosing category $k$ is given by

$$
P(Y_i = k|X_i) = \frac{p_k \cdot \phi(f_{ik1}, \ldots, f_{ikm}; \mu_k, \Sigma)}{\sum_{l=1}^K p_l \cdot \phi(f_{il1}, \ldots, f_{ilm}; \mu_l, \Sigma)},
$$

(76)

where $\phi(f_1, \ldots, f_m; \mu, \Sigma)$ is the multivariate normal density function with mean $\mu$ and covariance $\Sigma$, and $p_l, l = 1, \ldots, K$, are the probability of choosing category $l$. The variables $f_{il1}, \ldots, f_{ilm}$ are the $m$ principal components of $X_i(t)$ in category $l$, where $l = 1, \ldots, K$. Define $f_{il} = (f_{il1}, \ldots, f_{ilm})^T$, where $i = 1, \ldots, N_l$, and $\sum_{l=1}^K N_l = N$. To estimate the mean $\mu_l$ for each category $l$, and the common covariance $\Sigma$ among all categories, we use the conjugate normal-inverse-Wishart prior with hyperparameters (Gelman et al., 2013) for $(\mu_l, \Sigma)$

$$
\Sigma \sim IW_{\nu_0}(A_0^{-1}), \quad \mu_l|\Sigma \sim N(\mu_{l0}, \Sigma/\kappa_0).
$$

(77)

Then the posterior distribution of $(\mu_l, \Sigma)$ can be obtained in the following order

$$
\Sigma|Y \sim IW_{\nu_n}(A_n^{-1}), \quad \mu_l|\Sigma, Y \sim N(\mu_{ln}, \Sigma/\kappa_n),
$$

(78)
where $\nu_n = \nu_0 + N, \bar{f}_l = \sum_{i=1}^{N_l} f_{il}/N_l$, $S = \sum_{l=1}^{K} \sum_{i=1}^{N_l} (f_{il} - \bar{f}_l)(f_{il} - \bar{f}_l)^T$, 

$$\Lambda_n = \Lambda_0 + S + \sum_{l=1}^{K} \frac{N_l}{\kappa_0 + N_l}(\bar{f}_l - \mu_0)(\bar{f}_l - \mu_0)^T,$$

(79)

and

$$\kappa_n = \kappa_0 + N, \mu_{ln} = \frac{\kappa_0 \mu_0 + N_l \bar{f}_l}{\kappa_0 + N_l}, l = 1, \ldots, K.$$

(80)

### 6.2. Quadratic Discriminant Analysis

Quadratic discriminant analysis is defined in a similar way, except that it has a different covariance matrix for each category. The probability of choosing category $k$ is given by

$$P(Y_i = k|X_i) = \frac{p_k \cdot \phi(f_{i1}, \ldots, f_{im}; \mu_k, \Sigma_k)}{\sum_{l=1}^{K} p_l \cdot \phi(f_{i1}, \ldots, f_{im}; \mu_l, \Sigma_l)}.$$

(81)

To estimate the mean $\mu_l$ and the covariance $\Sigma_l$ for each category $l$, where $l = 1, \ldots, K$, we use the conjugate normal-inverse-Wishart prior with hyperparameters for $(\mu_l, \Sigma_l)$

$$\Sigma_l \sim IW_{\nu_0}(\Lambda_0^{-1}), \mu_l|\Sigma_l \sim N(\mu_0, \Sigma_l/\kappa_0),$$

(82)

for $l = 1, \ldots, K$. Then the posterior distribution of $(\mu_l, \Sigma_l)$ can be obtained in the following order

$$\Sigma_l|Y \sim IW_{\nu_{ln}}(\Lambda_{ln}^{-1}), \mu_l|\Sigma_l, Y \sim N(\mu_{ln}, \Sigma_l/\kappa_{ln}),$$

(83)

where $\nu_{ln} = \nu_0 + N_l, \bar{f}_l = \sum_{i=1}^{N_l} f_{il}/N_l$, $S_l = \sum_{i=1}^{N_l} (f_{il} - \bar{f}_l)(f_{il} - \bar{f}_l)^T$, 

$$\Lambda_{ln} = \Lambda_0 + S_l + \frac{N_l}{\kappa_0 + N_l}(\bar{f}_l - \mu_0)(\bar{f}_l - \mu_0)^T,$$

(84)

and

$$\kappa_{ln} = \kappa_0 + N_l, \mu_{ln} = \frac{\kappa_0 \mu_0 + N_l \bar{f}_l}{\kappa_0 + N_l}, l = 1, \ldots, K.$$

(85)

### 7. Simulation

#### 7.1. Data Generation

The simulated data are generated following different data generating process. All of the simulated data have three categories.

For the ordered multinomial probit data, we use Gaussian process with mean function $\sin(x)$ and covariance function $100 \exp\{-100(x - y)^2\}$ to generate functional data
at discrete time points $0, 0.01, \ldots, 0.09, 1$. The coefficient function was generated by Gaussian process with mean function $\cos(x)$ and variance function $100 \exp\{-100(x-y)^2\}$ at discrete time points $0, 0.01, \ldots, 0.09, 1$. The four threshold points were chosen to be $-\infty, 0, 8, \infty$. The four cut-off points construct three interval. If the inner product of a functional data and the coefficient function plus a standard normal variable falls in the $k$th interval $(\gamma_{k-1}, \gamma_k)$, then the functional data attributes to the category $k$.

For unordered multinomial probit data, we generate three coefficient functions by Gaussian process with mean function $\cos(x)$ and variance function $100 \exp\{-100(x-y)^2\}$ at discrete time points $0, 0.01, \ldots, 0.09, 1$. The functional data were generated by Gaussian process with mean function $\sin(x)$ and covariance function $100 \exp\{-100(x-y)^2\}$ at the same discrete time points. The inner product of a functional data and the three coefficient functions are added with standard normal variables, respectively. Sample these three normal variables, and obtain the corresponding probabilities. Then the functional data belongs to the category with the largest sampled value.

For the multinomial logistic data, we generate two coefficient functions by Gaussian process with mean function $\cos(x)$ and variance function $100 \exp\{-100(x-y)^2\}$ at discrete time points $0, 0.01, \ldots, 0.09, 1$, and the third coefficient function can be assumed to be zero everywhere. The functional data were generated by Gaussian process with mean function $\sin(x)$ and covariance function $100 \exp\{-100(x-y)^2\}$ at $0, 0.01, \ldots, 0.09, 1$. We compute the probability of a functional data falls in each category. Then the data attributes to the category with the largest probability.

For data satisfying the assumption of the linear discriminant analysis, we generate them from three Gaussian processes with different mean functions $\sin(x) + 2 \cos(x)$, $\sin(x)$, and $\sin(x) - 3 \cos(x)$, but the same covariance function $\exp\{-30(x-y)^2\}$.

For data satisfying the assumption of the quadratic discriminant analysis, we generate them from three Gaussian processes with different mean functions and different covariance functions. The mean functions were $\sin(x) + 2 \cos(x)$, $\sin(x)$, and $\sin(x) - 3 \cos(x)$, and the covariance functions were $\exp\{-2 \sin^2(\pi(x-y))\}$, $\exp\{-30(x-y)^2\}$, and $\exp\{-|x-y|\}$, respectively.

In this simulation study, we generated total 900 (300 for each category) functional data for each type of dataset. Construct the training data with 720 (240 for each category) of them, and the testing data with the remaining 180 (60 for each category) of them.

7.2. Basis Functions

For models using the finite random series prior, we consider B-spline basis. The B-spline basis functions on interval $[0, 1]$ can be created using the R package fda. In this simulation study, consider the possible number of B-spline basis functions to be $J = 5, \ldots, 15$. Those B-spline basis functions are generated at the same discrete time points as the functional data, that is $0, 0.01, \ldots, 0.09, 1$.

7.3. Results

Under the chosen models, we apply Bayesian estimation methods described in Section 3 on the training data. In this study, 5000 MCMC iterations were obtained, and the
first 1000 of them were discarded as burn-in. We use the last 4000 MCMC output of the parameter $B$ to classify the 180 transformed testing data, where $B = (\theta, \gamma_2, \gamma_3)$ for the ordered multinomial probit model, $B = (\theta_1, \theta_2)$ for the logistic model, $B = (\mu_1, \mu_2, \mu_3, \Sigma)$ for the linear discriminant analysis model, and $B = (\mu_1, \mu_2, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3)$ for the quadratic discriminant analysis model. A transformed testing data $z_i$ or $f_i$ is in category $k$ if 
\[ \sum_{g=1}^{4000} 1(Y_i = k | z_i or f_i, B(g)) > \sum_{g=1}^{4000} 1(Y_i = l | z_i or f_i, B(g)), \]
where $l \neq k$. Then we use the techniques described in Section 4 to average the results from the multinomial models.

As a comparison with the Bayesian method, the linear support vector machine (SVM) was also applied to the principal components of these training data, and made predictions on the testing data. To apply SVM, we use the R package e1071.

Table 1 shows the averaged misclassification rates for each data type under different models.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>OMP Model</th>
<th>UMP Model</th>
<th>MLO Model</th>
<th>LDA</th>
<th>QDA</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMP</td>
<td>7.72%</td>
<td>31.21%</td>
<td>30.56%</td>
<td>38.89%</td>
<td>48.89%</td>
<td>15.00%</td>
</tr>
<tr>
<td>UMP</td>
<td>39.03%</td>
<td>7.22%</td>
<td>6.11%</td>
<td>21.11%</td>
<td>21.11%</td>
<td>10.56%</td>
</tr>
<tr>
<td>MLO</td>
<td>49.44%</td>
<td>1.11%</td>
<td>2.22%</td>
<td>32.22%</td>
<td>36.11%</td>
<td>7.78%</td>
</tr>
<tr>
<td>LDA</td>
<td>25.95%</td>
<td>25.36%</td>
<td>26.11%</td>
<td>5.00%</td>
<td>5.00%</td>
<td>7.78%</td>
</tr>
<tr>
<td>QDA</td>
<td>24.05%</td>
<td>21.80%</td>
<td>21.67%</td>
<td>10.56%</td>
<td>9.44%</td>
<td>8.33%</td>
</tr>
</tbody>
</table>

8. Real Data

We also test our models on Phoneme data. This dataset can be found in the R package fds, and can also be found at https://www.math.univ-toulouse.fr/staph/npfda/. The original data has 2000 $(X, Y)$ pairs, and five categories. For computational efficiency, we only choose 900 of them from three categories. We obtained 5000 MCMC iterations, and discarded the first 1000 of them as burn-in.

According to Table 2, the unordered multinomial probit model is the best model for the Phoneme data. For this data, the categories are not naturally ordered, and hence ordered multinomial probit model is not natural for this problem, but we include it in the analysis for comparison. We also want to know the accuracy of the posterior mean. Table 3, 4, and 5 show the estimate and standard error of the posterior mean of the Phoneme data under unordered multinomial probit model, ordered multinomial probit model, and multinomial logistic model, when $J = 14$, $J = 15$, and $J = 15$, respectively. We choose these $J$ values because under these values the model has the largest marginal likelihood $m(Y | J)$. Although ordered multinomial probit model isn’t intuitive in this context, its performance is not too inferior.
Table 2: Averaged misclassification rates for Phoneme data

<table>
<thead>
<tr>
<th>OMP Model</th>
<th>UMP Model</th>
<th>MLO Model</th>
<th>LDA</th>
<th>QDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.51%</td>
<td>0.56%</td>
<td>4.58%</td>
<td>7.78%</td>
<td>5.00%</td>
</tr>
</tbody>
</table>

Table 3: Estimate and standard error of the posterior mean for the unordered multinomial model (J=14)

<table>
<thead>
<tr>
<th>Θ</th>
<th>estimate</th>
<th>standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>−15.40</td>
<td>49.92</td>
<td>0.93 0.85</td>
</tr>
<tr>
<td>35.78</td>
<td>79.79</td>
<td>0.52 0.81</td>
</tr>
<tr>
<td>45.94</td>
<td>32.11</td>
<td>0.58 0.75</td>
</tr>
<tr>
<td>4.97</td>
<td>−0.76</td>
<td>0.66 0.73</td>
</tr>
<tr>
<td>−23.23</td>
<td>−12.58</td>
<td>0.60 0.71</td>
</tr>
<tr>
<td>−15.09</td>
<td>−14.88</td>
<td>0.62 0.67</td>
</tr>
<tr>
<td>23.43</td>
<td>−21.71</td>
<td>0.68 0.73</td>
</tr>
<tr>
<td>−11.87</td>
<td>1.67</td>
<td>0.74 0.80</td>
</tr>
<tr>
<td>−0.96</td>
<td>−5.06</td>
<td>0.63 0.64</td>
</tr>
<tr>
<td>−0.27</td>
<td>−9.82</td>
<td>0.63 0.69</td>
</tr>
<tr>
<td>1.58</td>
<td>−7.46</td>
<td>0.70 0.78</td>
</tr>
<tr>
<td>−12.43</td>
<td>−14.68</td>
<td>0.57 0.60</td>
</tr>
<tr>
<td>−28.97</td>
<td>−7.74</td>
<td>0.64 0.69</td>
</tr>
<tr>
<td>−28.49</td>
<td>−3.38</td>
<td>0.53 0.57</td>
</tr>
</tbody>
</table>
Table 4: Estimate and standard error of the posterior mean for the ordered multinomial model (J=15)

<table>
<thead>
<tr>
<th>OMP</th>
<th>γ₂</th>
<th>θ</th>
</tr>
</thead>
<tbody>
<tr>
<td>estimate</td>
<td>3.817704</td>
<td>(27.90, 62.79, 1.27, -11.95, 0.26, 0.07, -14.34, -5.21, 1.28, -3.34, -3.88, -1.80, -1.74, 0.71, 0.92)</td>
</tr>
<tr>
<td>standard error</td>
<td>0.02914574</td>
<td>(0.29, 0.28, 0.20, 0.20, 0.19, 0.15, 0.16, 0.14, 0.14, 0.13, 0.11, 0.14, 0.15, 0.16)</td>
</tr>
</tbody>
</table>

Table 5: Estimate and standard error of the posterior mean for the multinomial logistic model (J=15)

<table>
<thead>
<tr>
<th>MLO</th>
<th>θ₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>estimate</td>
<td>(4.37, 14.87, 12.07, 16.30, -18.30, -5.73, 11.88, 6.82, -7.12, -6.80, 7.15, -1.53, -8.32, -17.19, -32.68)</td>
</tr>
<tr>
<td>standard error</td>
<td>(0.26, 0.45, 0.91, 1.17, 0.67, 0.98, 0.49, 0.69, 0.52, 0.48, 0.57, 0.42, 0.62, 0.71, 1.11)</td>
</tr>
<tr>
<td>MLO</td>
<td>θ₃</td>
</tr>
<tr>
<td>estimate</td>
<td>(16.69, 45.69, 18.26, 16.92, -8.00, 2.70, -26.06, 8.89, -4.91, -17.60, -5.70, 0.63, -10.30, -6.66, -11.12)</td>
</tr>
<tr>
<td>standard error</td>
<td>(0.81, 2.26, 0.91, 0.33, 0.33, 0.40, 0.81, 0.69, 0.67, 0.42, 0.56, 0.36, 0.58, 0.49, 0.54)</td>
</tr>
</tbody>
</table>
References


REFERENCES