Biometrics 72, 1325–1335
December 2016
DOI: 10.1111/biom.12504

Label-Noise Resistant Logistic Regression for Functional Data Classification with an Application to Alzheimer's Disease Study

Seokho Lee,[1,*] Hyejin Shin,[2] and Sang Han Lee[3]

[1]Department of Statistics, Hankuk University of Foreign Studies, Yongin, Gyeonggi, Korea
[2]Frontier CS Lab, Software R&D Center, Samsung Electronics, Seoul, Korea
[3]The Nathan S. Kline Institute for Psychiatric Research, New York, U.S.A.
email: lees@hufs.ac.kr

Summary. Alzheimer's disease (AD) is usually diagnosed by clinicians through cognitive and functional performance test with a potential risk of misdiagnosis. Since the progression of AD is known to cause structural changes in the corpus callosum (CC), the CC thickness can be used as a functional covariate in AD classification problem for a diagnosis. However, misclassified class labels negatively impact the classification performance. Motivated by AD–CC association studies, we propose a logistic regression for functional data classification that is robust to misdiagnosis or label noise. Specifically, our logistic regression model is constructed by adopting individual intercepts to functional logistic regression model. This approach enables to indicate which observations are possibly mislabeled and also lead to a robust and efficient classifier. An effective algorithm using MM algorithm provides simple closed-form update formulas. We test our method using synthetic datasets to demonstrate its superiority over an existing method, and apply it to differentiating patients with AD from healthy normals based on CC from MRI.

Key words: Alzheimer's disease; Fisher consistency; Functional data classification; Label noise; MM algorithm; Outlier detection; Robust classification.

1. Introduction

In recent years, there has been a growing interest in studying variation in brain physiology between normal adults and subjects with progression on Alzheimer's disease. Alzheimer's disease (AD) is an age-related, irreversible brain disorder with a relatively long development cycle that occurs gradually and results in episodic memory loss, behavior and personality changes, and a decline in thinking abilities. These losses are related to pathological amyloid deposition and hyperphosphorylation of structural proteins in brain which lead to progressive loss of connectivity between regions and aberrant changes in brain's structural topology. AD advances progressively from mild forgetfulness, also referred as predementia or mild cognitive impairment (MCI), to total dementia and severe loss of mental function.

Scientists are now working in several areas that may improve the ability of clinicians to make accurate diagnoses of AD even earlier. It is likely that additional studies in structural imaging will help to unravel the long-term consequences of AD, and aid in further diagnosis, treatment, and rehabilitation. Magnetic resonance imaging (MRI) is a relatively widespread method that allows visualization of brain biochemistry, metabolism, and anatomy. Structural MRI imaging analyses have successfully found various abnormalities in brain regions for AD. Albeit most of the studies have been focused on memory-related brain regions in gray matter such as medial temporal lobe and hippocampus, structural changes of the corpus callosum (CC) in AD have been also reported in a number of studies (Di Paola et al., 2010; Frederiksen et al., 2011). The CC is the largest white matter structure of bundle of neural fibers, which connects the two cerebral hemispheres. It is relatively easy to measure and there is a fully automated procedure (Ardekani et al., 2012).

In this article, we consider callosal thickness for AD classification. Typically, the midsagittal callosal area is considered for analysis. For our analysis, regional callosal thickness is measured at 99 interior points which make 100 equal intervals along the medial axis of the CC (See Figure 2). Subsequently, we have 99 positive values for each CC. Most of the studies are comparing patients to healthy controls to find the abnormal structure in the CC by statistical testing procedures. Various testing procedures are helpful to find the differences between two groups in some regions of the CC, but are less helpful to diagnosis each subject. For the purpose of diagnosis, classification analysis with CC thickness would be more helpful. Since CC images are typically in smooth shapes, AD classification with CC thickness profiles is a function data classification problem. This motivates us to develop a logistic regression model with a functional covariate for our analysis.

A potential issue in this data is that AD diagnosis based on clinicians' decision may not be accurate, especially when patients show mild symptoms in their behavior. The quality of a classifier depends on the accurate labeling of the training data. In classification, mislabeling in class variable is called the label noise. A comprehensive introduction to label noise under general applications can be found in Frenáy and Kabán (2014). Some mislabeled observations appear as outliers in classification. A simple way to address this problem is to filter out potential outliers before learning stage (Brodley and Friedl, 1999; Muhlenbach, Lallich, and Zighed, 2004), or to

put small weights on them in analysis (Carroll and Pederson, 1993). To identify outliers in high-dimensional data, there are numerous outlier detection techniques in traditional statistics (Aggarwal and Yu, 2001). However, this approach has a weakness that a portion of valuable sample is removed from the analysis. Moreover, distance-based outlier identification is not effective in high-dimensional case (Malossini, Blanzieri, and Ng, 2006). Another robust approach is to use robust loss minimization in the classifier learning process. Several studies show usefulness of nonconvex loss functions for robustness (Copas, 1988; Bianco and Yohai, 1996). Wu and Liu (2007) and Park and Liu (2011) use truncated loss functions for SVM and logistic regression, respectively. However, the resulting optimizations become nonconvex, so their solutions are not guaranteed to be the global minimizer. All robust approaches focus on reducing the effects from severe outliers, while most label noises frequently appear near the decision boundary due to ambiguous decision in diagnosis. A direct approach for label noise is to consider a label-flipping mechanism in developing a classifier (Bootkrajang and Kabán, 2012). This approach models the overall label-flipping probability, but it does not consider individual-specific characteristics in label noise.

She and Owen (2011) proposed an individual intercept model to identify outliers in regression problem. In their approach, the individual intercept estimates having nonzero values indicate that the corresponding observations are potential outliers. They showed that the resulting estimation becomes robust to outliers when the individual intercepts are included in the model. Tibshirani and Manning (2014) extended individual intercept model to logistic regression for incorporating mislabeling. Lee, MacEachern, and Jung (2012) also proposed the same model for regression and margin-based classification. They demonstrated that inclusion of individual intercepts in the procedure increases not only robustness but also efficiency of the procedure. In this article, we extend their idea to the logistic regression for functional data classification. We add individual intercepts to a discriminant function in the logistic regression with a functional covariate. Consequently, the discriminant function estimate becomes robust to label noise. Also, we consider regularization for intercepts, which is expected to enhance the efficiency of classification procedure under training data having label noise.

This article is organized as follows. In Section 2, we explain how mislabeled observations appear as outliers in the classification problem, and the individual-intercept model leads to a robust estimate of the discriminant function. Furthermore, logistic regression is modified to include a functional covariate. An efficient estimating formula using Majorization–Minimization algorithm is given in Section 3. Numerical study on the performance of the proposed method is provided in Section 4. Section 5 presents the application to AD classification based on the CC thickness profiles. Finally, this article concludes with some remarks in Section 6.

2. Robust Logistic Regression Model for Functional Data

2.1. *Logistic Regression and Outliers*

Logistic regression (LR) for classification builds a classifier from training data $\{(x_n, y_n) : n = 1, 2, \ldots, N\}$, where $x_n \in \mathbb{R}^p$ is a vector of p predictors and $y_n \in \{-1, +1\}$ is a class label. A linear discriminant function $f(x)$ is learned by finding α and β which minimize the negative Bernoulli log-likelihood

$$\min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^p} \sum_{n=1}^{N} \ell(y_n f(x_n)), \tag{1}$$

with the loss function $\ell(u) = -\log\sigma(u)$ and the discriminant function $f(x) = \alpha + x^T\beta$, where $x_n^T = (x_{n1}, \ldots, x_{np})^T$, $\beta = (\beta_1, \ldots, \beta_p)^T$, and $\sigma(u) = \{1 + \exp(-u)\}^{-1}$. Here, $\sigma(f(x)) = p(y = +1|x)$ is the probability that x belongs to the class $+1$. Once we obtain an estimate of the discriminant function, $\hat{f}(x) = \hat{\alpha} + x^T\hat{\beta}$, from the training data, $\text{sign}(\hat{f}(x^*))$ can be used for classifying a new observation x^*, i.e., $\hat{y} = +1$ if $\hat{f}(x^*) \geq 0$, and $\hat{y} = -1$ if $\hat{f}(x^*) < 0$. The hyperplane $f(x) = 0$ is called the decision boundary. Thus, LR for classification finds the decision boundary minimizing the loss function (1) whose argument, $yf(x)$, is called the margin (Lin, 2004).

When outlying observations are included in the training data, the resulting classifier is affected by them. In generative models for classification, the class-specific conditional distribution, $p(x|y)$, is learned from the training data. Subsequently, $p(y|x)$ is inferred from $p(x|y)$ and used for classification. Typical examples of generative model are linear/quadratic discriminant analysis, where $p(x|y)$ is modeled under Gaussian assumption. Since Gaussian distribution is specified by the first two moments which are highly sensitive to outliers, classification performance from generative models easily deteriorates due to some observations lying far from their group. Unlike generative models, LR directly constructs the conditional distribution $p(y|x)$ from the training data by minimizing the margin-based loss function in (1). This feature distinguishes LR from generative models in the way of outliers' influence.

Because negative Bernoulli log-likelihood function is decreasing in the margin $yf(x)$, a large positive value of $yf(x)$ takes a small value of the loss. Thus, outlying observation x on the correct side of the decision boundary (i.e., $yf(x) > 0$) does not seriously affect the classification procedure, even though x is located far from the majority of data with the same class. This means that LR has low sensitivity to observations far from the decision boundary, provided that those observations are on the correct side of the decision boundary. On the other hand, a relatively large value of the loss is taken with x having $yf(x) < 0$, which is the case that x is located on the wrong side of the decision boundary. Since class prediction is made by the sign of $f(x)$, the negative value of $yf(x)$ means misclassification. Thus, the loss can be large when x is misclassified, even if x is not far from the class where it belongs. This implies that influential outliers in the logistic regression always appear in the misclassification and, in other words, label noise can be properly treated under robust classification framework.

While incorrect class labeling in training data is suspected to be detrimental to classification performance, symmetric label noise is known to be less harmful (Frenáy and Kabán, 2014). Symmetric label noise means that a proportion of label noise in one class is the same as that in the other class. On the other hand, asymmetric label noise is more problematic in classification. The following propositions suggest this in terms

of Fisher consistency. Fisher consistency in classification task requires that the population minimizer of the loss function leads to the Bayes optimal rule of classification (Lin, 2004). Let $\tilde{Y}$ be the unobserved true class label and Y be the observed label-noised class label. Proofs of propositions are provided in web-based Supplementary Material available online.

PROPOSITION 1. *Assume that label noise is symmetric with label-switching probability* $0 \le \pi(x) < 1/2$ *for any* x. *Then, the logistic loss in* (1) *satisfies Fisher consistency. In other words, the minimizer* $f^*(x)$ *of* $E[\ell(Yf(X))|X = x]$ *has the same sign as* $\Pr(\tilde{Y} = +1|X = x) - 1/2$ *at any* x.

PROPOSITION 2. *Assume that label noise is asymmetric with label-switching probabilities* $\pi_0(x) = \Pr(Y = +1|\tilde{Y} = -1, X = x)$ *and* $\pi_1(x) = \Pr(Y = -1|\tilde{Y} = +1, X = x)$, *where* $0 \le \pi_0(x) < 1/2$, $0 \le \pi_1(x) < 1/2$, *and* $\pi_0(x) \ne \pi_1(x)$ *for some* x. *Let* $p(x) = \Pr(\tilde{Y} = +1|X = x)$. *The logistic loss in* (1) *satisfies Fisher consistency if and only if* $(1 - 2\pi_0(x))(1 - p(x)) < (1 - 2\pi_1(x))p(x)$ *for* $p(x) > 1/2$, *and* $(1 - 2\pi_0(x))(1 - p(x)) > (1 - 2\pi_1(x))p(x)$ *for* $p(x) < 1/2$. *A simple sufficient condition for Fisher consistency is* $\pi_0(x) > \pi_1(x)$ *for* $p(x) > 1/2$, *and* $\pi_0(x) < \pi_1(x)$ *for* $p(x) < 1/2$.

Proposition 1 shows that LR ensures Fisher consistency in symmetric label noise. However, in asymmetric label noise case, Fisher consistency is not guaranteed unless the condition on label noise rates in Proposition 2 is satisfied. Bootkrajang and Kabán (2012) consider label-flipping model in LR to deal with asymmetric label noise, but they treat the switching probabilities as constant parameters over all the observations while these switching probabilities may depend on the individual's covariates. This motivates us to consider an alternative method, where observations are treated individually for robustness and efficiency under the general label noise.

2.2. *Logistic Regression Model with Individual Intercepts*

To diminish the influence of outliers and enhance the efficiency of estimation with the presence of label noise, we use a logistic regression model with individual intercepts (Lee, MacEachern, and Jung, 2012; Tibshirani and Manning, 2014). Since the loss function (1) is a function of the margin $yf(x)$ and the effect from outliers and label-noised observations appears only through the margin, individual intercepts are included in the discriminant function to explain the individual abnormality. The resulting criterion is

$$\min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^p,\gamma\in\mathbb{R}^N} \sum_{n=1}^{N} \ell(y_n(f(x_n) + \gamma_n)) + P_\lambda(\gamma), \tag{2}$$

where $\gamma = (\gamma_1, \ldots, \gamma_N)^T$ is a vector of individual intercepts. The penalty function on γ, $P_\lambda(\gamma) = \sum_{n=1}^{N} p_\lambda(|\gamma_n|)$, regulates the size of γ with a tuning parameter λ. Since the number of intercepts equals the sample size, the regularization on γ is indispensable. We use L_1 penalty in our implementation, as in Tibshirani and Manning (2014), although other types of sparsity-inducing penalty can be considered. Here, we assume that $p_\lambda(u)$ is symmetric about 0, increasing in the positive domain, and not differentiable at 0. These kinds of penalties promote γ_n of small value to be exactly zero. The tuning parameter λ controls the size of γ: if $\lambda = 0$ then γ_n is estimated without constraint, and if $\lambda = \infty$ then all γ_n's will be zero, resulting in the traditional logistic regression. In Section 3, we show that the individual intercepts are efficiently estimated by a thresholding rule associated with the penalty function used. Once we learn $\hat{\alpha}$ and $\hat{\beta}$ from training data, a new observation x^* will be classified according to the sign of $f(x^*) = \hat{\alpha} + x^{*T}\hat{\beta}$.

The individual intercept model was suggested by She and Owen (2011) for outlier detection in regression. She and Owen (2011) and Lee, MacEachern, and Jung (2012) show that penalized least squares criterion with L_1 penalty on the individual intercepts is equivalent to Huber's criterion, leading to a robust regression coefficient estimate. The same idea was carried over to LR as in (2) (Tibshirani and Manning, 2014). We present the following propositions to understand how the margin-based loss minimization with the individual intercept model leads to robust estimation in classification problem.

PROPOSITION 3. *The estimate of an individual intercept,* $\hat{\gamma}_n$, *in* (2) *is zero or has the same sign as* y_n. *That is,*

$$y_n\hat{\gamma}_n \ge 0$$

for $n = 1, 2, \ldots, N$. *Also, if there is no penalty on* γ *in* (2), *then* $\hat{\gamma}_n = +\infty$ *or* $-\infty$, *according to the sign of* y_n.

PROPOSITION 4. *Let* $g_i(r) = \ell(x_i + r) + p_\lambda(r)$ *with a nonnegative* r *and* $\hat{r}_i = \arg\min_r g_i(r)$ *for* $i = 1, 2$. *Then,* $\hat{r}_1 \ge \hat{r}_2$ *if* $x_1 \le x_2$.

For an outlier, $yf(x)$ tends to have a large negative value when the discriminant function $f(\cdot)$ is Bayes optimal. Since $\ell(yf(x))$ contributes a large value to the objective function, minimization process tries to find an $\hat{f}(\cdot)$ that gives a value of $\ell(y\hat{f}(x))$ as small as possible. Thus, a discriminant function estimate $\hat{f}(\cdot)$ from the training data having outlying observations is likely to deviate from the true discriminant function $f(\cdot)$. In (2), the margin is split into two parts as $yf(x) + y\gamma$. For an outlier, since $y\hat{\gamma}$ is allowed to have a positive value as in Proposition 3, $y\hat{f}(x)$ can be relaxed to have a large negative value. Thus, minimization process allows $\ell(y\hat{f}(x) + y\hat{\gamma})$ to be small, while $y\hat{f}(x)$ can be a large negative. This mechanism increases the possibility that $\hat{f}(\cdot)$ is close to $f(\cdot)$ under the presence of outliers. Proposition 4 states that $|\hat{\gamma}|$ is decreasing in $yf(x)$, implying that individual intercept estimate can be used as a measure of outlyingness of an observation.

Robust discriminant function estimation in (2) can be understood under the weighted logistic regression framework. Note that the loss in (2) can be viewed as

$$\min_{\alpha,\beta} \sum_{n=1}^{N} w_n \ell(y_n f(x_n)),$$

with the weight $w_n = w(x_n) = \ell(y_n f(x_n) + y_n\gamma_n)/\ell(y_n f(x_n))$. If $\gamma_n = 0$ then $w_n = 1$, which implies that the nth observation fully exerts its influence on $f(\cdot)$ estimation. Otherwise, if $\gamma_n \ne 0$, then $0 < w_n < 1$ and its influence is reduced according to $|\gamma_n|$. Unlike our approach, typical weighted logistic regressions

use the distance-based weight, which is undesirable in high-dimensional case (Carroll and Pederson, 1993; Muhlenbach, Lallich, and Zighed, 2004).

Note that introducing individual intercept shifts the loss $\ell(y_n f(x_n))$ to the left by $y_n\gamma_n$ since $y_n\gamma_n$ is nonnegative. Because $\ell(\cdot)$ is strictly decreasing, this left-translation lowers the loss function when $\gamma_n \neq 0$. Amount of loss reduction varies across the observations and depends on the size of γ_n. Therefore, outliers take the lowered loss function values so that their harmful effects on estimation are relieved. Truncated loss approaches (Wu and Liu, 2007; Park and Liu, 2011) take a different strategy for loss reduction by adopting a new loss $l(y_n f(x_n)) = \min\{\ell(y_n f(x_n)), \ell(s)\}$ with a prespecified negative value of s. Therefore, observations far from the decision boundary on the wrong side take the lowered value of loss, $\ell(s)$, uniformly regardless of their positions. This approach is less attractive because most label noises commonly appear near the decision boundary due to uncertain decision. Unlike this approach, the shift-parameter γ_n in (2) is allowed to take a value adaptively according to the location of an individual observation, and take a nonzero value although the observation is near the boundary or even located on the correct side. Individual intercept plays a similar role of slack variable in SVM (Lee, MacEachern, and Jung, 2012). In SVM, slack variables allow individual observations to be on the wrong side of the margin or the decision boundary. Likewise, individual intercepts also allow some observations to be on the wrong side of the decision boundary, which reduces sensitivity to the perturbation in training data due to label noise. From our experience, LR with individual intercept model often improves the classification performance even in the case of no label noise.

2.3. *Robust Logistic Regression for Functional Data*

We now extend the penalized logistic regression model with individual intercepts in (2) to the classification problem of functional data. When a functional predictor $x(\cdot)$ is considered, the linear discriminant function should be changed accordingly. LR with a functional covariate is to find the linear discriminant function $f(u, x) = \alpha_0 + u^T\alpha + \int_{\mathcal{T}} x(t)\beta(t)dt$ with a vector-valued covariate $u = (u_1, \ldots, u_p)^T$ and its coefficient $\alpha = (\alpha_1, \ldots, \alpha_p)^T$. A functional covariate $x(\cdot)$ is assumed to be square integrable on the compact interval $\mathcal{T}$ and the coefficient function $\beta(\cdot)$ is assumed to be in an RKHS $\mathcal{H}$, which is a subspace of the Hilbert space of square integrable functions on $\mathcal{T}$. Then, the penalized negative Bernoulli log-likelihood minimization (2) is modified to

$$\min_{\alpha_0\in\mathbb{R},\alpha\in\mathbb{R}^p,\beta\in\mathcal{H},\gamma\in\mathbb{R}^N} \sum_{n=1}^{N} \ell(y_n(f(u_n, x_n) + \gamma_n)) + P_{\lambda_1}(\beta) + P_{\lambda_2}(\gamma), \tag{3}$$

under the training set $\{(u_n, x_n(\cdot), y_n) : n = 1, 2, \ldots, N\}$. The penalty functional on β is the form of $P_{\lambda_1}(\beta) = \lambda_1 J(\beta)$, where $J(\beta)$ imposes a smoothness-inducing penalty on β. Suppose that $\mathcal{H}$ has a decomposition $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ is a finite dimensional linear subspace of $\mathcal{H}$ with $\dim(\mathcal{H}_0) = L \leq N$, and $J(\beta) = \|P_1\beta\|^2_{\mathcal{H}}$, where P_1 is the orthogonal projection of β in $\mathcal{H}$ onto a subspace $\mathcal{H}_1$.

From the representer theorem (Wahba, 1990), the solution of (3) becomes

$$\beta_{\lambda_1}(t) = \sum_{l=1}^{L} d_l\phi_l(t) + \sum_{n=1}^{N} c_n\xi_n(t),$$

where $\{\phi_l : l = 1, \ldots, L\}$ is orthonormal basis of $\mathcal{H}_0$, K_1 is the reproducing kernel of $\mathcal{H}_1$, and $\xi_n(t) = \int_{\mathcal{T}} x_n(s)K_1(s, t)ds$. Thus, the discriminant function takes the form of $f(u_n, x_n) = \alpha_0 + u_n^T\alpha + \sum_{l=1}^{L} d_l \int_{\mathcal{T}} x_n(t)\phi_l(t)dt + \sum_{m=1}^{N} c_m\langle\xi_n, \xi_m\rangle_{\mathcal{H}}$. Letting $d = (\alpha_0, \alpha_1, \ldots, \alpha_p, d_1, \ldots, d_L)^T$, $z_n^T = (1, u_{n1}, \ldots, u_{np}, z_{n1}, \ldots, z_{nL})^T$ with $z_{nl} = \int_{\mathcal{T}} x_n(t)\phi_l(t)dt$, and $\Sigma = (\Sigma_{mn}) = (\Sigma_1, \cdots, \Sigma_N)^T$ with $\Sigma_{mn} = \langle\xi_m, \xi_n\rangle_{\mathcal{H}} = \int_{\mathcal{T}}\int_{\mathcal{T}} x_m(s)x_n(t)K_1(s, t)dsdt$, the criterion (3) is written in the vector form of

$$\min_{d\in\mathbb{R}^{1+p+L},c\in\mathbb{R}^N,\gamma\in\mathbb{R}^N} \sum_{n=1}^{N} \ell(y_n(z_n^T d + \Sigma_n^T c + \gamma_n)) + \lambda_1 c^T\Sigma c + P_{\lambda_2}(\gamma). \tag{4}$$

Note that, instead of an RKHS estimate of $\beta(\cdot)$, one can also develop the functional model using other fixed basis set approach.

For implementation, we take $\mathcal{H} = \mathcal{W}_2^m[0, 1] = \{\beta : \beta^{(j)}$ is absolutely continuous, $j = 0, \ldots, m-1$, and $\int_0^1 [\beta^{(m)}(t)]^2 dt < \infty\}$ with the inner product

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{k=0}^{m-1} f^{(k)}(0)g^{(k)}(0) + \int_0^1 f^{(m)}(t)g^{(m)}(t)dt$$

and $J(f) = \int_0^1 [f^{(m)}(t)]^2 dt$, so that the resulting estimate of β is a smoothing spline. Then, the reproducing kernel of $\mathcal{H}_1 = \{f \in \mathcal{H} : f^{(k)}(0) = 0, \ 0 \leq k \leq m-1\}$ is $K_1(s, t) = \{(m-1)!\}^{-2}\int_0^1 (s-u)_+^{m-1}(t-u)_+^{m-1}du$ with $u_+ = \max(u, 0)$ and $\phi_l(t) = t^{l-1}/(l-1)!, 1 \leq l \leq m$, are the orthonormal basis of $\mathcal{H}_0$. In the case of $m = 2$, $\beta_\lambda(t) = d_1 + d_2 t + \sum_{n=1}^{N} c_n \int_0^1 x_n(s)K_1(s, t)ds$ with $K_1(s, t) = \int_0^1 (s-u)_+(t-u)_+du$. Theoretical details on reproducing kernel Hilbert space can be found in Wahba (1990).

Inclusion of individual intercepts in (3) makes the estimation of β insensitive to outliers and label noise, in a similar way to Section 2.2. It is often difficult to discern or visualize which function has a different feature from the group where it belongs. In such situation, the estimate of intercept $\hat{\gamma}_n$ can be used for outlier detection.

3. Estimating Algorithm

We use the MM algorithm for stable and convenient optimization. The MM algorithm is an iterative scheme of optimization, consisting of two steps. The majorizing step is to construct a surrogate function that is an upper-bound of the original target function and is tangent to it at the current solution. The minimizing step is to minimize the surrogate function instead of the original function. Two steps alternate in turn until convergence is met. After convergence, the final solution achieves a local minimum of the original function (or

the global minimum if the original function is strictly convex). The effectiveness of the MM algorithm relies on the choice of surrogate function. In this problem, we use the quadratic surrogate function with the constant curvature, which leads to a very simple parameter updating formula. The same approach was used for various statistical models for binary variables (Jaakkola and Jordan, 2000; de Leeuw, 2006; Lee, Huang, and Hu, 2010; Lee and Huang, 2014).

To find a quadratic surrogate function to the negative Bernoulli log-likelihood, we use the following key inequality: at any x^o,

$$-\log\sigma(x) \le -\log\sigma(x^o) - 2\sigma(-x^o)^2 + \frac{1}{8}\left[x - \{x^o + 4\sigma(-x^o)\}\right]^2, \tag{5}$$

for all x. The equality holds only when $x = x^o$. Let $\Theta = \{d, c, \gamma\}$ denote all model parameters collectively and $\Theta^o = \{d^o, c^o, \gamma^o\}$ denote its current estimate. Using (5), a quadratic function majorizing the negative log-likelihood $l(\Theta) = \sum_{n=1}^N \ell(y_n(z_n^T d + \Sigma_n^T c + \gamma_n))$ can be constructed as

$$\tilde{l}(\Theta|\Theta^o) = \frac{1}{8}\sum_{n=1}^N (t_n - z_n^T d - \Sigma_n^T c - \gamma_n)^2 + C, \tag{6}$$

with $t_n = (z_n^T d^o + \Sigma_n^T c^o + r_n^o) + 4y_n\sigma(-y_n(z_n^T d^o + \Sigma_n^T c^o + r_n^o))$ and the constant $C = -\sum_{n=1}^N \{\log\sigma(y_n(z_n^T d^o + \Sigma_n^T c^o + \gamma_n^o)) + 2\sigma(-y_n(z_n^T d^o + \Sigma_n^T c^o + \gamma_n^o))^2\}$. The majorizing properties, $\tilde{l}(\Theta|\Theta^o) \ge l(\Theta)$ and $\tilde{l}(\Theta^o|\Theta^o) = l(\Theta^o)$, hold for any Θ in the parameter space.

Denote the objective function in (4) by $l_p(\Theta) = l(\Theta) + \lambda_1 c^T \Sigma c + P_{\lambda_2}(\gamma)$. Then, the majorizing function of $l_p(\Theta)$ at Θ^o can be immediately obtained by $\tilde{l}_p(\Theta|\Theta^o) = \tilde{l}(\Theta|\Theta^o) + \lambda_1 c^T \Sigma c + P_{\lambda_2}(\gamma)$. Note that minimizing $\tilde{l}_p(\Theta|\Theta^o)$ over Θ is equivalent to minimizing

$$g(\Theta|\Theta^o) = \|t - Zd - \Sigma c - \gamma\|_2^2 + 8\lambda_1 c^T \Sigma c + 8P_{\lambda_2}(\gamma), \tag{7}$$

where $t = (t_1, \ldots, t_N)^T$ and $Z = (z_1, \ldots, z_N)^T$. Therefore, minimization of the majorizing function becomes the penalized least squares problem. MM property says that the global minimum of (4), if it is convex, can be achieved by successively solving the penalized least squares until convergence is met. At each iteration, the current parameter estimate is updated by the minimizer of $g(\Theta|\Theta^o)$. The current estimate $\hat{\Theta}$ serves as Θ^o in the next iteration step.

Estimation through (7) can be decoupled into blockwise minimization. Given $\gamma = \hat{\gamma}$, parameters d and c are estimated as the minimizer of $\|t^* - Zd - \Sigma c\|_2^2 + 8\lambda_1 c^T \Sigma c$. It is given as the following closed-form solution

$$\hat{d} = (Z^T M_{\lambda_1}^{-1} Z)^{-1} Z^T M_{\lambda_1}^{-1} t^*,$$
$$\hat{c} = M_{\lambda_1}^{-1}(t^* - Z\hat{d}) \;=\; M_{\lambda_1}^{-1}(I_N - Z(Z^T M_{\lambda_1}^{-1} Z)^{-1} Z^T M_{\lambda_1}^{-1}) t^*,$$

where $t^* = t - \hat{\gamma}$, $M_{\lambda_1} = \Sigma + 8\lambda_1 I_N$, and I_N is the identity matrix of size N. And, given $(d, c) = (\hat{d}, \hat{c})$, estimation of the intercept vector γ can be done elementwise by a thresholding rule depending on the choice of penalty function $P_{\lambda_2}(\gamma) = \sum_{n=1}^N p_{\lambda_2}(\gamma_n)$. Note that the estimate of γ_n is given by the solution of

$$\min_{\gamma_n \in \mathbb{R}} (t_n^\dagger - \gamma_n)^2 + 8p_{\lambda_2}(\gamma_n), \tag{8}$$

for $n = 1, 2, \ldots, N$, where $t_n^\dagger = t_n - z_n^T \hat{d} - \Sigma_n^T \hat{c}$. This is a penalized least squares problem with the response variable $t_n^\dagger$. When L_1 penalty, $p_{\lambda_2}(|\gamma|) = \lambda_2|\gamma|$, is used, the solution of (8) is given by soft-thresholding rule:

$$\hat{\gamma}_n = \Theta^{\text{soft}}(t_n^\dagger, 8\lambda_2) = \text{sign}(t_n^\dagger)(|t_n^\dagger| - 4\lambda_2)_+ \quad \text{for } n = 1, 2, \ldots, N.$$

We list other types of penalty functions and the associated thresholding-based solutions of (8) in the web-based Supplementary Material available online. Since such penalties are nonconvex, there can exist multiple local minima. One can utilize typical global minimum search strategy, e.g., multiple initial trials.

We implement the above algorithm such that β resides in $\mathcal{H} = \mathcal{W}_2^2$. In that case, recall that $\phi_1(t) = 1, \phi_2(t) = t$, and $K_1(s,t) = \int_{\mathcal{T}} (s-u)_+(t-u)_+ du$. The matrices Z and Σ in (7) require evaluation of integrals. In our implementation, all necessary integrations are computed in the numerical manner. After obtaining $\hat{d}$ and $\hat{c}$, we evaluate the discriminant function for a new input data $(u_*, x_*(\cdot))$ as $f(u_*, x_*) = z_*^T \hat{d} + \Sigma_*^T \hat{c}$, where $z_* = (1, u_{*1}, \ldots, u_{*p}, z_{*1}, z_{*2})^T$ and $\Sigma_* = (\Sigma_{*n})_{n=1,\ldots,N}$ with $z_{*1} = \int_{\mathcal{T}} x_*(t)dt$, $z_{*2} = \int_{\mathcal{T}} x_*(t)tdt$, and $\Sigma_{*n} = \int_{\mathcal{T}}\int_{\mathcal{T}} x_*(s)x_n(t)K_1(s,t)dsdt$. For the regularization parameters λ_1 and λ_2, we select the optimal regularization parameters that achieve the minimum of validated log-likelihood among grid points using two-dimensional grid search. Since we do not often have a validation set free from label noise, 10% upper trimmed mean is used for validation score computation in order to reduce outlier effect on the model selection. Possible missing values on discretized function evaluation can be treated in the presmoothing step, and moderate level of missingness does not seriously affect the performance.

For comparison purpose, we also implemented the label-flipping logistic regression proposed by Bootkrajang and Kabán (2012). Its detailed implementation with a functional covariate is described in the web-based Supplementary Material available online.

4. Simulation Study

Following Shin (2008), we generated two groups of functions with a domain $\mathcal{T} = [0, 1]$ as

$$x_n(t) = \sum_{c\in\mathcal{C}} \mu_c(t) I(y_n = c) + \sum_{k=1}^{50} k^{-1/2} U_{nk} \sqrt{2}\cos(k\pi t)$$
$$\text{for } t \in \mathcal{T},\ n = 1, 2, \ldots, N,\ \mathcal{C} = \{+1, -1\},$$

where $\mu_{+1}(t) = 3\sqrt{2}\cos(\pi t) + \sqrt{2}\cos(2\pi t)$, $\mu_{-1}(t) = \sqrt{2}\cos(2\pi t)$, and U_{nk} are independent draws from the standard normal distribution. We labeled $y_n = +1$ for $n = 1, 2, \ldots, N_1$ and $y_n = -1$ for $n = N_1 + 1, N_1 + 2, \ldots, N_1 + N_2 (= N)$. In this simula-

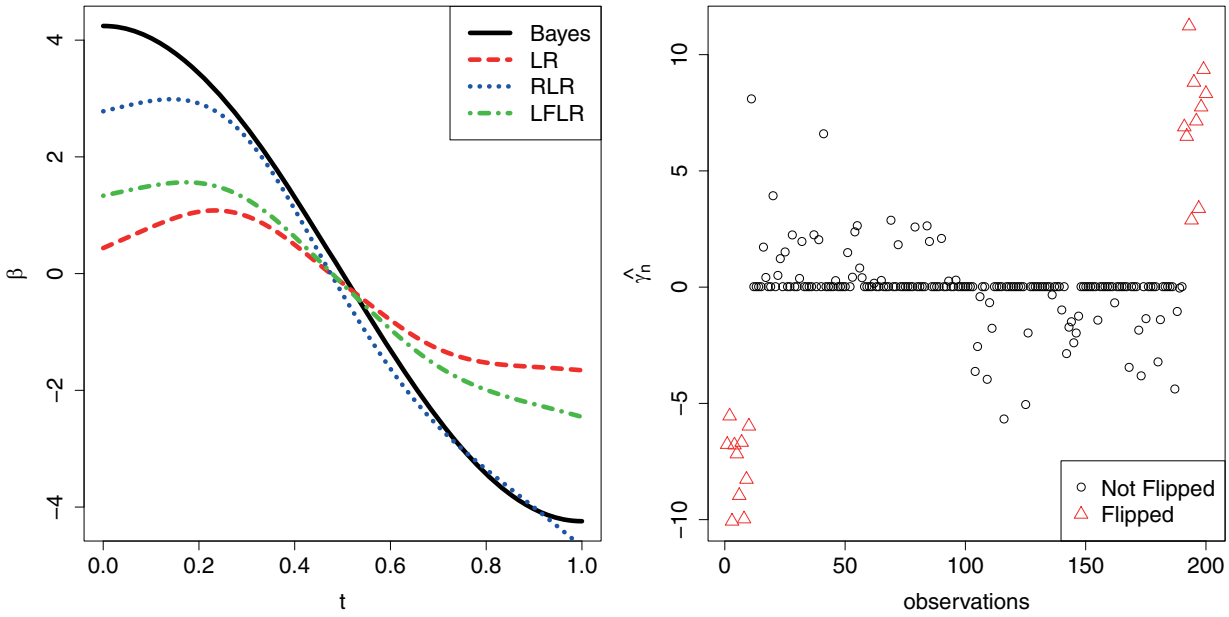

Figure 1. Results from a simulated data with 10% label noise. Left: Bayes coefficient function (Bayes) and its estimates using logistic regression (LR), robust logistic regression with L_1 penalty (RLR), and label-flipping logistic regression (LFLR). Right: individual intercept estimates from robust logistic regression with L_1 penalty. Triangles indicate the observations with flipped label. This figure appears in color in the electronic version of this article.

tion design, we can show that the coefficient function corresponding to the Bayes classifier is given as $\beta(t) = 3\sqrt{2}\cos(\pi t)$. We took the same sample size for both groups (i.e., $N_1 = N_2 = N/2$) in the simulation. We generated training data of size $N = 200$ for classification learning. We also generated validation data of the size 500 for the regularization parameter selection. To see label-noise effect on estimation, we flipped over the class labels for 10% observations in both classes. We estimated coefficient function from the noisy training data under each grid point of regularization parameters, and then found the optimal regularization parameter using validation data. The final estimate of the coefficient function was obtained using this optimal regularization parameter.

The left panel of Figure 1 presents estimates of the coefficient function from three methods. The coefficient estimate from the traditional functional logistic regression (LR in the plot legend) looks quite different from the true coefficient function (Bayes) by showing shrinkage toward zero. This suggests that label noise tends to diminish the importance of the functional covariate on the classification unless we treat the label noise properly in LR. On the other hand, a robust logistic regression (RLR) with L_1 penalty yields a coefficient function estimate reasonably close to the true one, while label-flipping logistic regression (LFLR) performed less satisfactorily. Individual intercept estimates from RLR are presented in the right panel of Figure 1, where the first 100 observations belong to the class of $y = +1$ and the remaining 100 observations belong to the class of $y = -1$. To indicate label-flipped observations, we highlighted them with the different color and shape in the plot. As shown in the plot, all label-flipped observations have nonzero intercept estimates. Thus, individual intercept estimate of nonzero can be regarded as a measure of decision uncertainty. We also observe that some observations locating on the wrong side or near the decision boundary have nonzero estimates while their labels are not flipped. This reduces the sensitivity to training data and enhances prediction performance, as shown in the below numerical study (See Tables 1 and 2).

We repeated the same simulation 1000 times for each case of training sample with sizes $N = 100, 200, 400$ and flipping rates $0, 0.05, 0.10, 0.15, 0.20$. Two different types of label noise were considered: symmetric and asymmetric label noise. In symmetric label noise scenario, the same proportion of observations in both classes was randomly selected and their labels were flipped over, while a proportion of label flipping in the class of $y = +1$ was given a half of label flipping proportion in the class of $y = -1$ under asymmetric label noise case. We measured estimation accuracy by $\|\hat{\beta} - \beta\|_2 = \sqrt{\int (\hat{\beta}(t) - \beta(t))^2 dt}$, which measures the distance between the true coefficient function β and its estimate $\hat{\beta}$. From 1000 repetitions, we calculated the average distance for each method and presented it in Table 1. In the case of no label noise, as we expected, LR yielded coefficient estimates closest to the true coefficient function on average. Interestingly, RLR outperformed LR even in no swapping case when $N = 100$. In learning from small training data, individual intercepts help reducing sensitivity for training set. As label noise level increases, however, the performance of LR gets worse. In contrast to this, the coefficient estimates from RLR and LFLR are more robust under label noise. We observed that RLR showed the best performance among them, while competitiveness of LFLR is observed in large training set ($N = 400$).

To measure test error rate, we generated test data of size 1000 and applied the classification rules derived from train-

Table 1

Average distance between the true coefficient function and its estimates based on the estimated decision boundaries from conventional logistic regression (LR), robust logistic regression with L_1 penalty (RLR), and label-flipping logistic regression (LFLR)

		Symmetric			Asymmetric		
N	Swapping rate	LR	RLR	LFLR	LR	RLR	LFLR
100	0.00	0.9223	0.8356	3.0339	-	-	-
	0.05	1.8885	0.9904	2.9436	1.8639	1.0118	2.9540
	0.10	3.7048	0.7499	3.6806	3.4095	0.8138	3.3522
	0.15	4.5765	1.2393	4.1480	4.7045	1.2495	3.9358
	0.20	5.6131	2.6386	4.5714	5.3399	2.5129	4.5933
200	0.00	0.5750	0.6830	0.9458	-	-	-
	0.05	2.3115	0.7639	1.3768	2.0744	0.8626	1.2918
	0.10	3.7240	0.5497	1.8475	3.5515	0.6076	1.7389
	0.15	4.7852	1.0980	2.4739	4.7117	1.1479	2.4124
	0.20	5.6218	2.5201	3.3770	5.4551	2.3600	3.1769
400	0.00	0.4542	0.6438	0.4466	-	-	-
	0.05	2.3213	0.8301	0.6566	2.1934	0.9004	0.6256
	0.10	3.7638	0.5858	0.9662	3.6579	0.6123	0.9539
	0.15	4.8048	1.1681	1.5242	4.7410	1.1355	1.4821
	0.20	5.6419	2.6348	2.4798	5.5187	2.4566	2.3716

ing data on them. We repeated this whole procedure 1000 times and provide average test error rates in Table 2. In prediction, unlike Table 1, LR shows better performance in the symmetric case than in the asymmetric case. This coincides the fact that symmetric label noise is less harmful than asymmetric label noise in the prediction performance (Frenáy and Kabán, 2014), which is partially supported by Proposition 2. However, this simulation suggests that symmetric label noise is not safe in coefficient function estimation, as shown in Table 1. Table 2 shows that RLR has the best predictive power for all levels of label noise, even in no label noise. LFLR shows a little gain in prediction only in the asymmetric label noise case.

5. Alzheimers Disease Application

5.1. *Data Description and Preparation*

We apply our method to build a classifier based on the CC thickness profiles, which differentiates the patients with mild AD from normal elderly people. To do this, we exploit images from the Open Access Series of Imaging Studies (OASIS) MRI database (Marcus et al., 2007). The OASIS cross-sectional dataset provides 3D MPRAGE MRI brain scans from 416 right-handed subjects. Out of 416, the 98 healthy normal subjects aged 60 or above without dementia (CDR = 0) and all 98 subjects aged 60 or above with very mild/mild AD (CDR = 0.5 or 1) are considered in this study. The clinical

Table 2

Average test error rates based on the estimated decision boundaries from conventional logistic regression (LR), robust logistic regression with L_1 penalty (RLR), and label-flipping logistic regression (LFLR)

		Symmetric			Asymmetric		
N	Swapping rate	LR	RLR	LFLR	LR	RLR	LFLR
100	0.00	0.0721	0.0706	0.0726	–	–	–
	0.05	0.0730	0.0719	0.0765	0.0745	0.0723	0.0761
	0.10	0.0751	0.0730	0.0850	0.0783	0.0741	0.0834
	0.15	0.0773	0.0749	0.0936	0.0890	0.0804	0.0966
	0.20	0.0815	0.0787	0.1100	0.1061	0.0920	0.1107
200	0.00	0.0702	0.0696	0.0704	–	–	–
	0.05	0.0709	0.0702	0.0727	0.0716	0.0704	0.0722
	0.10	0.0718	0.0707	0.0763	0.0767	0.0725	0.0758
	0.15	0.0737	0.0720	0.0827	0.0850	0.0771	0.0823
	0.20	0.0757	0.0739	0.0908	0.0985	0.0860	0.0912
400	0.00	0.0690	0.0688	0.0689	–	–	–
	0.05	0.0696	0.0693	0.0700	0.0707	0.0698	0.0698
	0.10	0.0702	0.0696	0.0715	0.0746	0.0713	0.0714
	0.15	0.0713	0.0706	0.0747	0.0829	0.0758	0.0741
	0.20	0.0729	0.0722	0.0794	0.0983	0.0851	0.0794

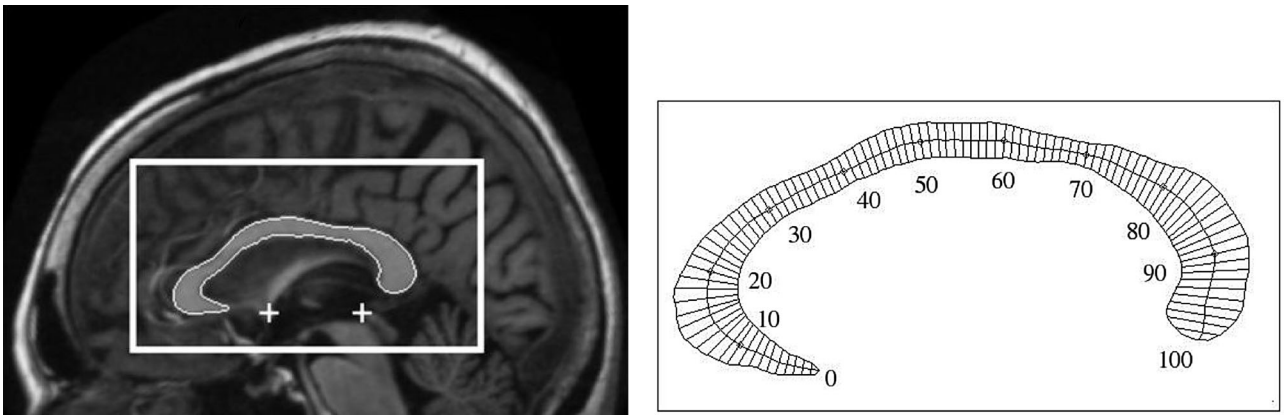

Figure 2. (Left) The outline shows the detected corpus callosum. The automatically detected anterior commissure (AC) and posterior commissure (PC) are shown by the plus signs. (Right) 99 thickness superimposed on a human corpus callosum.

dementia rating (CDR) is a clinical rating for dementia applicable to AD, which is widely used by clinicians (Morris, 1993). A global CDR score is derived from rates of a patient's cognitive and functional performance. It is ranged from 0 to 3: 0 = no dementia, 0.5 = very mild, 1 = mild, 2 = moderate, 3 = severe. We summarize the demographics of the subjects by CDR included in this study in the web-based Supplementary Material available online.

A multi-atlas model-based segmentation method using the Automatic Registration Toolbox (ART) nonlinear registration algorithm (Ardekani et al., 2012) was used to locate the CC, as shown in the left panel of Figure 2. After locating the CC, a module, *yuki* in ART, was used to compute callosal thicknesses at 99 equally spaced interior points along the medial axis of the CC (right panel of Figure 2). A CC thickness profile was generated for each of the 196 subjects in our study, each profile consisting of 99 nonzero thicknesses, with no missing values, sampled starting from the tip of the rostrum to the most inferior section of the splenium.

The CC thickness profiles of 196 subjects and their group mean curves (thicker lines) are displayed in the left panel of Figure 3, where black curves represent 98 normal subjects of CDR = 0, and red curves represent 70 very mild AD subjects of CDR = 0.5 and 28 mild AD subjects of CDR = 1. Although two groups are not easy to be discriminated visually, the average thickness of CC for non-AD subjects is thicker than that for AD-suspicious subjects over all 99 positions. It is known that the cross-sectional areas of the CC decreases with age and the volume of the CC is also correlated with the intra-cranial volume. Also, there is a potential of sexual dimorphism in the CC (Holloway et al., 1993). The OASIS dataset contains information on age, sex, and eTIV (estimated total intra-cranial volume) as well as CC images. To remove the effects of those variables, we took the residuals from the regression model of CC against age, sex, and $\text{eTIV}^{1/3}$ at each location (Lee et al., 2014). The right panel of Figure 3 presents the adjusted CC thickness, which was used as a functional covariate in the subsequent classification task.

5.2. *Classification Results*

We coded $y_n = -1$ for subjects having CDR = 0, and $y_n = +1$ for subjects of both CDR = 0.5 and CDR = 1. The OASIS data also provide a measure of cognitive impairment, the Mini-Mental State Examination (MMSE) score that is widely

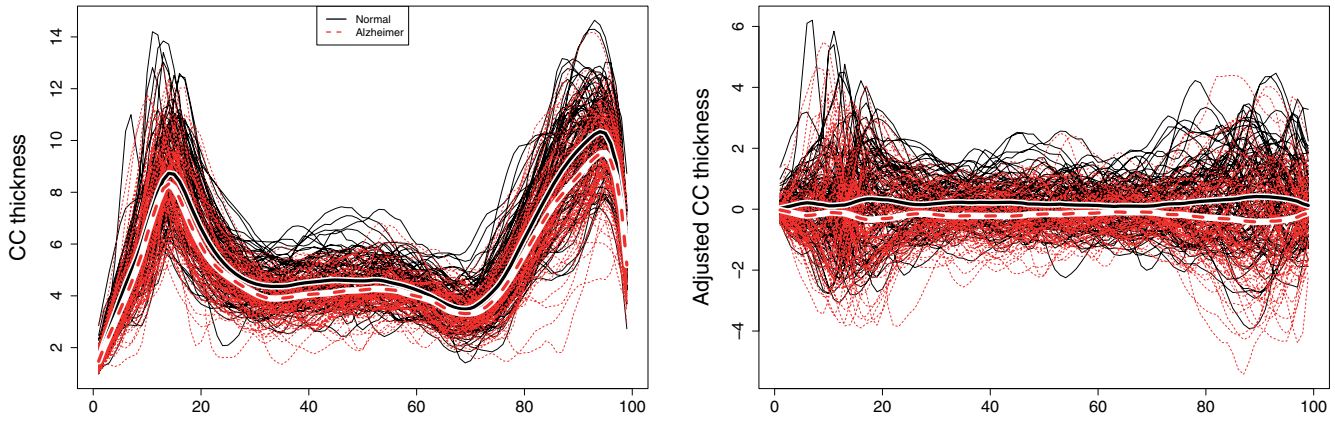

Figure 3. (Left) 196 CC thickness profiles with group means. Solid and dashed curves are, respectively, for normal subjects and very mild/mild Alzheimer subjects. (Right) CC profiles after adjusting SEX, AGE, and $\text{eTIV}^{1/3}$. This figure appears in color in the electronic version of this article.

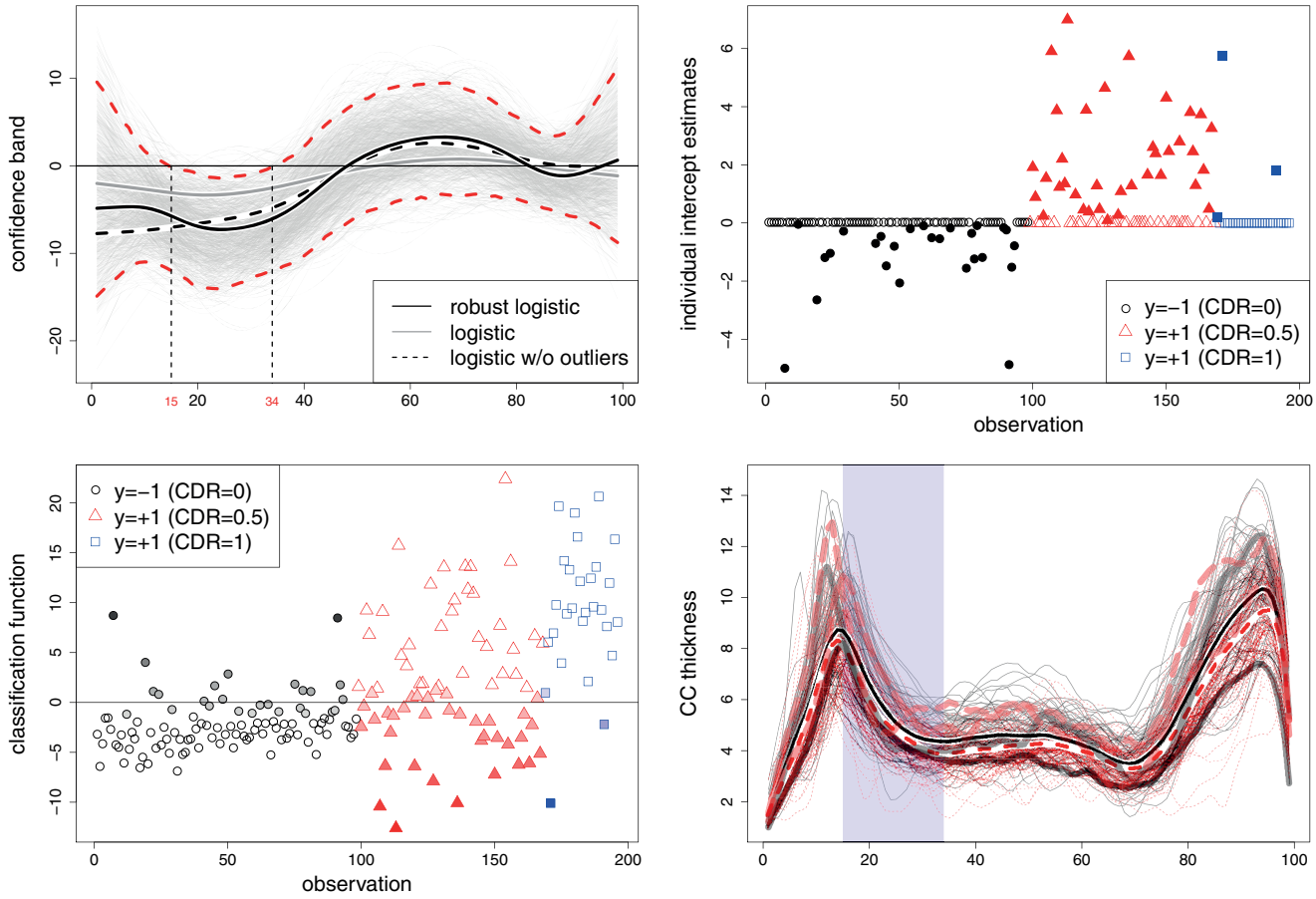

Figure 4. (Top-left) The estimated coefficient functions and bootstrap confidence band (outer dashed lines) are presented. (Top-right) The individual intercept estimates across all subjects are shown. Solid points indicate nonzero estimates. (Bottom-left) Discriminant function values for all subjects are displayed. Empty points represent the observations having zero intercept estimates and the observations having nonzero intercept estimate is highlighted as filled points whose intensity is proportional to the magnitude of intercept estimates. (Bottom-right) CC thickness profiles for potential outliers are highlighted with wide curves. This figure appears in color in the electronic version of this article.

used by clinicians to screen for dementia. To increase the prediction power, we included MMSE in the model. Thus, discriminant function is modeled by

$$f(u_n, x_n) = \alpha_0 + \alpha_1 \mathrm{MMSE}_n + \int \beta(t) x_n(t) dt.$$

Here, $x_n(t)$ denotes the adjusted CC thickness profile curve. The penalty parameters (λ_1, λ_2) were tuned by a fivefold cross-validation under two-dimensional grid search. Using the estimate of the above discriminant function, the accuracy of classification was 85% (166/196) with 77% (75/98) sensitivity and 93% (91/98) specificity.

The coefficient function estimate is depicted in the top-left panel of Figure 4 with the solid black curve. The solid gray curve represents the coefficient function estimate from LR without considering potential label noise existence. The latter is closer to zero than the former, indicating that there may exist label noise in this data as we observed from simulation study. To look at the significant difference in CC thickness, we constructed a pointwise 95% bootstrap confidence band (dashed red curves) from 1000 bootstrap simulations. Region between 15 and 34 shows significant negative coefficients, suggesting that these frontal CC regions are narrower in the mild AD group than in the normal group. The individual coefficient estimates from RLR are depicted as solid points in the top-right panel of Figure 4. In the AD group, most nonzero estimates appear among the very mild subjects (red points in the plot) as we expect that clinical decision for AD could be less accurate for them. We fit LR after removing the subjects of having nonzero intercepts and draw its coefficient function estimate (dashed black curve) in the top-left panel of Figure 4. It is close to the coefficient function estimate from RLR, indicating that our method yields a coefficient function estimate robust to inaccurate labeling.

The bottom-left panel of Figure 4 presents the discriminant function estimates from RLR, showing how far the subjects are located from the decision boundary. The subjects having nonzero intercept estimates are highlighted by filling colors whose densities are proportional to absolute values of their intercept estimates. We can see that the subjects locating on the

wrong side have nonzero intercepts, based on the estimated decision boundary. We also observe that the subjects near the decision boundary have nonzero intercepts, while they are correctly classified, indicating their uncertain diagnosis. Thus, nonzero intercept plays a useful measure for identifying subjects who require further investigation. Finally, we provide the plot in the bottom-right panel of Figure 4, where CC thickness profiles for observations having two largest and two smallest intercepts are highlighted with thicker curves. The shaded blue area indicates the region where the CC thickness is significantly different between AD and non-AD groups based on a bootstrap confidence band. Two subjects having CC thickness shown as wide black curves are diagnosed as normal, but they have large negative intercepts. This suggests that they are possibly misdiagnosed as normals while their CC thickness profiles are similar to mild AD patients. We can carry over this interpretation to CC profiles shown as wide red curves in a similar manner.

6. Discussion

In this article, motivated by MRI data on Alzheimer's disease, we propose to use a functional data classification method which is robust to label noise. Inclusion of individual intercept to the logistic regression is proven to be useful to derive a robust and efficient classification under the existence of label noise. We applied our robust classification to Alzheimer's disease classification using the CC thickness profiles derived from brain MRI images.

In this work, we develop a robust logistic regression for functional data classification by placing individual intercepts in the classification model. However, the same approach can be directly and easily applied to other margin-based classification methods (for example, boosting and support vector machines) for the purpose of robust classification and label noise treatment. Most margin-based classification methods rely on the convex loss minimization. For example, exponential loss, $\exp(-yf(x))$ is used for boosting and hinge loss, $(1 - yf(x))_+$ is for support vector machines. Since our arguments discussed in the article can be generalized to any convex loss function for classification, robust classifications similar to the proposed one in this article can be also derived under other margin-based classifications.

7. Supplementary Materials

Web Appendices referenced in Sections 2, 3, and 5 and R codes for implementation are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

Open Access Series of Imaging Studies (OASIS) project was supported under grants: P50 AG05681, P01 AG03991, R01 AG021910, P20 MHO71616, and U24 RR021382. We thank Drs Ardekani and Bachman for processing MRI images. S. Lee is financially supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1002779).

References

Aggarwal, C. C. and Yu, P. S. (2001). Outlier detection for high dimensional data. In *Proceedings of ACM SIGMOD 2001*, 37–46. Santa Barbara: ACM press.

Ardekani, B. A., Toshikazu, I., Bachman, A. H., and Szeszko, P. R. (2012). Multi-atlas corpus callosum segmentation with adaptive atlas selection. *Proceedings of the ISMRM*. Australia: Melbourne, International Society for Magnetic Resonance in Medicine (ISMRM).

Bianco, A. M. and Yohai, V. J. (1996). Robust estimation in the logistic regression model. In *"Robust Statistics, Data Analysis, and Computer Intensive Methods, Volume 109 of Lecture Notes in Statistics,"* H. Rieder (ed). New York: Springer-Verlag.

Bootkrajang, J. and Kabán, A. (2012). Label-noise robust logistic regression and its applications. *Machine Learning and Knowledge Discovery in Database Lecture Notes in Computer Sciences*, Part I, 143–158.

Brodley, C. E. and Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research* **11**, 131–166.

Carroll, R. J. and Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society, Series B* **55**, 693–706.

Copas, J. B. (1988). Binary regression models for contaminated data (with discussion). *Journal of the Royal Statistical Society, Series B* **50**, 225–265.

de Leeuw, J. (2006). Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data Analysis* **50**, 21–39.

Di Paola, M., Luders, E., Di Iulio, F., Cherubini, A., Passafiume, D., Thompson, P. M., et al. (2010). Callosal atrophy in mild cognitive impairment and Alzheimer's disease: Different effects in different stages. *Neuroimage* **49**, 141–149.

Frederiksen, K. S., Garde, E., Skimminge, A., Ryberg, C., Rostrup, E., Baaré, W. F. C., et al. (2011). Corpus callosum atrophy in patients with mild Alzheimer's disease. *Neurodegenerative Diseases* **8**, 476–482.

Frenáy, B. and Kabán, A. (2014). A comprehensive introduction to label noise. *Proceedings of the European Symposium on Artificial Neural Network*, 23–25. Bruges: Elsevier.

Holloway, R. L., Anderson, P. J., Defendini, R., and Harper, C. (1993). Sexual dimorphism of the human corpus callosum from three independent samples: Relative size of the corpus callosum. *American Journal of Physical Anthropology* **92**, 481–498.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter selection via variational methods. *Statistics and Computing* **10**, 25–37.

Lee, S. and Huang, J. Z. (2014). A biclustering algorithm for binary matrices based on penalized Bernoulli likelihood. *Statistics and Computing* **24**, 429–441.

Lee, S., Huang, J. Z., and Hu, J. (2010). Sparse logistic principal components analysis for binary data. *The Annals of Applied Statistics* **4**, 1579–1601.

Lee, Y., MacEachern, S. N., and Jung, Y. (2012). Regularization of case-specific parameters for robustness and efficiency. *Statistical Science* **27**, 350–372.

Lee, S. H., Yu, D., Bachman, A. H., Lim, J., and Ardekani, B. A. (2014). Application of fused lasso logistic regression to the study of corpus callosum thickness in early Alzheimer's disease. *Journal of Neuroscience Methods* **221**, 78–84.

Lin, Y. (2004). A note on margin-based loss functions in classification. *Statistics and Probability Letters* **68**, 73–82.

Malossini, A., Blanzieri, E., and Ng, R. T. (2006). Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics* **22**, 2114–2121.

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience* **19**, 1498–1507.

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology* **43**, 2412–2414.

Muhlenbach, F., Lallich, S., and Zighed, D. A. (2004). Identifying and handling mislabeled instances. *Journal of Intelligence Information Systems* **22**, 89–109.

Park, S. Y. and Liu, Y. (2011). Robust penalized logistic regression with truncated loss functions. *The Canadian Journal of Statistics* **39**, 300–323.

She, Y. and Owen, A. B. (2011). Outlier detection using non convex penalized regression. *Journal of the American Statistical Association* **106**, 626–639.

Shin, H. (2008) An extension of Fisher's discriminant analysis for stochastic processes. *Journal of Multivariate Analysis* **99**, 1191–1216.

Tibshirani, J. and Manning, D. C. (2014). Robust logistic regression using shift parameters. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* **2**, 124–129.

Wahba, G. (1990). *Spline Models for Observational Data*, Philadelphia: SIAM.

Wu, Y. and Liu, Y. (2007). Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association* **102**, 974–983.

Received August 2015. *Revised January* 2016.
Accepted January 2016.