Computational Statistics and Data Analysis 131 (2019) 37–49

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust probabilistic classification applicable to irregularly sampled functional data

Yeonjoo Park [1], Douglas G. Simpson *

Department of Statistics, University of Illinois at Urbana-Champaign, 725 S Wright St., Champaign, IL 61820, USA

ARTICLE INFO

Article history:
Received 14 November 2017
Received in revised form 12 May 2018
Accepted 3 August 2018
Available online 11 August 2018

Keywords:
Bayes classifier
Mixed effects model
Probabilistic classification
Robustness
t-model

ABSTRACT

A robust probabilistic classifier for functional data is developed to predict class membership based on functional input measurements and to provide a reliable probability estimate for class membership. The method combines a Bayes classifier and semi-parametric mixed effects model with robust tuning parameter to make the method robust to outlying curves, and to improve the accuracy of the risk or uncertainty estimates, which is crucial in medical diagnostic applications. The approach applies to functional data with varying ranges and irregular sampling without making parametric assumptions on the within-curve covariance. Simulation studies evaluate the proposed method and competitors in terms of sensitivity to heavy tailed functional distributions and outlying curves. Classification performance is evaluated by both error rate and logloss, the latter of which imposes heavier penalties on highly confident errors than on less confident errors. Run-time experiments on the R implementation indicate that the proposed method scales well computationally. Illustrative applications include data from quantitative ultrasound analysis and phoneme recognition.

1. Introduction

A common goal in medical image analysis is to use the information in the image from a new subject to classify type or stage of a region of interest based on prior training data. For example, a classification tool might be utilized during surgery to evaluate areas with abnormal tissues or to access tumor margins using biomedical imaging techniques (e.g., Nolan et al., 2016). In these scenarios, reporting a diagnostic probability estimate for the classification is more informative than simply reporting a most likely class, particularly in settings where the classes might be malignant or benign tumors. In such settings probabilistic classifiers, which predict class probability in addition to class membership, provide critical extra information (Hastie et al., 2009). In this paper, we build a probabilistic classifier for functional data, which provides class prediction and a probability estimate robust to outlying curves or regions of unduly large observations in the data. We demonstrate that the robust classifiers provide competitive error rate performance and more accurate probabilistic risk estimates than those provided by Gaussian process classifiers and competing methods.

This work was motivated in part by research on quantitative ultrasound (QUS) methodology in which biological tissue is evaluated using the information in the ultrasound backscatter signals. The frequency-dependent backscatter coefficients (BSC), constructed as attenuation-adjusted power spectra, comprise functional data that can be used for classification and testing. For example, Wirtzfeld et al. (2013) employed mouse and rat tumor models to investigate the efficacy of QUS

* Corresponding author.
E-mail addresses: yeonjoo.park@utsa.edu (Y. Park), dgs@illinois.edu (D.G. Simpson).
[1] Present address: Department of Management Science and Statistics, University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, USA.

https://doi.org/10.1016/j.csda.2018.08.001
0167-9473/

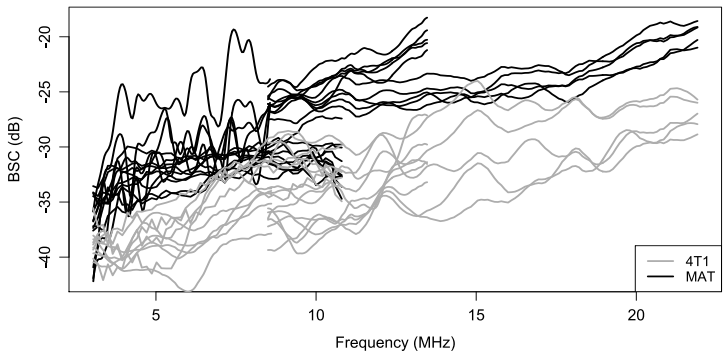

Fig. 1. BSC curves for mammary tumors from mice (4T1) and rats (MAT).

measurements for detecting the difference between the two types of tumors. The study produced irregular functional data in which the ultrasound frequency ranges varied across the collection of measured BSC curves. Fig. 1 illustrates the type of data produced. In the experiment, functional BSC were collected in several laboratories using different transducers which cover distinct ranges of frequency. Each tumor region of interest was scanned with multiple shifted scan lines, resulting in multiple levels of correlation within and between curves. Furthermore, noninvasive scanning raises the potential for outlying measurements due to unexpected contamination, e.g., heterogeneity due to the inclusion of neighboring tissue in the scanned region. The statistically distinct behaviors of BSC on different types of the tumor were demonstrated by Wirtzfeld et al. (2015). A subsequent goal is to develop a classifier to predict the class for a new measured BSC curve or set of curves. For medical diagnostic use, it is important to provide an accurate diagnostic probability along with the class membership.

A variety of techniques have been developed for classifying functional data. Existing methods include density-based classifiers (Hall et al., 2001; James and Hastie, 2001; Ferraty and Vieu, 2003; Delaigle and Hall, 2013; Dai et al., 2017), regression-based methods (James, 2002; Müller and Stadtmüller, 2005; Araki et al., 2009; Goldsmith et al., 2011; Zhu et al., 2012), k-nearest neighborhood classifiers (Biau et al., 2005; Cérou and Guyader, 2006; Biau et al., 2010; Fuchs et al., 2015), and Bayesian methods (Wang et al., 2007). Although there has been extensive development, many existing classifiers require data to be observed over the same interval, which cannot be applied to our motivating example without further processing or imputation.

Among methods that apply to irregular functional data, the procedures often entail a dimension reduction step such as functional principal component analysis (FPCA), exploiting a second-order based data-driven eigenanalysis to represent high-dimensional data on finite dimensional feature space. Functional classifiers have also been derived under Gaussian assumptions on the noise and random-effect terms in the functional mixed model (James and Hastie, 2001). These Gaussian or second order methods might be vulnerable to outlying curves and heavy tailed response distributions.

Robust versions of FPCA (Bali et al., 2011; Boente and Salibian-Barrera, 2015) are available for regular functional data but not for irregular functional data. Zhu et al. (2012) proposed a robust classifier based on the functional mixed model framework in wavelet space by allowing heavy tailed distributions for wavelet coefficients. The approach can incorporate multiple correlated functions, but the wavelet transformation is more suited for stationary regular, spiked functional data rather than the irregular smooth functional data in our applications.

Fuchs et al. (2015) developed a nonparametric ensemble method that estimates posterior probabilities by combining nearest neighbor posterior probabilities from multiple distance measures in an ensemble. However, their metrics are defined on regularly sampled functional samples. The application to irregular functional data does not appear straightforward, especially when the data are sparse. Nonetheless, a simplified version of this method is included in the semi-regular phoneme example below and performs well in that context.

Here we develop a robust probabilistic Bayes classifier based on a semiparametric mixed effects model with robust tuning parameter. Discretized training data are modeled using a mixed effects framework, as developed by Shi et al. (1996) and Rice and Wu (2001). We modify the approach by incorporating heavy-tailed distributions on the random coefficients as well as the noise errors. The resulting empirical posterior probabilities exhibit improved robustness and accuracy compared with the original methods. The approach also enables robust analysis of irregularly collected samples with flexible within-curve covariance structure. In addition to reducing the impact of outlying curves on the classification, our approach has the benefit of providing more accurate assessment of the classification uncertainty than the Gaussian methods. It is also straightforward to extend to multi-level data to build a cluster-level classification by adding cluster-specific random effect terms. In addition to error rates we employ logloss to measure accuracy of probability estimates, because it imposes more penalties in being very confident but incorrect prediction. With a similar concern, Araki et al. (2009) examined the posterior probabilities from functional logistic regression, but the evaluation considered prediction error rate only.

The rest of this article is organized as follows. The classification procedure is developed in Section 2 including details such as model selection and computational procedures. Section 3 presents comparative simulation studies under different

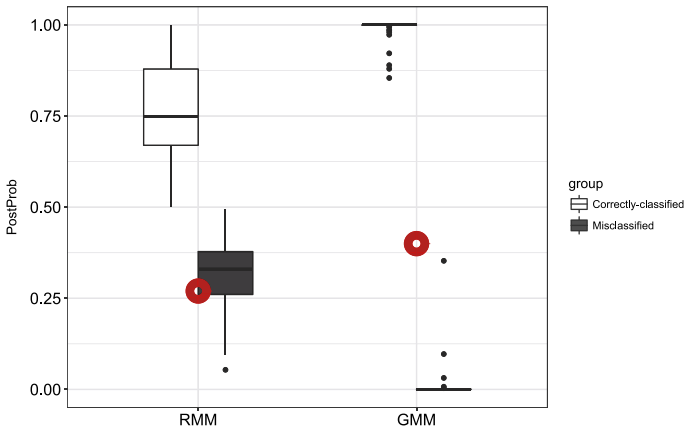

Fig. 2. Box plots of posterior probabilities for correctly classified (white) and incorrectly classified (black) functional data using from RMM and GMM. Thick red circles indicate overall misclassification rates.

scenarios. Section 4 provides comparative evaluations of the methods for data from quantitative ultrasound research and speech recognition. The paper concludes with a discussion in Section 5. Computational implementation of the method in R is provided in the supplementary material.

2. Functional classification rule

For $i = 1, \ldots, N$, let $y_i(s)$ denote a function from the training sample measured on a subset $\mathcal{S}_i$ of a common interval $\mathcal{S}$, and let $c_i \in \{1, \ldots, G\}$ be the corresponding class label. The observations from a given class g are assumed to share the same underlying mean function $\eta_g(s)$ and covariance function $\gamma_g(s, t)$, $s, t \in \mathcal{S}$, not necessarily stationary, although the individual functions might be observed over different subsets of $\mathcal{S}$. Let π_g denote the prior probability that an observation falls into group g. Our goal is to find a model to assign a new observation y to one of the G classes. The optimal Bayes classification rule maximizes the posterior probability of the class g,

$$P(class = g|y) = \frac{p(y|g)\pi_g}{\sum_{j=1}^{G} p(y|j)\pi_j}, \tag{1}$$

where $p(y|g)$ denotes the likelihood of y conditional on the class label g, $g = 1, \ldots, G$.

In practice, functional curves are recorded over finite grids rather than being observed continuously. Let $y_{it} = y(s_{it})$ represent observed functional measurements of ith curve at location s_{it}, $t = 1, \ldots, n_i$, $i = 1, \ldots, N$, a realization of a stochastic process from corresponding group. Then for a new functional observation in the form of a vector, $\mathbf{y}$, we can approximate (1) by replacing $p(y|g)$ with $f(\mathbf{y}|g)$, where $f(\cdot|g)$ denotes the density of the distribution of group g, estimated through discretized training information.

The posterior probability of class membership in (1) provides important information about the uncertainty of the classification beyond merely indicating the most probable class. We contend that this probability or risk information is critical in diagnostic applications. Traditional Gaussian classification methods may be vulnerable to overconfidence in the case of heavier tailed biomarker distributions. To illustrate, consider Fig. 2, which presents posterior probabilities from a robust classifier (RMM) of the type developed here and from a more traditional Gaussian process classifier (GMM). The data were generated under a heavy tailed distribution as part of the simulation study in Section 3. From the overall error rates, indicated by the heavy red circles, it is clear that these data are inherently difficult to classify. The distributions of the RMM posterior probability estimates accurately reflect this uncertainty in the classification. The Gaussian classifier, on the other hand, produces extreme posterior probabilities for the individual curves, indicating grossly over-optimistic confidence in the erroneous classifications. Our proposed robust classifier (RMM) not only gives a lower error rate, but it also gives more realistic and informative posterior probabilities on misclassified observations.

In the overall evaluation of methods we employ both the classification error rate and the logarithmic loss or LogLoss. The latter measure is commonly used to capture the accuracy of the probability estimates in the classification; see for example, the Kaggle competition web site (https://www.kaggle.com). The LogLoss is the negative log-likelihood of the Bernoulli model, which provides penalties for being confident in wrong classifications, defined as

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G} I_{ig}\log(p_{ig}), \tag{2}$$

where i indexes individual curves, $i = 1, \ldots, N$, and I_{ig} equals 1 if observation i is the member of class g, and equals 0 otherwise. p_{ig} denotes posterior or predicted probability of ith curve to be assigned to class g. The worst case is when a single observation is predicted as definitely false but it is actually true. To avoid infinite loss, the probability is usually lower bounded in this metric with small positive probability close to zero. We employ a lower bound of $\epsilon = 10^{-15}$.

2.1. Model specification

To obtain more robust classification and posterior densities in (1), we extend the Gaussian nonparametric linear mixed effects model of Rice and Wu (2001) to a robust semiparametric mixed effects model (RMM) for training data, with M_g observations in Class g, as follows:

$$\begin{aligned} &\mathbf{y}_i = \mathbf{B}_i\boldsymbol{\beta} + \mathbf{R}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, M_g, \\ &\boldsymbol{\gamma}_i \overset{ind}{\sim} t_q(\mathbf{0}, \Gamma, \nu),\ \boldsymbol{\epsilon}_i \overset{ind}{\sim} t_{n_i}(\mathbf{0}, \Lambda_i, \nu), \end{aligned} \tag{3}$$

where $\mathbf{y}_i$ denotes length n_i discretized response from ith curve, $\mathbf{B}_i$ and $\mathbf{R}_i$ are corresponding $n_i \times p$ and $n_i \times q$ spline basis matrices evaluated at corresponding grid points on $\{B_k(\cdot)\}$, a basis spline functions with a fixed knot sequence, and $\{R_l(\cdot)\}$, a possibly different basis for random coefficients, respectively. As a working model we employ a multivariate t-distribution with the same known or unknown degrees of freedom (ν) on random coefficients and measurement errors, where Γ and Λ_i represent their scale matrices, respectively.

This extension not only preserves advantages of Rice and Wu (2001), such as nonparametric fit of a mean function, a flexible approximation of within-curve covariance structure through unstructured covariance matrix Γ, and broad applicability to general functional data not limited to the regular structure, but also provides a base to construct the robust probabilistic classifier. With infinite degrees of freedom, the proposed model (3) includes Rice and Wu (2001). Besides nonparametric basis functions, the effect of other factors, such as length and weight of tumors in the motivating example, can be considered by adding corresponding terms in the model.

Based on the trained robust model for each class, the class likelihoods of a new observation are calculated to approximate (1). Using a gamma-normal mixture representation of the multivariate t model (Pinheiro et al., 2001), the marginal distribution of $\mathbf{y}_i$ is multivariate t-distribution with df ν, given by

$$\mathbf{y}_i \overset{ind}{\sim} t_{n_i}(\mathbf{B}_i\boldsymbol{\beta},\ \mathbf{R}_i\Gamma\mathbf{R}_i^T + \Lambda_i,\ \nu),\ i = 1, \ldots, M_g, \tag{4}$$

given the conditional independence of $\boldsymbol{\gamma}_i|\tau_i \overset{ind}{\sim} N_q(\mathbf{0}, \Gamma/\tau_i)$ and $\boldsymbol{\epsilon}_i|\gamma_i, \tau_i \overset{ind}{\sim} N_{n_i}(\mathbf{0}, \Lambda_i/\tau_i)$, with $\tau_i \overset{ind}{\sim}$ Gamma $(\nu/2, \nu/2)$. This result enables incorporation of the EM algorithm considering that marginal distribution of $\mathbf{y}_i$ with general non-normal assumptions on $\boldsymbol{\gamma}_i$ and $\boldsymbol{\epsilon}_i$ does not have closed form likelihood expression.

Whereas Dai et al. (2017) approximate density ratios (1) by the projection of functional data into a sequence of eigenfunctions that are common to the groups, our method fits discretized measurements over fine grids using basis splines. By doing so, it can be readily applied to general structures including irregularly sampled data and classification of multiple curves from the same subject at the same time as will be discussed in the next section.

The proposed classifier may use high-dimensional information for density modeling, but we regularize the functional curves by the truncation inherent in the knot selection on spline functions. Possible alternatives are roughness penalties (L_2 penalization) or sparsity conditions (L_1 penalization); see Morris (2015) for more details about regularization. Throughout this paper, we employ B-spline bases, which have been extensively used for nonparametric modeling, e.g., Rice and Wu (2001), James (2002) and Berhane and Molitor (2008). Compared to penalizing a high-dimensional approximation or using kernel smoothing, it is computationally efficient and fast, which is crucial for application on real-time analysis, for example, tumor margin assessment during the surgery, as mentioned in the introduction. The choice of optimal number and positions of knots is described in Section 2.3.

2.2. Classification based on correlated functions

In practice, medical or biological data are often collected with repetition from distinct subjects (or clusters) which induces correlation between curves. For example, in QUS, it is common to obtain multiple scans from each region of interest (Wirtzfeld et al., 2015). To account for dependency structure among curves, we extend (3) to include cluster-level random function with coefficients assumed to follow t-distribution with the same df ν in (3). Let $\mathbf{y}_{ij}$ be the n_{ij}-dimensional response vector from jth repetition of ith subject (or cluster). If each subject includes more than two repetitions from the same group, the multi-level RMM can be written as

$$\begin{aligned} &\mathbf{y}_{ij} = \mathbf{B}_{ij}\boldsymbol{\beta} + \mathbf{R}_{ij}\boldsymbol{\gamma}_{ij} + \mathbf{D}_{ij}\boldsymbol{\delta}_i + \boldsymbol{\epsilon}_{ij}, \quad j = 1, \ldots, m_i,\ i = 1, \ldots, M, \\ &\boldsymbol{\gamma}_{ij} \overset{ind}{\sim} t_q(\mathbf{0}, \Gamma, \nu),\ \boldsymbol{\delta}_i \overset{ind}{\sim} t_r(\mathbf{0}, \Psi, \nu),\ \boldsymbol{\epsilon}_{ij} \overset{ind}{\sim} t_{n_{ij}}(\mathbf{0}, \Lambda_{ij}, \nu), \end{aligned} \tag{5}$$

where $\mathbf{D}_{ij}$ is a $n_{ij} \times r$ spline basis matrix and $\boldsymbol{\delta}_i$ is a random vector on cluster level i. Other terms play the same role as in (3). If distinct measurements from each subject may have different class labels, then we can perform classification as in Section 2.1 using the marginal multivariate t-distribution of each individual, $\mathbf{y}_{ij} \sim t_{n_{ij}}(\mathbf{B}_{ij}\boldsymbol{\beta},\ \mathbf{R}_{ij}\gamma\mathbf{R}_{ij}^T + \mathbf{D}_{ij}\Psi\mathbf{D}_{ij}^T + \Lambda_{ij}, \nu)$.

Focusing on the setting in which all repetitions from each subject should have the same label, as in our QUS motivating example, we can perform cluster-level classification by combining information from the multiple measurements. Let $\mathbf{y}_i^* = [\mathbf{y}_{i1}^T, \ldots, \mathbf{y}_{im_i}^T]^T$, $\mathbf{B}_i^* = [\mathbf{B}_{i1}^T, \ldots, \mathbf{B}_{im_i}^T]^T$, and $\sum_{j=1}^{m_i} n_{ij} = n_i^*$. Given the mutual conditional independence among $\boldsymbol{\delta}_i|\tau_i$, $\boldsymbol{\gamma}_{ij}|\tau_i$ and $\boldsymbol{\epsilon}_i|\tau_i$, with τ_i specified below (4), the marginal distribution of $\mathbf{y}_i^*$ is written analogous to (4) as

$$\mathbf{y}_i^* \overset{ind}{\sim} t_{n_i^*}(\mathbf{B}_i^*\boldsymbol{\beta},\ \text{Diag}(\{\mathbf{R}_{ij}\Gamma\mathbf{R}_{ij}^T + \mathbf{D}_{ij}\Psi\mathbf{D}_{ij}^T + \Lambda_{ij}\}_{j=1}^{m_i}) + \text{Off-Diag}(\{\mathbf{D}_{ij}\Psi\mathbf{D}_{ij'}^T\}_{j,j'=1}^{m_i}),\ \nu), \tag{6}$$

where $\text{Diag}(\{A_jA_j^T\}_{j=1}^J)$ with $(n_j \times k)$ matrix A_j denotes $(\sum_{j=1}^J n_j \times \sum_{j=1}^J n_j)$ block diagonal matrix with diagonals A_{jj}, and we define

$$\text{Off-Diag}(\{A_jA_{j'}^T\}_{j,j'=1}^J) = \begin{bmatrix} \mathbf{0}_{n_1\times n_1} & A_{12} & \ldots & A_{1J} \\ A_{21} & \mathbf{0}_{n_2\times n_2} & \ldots & A_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ A_{J1} & A_{J2} & \ldots & \mathbf{0}_{n_J\times n_J} \end{bmatrix}, \tag{7}$$

where $\mathbf{0}_{n_j\times n_j}$ represents a matrix with zero's and $A_{jj'} = A_jA_{j'}^T$, $j, j' = 1, \ldots, J$. The cluster-level random function contributes to model between-curve correlation through basis approximations, specified as Off-Diag$(\cdot)$ in (6), and it improves the estimation of within-curve covariance structure with additional smoothed low-frequency components.

2.3. Model selection and computation procedures

To estimate posterior probabilities, first, the degrees of B-splines and the numbers and locations of knots should be specified to fit (3). As noted in Hall et al. (2001), it is sensible to employ the same basis for each group when comparing the conditional probabilities, thus we aggregate all training samples from each group for model selection. Originally the knots should be placed at grid points where many data are collected, but we use equally spaced knots by assuming that the ensemble of irregularly sampled curves covers the entire domain although individual curves do not, as in the motivating example. To determine the degree of splines and the number of knots, Rice and Wu (2001) suggested model selection criteria, cross-validated log likelihood, AIC or BIC, and last two are empirically proved successful with less computation compared to cross-validation approach. Another possible criterion is via cross-validated prediction error. Here we use modified BIC (Delattre et al., 2014), derived under general mixed effects framework, defined as below under single-level mixed effects model of (3),

$$\text{BIC} = -2\log L + q\log N + p\log\sum_{i=1}^{M} n_i, \tag{8}$$

where L is the maximized likelihood, q and p represent the dimension of random and fixed effect terms, respectively, with n_i and M denoted below (3).

Note, however, that the unstructured Γ in (3) has $q(q+1)/2$ parameters. Large q may lead to poor prediction error due to over-fitting or unstable parameter estimates with local maxima. Based on our experience, we recommend to set the maximum number of spline basis functions, Q, to compare BIC's, where Q can vary depending on the size of the data. In simulations and real data analyses, classification based on (3) or (5) fitted with larger than six splines gives poor performance. Thus, we set $Q = 6$ in the rest of the paper. As an extreme situation, James and Hastie (2001) considered estimation of the covariance function Γ under very sparse data where each curve has only a few measurements. To stabilize the estimation, the reduced rank mixed effects framework that imposes structural assumption on Γ was proposed. It can be an alternative to avoid practical restriction on q; however, we keep unstructured Γ to take advantage of flexible modeling. Also, we recommend using the same basis functions for $B(\cdot)$ and $R(\cdot)$ (or $D(\cdot)$), which leads better performance in our simulation studies compared to using different splines for each.

Regarding the choice of robust parameter, although ν can be treated as a parameter to be estimated, we recommend fixing ν a priori for computational efficiency because our experiments in Section 3.4 indicate that the performance is robust to the choice of ν provided it is not too large, to avoid Gaussian behavior. Further, the application of Section 4.1 demonstrates that the use of an estimated robustness parameter can lead to inaccurate estimates of the diagnostic probability. For these reasons, we use fixed conservative ν equal to 3 for all classes, not only to avoid computational cost but also to maximize the power of robustness. Furthermore, other simulation studies show that this robust classifier achieves good classification performance even under Gaussian functional data.

We fit (3) for each group with the same fixed robust tuning parameter ν over individuals on each group $g, g = 1, \ldots, G$, assuming a diagonal matrix for error term, $\Lambda_i = \sigma_\lambda^2 I_{n_i\times n_i}$, where $I_{n_i\times n_i}$ denotes identity matrix. Then the estimation of parameters of $\boldsymbol{\theta}_g = (\boldsymbol{\beta}_g, \boldsymbol{\Gamma}_g, \sigma_{\lambda,g}^2)$ is accomplished through efficient algorithm for robust estimation in mixed effects model (Pinheiro et al., 2001) and published package `'heavy'` in R (Osorio, 2018). The parameters for cluster-level mixed model (5) can be estimated similarly.

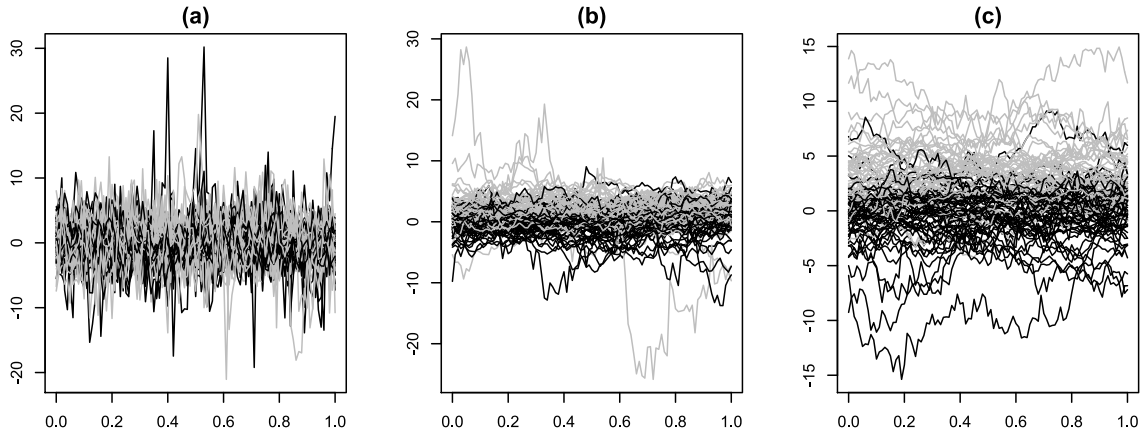

Fig. 3. Simulation results illustrating heavy-tailed functional data under different degrees of spatial dependency: (a) weak, (b) moderate, and (c) strong.

3. Numerical simulation studies

In this section, we evaluate the performance of the proposed robust classifier and compare it with existing approaches via simulation studies. These simulation experiments also included evaluation of computational complexity via the run times for training. The resulting comparisons between the proposed and competing methods are in the online supplementary material, which also provides a part of sensitivity analysis of the robust tuning parameter ν.

3.1. Simulation models

We examine the performances of the competing methods under two major types of outlier generating models. The first type includes heavy tailed stochastic process models with several different levels of within-curve dependency. The resulting simulated data with different levels of spatial dependence display different types of outlier patterns including pointwise (local) outliers, curve-level (global) and partial curve (semi-global) outlier behavior; see Fig. 3 for illustration. The second major type is Gaussian contamination models. Gaussian data are contaminated by a proportion of outlier data generated from a different model. The study includes (i) no contamination, (ii) mild contamination, and (ii) strong contamination. The uncontaminated model provides a comparison of the robust, Gaussian and competing classifiers under ideal circumstances for the Gaussian methods.

We first consider regular functional data, where each function shares a common range, to investigate the robustness and efficiency issues. Then we examine the performance of the various classifiers for irregular, densely sampled data, generated under the heavy-tailed distributional assumption, which mimics the motivating example. Predictive probabilities are evaluated in addition to misclassification rates.

In the simulations of regular functional data, we consider binary classification with 50 discrete response curves over 100 equally spaced grid points on $s_t \in [0, 1]$. In the first scenario, a mean shift model, independent curves are generated from a multivariate t distribution on 3 degrees of freedom with the structure, $y_{gi}(s_t) = \mu_g(s_t) + \epsilon_{gi}(s_t)$, $t = 1, \ldots, 100$, for $i = 1, \ldots, 50$ and $g = 1,2$, where the mean functions within classes are constant with $\mu_1 = 0$, $\mu_2 = \delta$, with $\delta > 0$. We used exponential spatial correlation, $\text{corr}\{\epsilon_{gi}(s_j), \epsilon_{gi}(s_k)\} = \exp(-|s_j - s_k|/d)$, and unit scale. The range parameter d determines the spatial dependence within a curve. The simulations used values for d of 0.01, 0.2 and 1.0, representing low, moderate, and strong dependence, respectively, as illustrated in Fig. 3.

In the second scenario, the discretized data for class g are generated based on an outlier generating mixture model of the form,

$$(1-\alpha)N_{100}(\mu_g, \Sigma) + \alpha N_{100}(\mu_g, \Sigma^c),\ g = 1, 2,$$

where α denotes the proportion of outliers included in the data. Three different levels of contamination are considered, $\alpha = 0,\ 0.1$ and 0.2, representing no contamination, mild contamination, and strong contamination. Here, $N_{100}(\mu_g, \Sigma)$ represents 100×1 multivariate Gaussian distribution with group mean μ_g and exponentially correlated covariance matrix Σ. Moderate within-curve dependency is considered in our experiment. The outlying curves are generated from Gaussian distribution under the same mean function, but different within-curve dependency structure, Σ^c. It is the plausible situation, for example, in motivating example, all collected curves originally contain the same information of target region, but a part of them are contaminated by environmental factors or unexpected noise. We consider various structures of Σ^c by varying the magnitude of noise error, σ_e^c, or dependency structure within a curve, d^c. In our experiment, we set $\sigma_e^c = 10$ and $d = 1$ for covariance matrix of outliers.

Table 1
Average test error and LogLoss for simulated Gaussian functional data.

	RMM	GMM	FPCA	GFLM	FLDA
Error rate	0.14	0.15	0.11	0.14	0.14
LogLoss	0.34	0.39	0.26	0.33	0.32

For the third simulation study under irregular sampling process, we artificially make irregular data under weak dependency structure. Specifically, [0, 1] is divided into three non-overlapped intervals and each one-third of curves are randomly assigned to one of three intervals. The functional values corresponding to the assigned domain are left, and the other two pieces are regarded as missing. The complete functional data are generated over 150 equally spaced grid points, thus 50 measurements remain at each curve.

All simulation results are based on 100 replications. For each replication, the number of basis functions is selected via modified BIC (8) with the restriction of the maximum dimension of splines as 6. We fit fixed and random effect terms in (3) using the same set of basis functions under $\nu = 3$. The performance is evaluated on test set containing 50 curves from each group. All of the simulations to generate multivariate normal and t distributed data were done using the R function `rmvt` in the '`mvtnorm`' package (Genz et al., 2017).

3.2. Competing methods

We compare the proposed classifier RMM with alternative approaches listed below. All of these competitors can provide predictive probabilities and are applicable to irregularly sampled data.

- GMM: the Bayes classifier based on the Gaussian linear mixed model.
- FPC: Functional Principal Component analysis followed by quadratic discriminant analysis (QDA). Hall et al. (2001) performed classification based on resulting coefficient through either kernel density estimation or quadratic discriminant analysis (QDA). We apply FPC technique by James et al. (2000), proposed for sparse functional data, and perform QDA based on obtained coefficients. Classification through kernel density estimation provides similar results. Internally it assumes mixed effects model with B-splines as our model for dimension reduction purpose, but it uses a reduced rank framework, described in Section 2.3, with normally distributed random coefficients. The same number of bases used on RMM model is employed for fair comparison.
- GFLM: Generalized Functional Linear Model (Müller, 2005). For the binary classification, functional logistic regression is fitted with the class label as a univariate response variable and FPC scores estimated in the sparse situation as predictors. It estimates mean, and covariance surfaces from non-parametric local linear smoother under Gaussian assumptions for the case of sparse data. For the classification with 3 or more classes, multinomial logistic regression is used.
- FLDA: Functional Linear Discriminant Analysis for irregularly sample curves (James and Hastie, 2001). It performs LDA by projecting functional curves into the subspace where the between-class covariance relative to within-curve covariance is maximized. Analogous to James et al. (2000), it adopts reduced rank mixed effects model with the Gaussian assumption on random and error terms.

Another promising method in the literature is the functional generalized additive model (McLean et al., 2014; Greven and Scheipl, 2017). It can be used to fit the model for binary class variables with functional predictors. Unlike some other methods, it does not depend on second-order based dimension reduction or parametric assumptions. However, as currently developed function in R is not designed to handle irregularly sampled functional predictors without additional pre-smoothing or imputation, which is beyond the scope of our study, so it is not included here.

Classification performance is evaluated through LogLoss and test error as described in Section 2. In practice, upper and lower bound of probabilities are used in the calculation of LogLoss to avoid infinity result under extreme probabilities. Specifically we set `eps` as `1e-15` and replace p_{ig} in (2) by `min(max(`p_{ig}`,eps), 1-eps)` (Yan, 2016).

3.3. Simulation results

We first consider the performance of all of the methods under the uncontaminated Gaussian model. The methods were compared based on 50 replications from a Gaussian a mean shift model with $\delta = 1.2$ and spatial correlation range $d = 0.2$. The results are summarized in Table 1. All five methods performed comparably, with a slight edge to FPCA in this setting. It is found that even our method with a conservative robust parameter, $\nu = 3$, yields very similar results as the Gaussian method, so there is little cost in performance under ideal Gaussian conditions in using the more robust method.

Next consider the heavy tailed simulation models of Scenario 1. Fig. 4 summarizes the results for RMM, FPCA, GFLM and FLDA. We find that GMM fails in all scenarios with extreme posterior probabilities and enormous LogLoss. Thus, corresponding results under Gaussian assumption are excluded from the visualization for better comparison among the other better performing classifiers. The simulations show that the classifier trained by weak dependency samples achieves

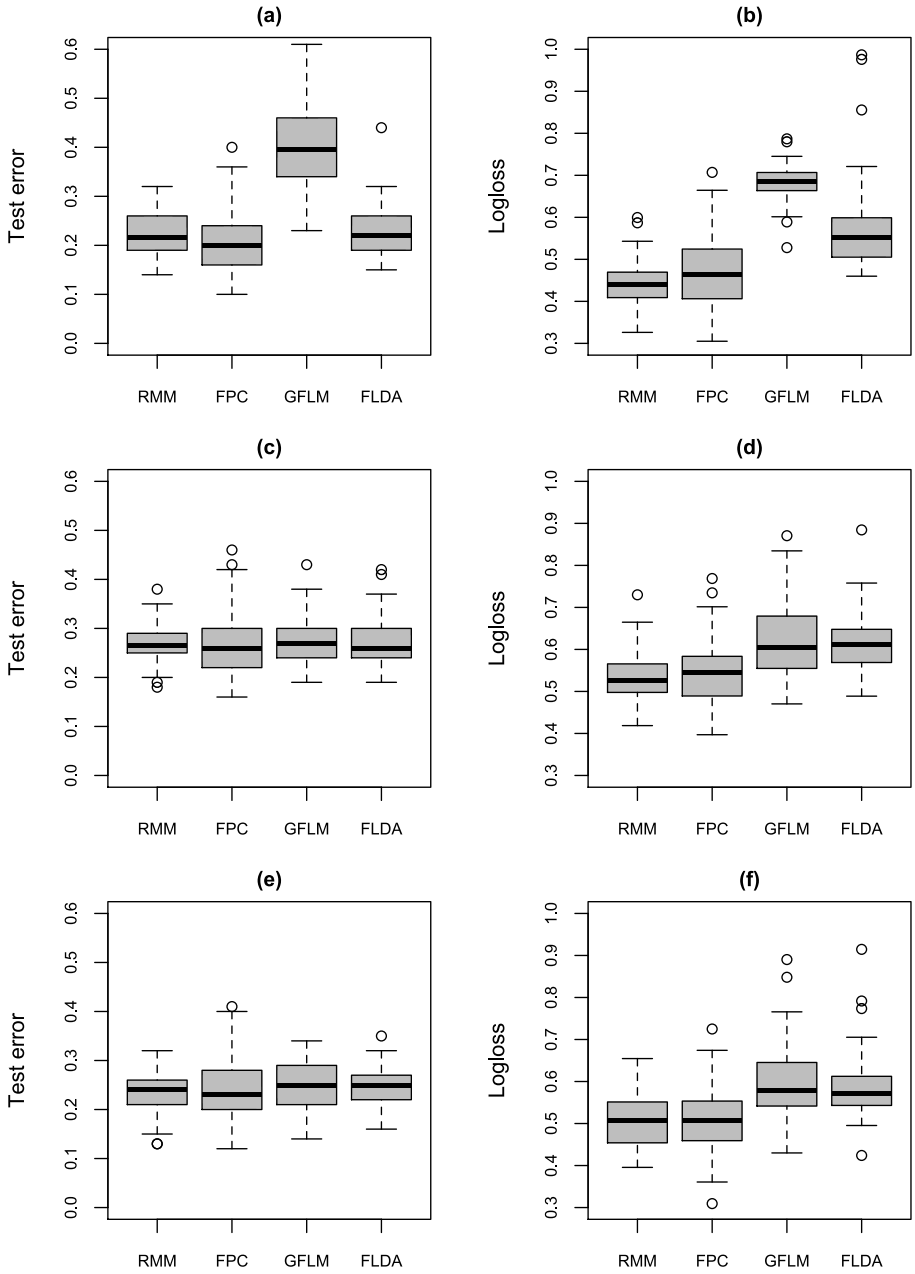

Fig. 4. Boxplots of test error over 50 replications in (a) weak, (c) moderate, and (e) strong within-curve dependency structure; boxplots of LogLoss over 50 replications in (b) weak, (d) moderate, and (f) strong within-curve dependency structure.

better performance from rich information at each grid, thus, δ's are varied with 0.5, 1.5 and 2.5, for each d, respectively, to obtain similar misclassification rates on each run. Other types of group means, such as smoothed functions via cubic B-splines or mean functions having some points of intersection are simulated as well, and they give similar results with no significant effect of the shape of mean functions on classification performance.

The top panel in Fig. 4 shows that GFLM lags behind the other three. It can be inferred that grid-level outliers make local smoothing technique ineffective. The middle and bottom panels illustrate results from simulated data under moderate and strong dependency, respectively. The GFLM and FLDA show larger LogLoss distributions compared to those of RMM and FPC. Regarding test error, all show comparable results. However, FPC displays larger variations in error rates compared to those of other three approaches. We see that RMM is relatively stable on both measures with the smallest variation between replications.

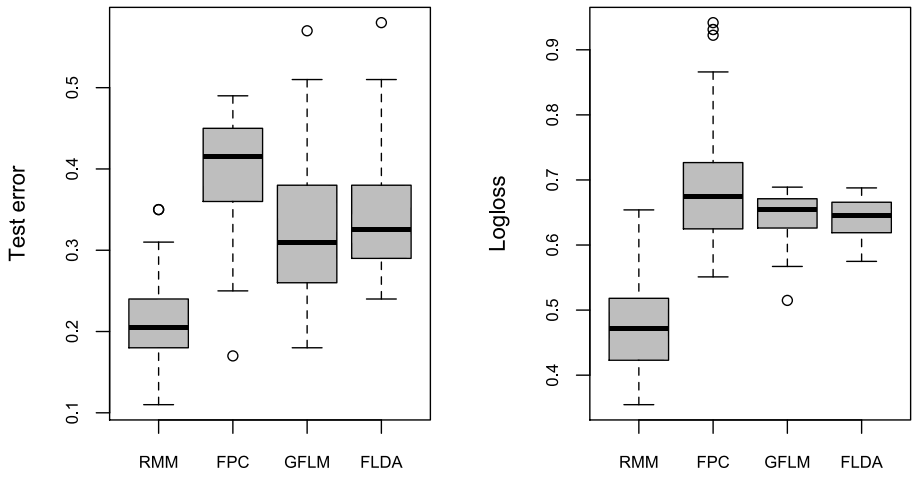

Fig. 5. Boxplots of test error (left) and LogLoss (right) over 50 replications in contaminated data (contamination rate = 20%).

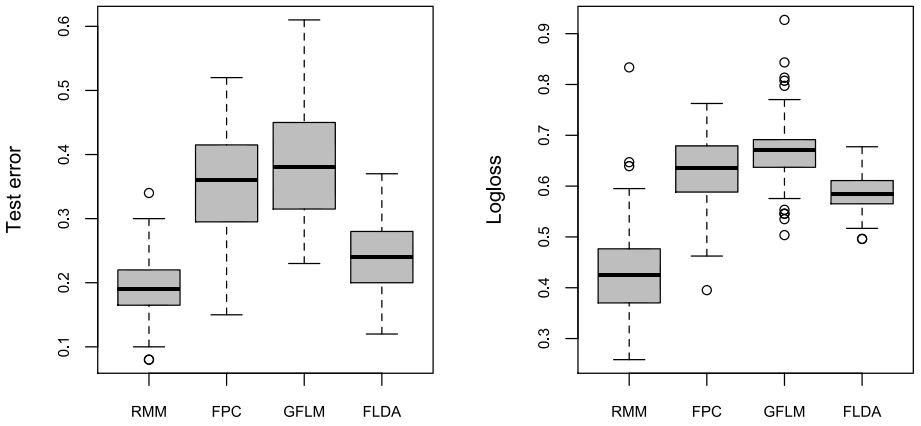

Fig. 6. Boxplots of test error and LogLoss over 50 replications in simulated irregular data under Scenario 3 with heavy tails.

Fig. 5 shows the performance of functional classifiers for regular data with strong contamination. We observe that RMM outperforms all competitors in error rate and LogLoss and result in mild contamination not shown here is very similar to Fig. 5. It implies that our proposed classifier built on heavy-distributional assumption also works well under a mixture of Gaussian distributions. Different parameter settings in Σ or Σ^c provide similar results.

Finally, we simulate heavy tailed irregular data collected over varying intervals of the domain on dense grids. Fig. 6 summarizes the results obtained from weak within-curve dependency and RMM is found to outperform others on both metrics. The FLDA, originally targeting discriminant analysis for sparse data, yields comparable test error, but ours provide more accurate and informative posterior probabilities with smaller LogLoss. Other results under the data with different dependency structures, not reported here, show the similar results.

3.4. Sensitivity analysis on robust tuning parameters

We investigate the behavior of the proposed classifier under different robust tuning parameters. Following heavy tailed models in Section 3.1, we vary distributional assumptions on an error with, t-distribution with df 1 (Cauchy), with df 5, and Gaussian distribution, representing heavy, moderate, and light tailed models, respectively. In each case, 50 simulations datasets were generated, and the RMM posterior probabilities were estimated across a range of values of the tuning parameter ν (3–20).

The graphical summary is provided in the supplementary material. Stable test errors and LogLoss values were observed when ν is small under the simulation sets of Cauchy and t-distribution with df 5. At the same time, as ν increase, LogLoss shows slowly increasing trend under those two sets with heavy tailed behavior. Under Gaussian data, there is no significant trend or noticeable differences among ν on both metrics. We conclude that the use of fixed robust tuning parameter as 3 is

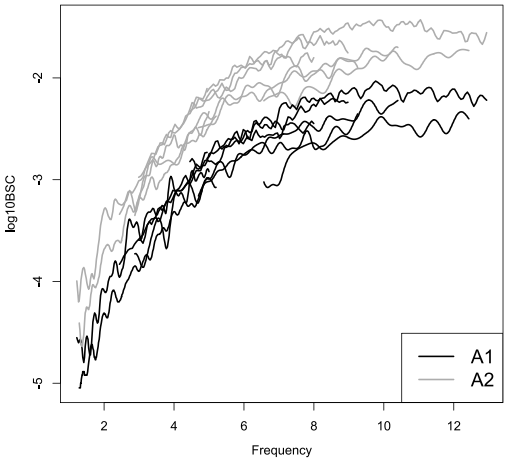

Fig. 7. Phantom data. A1A2.

Table 2
10-fold cross validated error and LogLoss for phantom A1A2 data.

	RMM	GMM	FPCA	GFLM	FLDA
Error rate	0.1	0.2	0.1	0.05	0.2
LogLoss	0.32	5.15	2.22	1.92	1.64

an effective choice that works well regardless of the magnitude of outlyingness in the data. Furthermore, fixing ν makes the classifier computationally efficient for real-time analysis while maintaining robustness and efficiency.

4. Irregular functional data applications

We present several functional data examples in which the functions are irregularly but piece-wise densely sampled. The performance of each method is evaluated through 10-fold cross-validated errors and LogLoss.

4.1. Phantom data

We consider phantom data, namely A1A2. The set of BSC curves are collected by scanning two types of phantom, A1, and A2, multiple times, in 9 laboratories with corresponding transducers which cover different frequency ranges. Two phantoms are embedded in different size of glass beads, thus they are physically and acoustically distinct. But multiple BSC curves from repeated scans for each phantom were averaged at each laboratory, thus, the A1A2 dataset contains only 9 curves for A1 and A2. By doing this, functional noise including outlying pattern might have been considerably diminished. Fig. 7 presents seemingly light tailed and well-separated BSC curves from two phantoms, which can be due to experimental design where target cells with glass beads are directly scanned under controlled environment or due to the use of the averaged curve.

We fit (3) for each group as training step using robust parameter $\nu = 3$ and the same spline functions for fixed and random parts, selected via modified BIC. The likelihood-based probabilistic classifier is derived based on estimated parameters. Also, we apply other competitors including GMM for comparison study. Especially, GFLM is modified by using spline-based FPC score, because original local-linear smoother approach turns out to yield an unstable result when only one (or a few) curve is observed over certain frequencies. We compute 10-fold cross validated error and LogLoss.

Table 2 shows similar cv error rates for all classifiers, but huge differences are found in LogLoss. It is remarkable because similar performances over methods are expected under this seemingly light tailed data. Our finding is that a specific curve in A2 is closely placed near group A1 and all classifiers misclassify this on cross-validation. It implies the situation when new observation has relatively heavy tailed noise compared to the collected information on the training set. Under this plausible situation, our classifier provides false label prediction but with weak confidence, while others provide a very strong degree of certainty on misclassification. The result illustrates the robustness of our proposed classifier to a new outlying observation.

4.2. The mouse and rat mammary tumor data

We now build a robust probabilistic classifier for the mouse and rat mammary tumor data. This experiment noninvasively scanned two types of induced mammary tumors, 4T1 and MAT tumors, from 13 mice and 8 rats using five different

Table 3
10-fold cross validated error and LogLoss for mammary tumor data.

	RMM	m-RMM	FPCA	GFLM	FLDA
Error rate	0.41	0.34	0.41	0.40	0.40
LogLoss	0.71	0.73	0.75	0.67	0.70

Table 4
10-fold cross validated error and LogLoss for original regular (Reg.) and contaminated irregular (Irr.) phoneme example.

		RMM	FPCA	GFLM	FLDA	kNN
Reg.	Error rate	0.10	0.12	0.12	0.57	0.11
	LogLoss	0.25	0.28	0.28	3.96	0.34
Irr.	Error rate	0.26	0.35	0.34	0.55	0.16
	LogLoss	0.70	0.84	0.83	8.03	0.42

transducers which cover a different range of frequency bandwidths. Two transducers, 9L4 and 18L6, from Siemens, cover 3–10.8 MHz, and L14-5 from Ultrasonix uses frequencies 3–8.5 MHz. Meanwhile, MS200 and MS400 from VisualSonics cover higher frequencies, 8.5–13.5 MHz, and 8.5–21.9 MHz, respectively (Wirtzfeld et al., 2015). The subset of data composed of animals having a large tumor (greater than 70 mm^3), 5 mice and 6 rats, is used for the analysis. Each large tumor in an animal is scanned by each transducer, 4 or 5 times, operated by shifting scan lines, accordingly multiple functional curves from each combination of tumor and transducer are correlated. To take into account of this information, we build up a cluster-level classification in addition to classifying each curve separately.

We model (3) for single-level classification which predicts the class of individual curves. For cluster-level classification, we consider simplified model that has only cluster-level random effect term to approximate both between- and within-curve covariance structure, instead of including two random terms as in (5). Fitting a data under (5) can be an alternative. Bayes classifier assigns multiple curves altogether to one of two groups based on joint likelihood.

The classification result is evaluated using 10-fold cross-validated misclassification error and LogLoss. We compare our single-curve based (RMM) and multiple-curve based (m-RMM) with FPCA, GFLM, and FLDA. Again, the optimal number of spline functions is selected via modified BIC (8), and the same selected B-spline bases are used to build up other three competitors for fair comparison. Again, GMM provides poor performance, and its result is not displayed.

Table 3 shows relatively large error rates around 0.4 for all single-curve based classification due to large noise errors. However, our multiple-curve based classifier outperforms others with improved prediction performance by making use of additional information of correlations between curves. The LogLoss is quite similar for all approaches, which means their degree of certainty on false classification is similar.

4.3. Phoneme data

The phoneme data have five classes in 4509 speech frames, each of them corresponding to selected five phonemes with "aa' (695)', "ao (1022)","dcl (757)","iy (1163)" and "sh (872)". Originally log-periodograms are measured at 256 frequencies with none missing. The dataset is available at http://statweb.stanford.edu/~tibs/ElemStatLearn/ and details are found in Hastie et al. (2009). In this paper, we artificially make the data irregular by discarding half of each curve at random. Specifically, we consider two frequency domains [1, 128] and [129, 256], then randomly assign each curve to one of two domains to leave only corresponding functional measures and discard others. Although it has an irregular form, it is a semi-regular set with each half set of samples collected over the same frequency range along with common grid points. Therefore, we can consider another competitor, k-Nearest-Neighbor ensemble method (kNN) by Fuchs et al. (2015), by using curves evaluated over the same frequencies as a training set to predict the group label of a new observation.

Fig. 8 shows a sample of 10 log-periodograms in each phoneme class from artificially contaminated data. Overall curves exhibit heavy-tailed behavior with similar patterns in simulated outlying curves of Fig. 3. Also, some groups, for example "iy" or "ao", show distinct trends in different frequency domains, which needs the potential role of basis functions to unify two trends. The proposed RMM, FPCA, modified GFLM as in Mammary tumor data analysis, FLDA and kNN are applied to both original and contaminated datasets.

Table 4 displays the 10-fold cv error and LogLoss. We see that the FLDA, which was proposed for sparse data, does poorly for these piece-wise dense data. On the other hand, RMM, two FPC based methods, and kNN give similar performances for the original balanced data set. For the induced irregular data, RMM outperforms the FPC based methods in both error rate and LogLoss. At the same time, we observe that the modified kNN outperforms all other methods in both performance measures,

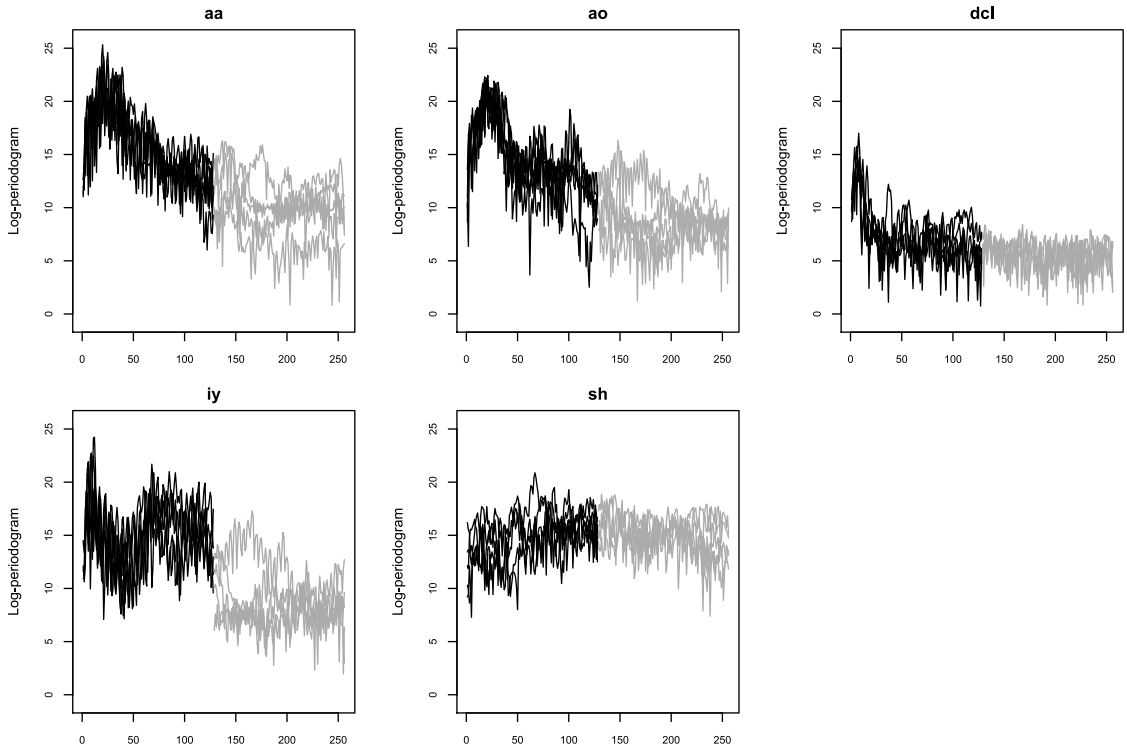

Fig. 8. A sample of 10 contaminated log-periodograms within each phoneme class, 5 black curves on frequency [1 : 128] and 5 gray curves on frequency [129 : 256].

although it is much more computationally expensive. Further work is needed in order to extend this method to less regular data like those in Sections 4.1 and 4.2.

5. Discussion

We have proposed robust probabilistic classifier for functional data, especially including irregularly sampled curves, based on a Bayes classifier under a robust semiparametric mixed effects model framework. The assumption of t-distribution in the mixed model makes classifier robust to outlying curves in terms of membership prediction as well as class probability estimates. In addition to robust prediction error, our proposed method provides honest risk uncertainty estimates especially for observations hard to be classified, which can be useful in medical diagnosis where decreasing misdiagnosis is important. Our proposed method has another advantage in its flexibility and efficiency by employing nonparametric spline functions and unstructured scale matrix of random coefficients as in Rice and Wu (2001). Also, by fitting a multi-level mixed effects model, cluster-level classification is possible where multiple curves from the same subject or cluster are assigned to one of G groups at once. A real data example illustrates improved performance of cluster-level classification by taking into account between correlation information in modeling.

In practice, if each curve is collected over an extremely fine grid, every kth grid points from the original set can be used in the analysis to relieve the computational load. The mammary tumor dataset empirically gives the similar result for the subset of data with $k = 2$ or 4. However, we should be cautious not to choose k as too large, especially when atypical local behavior is detected over a specific area.

Acknowledgments

The authors thank William D. O'Brien, Jr. and the Bioacoustics Research Lab at the University of Illinois at Urbana-Champaign for sharing the phantom data and the mouse and rat mammary tumor data. The research was supported in part by NIH grants R37EB002641, R01CA111289, and R01HD089935.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2018.08.001.

References

Araki, Y., Konishi, S., Kawano, S., Matsui, H., 2009. Functional logistic discrimination via regularized basis expansions. Commun. Stat.–Theory Methods 38 (16–17), 2944–2957. http://dx.doi.org/10.1080/03610920902947246.

Bali, J., Boente, G., Tyler, D., Wang, J.L., 2011. Robust functional principal components: A projection-pursuit approach. Ann. Statist. 39 (6), 2852–2882. http://dx.doi.org/10.1214/11-AOS923.

Berhane, K., Molitor, N.-T., 2008. A Bayesian approach to functional-based multilevel modeling of longitudinal data: Applications to environmental epidemiology. Biostatistics 9 (4), 686–699. http://dx.doi.org/10.1093/biostatistics/kxm059.

Biau, G., Bunea, F., Wegkamp, M., 2005. Functional classification in Hilbert spaces. IEEE Trans. Inform. Theory 51 (6), 2163–2172. http://dx.doi.org/10.1109/TIT.2005.847705.

Biau, G., Cérou, F., Guyader, A., 2010. Rates of convergence of the functional k-nearest neighbor estimate. IEEE Trans. Inform. Theory 56 (4), 2034–2040. http://dx.doi.org/10.1109/TIT.2010.2040857.

Boente, G., Salibian-Barrera, M., 2015. S-estimators for functional principal component. J. Amer. Statist. Assoc. 110 (511), 1100–1111. http://dx.doi.org/10.1080/01621459.2014.946991.

Cérou, F., Guyader, A., 2006. Nearest neighbor classification in infinite dimension. ESAIM Probab. Stat. 10, 340–355. http://dx.doi.org/10.1051/ps:2006014.

Dai, X., Müller, H.-G., Yao, F., 2017. Optimal bayes classifiers for functional data and density ratios. Biometrika 104 (3), 545–560. http://dx.doi.org/10.1093/biomet/asx024.

Delaigle, A., Hall, P., 2013. Classification using censored functional data. J. Amer. Statist. Assoc. 108 (504), 1269–1283. http://dx.doi.org/10.1080/01621459.2013.824893.

Delattre, M., Lavielle, M., Poursat, M.-A., 2014. A note on BIC in mixed-effects models. Electron. J. Stat. 8 (1), 456–475. http://dx.doi.org/10.1214/14-EJS890.

Ferraty, F., Vieu, P., 2003. Curves discrimination: A nonparametric functional approach. Comput. Statist. Data Anal. 44 (1–2), 161–173.

Fuchs, K., Gertheiss, J., Tutz, G., 2015. Nearest neighbor ensembles for functional data with interpretable feature selection. Chemometr. Intell. Lab. Syst. 146, 186–197. http://dx.doi.org/10.1016/j.chemolab.2015.04.019.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T., 2017.mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-6. URL http://CRAN.R-project.org/package=mvtnorm.

Goldsmith, J., Bobb, J., Crainiceanu, C., Caffo, B., Reich, D., 2011. Penalized functional regression. J. Comput. Graph. Statist. 20 (4), 830–851.

Greven, S., Scheipl, F., 2017. A general framework for functional regression modelling. Stat. Model. 17 (1–2), 1–35. http://dx.doi.org/10.1177/1471082X16681317.

Hall, P., Poskitt, D., Presnell, B., 2001. A functional data-analytic approach to signal discrimination. Technometrics 43 (1), 1–9. http://dx.doi.org/10.1198/00401700152404273.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer, New York.

James, G., 2002. Generalized linear models with functional predictors. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (3), 411–432. http://dx.doi.org/10.1111/1467-9868.00342.

James, G., Hastie, T., 2001. Functional linear discriminant analysis for irregularly sampled curves. J. R. Stat. Soc. Ser. B Stat. Methodol. 63 (3), 533–550. http://dx.doi.org/10.1111/1467-9868.00297.

James, G., Hastie, T., Sugar, C., 2000. Principal component models for sparse functional data. Biometrika 87 (3), 587–602. http://dx.doi.org/10.1093/biomet/87.3.587.

McLean, M., Hooker, G., Staicu, A.M., Scheipl, F., Ruppert, D., 2014. Functional generalized additive models. J. Comput. Graph. Statist. 23 (1), 249–269. http://dx.doi.org/10.1080/10618600.2012.729985.

Morris, J., 2015. Functional regression. Annu. Rev. Stat. Appl. 2 (1), 321–359. http://dx.doi.org/10.1146/annurev-statistics-010814-020413.

Müller, H.-G., 2005. Functional modeling and classification of longitudinal data. Scand. J. Stat. 32 (2), 223–240. http://dx.doi.org/10.1111/j.1467-9469.2005.00429.x.

Müller, H.-G., Stadtmüller, U., 2005. Generalized functional linear models. Ann. Statist. 33 (2), 774–805. http://dx.doi.org/10.1214/009053604000001156.

Nolan, R., Adie, S., Marjanovic, M., Chaney, E., South, F., Monroy, G., S., N.D., Erickson-Bhatt, S., Shelton, R., Bower, A., Simpson, D., Cradock, K., Liu, Z., Ray, P., Boppart, S., 2016. Intraoperative optical coherence tomography for assessing human lymph nodes for metastatic cancer. BMC Cancer 16 (1), 144. http://dx.doi.org/10.1186/s12885-016-2194-4.

Osorio, F., 2018.Heavy: Robust Estimation Using Heavy-Tailed Distributions. R package version 0.38.19. URL https://CRAN.R-project.org/package=heavy.

Pinheiro, J., Liu, C., Wu, Y., 2001. Efficient algorithm for robust estimation in linear mixed-effects models using the multivariate t-distribution. J. Comput. Graph. Statist. 10 (2), 249–276. http://dx.doi.org/10.1198/106186001152628059.

Rice, J., Wu, C., 2001. Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 57 (1), 253–259. http://dx.doi.org/10.1111/j.0006-341X.2001.00253.x.

Shi, M., Weiss, R., Taylor, J., 1996. An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. J. R. Stat. Soc. Ser. C. Appl. Stat. 45 (2), 151–163. http://dx.doi.org/10.2307/2986151.

Wang, K., Liang, M., Tian, L., Zhang, X., Li, K., Jiang, T., 2007. Altered functional connectivity in early alzheimer's disease: A resting-state fmri study. Hum. Brain Mapp. 28 (10), 967–978. http://dx.doi.org/10.1002/hbm.20324.

Wirtzfeld, L., Ghoshal, G., Rosado-Mendez, I., Nam, K., Kumar, V., Park, Y., Pawlicki, A., Miller, R., Simpson, D., Zagzebski, J., Oelze, M., Hall, T., OBrien Jr., W., 2015. Quantitative ultrasound comparison of MAT and 4T1 mammary tumors in mice and rats across multiple imaging systems. J. Ultrasound Med. Off. J. Am. Inst. Ultrasound Med. 34 (8), 1373–1383. http://dx.doi.org/10.7863/ultra.34.8.1373.

Wirtzfeld, L., Nam, K., Labyed, Y., Ghoshal, G., Haak, A., Sen-Gupta, E., He, Z., Hirtz, N., M., R.J., Sarwate, S., Simpson, D., Zagzebski, J., Bigelow, T., Oelze, M., Hall, T., OBrien Jr., W., 2013. Techniques and evaluation from a cross-platform imaging comparison of quantitative ultrasound parameters in an in vivo rodent fibroadenoma model. IEEE Trans. Ultrason. Ferroelectr. Freq. Control 60 (7), 1386–1400. http://dx.doi.org/10.1109/TUFFC.2013.2711.

Yan, Y., 2016.MLmetrics: Machine learning evaluation metrics. R package version 1.1.1. URL https://CRAN.R-project.org/package=MLmetrics.

Zhu, H., Brown, P., Morris, J., 2012. Robust classification of functional and quantitative image data using functional mixed models. Biometrics 68 (4), 1260–1268. http://dx.doi.org/10.1111/j.1541-0420.2012.01765.x.