Frequentist Model Averaging For Functional Logistic Regression Model

By Jun Shi

Department of Statistics
Uppsala University

Supervisor: Shaobo Jin

2018
Abstract

Frequentist model averaging as a newly emerging approach provides us a way to overcome the uncertainty caused by traditional model selection in estimation. It acknowledges the contribution of multiple models, instead of making inference and prediction purely based on one single model. Functional logistic regression is also a burgeoning method in studying the relationship between functional covariates and a binary response. In this paper, the frequentist model averaging approach is applied to the functional logistic regression model. A simulation study is implemented to compare its performance with model selection. The analysis shows that when conditional probability is taken as the focus parameter, model averaging is superior to model selection based on BIC. When the focus parameter is the intercept and slopes, model selection performs better.

Keywords: Model selection, classification, mean squared error, Monte Carlo simulation, quadratic programming
Contents

1. Introduction .............................................................................................................................................. 1

2. Functional logistic regression model and frequentist model averaging ............. 3
   2.1. Functional logistic regression model................................................................................................. 3
   2.2. Frequentist model averaging............................................................................................................ 7
   2.3. Functional logistic regression model with frequentist model averaging ................................. 11

3. Simulation studies ..................................................................................................................................... 12
   3.1. Data generation.................................................................................................................................. 13
   3.2. Conditional probability as focus parameters...................................................................................... 15
   3.3. Intercept and coefficients as focus parameters.................................................................................. 16

4. Conclusions ................................................................................................................................................ 19

References ..................................................................................................................................................... 22
1. Introduction

Model selection based on finding an optimal model from a set of candidate models is commonly used in many statistical fields. Traditional statistical inference is purely based on the selected model. To find the optimal model from several candidate models, an impressive number of model selection criteria have been proposed, such as methods based on AIC or BIC along with forward and backward stepwise selection method. However, the information contained in other models is completely discarded after the optimal model being selected. Besides, this procedure will ignore some uncertainty originated from the selecting step, leading to report too short confidence intervals and high instability stemming from a small perturbation in computing the information criteria (Hjort & Claeskens, 2003; Leeb & Pötscher, 2008). For example, if the BIC of model 1 is only slightly lower than that of model 2, a small perturbation of the data may cause model 2 to have a lower BIC than model 1. In such a case, model 2 is discarded before perturbation, but model 1 is ignored after perturbation. In contrast, model averaging does not try to select a model, but incorporates all the candidate models by implementing weighted averaging over them. In Bayesian model averaging (BMA), priors are assigned to the candidate models, then the Bayesian procedure can provide the posterior distribution of the parameter of interest (Hoeting et al., 1999). Hjort and Claeskens (2003) proposed the frequentist model averaging (FMA) approach that can be employed in likelihood-based models. FMA has been applied to the binary logit model (Claeskens et al., 2006), the multinomial
and ordered logit model (Wan, Zhang & Wang, 2014) and the hazard regression model (Hjort & Claeskens, 2006). Attention is also paid to FMA for the linear regression model with ordinary least squares by, for example, Hansen (2007), Liu (2015) and Liu and Kuo (2016). However, they are different from the likelihood-based FMA.

Because of the development of new technologies in recent years, data sets that follow an individual over a long time period emerge. For example, a smart watch is able to automatically record the heart beats multiple times per day, which can be essentially viewed as a curve of heart beats. Functional data analysis (FDA) is commonly used to analyze data providing information about curves, surfaces or any objects that vary over a continuum. In most cases, each sample element can be considered as a function of time in FDA. For regression models in FDA, different from their traditional counterparts, both covariates and response can be functional or not functional. FDA can be viewed as response regressing on curves and surfaces. Reader is directed to Ramsay and Silverman (2005) for a detailed description of FDA.

In FDA, there is an extension of logistic regression model called functional logistic regression model that is used to analyze the relationship between a binary response and one or more functional covariates. Hence, functional logistic regression can be viewed as classification of curves. In contrast, traditional logistic regression is used to classify points into two classes. Some researchers have applied functional logistic regression to the foetal heart rate data (Ratcliffe et al., 2002), the white-matter tract profiles in multiple sclerosis (Goldsmith et al., 2011) and analysis for gene detection in pancreatic cancer (Wei et al., 2014).
Until now, FMA has not been applied to functional logistic regression, not even FDA. The purpose of this thesis is to apply the FMA of Hjort and Claeskens (2003) to the functional logistic regression model. We hope that FMA may provide better estimation and prediction than the traditional model selection in functional logistic regression by taking the discarded models into consideration.

The rest of this thesis is structured as follows. In section 2, functional logistic regression model and the machinery of the FMA are introduced, and FMA is applied to functional logistic regression. A simulation study designed to investigate the performance of the proposed method is included in section 3. The conclusions and discussions are provided in section 4.

2. Functional logistic regression model and frequentist model averaging

In this section the functional logistic regression model and the machinery of FMA established by Hjort and Claeskens (2003) are introduced. FMA is then applied to functional logistic regression.

2.1. Functional logistic regression model

The logistic regression model is usually used to analyze the relationship between the binary response $y$ and a vector of covariates $x$ with its form being expressed as

$$
\pi(x) = \frac{\exp\{\alpha + \sum_{j=1}^{m} \beta_j x_j\}}{1 + \exp\{\alpha + \sum_{j=1}^{m} \beta_j x_j\}}.
$$
or equivalently
\[
\log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \alpha + \sum_{j=1}^{m} \beta_j x_j,
\]
where \( \pi(x) = P(y = 1| x) \) is the conditional probability, \( \alpha \) and \( \beta_j \) are the intercept and slopes respectively, and \( m \) is the number of covariates. In contrast, the covariates \( x \) in functional logistic regression model are assumed as square-integrable functions of \( t \) with a support \( T \subset \mathbb{R} \). Thus, the conditional probability can be written into the following form
\[
\pi(x) = \frac{\exp\{\alpha + \int_T \beta_1(t)x_1(t) dt + \ldots + \int_T \beta_m(t)x_m(t) dt\}}{1 + \exp\{\alpha + \int_T \beta_1(t)x_1(t) dt + \ldots + \int_T \beta_m(t)x_m(t) dt\}},
\]
where \( \alpha \) is a t-invariant intercept parameter and \( \beta_j(t) \) is a square-integrable slope function. Hence, the functional logistic regression model is of the form
\[
\log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \alpha + \int_T \beta_1(t)x_1(t) dt + \ldots + \int_T \beta_m(t)x_m(t) dt. \tag{1}
\]

In practice, the continuous function \( x_j(t) \) can only be observed at discrete points. To smooth the observed discrete data back to the continuous function \( x_j(t) \), a set of functions known as the basis function system is employed to represent \( x_j(t) \). The advantage of this procedure is that it can maintain the continuously functional nature of \( x_j(t) \) (Ramsay & Silverman, 2005). The basis can be a B-spline basis, a Fourier basis, an exponential basis or a wavelet basis etc.. In this thesis a fourth order B-spline basis is used to represent the \( i \)th observation of \( x_j(t) \), \( x_{ij}(t) \). B-splines of order \( n \) are basis functions for spline functions that are polynomial functions of degree \( n - 1 \). The order four is a common choice in the B-spline literature. Reader is directed to de Boor (1978) for the shapes and properties of B-spline basis functions.
Thus, \( x_{ij}(t) \) can be expressed as

\[
x_{ij}(t) = \sum_{k=1}^{K_{x_j}} c_{ijk} \phi_{jk}(t), \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m,
\]

where \( \phi_{jk}(t) \) denotes the \( k \)th basis function of the \( j \)th covariate function at time \( t \), \( n \) is the sample size, and \( K_{x_j} \) is the number of basis functions of the \( j \)th covariate function. The coefficient \( c_{ijk} \) is estimated such that the estimated curve \( \hat{x}_{ji}(t) = \sum_{k=1}^{K_{x_j}} c_{ijk} \phi_{jk}(t) \) is sufficiently close to the observed discrete data by minimizing

\[
\sum_i (x_{ij}(t) - \sum_{k=1}^{K_{x_j}} c_{ijk} \phi_{jk}(t))^2 \quad \text{with respect to} \quad c_{ijk}.
\]

Similar to \( x_{ij}(t) \), the slope function \( \beta_j(t) \) is also represented by a fourth order B-spline basis

\[
\beta_j(t) = \sum_{l=1}^{K_{\beta_j}} b_{jl} \psi_{jl}(t),
\]

where \( \psi_{jl}(t) \) denotes the \( l \)th basis function of the \( j \)th coefficient function evaluated at time \( t \), \( b_{jl} \) is the coefficient of the expansion, and \( K_{\beta_j} \) is the number of basis functions of the \( j \)th coefficient function.

Let \( c_{ij}^T = (c_{i1j}, c_{i2j}, \ldots, c_{iK_{x_j}j}) \), \( b_j^T = (b_{j1}, b_{j2}, \ldots, b_{jK_{\beta_j}}) \), \( \phi_j^T(t) = (\phi_{j1}(t), \phi_{j2}(t), \ldots, \phi_{jK_{x_j}}(t)) \) and \( \psi_j(t)^T = (\psi_{j1}(t), \psi_{j2}(t), \ldots, \psi_{jK_{\beta_j}}(t)) \). Then \( x_{ij}(t) \) and \( \beta_j(t) \) can be expressed in matrix notation as

\[
x_{ij}(t) = c_{ij}^T \phi_j(t),
\]

\[
\beta_j(t) = \psi_j^T(t) b_j.
\]

Let \( c_i^T = (c_{i1}^T, c_{i2}^T, \ldots, c_{im}^T) \), \( b_j^T = (b_{j1}^T, b_{j2}^T, \ldots, b_{jm}^T) \) and \( J_{\phi\psi} = \int_\tau \phi_j(t) \psi_j^T(t) dt \), (1) for individual \( i \) can be rewritten as
\[
\log \left( \frac{\pi_i(x)}{1 - \pi_i(x)} \right) = \alpha + c_i^T J_{\phi\psi} b, \tag{2}
\]

where

\[
J_{\phi\psi} = \begin{pmatrix}
J_{\phi\psi 1} & 0 & \cdots & 0 \\
0 & J_{\phi\psi 2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & J_{\phi\psi m}
\end{pmatrix}_{K_x \times K_\beta},
\]

\[K_x = \sum_{j=1}^{m} K_{x_j} \quad \text{and} \quad K_\beta = \sum_{j=1}^{m} K_{\beta_j} \quad \text{.}
\]

Equation (2) also implies that estimation of the slope function \(\beta(t)\) has been simplified to the estimation of the scalar slope \(b\). The parameters in (2) include \(\alpha\), \(b\) and \(c_i\), where \(c_i\) is often replaced by its estimate \(\hat{c}_i\) obtained above. Hence, the unknown parameters after the basis expansion are \(\alpha\) and \(b\), which will be estimated conditional on \(\hat{c}_i\).

Under the independence assumption, the likelihood function is given by

\[
L(\alpha, b) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1 - y_i}
\]

\[
= \prod_{i=1}^{n} \frac{e^{y_i(\alpha + \int f_{\phi} x_{i1}(t) \beta_1(t) dt + \cdots + f_{\phi} x_{im}(t) \beta_m(t) dt)}}{1 + e^{\alpha + \int f_{\phi} x_{i1}(t) \beta_1(t) dt + \cdots + f_{\phi} x_{im}(t) \beta_m(t) dt}}
\]

\[
= \prod_{i=1}^{n} \frac{e^{y_i(\alpha + \hat{c}_i^T J_{\phi\psi} b)}}{1 + e^{\alpha + \hat{c}_i^T J_{\phi\psi} b}}.
\]

Here, we have replaced \(c_i\) by its estimate \(\hat{c}_i\). As the traditional logistic regression model, the numerical methods such as the Fisher scoring approach can be applied to obtain the maximum likelihood estimator. The reader can find the details of Fisher scoring from Ratcliffe, Heller and Leader (2002).

To deal with the data-smoothing problem in fitting the functional data \(x(t)\) with a basis system, usually the integrated squared second derivative of \(x(t)\) multiplied by a smoothing parameter is incorporated in the fitting process as a penalty. However,
this situation will not be considered in this thesis for simplicity. The reader is directed
to Ramsay and Silverman (2005) for a description of the implementation and conse-
quence of the smoothing parameter.

2.2. Frequentist model averaging

In this subsection the notations are taken from Hjort and Claeskens (2003) to be
consistent with the FMA literature. Consider a traditional logistic regression example
with a vector of four covariates \( x = (x_1, x_2, x_3, x_4)^T \). For the purpose of this thesis,
consider the following form

\[
\log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = x_1^T \theta + x_2^T \gamma,
\]

where \( x_1 = (1, x_1)^T \) is always included in the model and subsets of \( x_2 = (x_2, x_3, x_4)^T \)
may or may not be included. Hence, there are eight potential submodels as shown in
Table 1 together with the elements in \( \gamma \). For all submodels, \( \theta = (\alpha, \beta_1)^T \), since the
intercept and \( x_1 \) are included in all models.

\textit{Table 1 All submodels and their corresponding } \( \gamma \)

<table>
<thead>
<tr>
<th>Submodel</th>
<th>( \gamma )</th>
<th>Submodel</th>
<th>( \gamma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, ( x_1 )</td>
<td>0,0,0</td>
<td>1, ( x_1, x_2, x_3 )</td>
<td>( \beta_2, \beta_3, 0 )</td>
</tr>
<tr>
<td>1, ( x_1, x_2 )</td>
<td>( \beta_2, 0,0 )</td>
<td>1, ( x_1, x_2, x_4 )</td>
<td>( \beta_2, 0, \beta_4 )</td>
</tr>
<tr>
<td>1, ( x_1, x_3 )</td>
<td>0, ( \beta_3, 0 )</td>
<td>1, ( x_1, x_3, x_4 )</td>
<td>0, ( \beta_3, \beta_4 )</td>
</tr>
<tr>
<td>1, ( x_1, x_4 )</td>
<td>0,0, ( \beta_4 )</td>
<td>1, ( x_1, x_2, x_3, x_4 )</td>
<td>( \beta_2, \beta_3, \beta_4 )</td>
</tr>
</tbody>
</table>
Traditionally, a model selection method based on AIC or BIC can be implemented to find the optimal model. In contrast, Hjort and Claeskens (2003) proposed the FMA framework which incorporates all the submodels by implementing weighted averaging to them. The machinery of the framework that is suitable for various likelihood based methods is introduced here.

Suppose independent data $Y_1, Y_2, ..., Y_n$ come from the density

$$f_{\text{true}}(y) = f(y; \theta, \gamma) = f(y; \theta_0, \gamma_0 + \delta/\sqrt{n}),$$

where $\theta$ is a $p \times 1$ vector including the parameters in all potential submodels with $\theta_0$ denoting its true value, $\gamma$ is a $q \times 1$ vector around $\gamma_0$ with perturbation $\delta/\sqrt{n}$, and $\gamma_0$ is a known vector. The submodel of the type $f(y; \theta, \gamma_0)$ is called the narrow model, whereas the others of the type $f(y; \theta, \gamma)$ are called extended models. That is, $\gamma = \gamma_0$ in the narrow model will not be estimated. In Table 1, $\theta = (\alpha, \beta_1)^T$, $\gamma_0 = (0,0,0)^T$, and the first submodel without $x_2$ is the narrow model.

Let the models under consideration be indexed by $s$ and $\mu = \mu(\theta, \gamma) = \mu(\theta, \gamma_0 + \delta/\sqrt{n})$ be the focus parameter, which is a smooth function of $\theta$ and $\gamma$. The assumption that $\gamma = \gamma_0 + \delta / \sqrt{n}$ is called the local asymptotic framework that is widely used in various FMA studies. The focus parameter in Hjort and Claeskens (2003) is a scalar, which has been generalized to a vector-valued focus parameter by Jin and Ankargren (2018) to covariance analysis models that is slightly modified here for our model. Let $\pi_s$ be a $q_s \times q$ projection matrix mapping the vector $\gamma = (y_1, y_2, ..., y_q)^T$ to $y_s = \pi_s \gamma$, where $y_s$ is a subset of $\gamma$ that is involved in the $s$th candidate submodel. The estimator of the $s$th submodel is
\[ \beta_s = \mu(\hat{\theta}_s, \hat{\gamma}_s, \gamma_{0,s}^c), \quad s = 1, 2, 3, \ldots, 2^q, \]

where \( \hat{\theta}_s \) and \( \hat{\gamma}_s \) are the maximum likelihood estimators of the \( s \)th submodel, and \( \gamma_{0,s}^c \) corresponds to the \((q - q_s) \times 1\) vector of values of \( \gamma_0 \) not included in the \( s \)th submodel. Thus, the weighted averaged estimator of focus parameter is \( \hat{\mu}(\omega) = \sum_{s=1}^{2^q} \omega_s \beta_s \), where \( \omega_s \) is the corresponding weight of the \( s \)th submodel and \( \omega = (\omega_1, \ldots, \omega_{2^q})^T \). When \( \pi_s = I_q \), the submodel is called full model with its estimator being denoted as \( \hat{\mu}_{full} = \mu(\hat{\theta}_{full}, \hat{\gamma}_{full}) \). For instance in Table 1, the last submodel is the full model, and \( \gamma_{0,8}^c = \emptyset \). Let \( L(\theta, \gamma) \) be the likelihood function under the full model. Its information matrix is

\[
\begin{align*}
J_{n,full} & = -\frac{1}{n} \frac{\partial^2 \log L(\theta, \gamma)}{\partial (\theta^T, \gamma^T) \partial (\theta^T, \gamma^T)} \\
& = \begin{pmatrix}
\frac{\partial^2 \log L(\theta, \gamma)}{\partial \theta \partial \theta^T} & \frac{\partial^2 \log L(\theta, \gamma)}{\partial \theta \partial \gamma^T} \\
\frac{\partial^2 \log L(\theta, \gamma)}{\partial \gamma \partial \theta^T} & \frac{\partial^2 \log L(\theta, \gamma)}{\partial \gamma \partial \gamma^T}
\end{pmatrix} \\
& = \begin{pmatrix}
J_{n,00} & J_{n,01} \\
J_{n,10} & J_{n,11}
\end{pmatrix},
\end{align*}
\]

where \( J_{n,00} \) is a \( p \times p \) matrix, \( J_{n,01} \) is a \( p \times q \) matrix, \( J_{n,10} \) is a \( q \times p \) matrix, and \( J_{n,11} \) is a \( q \times q \) matrix. Let

\[
J_{full} = \begin{pmatrix}
J_{00} & J_{01} \\
J_{10} & J_{11}
\end{pmatrix}
\]

be the limiting information matrix. For the \( s \)th submodel, its information matrix is

\[
J_s = \begin{pmatrix}
J_{00} & J_{01,s} \\
J_{10,s} & J_{11,s}
\end{pmatrix} = \begin{pmatrix}
J_{00} & \pi_s J_{01} \\
\pi_s J_{10} & \pi_s J_{11,s}
\end{pmatrix}.
\]

Let \( K = (J_{11} - J_{10} J_{00}^{-1} J_{01})^{-1} \), \( K_s = (\pi_s K^{-1} \pi_s^T)^{-1} \), and \( W = \frac{\partial \mu}{\partial \theta^T} J_{00}^{-1} J_{01} - \frac{\partial \mu}{\partial \gamma^T} \). The par-
tial derivatives are evaluated at $\theta_0$ and $\gamma_0$. Let $\hat{\delta}_{full} = \sqrt{n}(\hat{\gamma}_{full} - \gamma_0)$, its asymptotic distribution can be derived from Lemma 3.2 in Hjort and Claeskens (2003) with a little algebra

$$\hat{\delta}_{full} \xrightarrow{d} D \sim N_q(\delta, K).$$

It implies that both $\theta$ and $\gamma$ can be consistently estimated. However, $\hat{\delta}$ can only be asymptotically unbiasedly estimated.

By Theorem 4.1 in Hjort and Claeskens (2003) the asymptotic distribution of the weighted averaged estimator is

$$\sqrt{n}(\hat{\mu}(\omega) - \mu_{true}) \xrightarrow{d} A = \left(\frac{\partial \mu}{\partial \theta^T} J^{-1}_{00} \right) M + W[\delta - \hat{\delta}(D)],$$

where $M \sim N_p(0, J_{00})$ is independent of $D$, $\hat{\delta}(D) = \sum_s \omega_s \pi_s^T K_s \pi_s^{-1} D$ and $\mu_{true}$ is the true value of the focus parameter. Thus the asymptotic mean squared error (MSE) of $\hat{\mu}(\omega)$ is

$$nE(\hat{\mu}(\omega) - \mu_{true})^T(\hat{\mu}(\omega) - \mu_{true}) = ntr \left[ \frac{\partial \mu}{\partial \theta^T} J^{-1}_{00} \left( \frac{\partial \mu}{\partial \theta^T} \right)^T + W \delta \delta^TW^T \right] - 2n \sum_{s=1}^{2^q} \sum_{\omega_s} \omega_s tr[W(\pi_s^T K_s \pi_s)K^{-1} \delta \delta^TW^T]$$

$$+ n \sum_{s=1}^{2^q} \sum_{t=1}^{2^q} \omega_s \omega_t tr[W(\pi_s^T K_s \pi_s)K^{-1}(K + \delta \delta^T)K^{-1}(\pi_t^T K_t \pi_t)W^T].$$

Hjort and Claeskens (2003) proposed that the weights $\{\omega_s\}$ can be obtained by minimizing an estimator of above asymptotic MSE, and it is equivalent to minimizing

$$Q(\{\omega_s\}) = - \sum_{s=1}^{2^q} \omega_s tr[W(\pi_s^T K_s \pi_s)K^{-1} \delta \delta^TW^T]$$

$$+ \frac{1}{2} \sum_{s=1}^{2^q} \sum_{t=1}^{2^q} \omega_s \omega_t tr[W(\pi_s^T K_s \pi_s)K^{-1}(K + \delta \delta^T)K^{-1}(\pi_t^T K_t \pi_t)W^T]$$

(3)
subject to the condition \( \{ \omega_s : \omega_s \geq 0 \text{ and } \sum_s \omega_s = 1 \} \). This is a standard quadratic programming problem and can be solved by using the R package \texttt{kernlab} (Karatzoglou et al. 2004).

2.3. Functional logistic regression model with frequentist model averaging

In this part, FMA is applied to the functional logistic regression model. All matrices needed to compute (2) are derived here. For the functional logistic regression model in (1), we let intercept to be always included in the model and the covariates may or may not be included. Hence, \( \theta = \alpha \) and \( y^T = b^T = (b_1^T, b_2^T, \ldots, b_m^T) \) of dimension \( K_\beta \times 1 \). Consequently, the information matrix is

\[
I_{n,full} = \frac{1}{n} \begin{pmatrix}
\sum_{i=1}^{n} \frac{e^{\alpha + c_i^T J \phi b}}{(1 + e^{\alpha + c_i^T J \phi b})^2} & \sum_{i=1}^{n} \frac{e^{\alpha + c_i^T J \phi b}}{(1 + e^{\alpha + c_i^T J \phi b})^2} c_i^T J \phi c_i \\
\sum_{i=1}^{n} \frac{e^{\alpha + c_i^T J \phi b}}{(1 + e^{\alpha + c_i^T J \phi b})^2} c_i^T J \phi c_i & \sum_{i=1}^{n} \frac{e^{\alpha + c_i^T J \phi b}}{(1 + e^{\alpha + c_i^T J \phi b})^2} c_i^T J \phi c_i
\end{pmatrix}
\]

where \( I_{n,full} \) is a \( (K_\beta + 1) \times (K_\beta + 1) \) matrix. Up to now, the estimation of functional logistic regression in R can only be solved by using the package \texttt{fda.usc} (Febreiro-Bande & Oviedo de la Fuente, 2012). In this package, the coefficient vector \( c_i \) is estimated when smoothing the covariate \( x_{ij}(t) \), which is independent from estimating \( \alpha \) and \( b \) as mentioned in section 2.1, so the information matrix does not include the second partial derivative of the likelihood function respect to \( c_i \) for simplicity.

Two different focus parameters are considered in this thesis. First, Wan, Zhang and Wang (2014) proposed to use the conditional probability as the focus parameter, if the main purpose is out-of-sample classification. Let \( \mu = \mu(\alpha, b) = \pi(x^{(oos)}) \),
where \( x^{(oos)} \) is out-of-sample observation and \( \pi(x^{(oos)}) \) is the out-of-sample forecasting conditional probability. In this setting we can show that the partial derivatives of \( \mu \) with respect to \( \alpha \) and \( b \) are

\[
\frac{\partial \mu}{\partial \theta} = \frac{\partial \mu}{\partial \alpha} = \frac{e^{\alpha + c^T J_{\phi \psi} b}}{(1 + e^{\alpha + c^T J_{\phi \psi} b})^T},
\]

\[
\frac{\partial \mu}{\partial y^T} = \frac{\partial \mu}{\partial b^T} = \frac{e^{\alpha + c^T J_{\phi \psi} b}}{(1 + e^{\alpha + c^T J_{\phi \psi} b})^T} c^T J_{\phi \psi}.
\]

Second, the out-of-sample forecasting may not be stable due to the possible extreme values contained in \( x^{(oos)} \). Thus Ankargren and Jin (2018) and Liu and Kuo (2016) proposed to use the intercept and slopes as the focus parameter. Let \( \mu^T = (\alpha, b^T) \), we can show that the partial derivatives of \( \mu \) with respect to \( \alpha \) and \( b \) can be expressed as

\[
\frac{\partial \mu}{\partial \theta} = \frac{\partial \mu}{\partial \alpha} = (1, 0, 0, \ldots, 0)^T,
\]

\[
\frac{\partial \mu}{\partial y^T} = \frac{\partial \mu}{\partial b^T} = \begin{pmatrix}
0 & 0 & 0 & \ldots & 0 \\
1 & 0 & 0 & \ldots & 0 \\
0 & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 1
\end{pmatrix},
\]

where \( \partial \mu / \partial \alpha \) is a \((K_\beta + 1) \times 1\) vector, and \( \partial \mu / \partial b^T \) is a \((K_\beta + 1) \times K_\beta\) matrix. Using the derived partial derivatives, we can compute \( K, K_s \) and \( W \) that are needed in (3). Then the model weights \( \{\omega_s\} \) can be easily obtained from the quadratic programming problem.

3. Simulation studies
In this section, two simulation studies are implemented to evaluate the performance of the proposed method for the focus parameters mentioned in section 2.3, respectively. A post-model selection method based on BIC and an equally weighted model averaging method are included to compare with the FMA method in each simulation.

3.1. Data generation

The full model includes three iid. explanatory functional covariates $x_{i1}(t), x_{i2}(t), x_{i3}(t)$. The response $y_i$ is generated from a logistic regression model with a fixed $b$. This data generating method is adapted from Mousavi and Sørensen (2017) and the work of Escabias et al. (2004). Let $T = [0,10]$ and $2^8 = 256$ equally spaced time points $\{t_z \in [0,10], z = 1, 2, \ldots, 256\}$, and assume all the covariates share the same basis function system. Thus the covariates are expressed as

$$x_{ij}(t_z) = \sum_{k=1}^{8} c_{ijk} \phi_{jk}(t_z), \ i = 1, 2, \ldots, n, \ j = 1, 2, 3, \ z = 1, 2, \ldots, 256,$$

(4)

where the basis functions $\phi_{jk}(t_z)$ are order four B-splines and $c_{ijk}$ are simulated from a $n \times 8$ matrix of iid. standard normal variables multiplied by an $8 \times 8$ matrix of iid. uniform variables distributed on $[0,1]$. The sample size $n$ is set to be 510 and 1010, respectively. $n - 10$ observations are used as training data and 10 observations as testing data. The response $y_i$ is generated from

$$\log \left( \frac{Pr(y_i = 1|x)}{1 - Pr(y_i = 1|x)} \right) = \alpha + \int_{0}^{10} x_{i1}(t) \beta_1(t) dt, \ i = 1, 2, \ldots, n,$$

(5)

where $\alpha = -0.05$. Hence, $x_{i2}(t)$ and $x_{i3}(t)$ have no effects to $y_i$ in the true data generating process. For $\beta_1(t)$, it is designed as the following
\[ \beta_1(t) = \sum_{i=1}^{5} b_{1i} \psi_{1i}(t) = \mathbf{\psi}_1^T \mathbf{b}_1, \]

where, \( \mathbf{b}_1^T = \kappa(sin\pi/30, sin103\pi/120, sin101\pi/60, sin301\pi/120, sin10\pi/3) \), the basis functions \( \psi_{1i}(t) \) are order four B-splines, and \( \kappa \in \{1/10\sqrt{n}, 2/10\sqrt{n}, \ldots, 9/10\sqrt{n}, 1/\sqrt{n}, 1\} \), which is used to control the magnitude of the coefficients. A random sample of \( x_1(t) \) and the shape of \( \beta_1(t) \) with \( \kappa = 1 \) are displayed in Figure 1 for illustration. When \( \kappa \neq 1 \), \( \mathbf{b}_1 \) follows the local asymptotic framework, whereas when \( \kappa = 1 \), \( \mathbf{b}_1 \) does not follow the local asymptotic framework. This allows us to investigate the consequence of violating the local asymptotic assumption. For each \( \kappa \) the simulation is based on 1000 replications.

**Figure 1**

Left: Sample of 6 random functional covariates \( x_1(t) \) generated from (4). Curves are coloured according to the outcome \( y \) (black for 0 and red for 1), generated from (5) with parameter function \( \beta_1(t) \). Right: The curve of \( \beta_1(t) \) when \( \kappa = 1 \).
3.2. Conditional probability as focus parameters

The MSE,

\[ \text{MSE}_{\text{fore}} = \frac{1}{1000} \sum_{r=1}^{1000} \sum_{i=1}^{10} (\hat{\pi}_i^r - \pi_i^r)^2, \]

is used to assess the accuracy of the out-of-sample forecasting probability, where \( \hat{\pi}_i^r \) is the forecast of the probability of the \( i \)th test observation \( \pi_i^r \) (the true conditional probability). The simulation results are shown in Table 2. For every \( \kappa \) the lowest MSE is shown in red. Table 2 illustrates that when \( \kappa \neq 1 \), model averaging strategies have better performance than model selection, if the conditional probability is taken as the focus parameter. In most cases, equally weighted model averaging approach produced the lowest MSE. But, equally weighted model averaging approach is better than FMA in this situation. For \( \kappa = 1 \), model selection produces the best estimation. This is because when \( \kappa = 1 \), the local asymptotic assumption is not satisfied.
Table 2 MSE of the forecasting probability when conditional probability is the focus parameter

<table>
<thead>
<tr>
<th>$\kappa$</th>
<th>$n = 510$</th>
<th>$n = 1010$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FMA</td>
<td>EW</td>
</tr>
<tr>
<td>$1/10\sqrt{n}$</td>
<td>0.0519</td>
<td>0.0078</td>
</tr>
<tr>
<td>$2/10\sqrt{n}$</td>
<td>0.0517</td>
<td>0.0075</td>
</tr>
<tr>
<td>$3/10\sqrt{n}$</td>
<td>0.0522</td>
<td>0.0074</td>
</tr>
<tr>
<td>$4/10\sqrt{n}$</td>
<td>0.0520</td>
<td>0.0074</td>
</tr>
<tr>
<td>$5/10\sqrt{n}$</td>
<td>0.0503</td>
<td>0.0074</td>
</tr>
<tr>
<td>$6/10\sqrt{n}$</td>
<td>0.0513</td>
<td>0.0074</td>
</tr>
<tr>
<td>$7/10\sqrt{n}$</td>
<td>0.0510</td>
<td>0.0074</td>
</tr>
<tr>
<td>$8/10\sqrt{n}$</td>
<td>0.0512</td>
<td>0.0072</td>
</tr>
<tr>
<td>$9/10\sqrt{n}$</td>
<td>0.0504</td>
<td>0.0072</td>
</tr>
<tr>
<td>$1/\sqrt{n}$</td>
<td>0.0505</td>
<td>0.0074</td>
</tr>
<tr>
<td>1</td>
<td>0.0369</td>
<td>0.0156</td>
</tr>
</tbody>
</table>

Note: FMA=frequentist model averaging, EW=equally weighted model averaging.

3.3. Intercept and slopes as focus parameters

The MSE,

$$MSE_{coef} = \frac{1}{1000} \sum_{r=1}^{1000} \left( \left( \hat{\alpha}^r, \hat{\beta}^r \right) - \left( \alpha, \beta \right) \right)^T \left( \left( \hat{\alpha}^r, \hat{\beta}^r \right) - \left( \alpha, \beta \right) \right),$$

is employed to evaluate the accuracy of the coefficients estimated from the models, where $\left( \hat{\alpha}^r, \hat{\beta}^r \right)^T$ is the estimated parameter of the $r$th replication. Different from the MSE in section 3.2, the computation of $MSE_{coef}$ only uses the $n - 10$ training data, so $MSE_{coef}$ is not out-of-sample. The out-of-sample misclassification rate (MCR) is evaluated as following.
\[ MCR = \frac{1}{1000} \sum_{r=1}^{1000} \sum_{i=1}^{10} (\hat{Y}_i^r - Y_i^r)^2, \]

where \( \hat{Y}_i^r \) is the forecast of the \( i \)th response variable \( Y_i^r \). \( \hat{Y}_i^r \) is assigned to 1, if \( \pi_i^r \geq 0.5 \) and 0 otherwise. The simulation results are shown in Tables 3-4. For every \( \kappa \) the method with lowest MSE or MCR is shown in red. Some conclusions can be drawn from these two tables.

First, in most cases, model selection strategy based on BIC has higher accuracy than the two model averaging strategies in terms of MSE and MCR. It is because there are only three explanatory variables in the full model, so model selection strategy usually has higher probability to choose the true model, which only has \( x_1(t) \) as its covariate. Second, the results show that the equally weighted model averaging strategy is superior to frequentist model averaging, especially when \( \kappa = 1 \), that is, the true coefficients do not follow the local misspecification framework.
Table 3 MSE of the estimator of the coefficients when intercept and slopes are the focus parameter

<table>
<thead>
<tr>
<th>$\kappa$</th>
<th>$n = 510$</th>
<th></th>
<th>$n = 1010$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FMA</td>
<td>EW</td>
<td>BIC</td>
<td>FMA</td>
</tr>
<tr>
<td>1/10$\sqrt{n}$</td>
<td>3.6484</td>
<td>1.9784</td>
<td>0.0083</td>
<td>1.2584</td>
</tr>
<tr>
<td>2/10$\sqrt{n}$</td>
<td>2.8399</td>
<td>1.7454</td>
<td>0.0081</td>
<td>1.2826</td>
</tr>
<tr>
<td>3/10$\sqrt{n}$</td>
<td>3.3071</td>
<td>1.8921</td>
<td>0.0080</td>
<td>1.3365</td>
</tr>
<tr>
<td>4/10$\sqrt{n}$</td>
<td>2.8942</td>
<td>1.7089</td>
<td>0.0091</td>
<td>1.2883</td>
</tr>
<tr>
<td>5/10$\sqrt{n}$</td>
<td>2.7261</td>
<td>1.6406</td>
<td>0.0094</td>
<td>1.2058</td>
</tr>
<tr>
<td>6/10$\sqrt{n}$</td>
<td>2.5406</td>
<td>1.5962</td>
<td>0.0095</td>
<td>1.2063</td>
</tr>
<tr>
<td>7/10$\sqrt{n}$</td>
<td>2.7578</td>
<td>1.6811</td>
<td>0.0111</td>
<td>1.5827</td>
</tr>
<tr>
<td>8/10$\sqrt{n}$</td>
<td>2.9176</td>
<td>1.7276</td>
<td>0.0108</td>
<td>1.2748</td>
</tr>
<tr>
<td>9/10$\sqrt{n}$</td>
<td>2.7414</td>
<td>1.6939</td>
<td>0.0125</td>
<td>1.2395</td>
</tr>
<tr>
<td>1/$\sqrt{n}$</td>
<td>2.8741</td>
<td>1.7077</td>
<td>0.0138</td>
<td>1.2057</td>
</tr>
<tr>
<td>1</td>
<td>4.4591</td>
<td><strong>2.7375</strong></td>
<td><strong>3.0109</strong></td>
<td>1.8289</td>
</tr>
</tbody>
</table>

Note: FMA=frequentist model averaging, EW=equally weighted model averaging.
### Table 4 Misclassification rates when intercept and slopes are the focus parameter

<table>
<thead>
<tr>
<th>$\kappa$</th>
<th>$n = 510$</th>
<th></th>
<th></th>
<th>$n = 1010$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FMA</td>
<td>EW</td>
<td>BIC</td>
<td>FMA</td>
<td>EW</td>
<td>BIC</td>
</tr>
<tr>
<td>$1/10\sqrt{n}$</td>
<td>0.4950</td>
<td>0.4982</td>
<td>0.4926</td>
<td>0.4938</td>
<td>0.4925</td>
<td>0.4841</td>
</tr>
<tr>
<td>$2/10\sqrt{n}$</td>
<td>0.4925</td>
<td>0.4920</td>
<td>0.4830</td>
<td>0.5001</td>
<td>0.5069</td>
<td>0.4854</td>
</tr>
<tr>
<td>$3/10\sqrt{n}$</td>
<td>0.5028</td>
<td>0.5071</td>
<td>0.4891</td>
<td>0.4982</td>
<td>0.4999</td>
<td>0.4903</td>
</tr>
<tr>
<td>$4/10\sqrt{n}$</td>
<td>0.4973</td>
<td>0.5010</td>
<td>0.4997</td>
<td>0.5041</td>
<td>0.4985</td>
<td>0.4814</td>
</tr>
<tr>
<td>$5/10\sqrt{n}$</td>
<td>0.4926</td>
<td>0.4931</td>
<td>0.4865</td>
<td>0.4960</td>
<td>0.4954</td>
<td>0.4813</td>
</tr>
<tr>
<td>$6/10\sqrt{n}$</td>
<td>0.5015</td>
<td>0.5036</td>
<td>0.4784</td>
<td>0.4982</td>
<td>0.5009</td>
<td>0.4813</td>
</tr>
<tr>
<td>$7/10\sqrt{n}$</td>
<td>0.5063</td>
<td>0.5017</td>
<td>0.4828</td>
<td>0.5017</td>
<td>0.5071</td>
<td>0.4798</td>
</tr>
<tr>
<td>$8/10\sqrt{n}$</td>
<td>0.4975</td>
<td>0.4932</td>
<td>0.4784</td>
<td>0.5023</td>
<td>0.4965</td>
<td>0.4901</td>
</tr>
<tr>
<td>$9/10\sqrt{n}$</td>
<td>0.4971</td>
<td>0.5029</td>
<td>0.4726</td>
<td>0.5017</td>
<td>0.5050</td>
<td>0.4763</td>
</tr>
<tr>
<td>$1/\sqrt{n}$</td>
<td>0.4769</td>
<td>0.4755</td>
<td>0.4851</td>
<td>0.4937</td>
<td>0.4909</td>
<td>0.4746</td>
</tr>
<tr>
<td>1</td>
<td>0.3508</td>
<td>0.3599</td>
<td>0.3471</td>
<td>0.3507</td>
<td>0.3505</td>
<td>0.3504</td>
</tr>
</tbody>
</table>

Note: FMA=frequentist model averaging, EW=equally weighted model averaging.

### 4. Conclusions

In this thesis the FMA machinery is applied to functional logistic regression model. Our work generalizes the work of Wan, Zhang and Wang (2014) from multinomial and ordered data to functional data. The simulation study reveals that for functional logistic regression model FMA is inferior to both equally weighted model averaging method and model selection based on BIC, no matter conditional probability or the coefficient is treated as the focus parameter. Hence, FMA with functional logistic regression should be used in caution.

When the focus parameter comes to the intercept and slopes, model selection
produces a higher accuracy, because it tends to find the true model under the condition that only three potential covariates are included. Our results also show that FMA produces a much higher MSE than model selection. Results not shown here reveal that the curve $\beta_1(t)$ is still well fitted, even though the coefficient estimators have a large MSE. Further, Table 4 indicates that the misclassification rates of FMA are not substantially worse than model selection either. On the other hand, if we only consider the conditional probability, two model averaging methods have obvious advantage over the model selection method, especially the equally weighted model averaging. This result is close to Wan, Zhang and Wang (2014). Their simulation results also show the superiority of equally weighted model averaging. Only when $\kappa = 1$, the performance of model selection is better than the others. This indicates that violating the local asymptotic assumption will negatively influence the performance of FMA.

This thesis demonstrates that for functional logistic regression model, frequentist model averaging method does not perform very well. The simulation study could be improved by increasing the number of potential covariates, since the uncertainty of model selection may increase. Adding roughness penalty in functional data analysis can contribute to increase the goodness of fit of the estimated curve without making the curve to be excessively twisty. Including penalization in future studies can provide a more general form of functional logistic regression with the FMA. In the R package fda.usc, the coefficients $c_{ij}$ are estimated independently from $\alpha$ and $b$. This leads to the uncertainty produced by estimating $c_{ij}$ not included in the framework, because the information matrix mentioned in section 2.3 does not
consider $c_{ij}$. Thus the MSEs computed are of the empirical Bayes type and cannot reflect the whole nature of the estimated focus parameter. Including the uncertainty in estimating $c_{ij}$ will be considered in a future study.
References


