Electronic Journal of Statistics
Vol. 12 (2018) 4669–4696
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1522

Bayesian classification of multiclass functional data

Xiuqi Li

Operations Research Graduate Program
North Carolina State University
Raleigh, NC 27695
U.S.A.
e-mail: xli35@ncsu.edu

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 Markov Chain Monte Carlo (MCMC) output. Posterior contraction rates for the three multinomial models are computed. We also consider Bayesian linear and quadratic discriminant analyses on the multivariate data obtained by applying a functional principal component technique on the original functional data. A simulation study is conducted to compare these methods on different types of data. We also apply these methods to a phoneme dataset.

Keywords and phrases: Multiclass functional data, multinomial probit models, B-splines, posterior contraction rate, discriminant analysis.

Received August 2018.

Contents

1. Introduction

Functional data analysis (FDA) deals with the analysis of data occurring in the form of functions. Wang et al. (2016) gave an overview of FDA including functional principal component analysis, functional linear regression, clustering and classification of functional data. FDA is increasingly drawing attention in many areas, such as biomedicine, environmental studies, and economics (Ullah and Finch, 2013). Mallor, Moler and Urmeneta (2018) proposed a model based on functional principal component analysis to predict household electricity consumption. Wagner-Muns et al. (2018) proposed a method that uses functional principal components analysis to forecast traffic volume. 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). The linear discriminant analysis is also used for functional data classification (James and Hastie, 2001). Preda, Saporta and Lévéder (2007) proposed the partial least squares regression on functional data for linear discriminant analysis. Rossi and Villa (2006) adapted support vector machines to functional data classification. Li and Yu (2008) proposed a functional segment discriminant analysis (FSDA), which combines the classical linear discriminant analysis and support vector machines. Wavelets approaches are also applied to classify and cluster functional data (Ray and Mallick, 2006; Antoniadis et al., 2013; Chang, Chen and Ogden, 2014; Suarez and Ghosal, 2016). There are also nonparametric approaches for functional data classification (Biau, Bunea and Wegkamp, 2005; Ferraty and Vieu, 2003). However, there are only a few approaches proposed in the context of Bayesian classification for functional data. Wang, Ray and Mallick (2007) developed a Bayesian hierarchical model which combines

the adaptive wavelet-based function estimation and the logistic classification. Zhu, Vannucci and Cox (2010) proposed a Bayesian hierarchical model that takes into account random batch effects and selects effective functions among multiple functional predictors. Stingo, Vannucci and Downey (2012) proposed a Bayesian conjugate normal discriminant model on the wavelet transform of the functional data. Zhu, Brown and Morris (2012) introduced two Bayesian approaches: the Gaussian, wavelet-based functional mixed model and the robust, wavelet-based functional mixed model.

In this paper, we consider a 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 $\mathrm{P}(Y = k|X)$, which can be conveniently modeled by a function of $\int \beta(t)X(t)dt$

$$\mathrm{P}(Y = k|X) = H_k\left(\int \beta(t)X(t)dt\right), \tag{1.1}$$

where H_k is a cumulative distribution function, and $\beta(\cdot)$ is an unknown (possibly vector of) coefficient function(s). Ordered multinomial probit, unordered multinomial probit and multinomial logistic models are considered in this paper which correspond to different choices of $H_k, k = 1, \ldots, K$. For an ordered multinomial probit model, there are additional order restrictions, and H_k is expressed as in (2.1). For the unordered multinomial probit model, $\beta(\cdot)$ is a vector of coefficient functions $\beta_1(\cdot), \ldots, \beta_K(\cdot)$, and H_k is in the form of $F(\int \beta_k(t)X(t)\,\mathrm{d}t - \int \beta_l(t)X(t)\,\mathrm{d}t)$, where F is the cumulative distribution function of $\varepsilon_l - \varepsilon_k$ for $l \neq k$. For the multinomial logistic model, $\beta(\cdot)$ is also a vector of coefficient functions $\beta_1(\cdot), \ldots, \beta_K(\cdot)$, and H_k is expressed as in (2.6). Finite random series priors (Shen and Ghosal, 2015) are applied to the three multinomial models. We compare these methods with Bayesian linear and quadratic discriminant analyses applied on the data reduced to multivariate form by a functional principal component technique. Following a Bayesian approach, the posterior distribution of the parameters are obtained using the training data, and then the classification rules are applied to the test data using the posterior probability of class membership.

The primary goal of a basis expansion method is to reduce a more complex problem to a simpler problem which has either a known solution or is likely to be easier to solve. A prior on the target function through a finite random series is a standard tool in nonparametric Bayesian inference, but in the context of functional data, the technique has not been utilized to its fullest potential, especially regarding the study of theoretical properties of Bayesian methods. Only one paper (Shen and Ghosal, 2015) has an example of functional linear regression treated using finite random series priors. We take that idea but develop it in the context of functional data classification. Characterizing contraction rates is a major goal of this paper. For this, we need to estimate the complexity of the model and the prior concentration. Even though, the model reduces to the finite dimensional setting from the computational point of view, the effect of the residual bias in the approximation of function must be properly addressed.

Hence the treatment substantially differs from that of a parametric problem. In particular, the dimension of the basis must be adapted with the smoothness and the sample size by using a prior on it. The two inference problems are fundamentally different even though the resulting rates are similar. In Shen and Ghosal (2015), (one of) the problems was nonparametric binary regression, with a scalar response but nonparametric response probability function. A finite random series prior was used on the response probability function. In contrast, in this paper, the predictor variable is functional and the response probability function is a parametric function of the linear combination. Depending on the choice of the parametric model (ordered multinomial probit, unordered multinomial probit, and multinomial logistic), we get different problems and those need to be treated differently. The linear coefficient function is an unknown smooth function which makes the problem infinite-dimensional in spite of the parametric response probability function. Thus other than some similarity in the looks and in the obtained rates, the two problems are very different.

The paper is organized as follow. In Section 2, the three functional multinomial models are described. Section 3 gives the description of applying the finite random series prior to these models. The marginal likelihood estimation is described in Section 4. In Section 5, the posterior contraction rates of the three functional multinomial models with finite random series priors are computed. Section 6 describes the Bayesian discriminant analysis of functional data, which is used to compare with the proposed models. In Section 7, a simulation study is conducted on various types of data. In Section 8, the three multinomial models and Bayesian discriminant analysis are tested on a phoneme dataset.

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 follows: there exists a latent variable W_i distributed as $\mathrm{N}(\int \beta(t)X_i(t)dt, 1)$, for $i = 1, \ldots, n$, and that $Y_i = k$ if $\gamma_{k-1} < W_i \leq \gamma_k$, where $k = 1, \ldots, K$. The latent variables W_i, $i = 1, \ldots, n$, are independent. The coefficient function $\beta(\cdot)$ is unkown. The cut-points γ_k are also unknown except that $\gamma_0 = -\infty$ and $\gamma_K = \infty$. To ensure identifiability, we set $\gamma_1 = 0$. Under the assumed model, the probability of choosing a category k is given by

$$\mathrm{P}(Y_i = k|X_i) = \Phi\left(\gamma_k - \int \beta(t)X_i(t)dt\right) - \Phi\left(\gamma_{k-1} - \int \beta(t)X_i(t)dt\right), \quad (2.1)$$

where Φ stands for the distribution function of the standard normal distribution.

2.2. Unordered multinomial probit model

Let $X_i(t)$, $i = 1, \ldots, n$, be the same as in the Section 2.1, and also same for Section 2.3.

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 variable, such that W_{il}' follows a linear model

$$W_{il}' = \int \beta_l'(t) X_i(t) dt + \varepsilon_{il}', \tag{2.2}$$

where $\varepsilon_{il}' \sim \mathrm{N}(0, 1)$, $i = 1, \ldots, n$, $l = 1, \ldots, K$, are i.i.d. standard normal random variables. Consider the latent variables $W_i = (W_{i1}, \ldots, W_{iK-1})^T$, $W_{il} = W_{il}' - W_{iK}'$,

$$W_{il} = \int \beta_l'(t) X_i(t) dt - \int \beta_K'(t) X_i(t) dt + \varepsilon_{il}, \tag{2.3}$$

where $\varepsilon_{il} = \varepsilon_{il}' - \varepsilon_{iK}'$, and $l = 1, \ldots, K-1$. Let $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iK-1})^T$. Then ε_i follows $\mathrm{N}(0, \Sigma)$, where Σ 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 kth ($k = 1, \ldots, K-1$) alternative is given by

$$\mathrm{P}(Y_i = k | X_i) = \mathrm{P}(W_{ik} > W_{il}, \text{ for all } l \neq k, \text{and } W_{ik} > 0), \tag{2.4}$$

and the probability of choosing alternative K is given by

$$\mathrm{P}(Y_i = K | X_i) = \mathrm{P}(W_{il} < 0 \text{ for all } l = 1, \ldots, K-1). \tag{2.5}$$

2.3. Multinomial logistic model

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

$$\mathrm{P}(Y_i = k | X_i) = \frac{\exp[\int \beta_k(t) X_i(t) dt]}{\sum_{l=1}^{K} \exp[\int \beta_l(t) X_i(t) dt]}. \tag{2.6}$$

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

$$\mathrm{P}(Y_i = k | X_i) = \frac{\exp[\int \beta_k(t) X_i(t) dt]}{1 + \sum_{l=1}^{K-1} \exp[\int \beta_l(t) X_i(t) dt]}, \tag{2.7}$$

and the probability of choosing category K is given by

$$\mathrm{P}(Y_i = K | X_i) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp[\int \beta_l(t) X_i(t) dt]}. \tag{2.8}$$

3. Finite random series prior

The functional coefficient $\beta(t)$ (or $\beta_1(t), \ldots, \beta_K(t)$ for unordered multinomial probit and multinomial logistic models) 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, formed by B-splines, Fourier functions, or wavelets. A prior is put on the unknown coefficients $(\theta_1, \ldots, \theta_J)$. The number of basis functions J is also unknown and should be given a hyperprior. Instead of sampling across the different dimensions using reversible jump MCMC (Green, 1995) which has computational difficulty for complicated models, we can implement MCMC for a given J value, and repeat it for relevant J values. Thus, we can compute the marginal likelihood $m(Y|J)$ for potentially interested values of J, and obtain the posterior probability of J, which are discussed in Section 4.

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

$$\int \beta(t) X_i(t) dt = \int \sum_{j=1}^{J} \theta_j \psi_j(t) X_i(t) dt = \sum_{j=1}^{J} \theta_j Z_{ij}, \tag{3.1}$$

where $Z_{ij} = \int \psi_j(t) X_i(t) dt$ is known, and can be computed by Simpson's rule.

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 (2.1) can be rewritten as

$$\mathrm{P}(Y_i = k|X_i) = \Phi\left(\gamma_k - \sum_{j=1}^{J} \theta_j Z_{ij}\right) - \Phi\left(\gamma_{k-1} - \sum_{j=1}^{J} \theta_j Z_{ij}\right). \tag{3.2}$$

Define $\theta = (\theta_1, \ldots, \theta_J)^T$, and $Z_i = (Z_{i1}, \ldots, Z_{iJ})^T$. Then (3.2) can be written compactly as

$$\mathrm{P}(Y_i = k|X_i) = \Phi(\gamma_k - Z_i^T \theta) - \Phi(\gamma_{k-1} - Z_i^T \theta). \tag{3.3}$$

Clearly the unobserved latent variable W_i follows $\mathrm{N}(Z_i^T \theta, 1)$. Assign a conjugate prior $\theta \sim \mathrm{N}_J(\theta_0, B_0)$, where N_J stands for the J-variate normal distribution, θ_0 is $J \times 1$ mean vector, and B_0 is a $J \times J$ covariance matrix. Then the posterior distribution of θ is given by

$$\theta|Y, W \sim \mathrm{N}_J(\theta_n, B_n),\ B_n = (B_0^{-1} + Z^T Z)^{-1},\ \theta_n = B_n(B_0^{-1}\theta_0 + Z^T W), \tag{3.4}$$

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

$$W_i|(\theta, \gamma, Y_i = k) \sim \mathrm{TN}(Z_i^T\theta, 1, \gamma_{k-1}, \gamma_k), \tag{3.5}$$

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

$$\alpha_1 = \log \gamma_2,\ \alpha_j = \log(\gamma_{j+1} - \gamma_j),\ 2 \le j \le K-2. \tag{3.6}$$

Note that $\gamma_1 = 0$ and by the inverse map

$$\gamma_j = \sum_{l=1}^{j-1} e^{\alpha_l},\ 2 \le j \le K-1. \tag{3.7}$$

Then γ can be reparameterized by $\alpha = (\alpha_1, \ldots, \alpha_{K-2})$. Assign a multivariate normal prior with mean α_0, and covariance A_0 on α. To sample γ, apply the following steps of Metropolis-Hastings algorithm.

1. Sample α' from a 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 α' from the current α with probability

$$\rho(\alpha, \alpha'|Y, \theta, W) = \min\Big\{\frac{f(Y|\alpha', \theta, W)\pi(\alpha', \theta)}{f(Y|\alpha, \theta, W)\pi(\alpha, \theta)}\frac{q(\alpha', \alpha|Y, \theta, W)}{q(\alpha, \alpha'|Y, \theta, W)}, 1\Big\}. \tag{3.8}$$

3. Compute γ by the inverse map (3.7).

To implement the MCMC sampling, first draw γ by the above steps. Then sample from the posterior distributions (3.5) and (3.4).

The values of γ sampled from the Metropolis-Hastings algorithm converges quickly. We demonstrate it on the real data in Section 8 by plotting the sampling values of γ.

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 (2.3) can be rewritten as

$$W_{il} = \sum_{j=1}^J \theta_{lj}' Z_{ij} - \sum_{j=1}^J \theta_{Kj}' Z_{ij} + \epsilon_{il} = \sum_{l=1}^J (\theta_{jl}' - \theta_{jK}')Z_{ij} + \varepsilon_{il}. \tag{3.9}$$

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 (3.9) is given by

$$W_{il} = Z_i^T \theta_l + \varepsilon_{il}, \tag{3.10}$$

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 $W_i = Z_i^T \Theta + \varepsilon_i$, where $W_i = (W_{i1}, \ldots, W_{iK-1})^T$, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iK-1})^T$, and $\varepsilon_i \sim \mathrm{N}(0, \Sigma)$.

In the model described in Section 2.2, Σ is known with 2 on diagonal entries and 1 on all off-diagonal entries. The only parameter needs to be estimated is Θ. In order to draw the matrix Θ using the Gibbs sampling, we can stack the data in a matrix form which is given by

$$W = Z\Theta + \varepsilon, \tag{3.11}$$

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 $\varepsilon = (\varepsilon_1^T, \ldots, \varepsilon_n^T)^T$ is an $n \times (K-1)$ matrix.

This results in a matrix normal distribution. The density function of matrix normal distribution $\mathrm{MN}_{n \times p}(M, U, V)$ is

$$(2\pi)^{-np/2} |V|^{-n/2} |U|^{-p/2} \exp\left(-\frac{1}{2} \mathrm{tr}[V^{-1}(X - M)^T U^{-1}(X - M)] \right), \tag{3.12}$$

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, tr stands for the trace of a matrix, and $|U|$ and $|V|$ denote the determinants of U and V respectively.

Thus $W|\Theta \sim \mathrm{MN}_{n \times (K-1)}(Z\Theta, I_n, \Sigma)$. 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 $\Theta \sim \mathrm{MN}_{J \times (K-1)}(U_0, V_0, \Sigma)$. By a standard conjugacy calculation, the posterior is given by

$$\begin{aligned} &\Theta|Y, W \sim \mathrm{MN}_{J \times (K-1)}(U_n, V_n, \Sigma), \\ &V_n = (Z^T Z + V_0^{-1})^{-1},\ U_n = V_n(Z^T W + V_0^{-1} U_0). \end{aligned} \tag{3.13}$$

To draw a sample of W, we use the method introduced by McCulloch and Rossi (1994). Let $W_{i,-l}$ denote $(W_{i1}, \ldots, W_{i,l-1}, W_{i,l+1}, \ldots, W_{iK-1})^T$, $Z_{i,\cdot}$ denote the ith row of Z, the vector $\Theta_{\cdot,l}$ denote the lth column of Θ, the matrix $\Theta_{\cdot,-l}$ denote Θ without the lth column, the scalar $\Sigma_{l,l}$ denote the (l,l)th entry of Σ, $\Sigma_{-l,-l}$ denote Σ without the lth row and the lth column, $\Sigma_{-l,l}$ denote the lth column of Σ without the lth entry, and $\Sigma_{l,-l}$ denote the lth row of Σ without the lth entry. We draw W_{il} from the conditional truncation of the normal distribution with the mean m_{il} and variance τ_{il}^2 to the interval (a, b) described

below:

$$
\begin{aligned}
&W_{il}|(W_{i,-l},\Theta,Y_i)\sim \mathrm{TN}(m_{il},\tau_{il}^2,a,b),\\
&m_{il}=Z_{i,\cdot}\Theta_{\cdot,l}+\Sigma_{-l,l}^T\Sigma_{-l,-l}^{-1}(W_{i,-l}-Z_{i,\cdot}\Theta_{\cdot,-l}),\\
&\tau_{il}^2=\Sigma_{l,l}-\Sigma_{l,-l}\Sigma_{-l,-l}^{-1}\Sigma_{-l,l},\\
&(a,b)=\begin{cases}(\max\{W_{i,-l},0\},\infty), & \text{if } Y_i=l,\ l=1,\dots,K-1,\\ (-\infty,\max\{W_{i,-l}\}), & \text{if } Y_i\neq l,\ l=1,\dots,K-1,\\ (-\infty,0), & \text{if } Y_i=K,\end{cases}\\
&i=1,\dots,n,\ l=1,\dots,K-1.
\end{aligned}
\tag{3.14}
$$

To implement the Gibbs sampling, we draw samples from (3.13) and (3.14).

3.3. Multinomial logistic model

Let $\beta_k(t)=\sum_{j=1}^J\theta_{kj}\psi_j(t)$. Then (2.7) and (2.8) can be rewritten as

$$
\mathrm{P}(Y_i=k|X_i)=\frac{\exp[\sum_{j=1}^J\theta_{kj}Z_{ij}]}{1+\sum_{l=1}^{K-1}\exp[\sum_{j=1}^J\theta_{lj}Z_{ij}]},\ k=1,\dots,K-1, \tag{3.15}
$$

$$
\mathrm{P}(Y_i=K|X_i)=\frac{1}{1+\sum_{l=1}^{K-1}\exp[\sum_{j=1}^J\theta_{lj}Z_{ij}]}. \tag{3.16}
$$

Define $\theta_k=(\theta_{k1},\dots,\theta_{kJ})^T, k=1,\dots,K-1$, and $Z_i=(Z_{i1},\dots,Z_{iJ})^T$. Then (3.15) and (3.16) are given by

$$
\mathrm{P}(Y_i=k|X_i)=\frac{\exp[Z_i^T\theta_k]}{1+\sum_{l=1}^{K-1}\exp[Z_i^T\theta_l]},\ k=1,\dots,K-1, \tag{3.17}
$$

$$
\mathrm{P}(Y_i=K|X_i)=\frac{1}{1+\sum_{l=1}^{K-1}\exp[Z_i^T\theta_l]}. \tag{3.18}
$$

For each θ_k, $k=1,\dots,K-1$, we assign a multivariate normal prior $\mathrm{N}_J(\mu_k,\Sigma_k)$, and apply Metropolis-Hastings algorithm to sample θ_k.

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

$$
\rho(\theta_k,\theta_k'|Y,\theta_{-k})=\min\Big\{\frac{f(Y|\theta_k',\theta_{-k})\pi(\theta_k',\theta_{-k})}{f(Y|\theta_k,\theta_{-k})\pi(\theta_k,\theta_{-k})}\frac{q(\theta_k',\theta_k|Y,\theta_{-k})}{q(\theta_k,\theta_k'|Y,\theta_{-k})},1\Big\}, \tag{3.19}
$$

where θ_{-k} denotes all the blocks except the kth one.

4. Marginal likelihood and model averaging

In Section 3, we described the MCMC sampling technique for a given J value, which we need to repeat it for all possible J values. In the actual computation, however, it is impossible to consider all values of J. With a given prior on J, for example, geometric or Poisson distribution, the posterior probabilities for very small or very large values of J decay to zero very quickly. Thus, we do not need to consider these J values. Let $J_1, \ldots, J_S$ denote the values of J we need to consider. If we can get the marginal likelihood $m(Y|J_s)$, then we can compute the posterior probability of J_s using Bayes's rule

$$\mathrm{P}(J_s|Y) = \frac{m(Y|J_s)p(J_s)}{\sum_{l=1}^{S} m(Y|J_l)p(J_l)}, \tag{4.1}$$

where $p(J_s)$, $s = 1, \ldots, S$, is the prior probability for $J = J_s$.

For each given J_s, we have a misclassification rate r_s, which is defined as the ratio of the number of falsely classified data to the total number of data. Then we can obtain the average misclassification rate $\bar{r}$ for each multinomial model:

$$\bar{r} = \sum_{s=1}^{S} \mathrm{P}(J_s|Y) \cdot r_s. \tag{4.2}$$

We call it the model averaging method.

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

$$m(Y|J_s) = \frac{f(Y|J_s, B)\pi(B|J_s)}{\pi(B|Y, J_s)}, \tag{4.3}$$

where B is a convenient value of the parameter in the context of the support of the posterior distribution such as the posterior mean, because (4.3) 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

$$\log \hat{m}(Y|J_s) = \log f(Y|J_s, B^*) + \log \pi(B^*|J_s) - \log \hat{\pi}(B^*|Y, J_s). \tag{4.4}$$

The details of $\pi(B^*|Y, J_s)$ estimation for each model are described in Appendix A.

5. Posterior contraction rate

For a classification problem, the most important object to study is the misclassification rate. By examining convergence to the true distribution, it follows that the Bayes procedure has misclassification rate close to that of the oracle procedure which uses the true values of the regression functions and other

parameters (if any), e.g., cut-points in the ordered multinomial probit model. In the Bayesian nonparametric setting, Hellinger convergence is established by applying the general theory (Ghosal and van der Vaart, 2017). Thus, in this section, we only consider the contraction rate of the posterior distribution with respect to a metric on the probability of categories, which is equivalent with the Hellinger distance on the joint distribution. The posterior contraction rates of the three multinomial models with finite random series priors can be obtained using calculation similar to those in Shen and Ghosal (2015) on posterior contraction rates for finite random series.

We use $\lesssim$ to denote an inequality up to a constant multiple, $f \asymp g$ 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 $1 \le p < \infty$, and $\|\theta\|_\infty = \max_{1\le i\le d}|\theta_i|$. Similarly, for a function f with respect to a measure G, we define $\|f\|_{p,G} = \{\int |f(x)|^p dG\}^{1/p}$, where $1 \le p < \infty$, and $\|f\|_{\infty,G} = \sup_x |f(x)|$. Let $\mathcal{N}(\epsilon, T, d)$ denote the ϵ-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_i, Y_i), $i = 1,\dots,n$, are the independent observations. Let p denote the joint probability of (X,Y), where Y takes values $1,\dots,K$, and p_0 denote the true joint probabilty. Let $(X^{(n)}, Y^{(n)})$ be the vector of n obeservations following the probability $p^{(n)}$. Let $\pi_k(X) = \mathrm{P}(Y = k|X)$ be the probability of the kth category conditioned on X, and π_{0k} be the true probablity of the kth category conditioned on X. Define the probability vector $\pi = (\pi_1,\dots,\pi_K)^T$, where $\pi_K = 1 - \sum_{k=1}^{K-1}\pi_k$, and $\pi_0 = (\pi_{01},\dots,\pi_{0K})^T$, where $\pi_{0K} = 1 - \sum_{k=1}^{K-1}\pi_{0k}$. Assume that the distribution of X is G, and ν denotes the counting measure on $\{1,\dots,K\}$. For these multinomial models, the KL divergences $K(p_0,p)$, and $V(p_0,p)$ can be reduced to

$$\begin{aligned} K(p_0,p) &= \int\int p_0(x,y)\log\frac{p_0(x,y)}{p(x,y)}\,\mathrm{d}\nu(y)\,\mathrm{d}x \\ &= \int\int \pi_0(y|x)\log\frac{\pi_0(y|x)}{\pi(y|x)}\,\mathrm{d}\nu(y)\,\mathrm{d}G(x) \\ &= \mathrm{E}_X\Big\{\sum_{k=1}^K \pi_{0k}(X)\log\frac{\pi_{0k}(X)}{\pi_k(X)}\Big\} \\ &= K(\pi_0,\pi),\ \text{say;} \end{aligned} \tag{5.1}$$

$$\begin{aligned} V(p_0,p) &= \int\int p_0(x,y)\log^2\frac{p_0(x,y)}{p(x,y)}\,\mathrm{d}\nu(y)\,\mathrm{d}x \\ &= \int\int \pi_0(y|x)\log^2\frac{\pi_0(y|x)}{\pi(y|x)}\,\mathrm{d}\nu(y)\,\mathrm{d}G(x) \\ &= \mathrm{E}_X\Big\{\sum_{k=1}^K \pi_{0k}(X)\log^2\frac{\pi_{0k}(X)}{\pi_k(X)}\Big\} \\ &= V(\pi_0,\pi),\ \text{say.} \end{aligned} \tag{5.2}$$

Similarly, the squared Hellinger distance $h^2(p_1, p_2)$ can be reduced to

$$\begin{aligned} h^2(p_1,p_2) &= \int\int \left(\sqrt{\pi_1(y|x)} - \sqrt{\pi_2(y|x)}\right) \mathrm{d}\nu(y)\, \mathrm{d}G(x) \\ &= \mathrm{E}_X\Big\{\sum_{k=1}^{K} \left(\sqrt{\pi_{1k}(X)} - \sqrt{\pi_{2k}(X)}\right)^2\Big\} \\ &= h^2(\pi_1,\pi_2), \text{ say.} \end{aligned} \tag{5.3}$$

We define a metric by

$$d(\pi,\pi_0) = \sqrt{\sum_{k=1}^{K} \mathrm{E}_X|\pi_k(X) - \pi_{0k}(X)|^2}. \tag{5.4}$$

Then we have the following general posterior contraction theorem suitable in our context.

Theorem 1. *Assume that π_0 is bounded away from zero. Let $\epsilon_n \geq \bar{\epsilon}_n$ be two sequences of positive numbers satisfying $\epsilon_n \to 0$ and $n\bar{\epsilon}_n^2 \to \infty$. Let $\mathcal{X}_0$ be such that $P(X \in \mathcal{X}_0) = 1$ and $\pi_k(x), k = 1, \ldots, K$ for $x \in \mathcal{X}_0$ is bounded away from 0. Let $\mathcal{W}_n$ be a subset of the parameter space such that the following conditions hold for some positive constants a_2 and $a_1 > a_2 + 2$:*

$$\log \mathcal{N}(\epsilon_n, \mathcal{W}_n, h) \lesssim n\epsilon_n^2, \tag{5.5}$$

$$\Pi(\pi \notin \mathcal{W}_n) \leq \exp\{-a_1 n\bar{\epsilon}_n^2\}, \tag{5.6}$$

$$-\log \Pi\left(\sum_{k=1}^{K} \|\pi_k - \pi_{0k}\|_{\infty,\mathcal{X}_0}^2 \leq \bar{\epsilon}_n^2\right) \leq a_2 n\bar{\epsilon}_n^2, \tag{5.7}$$

where $\|\pi_k - \pi_{0k}\|_{\infty,\mathcal{X}_0} = \sup_{x\in\mathcal{X}_0} |\pi_k(x) - \pi_{0k}(x)|$. 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.

The proof follows from Theorem 4 of Ghosal and van der Vaart (2007a), by observing that

$$h^2(\pi,\pi_0) = \mathrm{E}_X \sum_{k=1}^{K} \frac{|\pi_k(X) - \pi_{0k}(X)|^2}{|\sqrt{\pi_k(X)} + \sqrt{\pi_{0k}(X)}|^2} \gtrsim \mathrm{E}_X \sum_{k=1}^{K} |\pi_k(X) - \pi_{0k}(X)|^2, \tag{5.8}$$

and by expanding in Taylor's expansion

$$\max\{K(\pi_0,\pi), V(\pi_0,\pi)\} \lesssim \sum_{k=1}^{K} \|\pi_k - \pi_{0k}\|_{\infty,\mathcal{X}_0}^2. \tag{5.9}$$

Let Π be a generic notation for priors on the number J of basis functions. As in 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 conditons (A1) and (A2). For the ordered multinomial probit model, we add condition (A3).

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

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

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

Geometric distribution with $t_1 = t_2 = 0$, and Poisson distribution with $t_1 = t_2 = 1$ on J satisfy (A1). The multivariate normal distribution on θ and γ satisfy (A2) and (A3) respectively.

To obtain the posterior contraction rate, we need to verify the conditions (5.5)–(5.7), 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 ∞. Assume that there exist a $\theta_0 \in \mathbb{R}^J$, $\|\theta_0\|_\infty \le H$ and $K_0 \ge 0$ such that

$$\|\beta_0(\cdot) - \theta_0^T \psi(\cdot)\|_r \lesssim J^{-\alpha}, \tag{5.10}$$

$$\|\theta_1^T \psi(\cdot) - \theta_2^T \psi(\cdot)\|_r \lesssim J^{K_0} \|\theta_1 - \theta_2\|_2, \ \theta_1, \theta_2 \in \mathbb{R}^J. \tag{5.11}$$

Remark 2 of Shen and Ghosal (2015) gave examples of bases satisfying relations (5.10) and (5.11). For B-splines, the relations hold when $K_0 = 1/2$ with $r = 2$, and $K_0 = 1$ with $r = \infty$.

Remark 1. Parameter estimation plays a secondary role here. The problem of estimating model parameters is interesting in its own right but is not necessary for good classifications. Cai and Hall (2006) and Yuan and Cai (2010) showed that the parameter function estimation and the prediction from an estimator of the parameter function have different characteristics.

5.1. Ordered multinomial probit model

Let $\gamma = (\gamma_1, \ldots, \gamma_K)^T$ be the vector of the threshold points, and $\gamma_0 = (\gamma_{01}, \ldots, \gamma_{0K})^T$ 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

$$\pi_k(X) = \Phi\Big(\gamma_k - \int \beta(t) X(t) dt\Big) - \Phi\Big(\gamma_{k-1} - \int \beta(t) X(t) dt\Big), \tag{5.12}$$

and

$$\pi_{0k}(X) = \Phi\Big(\gamma_{0k} - \int \beta_0(t) X(t) dt\Big) - \Phi\Big(\gamma_{0k-1} - \int \beta_0(t) X(t) dt\Big). \tag{5.13}$$

Theorem 2. *Assume that* $\|X\|_1 = \int |X(t)| \, \mathrm{d}t$ *is a bounded random variable, the priors satisfy the conditons (A1), (A2) and (A3), and that the basis* $\psi(t)$

satisfies (5.10) and (5.11) with $r = \infty$. Then the posterior contraction rate of the ordered multinomial probit model is $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_2)/2}$ relative to $d(\pi, \pi_0)$. More explicitly, for every $M_n \to \infty$, $\Pi(\beta : \rho(\beta, \beta_0) \geq M_n\epsilon_n | X^{(n)}, Y^{(n)}) \to 0$ in probability, where $\rho(\beta, \beta_0) = \mathrm{E}_X |\int (\beta(t) - \beta_0(t))X(t)\,\mathrm{d}t|$, and $\Pi(\gamma : \max_j |\gamma_j - \gamma_{0j}| \geq M_n\epsilon_n | X^{(n)}, Y^{(n)}) \to 0$ in probability.

Proof. For any $x \in \mathcal{X}_0 = \{\int |X(t)|\,\mathrm{d}t \leq M\}$, say, by the Lipschitz continuity of Φ, we have

$$\begin{aligned} |\pi_k(x) - \pi_{0k}(x)| &\lesssim \max_k |\gamma_k - \gamma_{0k}| + \left| \int (\beta(t) - \beta_0(t)x(t)dt \right| \\ &\lesssim \|\gamma - \gamma_0\|_\infty + \|\beta(\cdot) - \beta_0(\cdot)\|_\infty \int |x(t)|\,dt \\ &\lesssim \|\gamma - \gamma_0\|_\infty + \|\beta(\cdot) - \beta_0(\cdot)\|_\infty. \end{aligned} \tag{5.14}$$

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

$$\begin{aligned} \|\beta(\cdot) - \beta_0(\cdot)\|_\infty &= \|\theta^T\psi(\cdot) - \theta_0^T\psi(\cdot) + \theta_0^T\psi(\cdot) - \beta_0(\cdot)\|_\infty \\ &\leq \|\theta^T\psi(\cdot) - \theta_0^T\psi(\cdot)\|_\infty + \|\theta_0^T\psi(\cdot) - \beta_0(\cdot)\|_\infty. \end{aligned} \tag{5.15}$$

Then we have

$$\begin{aligned} &\Pi\Big(\sum_{k=1}^K \|\pi_k - \pi_{0k}\|_{\infty,\mathcal{X}_0}^2 \leq \bar{\epsilon}_n^2\Big) \\ &\geq \Pi(\|\gamma - \gamma_0\| \leq \bar{\epsilon}_n/\sqrt{2})\Pi(\|\beta(\cdot) - \beta_0(\cdot)\|_\infty \leq \bar{\epsilon}_n/\sqrt{2}) \\ &\geq \Pi(\|\gamma - \gamma_0\| \leq \bar{\epsilon}_n/\sqrt{2})\Pi(\|\theta - \theta_0\| \leq \bar{\epsilon}_n/(2\sqrt{2}\bar{J}_n^{K_0})) \\ &\gtrsim \exp\{-K\log(\sqrt{2}/\bar{\epsilon}_n)\}\exp\{-\bar{J}_n\log(2\sqrt{2}\bar{J}_n^{K_0}/\bar{\epsilon}_n)\}. \end{aligned} \tag{5.16}$$

To satisfy the relation (5.7), we need $\bar{J}_n^{-\alpha} \lesssim \bar{\epsilon}_n$ and

$$K\log(\sqrt{2}/\bar{\epsilon}_n) + \bar{J}_n\log(2\sqrt{2}\bar{J}_n^{K_0}/\bar{\epsilon}_n) \lesssim n\bar{\epsilon}_n^2. \tag{5.17}$$

Thus (5.17) leads to the conditions that $\bar{J}_n \log n \lesssim n\bar{\epsilon}_n^2$. Then we obtain the preliminary contraction rate $\bar{\epsilon}_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)}$, for $\bar{J}_n \asymp (n/\log n)^{1/(2\alpha+1)}$.

Using (5.14), we obtain

$$\log \mathcal{N}(\epsilon_n, \mathcal{W}_n, h) \lesssim \log \mathcal{N}(\epsilon_n, \mathcal{W}_n, \|\cdot\|_\infty) \lesssim n\epsilon_n^2. \tag{5.18}$$

According to Theorem 2 of Shen and Ghosal (2015), to satify (5.18), we need

$$J_n\{(K_0+1)\log J_n + \log M_n + C_0\log n\} \leq n\epsilon_n^2, \tag{5.19}$$

for some positive constant C_0. To satify (5.6), we need

$$bn\bar{\epsilon}_n^2 \leq J_n\log^{t_2} J_n,\ \log J_n + n\bar{\epsilon}_n^2 \leq M_n^{t_3}, \tag{5.20}$$

for some $b > 0$. For $M_n = n^{1/t_3}$, (5.20) implies that $J_n \log^{t_2} n \gtrsim n\bar{\epsilon}_n^2$. Thus $J_n \asymp n^{1/(2\alpha+1)}(\log n)^{2\alpha/(2\alpha+1)-t_2}$. Relation (5.19) implies that $J_n \log n \lesssim n\epsilon_n^2$. As a result, 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 $d(\pi, \pi_0)$.

Further, by Jensen's inequality, we have

$$\mathrm{E}_X|\pi_k(X) - \pi_{0k}(X)|^2 \geq \Big\{\mathrm{E}_X|\pi_k(X) - \pi_{0k}(X)|\Big\}^2. \tag{5.21}$$

If $k = 1$, by the mean value theorem and the uniform positivity of Φ on compact interval, then

$$\begin{aligned}\mathrm{E}_X|\pi_1(X) - \pi_{01}(X)| &= \mathrm{E}_X\left|\Phi(-\int \beta(t)X(t)dt) - \Phi(-\int \beta_0(t)X(t)dt)\right| \\ &\gtrsim \mathrm{E}_X\left|\int \beta(t)X(t)dt - \int \beta_0(t)X(t)dt\right|.\end{aligned} \tag{5.22}$$

Hence if $\mathrm{E}_X|\pi_1(X) - \pi_{01}(X)|^2 \leq \epsilon_n^2$, then $\mathrm{E}_X\left|\int \beta(t)X(t)dt - \int \beta_0(t)X(t)dt\right| \lesssim \epsilon_n$. If $k = 2$, we have

$$\begin{aligned}\mathrm{E}_X|\pi_2(X) - \pi_{02}(k)| &= \mathrm{E}_X\left|\Phi(\gamma_2 - \int \beta(t)X(t)dt) - \Phi(\gamma_{02} - \int \beta_0(t)X(t)dt)\right. \\ &\qquad \left. - \Phi(-\int \beta(t)X(t)dt) + \Phi(-\int \beta_0(t)X(t)dt)\right| \\ &\gtrsim \mathrm{E}_X\left|\Phi(\gamma_2 - \int \beta(t)X(t)dt) - \Phi(\gamma_{02} - \int \beta_0(t)X(t)dt)\right| \\ &\qquad - \mathrm{E}_X\left|\Phi(-\int \beta(t)X(t)dt) - \Phi(-\int \beta_0(t)X(t)dt)\right|.\end{aligned} \tag{5.23}$$

From (5.22), we know that $\mathrm{E}_X\left|\Phi(-\int \beta(t)X(t)dt) - \Phi(-\int \beta_0(t)X(t)dt)\right| \lesssim \epsilon_n$, and if $\mathrm{E}_X|\pi_2(X) - \pi_{02}(X)|^2 \leq \epsilon_n^2$, then

$$\mathrm{E}_X\left|\Phi(\gamma_2 - \int \beta(t)X(t)dt) - \Phi(\gamma_{02} - \int \beta_0(t)X(t)dt)\right| \lesssim \epsilon_n \tag{5.24}$$

By the mean value theorem and the uniform positivity of Φ on compact interval, we have

$$\begin{aligned}&\mathrm{E}_X\left|\Phi(\gamma_2 - \int \beta(t)X(t)dt) - \Phi(\gamma_{02} - \int \beta_0(t)X(t)dt)\right| \\ &\gtrsim \mathrm{E}_X\left|\gamma_2 - \gamma_{02} - \int \beta(t)X(t)dt + \int \beta_0(t)X(t)dt\right| \\ &\gtrsim |\gamma_2 - \gamma_{02}| - \mathrm{E}_X\left|\int \beta(t)X(t)dt - \int \beta_0(t)X(t)dt\right|.\end{aligned} \tag{5.25}$$

Hence $|\gamma_2 - \gamma_{02}| \lesssim \epsilon_n$. Similarly, we can prove that $|\gamma_k - \gamma_{0k}| \lesssim \epsilon_n$ for any k. □

5.2. Unordered multinomial probit model

Note that by (A.10)

$$\pi_k(X) = \frac{1}{\sqrt{\pi}} \int_0^\infty \Big\{ \prod_{l=1}^{K-1} \Phi(-z\sqrt{2} - \int \beta_l(t)X(t)dt) + \prod_{l=1}^{K-1} \Phi(z\sqrt{2} - \int \beta_l(t)X(t)dt) \Big\} e^{-z^2} dz \tag{5.26}$$

Theorem 3. *Assume that $\|X\|_1 = \int |X(t)|\,\mathrm{d}t$ is a bounded random variable, the priors satisfy the conditons (A1) and (A2), and that the basis $\psi(t)$ satisfies (5.10) and (5.11) with $r = \infty$. Then 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 $d(\pi, \pi_0)$.*

Proof. For some $M > 0$, $\mathrm{P}(\mathcal{X}_0) = 1$ for $\mathcal{X}_0 = \{\int |X(t)|\,\mathrm{d}t \le M\}$. For any $x \in \mathcal{X}_0$, by the Lipschitz continuity of the function Φ, we have

$$\begin{aligned} |\pi_k(x) - \pi_{0k}(x)| &\lesssim \left| \int \beta_k(t)x(t)dt - \int \beta_{0k}(t)x(t)dt \right| \\ &\lesssim \int |\beta_k(t) - \beta_{0k}(t)|\,|x(t)|\,dt \\ &\lesssim \|\beta_k(\cdot) - \beta_{0k}(\cdot)\|_\infty. \end{aligned} \tag{5.27}$$

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

$$\begin{aligned} \|\beta_k(\cdot) - \beta_{0k}(\cdot)\|_\infty &= \|\theta_k^T\psi(\cdot) - \theta_{0k}^T\psi(\cdot) + \theta_{0k}^T\psi(\cdot) - \beta_{0k}(\cdot)\|_\infty \\ &\le \|\theta_k^T\psi(\cdot) - \theta_{0k}^T\psi(\cdot)\|_\infty + \|\theta_{0k}^T\psi(\cdot) - \beta_{0k}(\cdot)\|_\infty. \end{aligned} \tag{5.28}$$

Then we have

$$\begin{aligned} \Pi\Big(\sum_{k=1}^K \|\pi_k - \pi_{0k}\|_{\infty,\mathcal{X}_0}^2 \le \bar{\epsilon}_n^2\Big) &\ge \Pi\Big(\sum_{k=1}^K \|\beta_k(\cdot) - \beta_{0k}(\cdot)\|_\infty^2 \le \bar{\epsilon}_n^2\Big) \\ &\ge \Pi(\|\theta_k - \theta_{0k}\| \le \bar{\epsilon}_n/(2\sqrt{K}\bar{J}_n^{K_0})) \\ &\gtrsim \exp\{-\bar{J}_n \log(2\sqrt{K}\bar{J}_n^{K_0}/\bar{\epsilon}_n)\}. \end{aligned} \tag{5.29}$$

To satisfy the relation (5.7), we need $\bar{J}_n^{-\alpha} \lesssim \bar{\epsilon}_n$ and

$$\bar{J}_n \log(2\sqrt{K}\bar{J}_n^{K_0}/\bar{\epsilon}_n) \lesssim n\bar{\epsilon}_n^2. \tag{5.30}$$

Thus, (5.30) leads to the conditions that $\bar{J}_n \log n \lesssim n\bar{\epsilon}_n^2$. Then we obtain the preliminary contraction rate $\bar{\epsilon}_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)}$, for $\bar{J}_n \asymp (n/\log n)^{1/(2\alpha+1)}$.

Following the same arguments as (5.18)–(5.20), 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 $d(\pi, \pi_0)$. □

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. *Assume that* $\|X\|_1 = \int |X(t)|\,\mathrm{d}t$ *is a bounded random variable, the priors satisfy the conditons (A1) and (A2), and that the basis* $\psi(t)$ *satisfies (5.10) and (5.11) with* $r = \infty$. *Then the posterior contraction rate of the multinomial logistic model is* $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_2)/2}$ *relative to* $d(\pi, \pi_0)$.

Proof. The proof is similar to that of Theorem 3. □

6. Discriminant analysis

As a comparison to those multinomial models, we use Bayesian discriminant analysis to classify the functional data. Instead of modeling the class probability directly, the discriminant analysis uses Bayes's rule to compute the marginal likelihood of Y_i (Gelman et al., 2013). The classical discriminant analysis applies only to multivariate data. For functional data, we can use certain orthogonal linear functions to determine the classification probabilities:

$$(f_{i1}, \ldots, f_{im})^T = \left(\int \beta_1(t) X_i(t) dt, \ldots, \int \beta_m(t) X_i(t) dt \right)^T \tag{6.1}$$

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 consider $\beta_1(t), \ldots, \beta_m(t)$ 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 categories, the set of linear function $(f_1, \ldots, f_m)$ follows a normal distribution with the same covarince matrix: $(f_{il1}, \ldots, f_{ilm})^T \sim \mathrm{N}(\mu_l, \Sigma)$, where μ_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

$$\mathrm{P}(Y_i = k | X_i) = \frac{p_k \cdot \phi(f_{ik1}, \ldots, f_{ikm}; \mu_k, \Sigma)}{\sum_{l=1}^K p_l \cdot \phi(f_{il1}, \ldots, f_{ilm}; \mu_l, \Sigma)}, \tag{6.2}$$

where $\phi(f_1, \ldots, f_m; \mu, \Sigma)$ is the multivariate normal density function with mean μ and covariace Σ, and p_l, $l = 1, \ldots, K$, is the probability of choosing category l.

The variables $f_{il1}, \ldots, f_{ilm}$ are the m principal components of $X_i(t)$ in categoty 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 μ_l for each category l, and the common covariance Σ among all categories, we use the conjugate normal-inverse-Wishart prior with hyperparameters (Gelman et al., 2013) for (μ_l, Σ)

$$\Sigma \sim \mathrm{IW}_{\nu_0}(\Lambda_0^{-1}),\ \mu_l|\Sigma \sim \mathrm{N}(\mu_{l0}, \Sigma/\kappa_0). \tag{6.3}$$

Then the posterior distribution of (μ_l, Σ) can be obtained in the following order

$$\Sigma|Y \sim \mathrm{IW}_{\nu_n}(\Lambda_n^{-1}),\ \mu_l|\Sigma, Y \sim \mathrm{N}(\mu_{ln}, \Sigma/\kappa_n), \tag{6.4}$$

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

$$\Lambda_n = \Lambda_0 + S + \sum_{l=1}^{K} \frac{\kappa_0 n_l}{\kappa_0 + n_l} (\bar{f}_l - \mu_{l0})(\bar{f}_l - \mu_{l0})^T, \tag{6.5}$$

and

$$\kappa_n = \kappa_0 + n,\ \mu_{ln} = \frac{\kappa_0 \mu_{l0} + n_l \bar{f}_l}{\kappa_0 + n_l},\ l = 1, \ldots, K. \tag{6.6}$$

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

$$\mathrm{P}(Y_i = k|X_i) = \frac{p_k \cdot \phi(f_{ik1}, \ldots, f_{ikm}; \mu_k, \Sigma_k)}{\sum_{l=1}^{K} p_l \cdot \phi(f_{il1}, \ldots, f_{ilm}; \mu_l, \Sigma_l)}. \tag{6.7}$$

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

$$\Sigma_l \sim \mathrm{IW}_{\nu_{l0}}(\Lambda_{l0}^{-1}), \quad \mu_l|\Sigma_l \sim \mathrm{N}(\mu_{l0}, \Sigma_l/\kappa_{l0}), \tag{6.8}$$

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

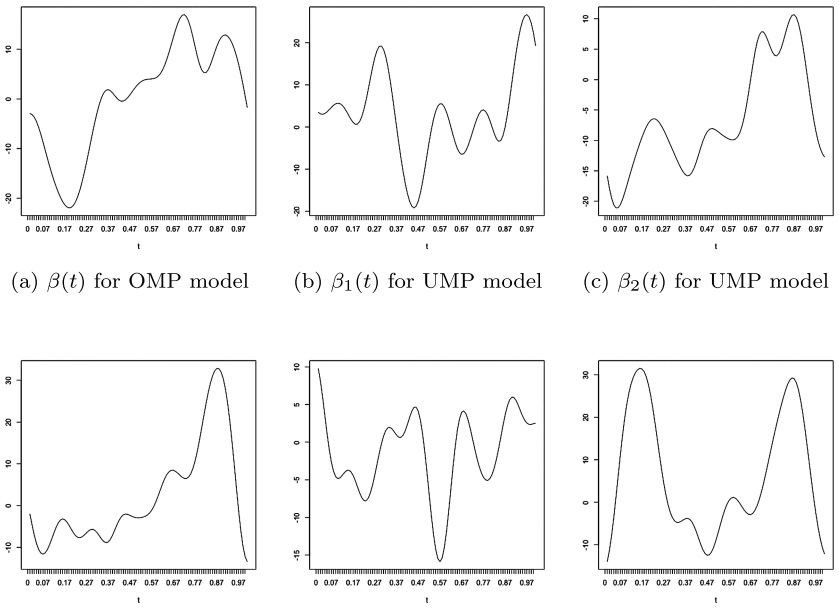
$$\Sigma_l|Y \sim \mathrm{IW}_{\nu_{ln}}(\Lambda_{ln}^{-1}), \quad \mu_l|\Sigma_l, Y \sim \mathrm{N}(\mu_{ln}, \Sigma_l/\kappa_{ln}), \tag{6.9}$$

where $\nu_{ln} = \nu_{l0} + n_l$, $\bar{f}_l = \sum_{i=1}^{n_l} f_{il}/n_l$, $S_l = \sum_{i=1}^{n_l} (f_{il} - \bar{f}_l)(f_{il} - \bar{f}_l)^T$,

$$\Lambda_{ln} = \Lambda_{l0} + S_l + \frac{\kappa_{l0} n_l}{\kappa_{l0} + n_l} (\bar{f}_l - \mu_{l0})(\bar{f}_l - \mu_{l0})^T, \tag{6.10}$$

and

$$\kappa_{ln} = \kappa_{l0} + n_l,\ \mu_{ln} = \frac{\kappa_{l0} \mu_{l0} + n_l \bar{f}_l}{\kappa_{l0} + n_l},\ l = 1, \ldots, K. \tag{6.11}$$



(a) $\beta(t)$ for OMP model (b) $\beta_1(t)$ for UMP model (c) $\beta_2(t)$ for UMP model

(d) $\beta_3(t)$ for UMP model (e) $\beta_1(t)$ for MLO model (f) $\beta_2(t)$ for MLO model

FIG 1. *Coefficient functions for the multinomial models.*

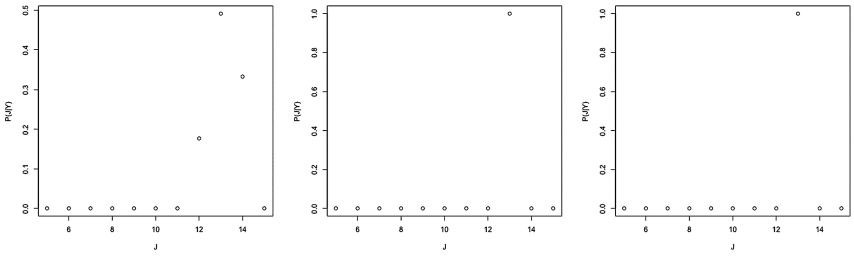
7. Simulation

7.1. Data generation

The simulated data are generated following different data generating process. All of the simulated data have three categories. In all cases considered below, we generate the functional data from a Gaussian process at discrete time points 0, 0.01, ..., 0.99, 1, with the mean function $\sin t$ and variation kernel $100\exp\{-100(t_i - t_j)^2\}$, where t_i and t_j were the discrete time point 0, 0.01, ..., 0.99, 1.

For the ordered multinomial probit data, the coefficient function $\beta(t)$ is plotted in Figure 1 (a), and the four threshold points are chosen to be $-\infty$, 0, 8, ∞. The four cut-off points construct three intervals. If the inner product of a functional data and the coefficient function plus a standard normal variable falls in the kth interval (γ_{k-1}, γ_k), then the functional data attributes to the category k.

For unordered multinomial probit data, the coefficient functions $\beta_1(t), \beta_2(t)$, $\beta_3(t)$ are plotted in Figure 1(b)–(d). The inner product of a functional data and the three coefficient functions are added with standard normal variables, respectively. We sample from these three normal variables, and obtain the corresponding probabilities. Then the functional data belongs to the category with the largest probability.



(a) Under OMP model (b) Under UMP model (c) Under MLO model

FIG 2. *Posterior probabilities of* $J = 5, \ldots, 15$*, for ordered multinomial probit data under different models.*

For the multinomial logistic data, the coefficient functions $\beta_1(t), \beta_2(t)$ are plotted in Figure 1(e)–(f), and the third coefficient function $\beta_3(t)$ can be assumed to be zero everywhere. We compute the probabilities of a functional data falling into each category. Then the data attributes to the category with the largest probability.

To generate data satisfying the assumption of the linear discriminant analysis, we use three Gaussian processes with different mean functions $\sin t + 2\cos t$, $\sin t$, and $\sin t - 3\cos t$, but the same variation kernel $\exp\{-30(t_i - t_j)^2\}$.

To generate data satisfying the assumption of the quadratic discriminant analysis, we use three Gaussian processes with different mean functions and different variation kernels. The mean functions are $\sin t + 2\cos t$, $\sin t$, and $\sin t - 3\cos t$, and the variation kernels are $\exp\{-2\sin^2(\pi(t_i - t_j))\}$, $\exp\{-30(t_i - t_j)^2\}$, and $\exp\{-|t_i - t_j|\}$, respectively.

In this simulation study, we generate total 900 (300 for each category) functional data for each type of dataset. We constructe 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 the B-spline basis. The B-spline basis functions on interval $[0, 1]$ can be created using the R package `fda`. In this simulation study, we put a geometric prior with $p = 0.5$ on J. We only consider the possible number of B-spline basis functions to be $J = 5, \ldots, 15$, since the probability outside this range is too small. Figure 2 shows the posterior probabilities of $J = 5, \ldots, 15$ for the simulated ordered multinomial probit data under ordered multinomial probit, unordered multinomial probit, and multinomial logistic models. We can see that the posterior probabilities for small and large values of J decay to zero very quickly. Similarly for other types of data under these models, the posterior probabilites of J also decay very quicky outside of this range. The B-spline basis functions are generated at the same discrete time points as the functional data, that is $0, 0.01, \ldots, 0.99, 1$.

TABLE 1
Averaged misclassification rates for simulated data

Dataset	OMP Model	UMP Model	MLO Model	LDA	QDA	SVM
OMP	7.69%	30.56%	28.33%	38.89%	48.89%	15.00%
UMP	38.96%	7.22%	7.78%	21.11%	21.11%	10.56%
MLO	49.44%	4.75%	3.89%	32.22%	36.11%	7.78%
LDA	26.32%	25.69%	26.11%	5.00%	5.00%	7.78%
QDA	24.28%	21.95%	21.67%	10.56%	9.44%	8.33%

7.3. Results

Under the chosen models, we apply Baysian estimation methods described in Section 3 on the training data. In this study, 5000 MCMC iterations are obtained, and the first 1000 of them are 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, $B = \Theta$ for the unorederd 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 $B = (\mu_1, \mu_1, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3)$ for the quadratic discriminant analysis model. A transformed testing data z_i or f_i is in categoty k if $\sum_{g=1}^{4000} \mathbb{1}(Y_i = k | z_i \text{ or } f_i, B^{(g)}) > \sum_{g=1}^{4000} \mathbb{1}(Y_i = l | z_i \text{ or } f_i, B^{(g)})$, 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) is 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.

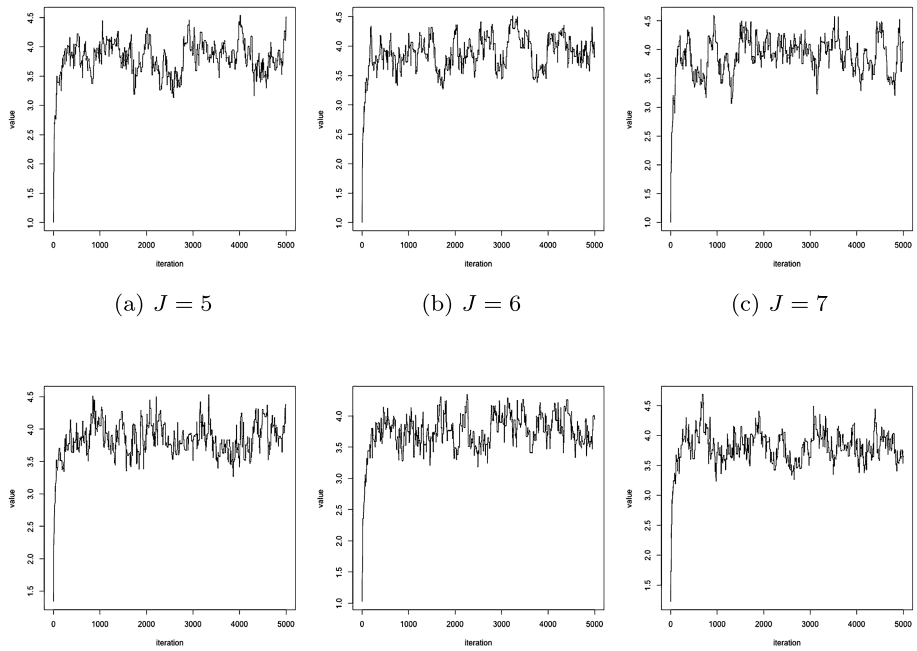
8. Application

We also test our models on a phoneme dataset. 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 use 900 of them from three categories. We split the data into training and testing set by randomly sampling from each class, and keeping the same percentage of samples of each class as the complete set. The size of the testing data is 20% of the total data size. That is we have 240 data for each class in the training set, and 60 data for each class in the testing set. We put a geometric prior with $p = 0.5$ on J, and it is enough for us to consider the number of B-spline basis functions to be $J = 5, \ldots, 15$. We obtain 5000 MCMC iterations and discard 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

TABLE 2
Averaged misclassification rates for phoneme data

OMP Model	UMP Model	MLO Model	LDA	QDA
9.84%	0.56%	5.56%	7.78%	5.00%



(a) $J = 5$ (b) $J = 6$ (c) $J = 7$

(d) $J = 13$ (e) $J = 14$ (f) $J = 15$

FIG 3. γ_2 *sampled from Metropolis-Hastings when* $J = 5 - 7$ *and* $13 - 15$.

TABLE 3
Estimate and standard error of the posterior mean for the ordered multinomial model ($J = 6$)

	γ_2	θ
Estimate	3.87	$(52.12, -9.60, -8.89, -0.19, -4.91, 2.85)$
Standard error	0.03	$(0.34, 0.11, 0.13, 0.08, 0.10, 0.10)$

problem, but we include it in the analysis for comparison. Figure 3 displays the cut-point γ_2 sampled by Metropolis-Hastings under different J, and we can tell that γ_2 converges around 500 iterations. Tables 3, 4, and 5 show the estimate and standard error of the posterior mean of the phoneme data under ordered multinomial probit model, multinomial logistic model, and unordered multinomial probit model, when $J = 6$, $J = 10$, and $J = 14$, respectively. We choose these J values because under these values the model has the largest posterior probability $\mathrm{P}(J|Y)$. Although ordered multinomial probit model is not intuitive in this context, its performance is not too inferior.

TABLE 4
Estimate and standard error of the posterior mean for the multinomial logistic model ($J = 10$)

	θ_2
Estimate	$(13.10, 18.25, 6.04, -15.29, 15.52, 1.30, -5.81, 4.65, -28.24, -16.91)$
Standard error	$(0.94, 1.08, 0.64, 0.63, 0.75, 0.66, 0.37, 0.48, 0.70, 1.03)$
	θ_3
Estimate	$(39.42, 34.30, -3.47, 5.26, -7.36, 0.38, -17.99, -4.43, -11.23, 2.44)$
Standard error	$(1.21, 1.44, 0.42, 0.31, 1.08, 0.27, 0.53, 0.33, 0.93, 0.32)$

TABLE 5
Estimate and standard error of the posterior mean for the unordered multinomial model ($J = 14$)

	estimate		standard error	
Θ	-15.40	49.92	0.93	0.85
	35.78	79.79	0.52	0.81
	45.94	32.11	0.58	0.75
	4.97	-0.76	0.66	0.73
	-23.23	-12.58	0.60	0.71
	-15.09	-14.88	0.62	0.67
	23.43	-21.71	0.68	0.73
	-11.87	1.67	0.74	0.80
	-0.96	-5.06	0.63	0.64
	-0.27	-9.82	0.63	0.69
	1.58	-7.46	0.70	0.78
	-12.43	-14.68	0.57	0.60
	-28.97	-7.74	0.64	0.69
	-28.49	-3.38	0.53	0.57

Appendix A: Posterior density estimation from MCMC output

A.1. Ordered multinomial probit model

There are two parameter blocks in this model, θ and α, where α is the transformation of γ as in (3.6). 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

$$\pi(\theta^*, \alpha^*|Y, J_s) = \pi(\alpha^*|Y, J_s)\pi(\theta^*|Y, J_s, \alpha^*), \tag{A.1}$$

where

$$\pi(\theta^*|Y, J_s, \alpha^*) = \int \pi(\theta^*|Y, J_s, \alpha^*, W)\pi(W|Y, J_s, \alpha^*)dW. \tag{A.2}$$

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

$$\hat{\pi}(\theta^*|Y, J_s, \alpha^*) = M^{-1}\sum_{m=1}^{M}\pi(\theta^*|Y, J_s, \alpha^*, W^{(m)}), \tag{A.3}$$

where $\{W^{(m)}\}_{m=1}^{M}$ are sampled from distribution $[W|Y, J_s, \alpha^*]$. The draws of W from the Gibbs sampler are from the distribution $[W|Y, J_s]$, so $\pi(\theta^*|Y, J_s, \alpha^*, W)$

cannot be averaged directly by the Gibbs sampling output. Additional sampling for W is needed. We sample $\{\theta^{(m)}\}$ from the density $\pi(\theta|Y, J_s, \alpha^*, W)$, and given that $\theta^{(m)}$, we sample $\{W^{(m)}\}$ from $\pi(W|Y, J_s, \theta, \alpha^*)$.

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

$$\pi(\alpha^*|Y, J_s) = \frac{\mathrm{E}_1\{\rho(\alpha, \alpha^*|Y, J_s, \theta, W)q(\alpha, \alpha^*|Y, J_s, \theta, W)\}}{\mathrm{E}_2\{\rho(\alpha^*, \alpha|Y, J_s, \theta, W)\}}, \tag{A.4}$$

where $\rho(\alpha, \alpha^*|Y, J_s, \theta, W)$ is defined in (3.8), $q(\alpha, \alpha^*|Y, J_s, \theta, W)$ is the proposal density, the expectation E_1 is with respect to the distribution $\pi(\theta, \alpha, W|Y, J_s)$, and E_2 is with respect to the distribution $\pi(\theta, W|Y, J_s, \alpha^*) \times q(\alpha^*, \alpha|Y, J_s, \theta, W)$.

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

$$\frac{G^{-1}\sum_{g=1}^{G}\rho(\alpha^{(g)}, \alpha^*|Y, J_s, \theta^{(g)}, W^{(g)})q(\alpha^{(g)}, \alpha^*|Y, J_s, \theta^{(g)}, W^{(g)})}{M^{-1}\sum_{m=1}^{M}\rho(\alpha^*, \alpha^{(m)}|Y, J_s, \theta^{(m)}, W^{(m)})}, \tag{A.5}$$

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

A.2. Unordered multinomial probit model

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

$$\pi(\Theta^*|Y, J_s) = \int \pi(\Theta^*|Y, J_s, W)\pi(W|Y, J_s)dW. \tag{A.6}$$

Then the Monte Carlo estimate of $\pi(\Theta^*|Y, J_s)$ is

$$\hat{\pi}(\Theta^*|Y, J_s) = \sum_{g=1}^{G}\pi(\Theta^*|Y, J_s, W^{(g)}), \tag{A.7}$$

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

For the unordered multinomial probit model, we also need to estimate the likelihood at some convenient values in the support of the posterior distribution. From Section 3.2, $\Theta = (\theta_1, \ldots, \theta_{K-1})$, where $\theta_l = \theta'_l - \theta'_K$, $l = 1, \ldots, K-1$. Then (2.5) can be rewritten as

$$\begin{aligned}&\mathrm{P}(Y = K)\\ &= \frac{1}{(2\pi)^{(K-1)/2}|\Sigma|^{1/2}}\int_{-\infty}^{-Z^T\Theta_{\cdot,1}}\cdots\int_{-\infty}^{-Z^T\Theta_{\cdot,K-1}}\exp\big(-\frac{1}{2}U^T\Sigma^{-1}U\big)dU,\end{aligned} \tag{A.8}$$

where $\Theta_{\cdot,l}$ denotes the lth column of Θ.

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

$$\begin{aligned} &\mathrm{P}(Y = l) \\ &= \frac{1}{(2\pi)^{(K-1)/2}|\Sigma|^{1/2}} \int_{-\infty}^{-Z^T\Theta^l_{\cdot,1}} \cdots \int_{-\infty}^{-Z^T\Theta^l_{\cdot,K-1}} \exp\big(-\frac{1}{2}U^T\Sigma^{-1}U\big)dU. \end{aligned} \tag{A.9}$$

Due to the exchangeable correlation structure of Σ, (A.9) can be reduced to a one dimensional integral (Dunnett, 1989) given by

$$\begin{aligned} &\mathrm{P}(Y = l) \\ &= \frac{1}{\sqrt{\pi}} \int_0^\infty \Big\{ \prod_{k=1}^{K-1} \Phi(-u\sqrt{2} - Z^T\Theta^l_{\cdot,k}) + \prod_{k=1}^{K-1} \Phi(u\sqrt{2} - Z^T\Theta^l_{\cdot,k}) \Big\} e^{-u^2} du. \end{aligned} \tag{A.10}$$

The expression in (A.8) can also be reduced to the same form as in (A.10). Then (A.10) can be approximated by a Gaussian quadrature as follows

$$\mathrm{P}(Y = l) \approx \frac{1}{2} w_q \Big\{ \prod_{k=1}^{K-1} \Phi(-\sqrt{2x_q} - Z^T\Theta^l_{\cdot,k}) + \prod_{k=1}^{K-1} \Phi(\sqrt{2x_q} - Z^T\Theta^l_{\cdot,k}) \Big\}, \tag{A.11}$$

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

Thus, the likelihood of this unordered multinomial probit model can be approxiamted using (A.11).

A.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 $\{\theta_k^{(g)}\}_{g=1}^G$ are from the Metropolis-Hastings sampling output, the joint posterior density can be written as

$$\pi(\theta_1^*, \ldots, \theta_{K-1}^* | Y, J_s) = \prod_{i=1}^{K-1} \pi(\theta_i | Y, J_s, \theta_1^*, \ldots, \theta_{i-1}^*). \tag{A.12}$$

By the local reversibility, each full conditional density can be written as

$$\begin{aligned} &\pi(\theta_i | Y, J_s, \theta_1^*, \ldots, \theta_{i-1}^*) \\ &= \frac{\mathrm{E}_1\{\rho(\theta_i, \theta_i^* | Y, J_s, \Psi_{i-1}^*, \Psi^{i+1}) q(\theta_i, \theta_i^* | Y, J_s, \Psi_{i-1}^*, \Psi^{i+1})\}}{\mathrm{E}_2\{\rho(\theta_i^*, \theta_i | Y, J_s, \Psi_{i-1}^*, \Psi^{i+1})\}}, \end{aligned} \tag{A.13}$$

where $\Psi_{i-1} = (\theta_1, \ldots, \theta_{i-1})$, $\Psi^{i+1} = (\theta_{i+1}, \ldots, \theta_{K-1})$, $\rho(\theta_i, \theta_i^* | Y, J_s, \Psi_{i-1}^*, \Psi^{i+1})$ is defined in (3.19), $q(\theta_i, \theta_i^* | Y, J_s, \Psi_{i-1}^*, \Psi^{i+1})$ is the proposal density, E_1 is the

expectaion with respect to the distribution $\pi(\theta_i, \Psi^{i+1}|Y, J_s, \Psi^*_{i-1})$, and E_2 is that with respect to $\pi(\Psi^{i+1}|Y, J_s, \Psi^*_{i-1}, \theta^*_i) \times q(\theta^*_i, \theta_i|Y, J_s, \Psi^*_{i-1}, \Psi^{i+1})$.

Then an estimate of $\pi(\theta_i|Y, J_s, \theta^*_1, \ldots, \theta^*_{i-1})$ is given by

$$\hat{\pi}(\theta_i|Y, J_s, \theta^*_1, \ldots, \theta^*_{i-1}) = \frac{G^{-1}\sum_{g=1}^{G} \rho(\theta_i^{(g)}, \theta^*_i|Y, J_s, \Psi^*_{i-1}, \Psi^{i+1,(g)}) q(\theta_i^{(g)}, \theta^*_i|Y, J_s, \Psi^*_{i-1}, \Psi^{i+1,(g)})}{M^{-1}\sum_{m=1}^{M} \rho(\theta^*_i, \theta_i^{(m)}|Y, J_s, \Psi^*_{i-1}, \Psi^{i+1,(m)})}, \tag{A.14}$$

where $\{\theta_i^{(g)}, \Psi^{i+1,(g)}\}_{g=1}^{G}$ are obtained from $\pi(\theta_i, \Psi^{i+1}|Y, J_s, \Psi^*_{i-1})$. $\{\Psi^{i+1,(m)}\}$ are obtained from $\pi(\Psi^{i+1}|Y, J_s, \Psi^*_{i-1}, \theta^*_i)$, and then for each $\{\Psi^{i+1,(m)}\}$, sample $\theta_i^{(m)}$ from $q(\theta^*_i, \theta_i|Y, J_s, \Psi^{i+1,(m)})$.

References

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88** 669-679. MR1224394

Albert, J. and Chib, S. (1997). Bayesian methods for cumulative, sequential, and two-step ordinal data regression models. Report, Department of Mathematics and Statistics, Bowling Green State University.

Antoniadis, A., Brossat, X., Cugliari, J. and Poggi, J. (2013). Clustering functional data using wavelets. *International Journal of Wavelets, Multiresolution and Information Processing* **11** 1350003-1350032. MR3038615

Biau, G., Bunea, F. and Wegkamp, M. H. (2005). Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory* **51** 2163-2172. MR2235289

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics* **34** 2159-2179. MR2291496

Chang, C., Chen, Y. and Ogden, R. T. (2014). Functional data classification: a wavelet approach. *Computational Statistics* **29** 1497-1513. MR3279004

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90** 1313-1321. MR1379473

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hasting output. *Journal of the American Statistical Association* **96** 270-281. MR1952737

Dunnett, C. W. (1989). Multivariate normal probability integrals with product correlation structure. *Journal of the Royal Statistical Society, Series C* **38** 564-579.

Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis* **44** 161-173. MR2020144

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2013). *Bayesian Data Analysis*. CRC Press, Boca Raton, FL. MR3235677

Ghosal, S. and van der Vaart, A. (2007a). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics* **35** 192-223. MR2332274

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge, UK. MR3587782

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711-732. MR1380810

James, G. M. (2002). Generalized linear models functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 411-432. MR1924298

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63** 533-550. MR1858401

Li, B. and Yu, Q. (2008). Classification of functional data: a segmentation approach. *Computational Statistics and Data Analysis* **52** 4790-4800. MR2521623

Mallor, F., Moler, J. A. and Urmeneta, H. (2018). Simulation of household electricity consumption by using functional data analysis. *Journal of Simulation* **12** 271-282.

McCulloch, R. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* **64** 207-240. MR1310524

Müller, H. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics* **33** 774-805.

Preda, C., Saporta, G. and Lévéder, C. (2007). PLS classification of functional data. *Computational Statistics* **22** 223-235. MR2318457

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer-Verlag, New York, NY. MR2168993

Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 305-332. MR2188987

Rossi, F. and Villa, N. (2006). Support vector machine for functional data classification. *Neurocomputing* **69** 730-742.

Shen, W. and Ghosal, S. (2015). Adaptive Bayesian procedures using random series priors. *Scandinavian Journal of Statistics* **42** 1194-1213. MR3426318

Stingo, F. C., Vannucci, M. and Downey, G. (2012). Bayesian wavelet-based curve classification via discriminant analysis with Markov random tree priors. *Statistica Sinica* **22** 465-488. MR2954348

Suarez, A. and Ghosal, S. (2016). Bayesian clustering of functional data using local features. *Bayesian Analysis* **11** 71-98. MR3447092

Ullah, S. and Finch, C. F. (2013). Applications of functioanl data analysis: a systematic review. *BMC Medical Research Methodology* **13** 43-54.

Wagner-Muns, I. M., Guardiola, I. G., Samaranayke, V. A. and Kayani, W. I. (2018). A functional data analysis approach to traffic volume forecasting. *IEEE Transactions on Intelligent Transportation Systems* **19** 878-888.

Wang, X., Ray, S. and Mallick, B. K. (2007). Bayesian curve classification using wavelets. *Journal of the American Statistical Association* **102** 962-973. MR2354408

Wang, J., Chiou, J., Zhu, J. and Müller, H. (2016). Functional data analysis. *Annual Review of Statistics and Its Application* **3** 257-295.

Yuan, M. and Cai, T. T. (2010). A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics* **38** 3412-3444. MR2766857

Zhu, H., Brown, P. J. and Morris, J. S. (2012). Robust classification of functional and quantitative image data using functional mixed models. *Biometrics* **68** 1260-1268. MR3040032

Zhu, H., Vannucci, M. and Cox, D. D. (2010). A Bayesian hierarchical model for classification with selection of functional predictors. *Biometrics* **66** 463-473. MR2758826