EFFECT SIZE ESTIMATION AND ROBUST CLASSIFICATION FOR IRREGULARLY SAMPLED FUNCTIONAL DATA

BY

YEONJOO PARK

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Douglas Simpson, Chair
Professor Jeffrey Douglas
Associate Professor Feng Liang
Professor Xiaofeng Shao

Abstract

Functional data arise frequently in numerous scientific fields with the development of modern technology. Accordingly, functional data analysis to extract information on curves or functions is an important area for investigation. In this thesis, we address two key issues: measuring an effect size of variable of the interest in functional analysis of variance (fANOVA) model and the development of robust probabilistic classifier in functional response model. We especially consider irregular functional data in our study, where curves are collected over varying or non-overlapping intervals.

First, we develop an approach to quantify the effect size on functional data, perform functional ANOVA hypothesis test, and conduct power analysis. We develop an approach to quantify the effect size on functional data, perform functional ANOVA hypothesis test, and conduct power analysis. We introduce the functional signal-to-noise ratio ($fSNR$), visualize the magnitude of effects over the interval of interest, and perform bootstrapped inferences. It can be applicable when the individual curves are sampled at irregularly spaced points or collected over varying intervals. The proposed methods are applied in the analysis of functional data from inter-laboratory quantitative ultrasound measurements, and in a reanalysis of Canadian weather data. Moreover, we represent the asymptotic power of functional ANOVA test as a function of proposed measure. The agreement between the asymptotic and empirical results is examined and found to be quite good even for small sample sizes. The asymptotic lower bound of power can be reasonably used to determine sample size in planning experimental design.

Secondly, we build a robust probabilistic classifier for functional data, which predicts the membership for given input as well as provides informative posterior probability distribution over a set of classes. This method combines Bayes formula and semiparametric mixed effects model with robust tuning parameter. We aim to make the method robust to outlying curves especially in providing robust degree of certainty in prediction, which is crucial in medical diagnosis. It can be applicable to various practical structures, such as unequally and sparsely collected samples or repeatedly measured curves retaining between-curve correlation, with very flexible spatial covariance function. As an illustration we conduct simulation studies to investigate the sensitivity behaviors of probability estimates to outlying curves under Gaussian assumption and compare our proposed classifier with other functional classification approaches. The performance is evaluated by imposing more penalty for being confident but false prediction. The value of the proposed approach hinges on its simple, flexible, and computational efficiency. We illustrate the issues and methodology in ultrasound quantitative ultrasound, backscatter coefficient vs. frequency functional data, commonly obtained as irregular form and public dataset with artificial contamination. We also show how to implement proposed classifier in R.

This dissertation is dedicated to

my parents, for yours steadfast support and love,

and my husband, for being an amazing partner and friend

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

With continual developments in instrumentation and advanced computing, there has been an increasing need for modeling and analyzing functional data that are collected nearly continuously over fine grids or regions of interest. In accordance with this growth, there is an extensive literature on methods for functional data analysis (Ramsay and Silverman, 2005). Less well developed, to our knowledge, are studies on irregularly sampled curves which are collected on varying or non-overlapped intervals. Our studies focus on the analysis of irregularly sampled functional data, especially the estimation of effect size to quantify the magnitude of the relationship between functional response and variable of interest in fANOVA and construction of robust classifier providing degree of certainty for diagnosis purpose.

This research is motivated by quantitative ultrasound (QUS) data which aim to extract diagnostically useful information from the ultrasound radio frequency signals, in particular the backscatter (BSC) and attenuation properties of the scanned material along different scan lines (Wirtzfeld *et al.,* 2013). Wirtzfeld *et al.* (2015) presented data and results from diagnostic ultrasound studies using multiple transducers to scan mammary tumors and fibroadenomas (benign fibrous masses) in rats and mice. The frequency dependent BSC curves derived from the power spectra of these scans took the form of functional measurements spanning the frequency range of the ultrasound transducer.

Due to the noninvasive nature of ultrasound imaging, diagnosis via ultrasound is widely used in medical applications. However, for BSC measurements to translate to the clinic, the detection of differences in the features of the BSC curves for different tumors needs to be statistically assessed. Beyond statistical inference, variation due to tumor effect needs to be compared with variation due to background noise to study the precision in diagnosis. In such settings it is worth examining tumor effect sizes on BSC curves both locally and globally. For example, clinicians might be interested in the overall effect size over the frequency range of interest to make sure about the precision in diagnosis, otherwise, the pointwise effect size evaluated at each grid is of interest for researchers to develop the measurement system achieving the most effective separation. Furthermore, by examining confidence intervals of local effect size, we can infer whether the change of effect size over frequencies is statistically significant or not.

As a next stage, if statistically distinct behaviors in BSC functions over different types of tumor are proved, then an immediate question we may have correspondingly is, whether the functional data classification method can diagnose the future observations into the correct classes. Especially providing stable and informative posterior probabilities to be assigned to each class is of interest in terms of diagnostic purpose.

However motivating data has a challenging structure as in Figure 1.1. It is irregularly collected over frequencies, seemingly heavy-tail distributed with large noise and has dependence structure between multiple curves. In experiment, functional BSC are collected in several laboratories using different transducers covering different ranges of frequency, scanning the target tumor in living animal multiple times. Thus curves retaining between-curve correlation have varying grid points and intervals. In addition, noninvasive scan causes potential outlying behaviors suffered from unexpected contamination by scanning neighboring other tissues or noise in environment.

In Chapter 2, we introduce functional Signal-to-Noise Ratio ($fSNR$) to measure

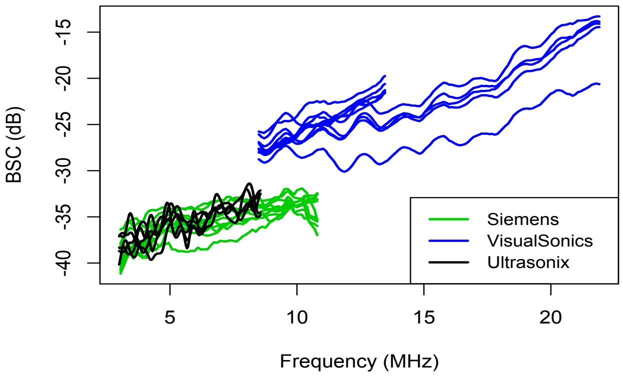


Figure 1.1: Backscatter functions for one of the scanned tumors

local effect size along the entire functional domain. It provides not only graphical visualization but also valuable information about which ranges have the largest effect size. Secondly, we define a globalized effect size that summarizes effect size over region of interest. Third, we represent the asymptotic power of fANOVA test as a function of proposed global measure. The agreement between the asymptotic and empirical results is examined via simulation studies under different scenarios and found to be quite good even for small sample sizes. This agreement and asymptotic lower bound of power enable to derive sample size estimation tool for planning experimental design.

In Chapter 3, we build a robust probabilistic classifier for functional grouped data, which provides a predicted class label as well as a probability distribution over a set of classes. It is based on spline based mixed-model with robust tuning parameter and Bayes rule, and especially mixed effects model approach enables to approximate covariance function in a flexible and efficient way. The key of our method is to impose heavy-tail distribution assumption with robustness parameter ν on random coefficients to yield robust result. We focus on the evaluation of functional data classifier in terms of accuracy for predicted posterior probability to be assigned to the correct class.

In Chapter 4, we sketch the foundation of asymptotic analysis for unbalanced functional data, for large sample inference and theoretic basis for use of the bootstrap.

Chapter 2

Effect Size and Power Analysis for Functional ANOVA

2.1 Introduction

Functional data in which the response measurements consist of functions observed continuously over a fine grid occur in many different fields more often in recent years. In accordance with this growth, there is an extensive literature on methods for functional data analysis. Ramsay and Silverman (2005) provide a comprehensive treatment. Also a number of authors have developed global testing and inference for k-group functional response data including the functional analysis of variance (fANOVA) methods of Cuevas *et al.* (2004), Shen and Faraway (2004), and Zhang and Liang (2014).

Less well developed, to our knowledge, are studies on the estimation of effect size to quantify the magnitude of the relationship between functional response and variable of interest. Indeed, characterizing an effect size is prominent in practical studies, because research findings can be clearly presented by this measure. It also facilities interpretation and performs a fair comparison among variables due to its robustness in scale and measurement units. Additionally, effect size is closely related to statistical power of a hypothesis test, which can be used for sample size determination in experimental design or for interpretation of test result.

In related work Yao, Muller and Wang (2005a) proposed the coefficient of determination in functional linear regression to define a global measure of the association.

They proposed two types of functional R^2 by integrating the pointwise $R^2(s)$ over the domain s and by integrating the numerator and the denominator separately. Those measures estimate global effect size over s, however, further statistical inference based on them was not concerned. Partial R^2 proposed by Edwards *et al.* (2008) measures such magnitude in mixed effect structure, especially for the longitudinal linear mixed model. However, some restrictive parametric assumptions are required and the connection between an effect size and statistical inference received less attention in the study.

We propose a general approach to estimate the effect size of the variable of interest and further analysis for the functional data. While many developed methodologies are restricted to regular structure where curves are collected over common grids and interval, our proposed analysis can be applicable to irregular structure where collections of curves are unequally and sparsely sampled over varying intervals. In this paper, the measures quantifying local and global effects of the variable are developed. A key idea is to extend the signal-to-noise ratio (SNR), a widely used measure in engineering defined as the ratio of the variance of the target signal to the variance of noise, to functional structure. Indeed, closely related concepts have developed in the statistics literature as well, for example, noncentrality parameter of the F-statistic in ANOVA. It is often used as a measure of effect size or as a planning tool in power analysis. The extensions developed in the present paper are designed to provide analogous types of analysis for functional response data. Specifically, the use of estimated global measure as an inferential statistic to detect significant effect over the domain is studied. We discuss the use of the proposed local measure for visualization and derivation of confidence intervals to find which parts of the function domain are most informative.

Much functional data analysis research have been centered around data analysis, with little attention paid to power analysis under finite-sample. Although Zhang (2011) and Zhang and Liang (2014) studied asymptotic powers of functional F-type tests in fANOVA model, the main goal was to show root-n consistency of tests. Shen and Faraway (2004) estimated power and size of test via simulation studies, but the purpose was to compare existing test methods. In this paper, we represent asymptotically approximated statistical power as a function of proposed effect size and study the agreement between the asymptotic and empirical powers under finite sample sizes. We also derive asymptotic lower bounds of the power of fANOVA tests based on the proposed effect size. The accuracy of the asymptotic approximation is found to be good for moderate sample sizes, which implies approximated lower bounds can be used for sample size determination. It enables sample size estimation in the design of experiment.

The rest of the paper is organized as follows. In Section 2, we introduce the functional SNR to measure the local effect size and extend it to fANOVA model. We define global measure over the interval of interest and present estimation of proposed measures. Also statistical inferences based on local and global effect sizes are introduced. In Section 3, the agreement between the asymptotic and empirical powers under finite sample size are investigated on various scenarios. Also it provides asymptotic approximation of lower bounds of power as a tool for sample size determination and its numerical implementation. We return to quantitative ultrasound data in Section 4 with an application to real data. The Canadian weather data example illustrates the usefulness of the proposed methods and it is relegated to the supplementary file. Discussions and concluding remarks are in Section 4.

2.2 Functional Signal-to-Noise Ratio

Let $y(s)$ denote functional response over a given interval of interest $\mathcal{S}$. The individual curve can be decomposed into the systematic signal and random noise components,

$$y(s) = \mu(s) + \epsilon(s), \quad s \in \mathcal{S}, \tag{2.1}$$

where $\mu(s)$ is a functional mean and $\epsilon(s)$ is a stochastic process with mean zero and covariance function $\gamma(s,t),\ s,t \in \mathcal{S}$. If $\gamma(s,t)$ is strictly positive definite and $\int_s \gamma(s,s)ds < \infty$, the spectral decomposition of $\gamma(s,t)$ leads to $\gamma(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t)$, where $\lambda_j \geq 0$ are the eigenvalues in descending order and $\phi_j(s)$ the corresponding orthonormal eigenfunctions. Letting $\sigma^2(s) = \gamma(s,s)$, it will be assumed that $\mu(\cdot)$ and $\sigma(\cdot)^{-1}$ are continuous, Riemann square-integrable functions on $\mathcal{S}$ so that integrals over the continuous domain can be approximated by summations over a fine grid.

Within this framework we focus on measuring the deviation of $\mu(\cdot)$ from a functional null space, Θ_0, of no effect, in comparison to the noise level. Thus we define the functional signal-to-noise ratio ($fSNR$),

$$fSNR(s) = \sqrt{\{|\mu(s) - \mu_0(s)|/\sigma(s)\}^2},\ s \in \mathcal{S}, \tag{2.2}$$

where

$$\mu_0(\cdot) = \underset{\eta \in \Theta_0}{\arg\min}\ \|\eta(\cdot) - \mu(\cdot)\|, \tag{2.3}$$

for an appropriate norm $\|\cdot\|$ such as a weighted L_2 norm over $\mathcal{S}$.

If the no-effect hypothesis implied by Θ_0 imposes only point-wise constraints, for example, $\Theta_0 = \{\mu : \mu(s_t) = \mu_0(s_t)$, for a known fixed function μ_0, where $\{s_t; t = 1, ..., T\} \in \mathcal{S}\}$, then $fSNR(s)$ is the function of pointwise signal-to-noise ratios. On the other hand, if Θ_0 imposes constraints defined across s, then $fSNR(s)$ is not necessarily equal to the pointwise signal to noise ratio. Consider, for example, $\Theta_0 = \{\mu : \mu(s) = c,\ s \in [a, b],\ c \text{ unspecified}\}$, and suppose we measure the distance from Θ_0 using the norm $\|f\|_\sigma = \{\frac{1}{(b-a)} \int_a^b f^2(s)/\sigma^2(s)ds\}^{\frac{1}{2}}$. The solution $\mu_0(s)$ is a constant function equal to the weighted mean of $\mu(s)$ over $s \in [a, b]$, and is given by $\bar{\mu}_\sigma := \int\{\mu(s)/\sigma^2(s)\}ds/\int\{1/\sigma^2(s)\}ds$. Another example is a smoothing constraint, such as $\Theta_0 = \{\mu : \int(\mu'')^2(s)ds < c,\ s \in [a, b]\}$. Each of these cases, μ_0 is jointly specified over s rather than pointwise.

In the remainder of this article, we focus on settings in which the no-effect hypothesis can be specified pointwise, which is the case in our motivating application.

2.2.1 Missing data framework for irregular functional data

In this section, we construct a missing data interval sampling framework for irregularly collected data motivated by our collaborative research (Wirtzfeld et al. 2015); cf. Figure 1.1. We differentiate two stochastic processes; the complete random process $y^c(s)$ on $\mathcal{S}$, and the observed incomplete random process $y(s)$ denoting $y^c(s)$ observed only on a random sub-interval in $\mathcal{S}$. The irregular functional data can be understood to be a collection of realizations of $y(s)$.

Let $y_i^c(s),\ i = 1, ..., n$ denote the complete-data random functions defined over the full range $\mathcal{S} = [a, b]$, and let $I_i,\ i = 1, ..., n$ denote random intervals in $\mathcal{S}$. Let $SP(\mu, \gamma)$ denote a stochastic process with mean function $\mu(s),\ s \in \mathcal{S}$ and covariance function $\gamma(s, t),\ s, t \in \mathcal{S}$, and let $[L, U]$ represent random interval with random lower and upper bounds satisfying $P(L < U) = 1$. We consider the following model assumptions:

$$\begin{cases} y_1^c(s), ..., y_n^c(s) \overset{i.i.d.}{\sim} SP(\mu, \gamma), \\ L_i \overset{i.i.d.}{\sim} F_L,\ U_i \overset{i.i.d.}{\sim} F_U \quad \text{with} \quad P([L_i, U_i] \subset \mathcal{S}) = 1, \\ \inf_{s \in \mathcal{S}} P(s \in [L_i, U_i]) > 0, \end{cases} \tag{2.4}$$

for $i = 1, \ldots, n$. Then $y_i(s) = y_i^c(s)\mathbb{1}_{[L_i,U_i]}(s)$ for $s \in [L_i, U_i]$ and is undefined elsewhere. Let $y_i(s),\ i = 1, ..., n$ denote random functional samples under this framework. Then for each $s \in \mathcal{S}$, the weak law of large numbers and the continuous mapping theorem imply that

$$\bar{y}(s) = \frac{\sum_{i=1}^n y_i^c(s)\mathbb{1}_{[L_i,U_i]}(s)}{\sum_{i=1}^n \mathbb{1}_{[L_i,U_i]}(s)} \xrightarrow{p} \mu(s),$$

$$\hat{\sigma}^2(s) = \frac{\sum_{i=1}^n (y_i^c(s) - \bar{y}(s))^2 \mathbb{1}_{[L_i,U_i]}(s)}{\sum_{i=1}^n \mathbb{1}_{[L_i,U_i]}(s) - 1} \xrightarrow{p} \sigma^2(s). \tag{2.5}$$

See supplementary materials for proof. For each s, sample mean and variance converge to mean and variance of $y^c(s)$. This result can be used to estimate consistent pointwise effect size for each s in section 2.3.

As an example, suppose $L = \min(V_1, V_2)$, $U = \max(V_1,\ V_2)$ with $V_h \overset{i.i.d.}{\sim} F_V,\ h = 1, 2$. The coverage probability is positive and bounded away from zero with random variable V defined on $\mathcal{S}$ satisfying $\inf_{s \in \mathcal{S}} F_V(s)\{1 - F_V(s)\} > 0$. By doing so, each s has rich information as sample n increases and the unobserved parts of each individual curve on $\mathcal{S}$ can be understood as Missing Completely at Random (MCAR).

2.2.2 Application to functional ANOVA Model

We now consider a functional ANOVA model with the goal of measuring the effect size of the grouping variable. Let $y_g(s),\ s \in \mathcal{S}$, denote the functional response data and $g = 1, ..., k$, be a group factor. The individual curve can be decomposed into overall mean, group mean and noise parts similar to ANOVA model as follows.

$$y_g(s) = \mu_0(s) + \beta_g(s) + \epsilon(s), \quad s \in \mathcal{S}, \tag{2.6}$$

where $\mu_0(s)$ is a group-independent mean function, $\beta_g(s)$ represents the group dependent effect with constraint $\sum_g n_g \beta_g(s) = 0$, and $\epsilon(s)$ denotes a stochastic process in (2.1). Under the null hypothesis of no group-effect, $\beta_g(s) = 0,\ g = 1, ..., k$. We measure functional deviations from the null by extending $fSNR$ to fANOVA model as,

$$fSNR(s) = \sqrt{WAVE\{|\beta_g(s)/\sigma(s)|^2\}}, \quad s \in \mathcal{S}, \tag{2.7}$$

where $WAVE\{x_g\} := N^{-1} \sum_g n_g x_g$ denotes the weighted average. Here $n_g,\ g = 1, ..., k$, denote the number of curves in each group and $N = \sum_g n_g$.

In order to develop test statistic as well as global measure of effect size over the interval of interest, we may summarize the effect size using various functions of $fSNR$. Assuming that both μ and σ^{-1} are square-integrable functions on $\mathcal{S}$, we define the summary measure as,

$$G_{fSNR} = \|fSNR(\cdot)\|, \tag{2.8}$$

where $\|f(\cdot)\| := \{\frac{1}{|\mathcal{S}|} \int_{\mathcal{S}} f^2(s)\ ds\}^{1/2}$ and $|\mathcal{S}|$ is the length of an interval. We can measure more refined effects by calculating values from subintervals of $\mathcal{S}$. Another way, not considered here, is to replace the L_2 norm by a sup norm.

Alternatively, if homoscedasticity is assumed over $\mathcal{S}$, we define the second type of global measure as,

$$G^*_{fSNR} = \frac{1}{\|\sigma(\cdot)\|}\sqrt{WAVE\{\|\beta_g(\cdot)\|^2\}}, \tag{2.9}$$

It compares the norm of the functional deviation in mean curve with the norm of the standard deviation curve.

2.2.3 Pointwise and smoothed estimates of $fSNR$

Let $y_{gi}(s) = y^c_{gi}\mathbb{1}_{[U_{gi},L_{gi}]}(s),\ g = 1, ..., k,\ i = 1, ..., n_g$ be observed functional data under irregular sampling framework. In order not to overload notation, throughout this paper we will write $n_g(s) = \sum_{i=1}^{n_g}\mathbb{1}_{[U_{gi},L_{gi}]}(s),\ g = 1, ..., k,$ and $N(s) = \sum_{g=1}^{k}\sum_{i=1}^{n_g}\mathbb{1}_{[U_{gi},L_{gi}]}(s)$. And we keep n_g and N to denote the number of curves in group g and total number of samples over G groups, respectively. Then the consistent and unbiased estimator of $fSNR$ can be derived via F-statistic *function* as,

$$f\hat{SNR}^2(s) = \begin{cases} (k-1)(F(s)-1)/N(s) & \text{if } F(s) \geq 1, \\ 0 & \text{elsewhere} \end{cases} \tag{2.10}$$

where $F(s) = MSB(s)/MSW(s)$, with $MSB(s) = \sum_{g=1}^{k} n_g(s)\{\bar{y}_{g\cdot}(s) - \bar{y}_{\cdot\cdot}(s)\}^2/(k-1)$ and $MSW(s) = \sum_{g=1}^{k}\sum_{i=1}^{n_g(s)}\{y_{gi}(s) - \bar{y}_{g\cdot}(s)\}^2/(N(s)-k)$ denote the functions of weighted mean square deviations between groups and the mean square deviations within groups for each s. Here $\bar{y}_{g\cdot}(s)$ and $\bar{y}_{\cdot\cdot}(s)$ are group mean and overall mean curves averaged over $n_g(s)$ and $N(s)$ for each s. It is derived analogous to group-effect size estimation in Wirtzfeld *et al.* (2013) and indeed, if $F(s) \geq 1$, it is a consistent estimator of $fSNR^2(s)$ by (2.5) as bias correction term $(k-1)/N(s) \xrightarrow{p} 0$. If $F(s) < 1$, we will replace $f\hat{SNR}^2(s)$ by 0. In practice, functional curves are recorded over finite number of grid points rather than being observed continuously. Suppose that we observe $y_{gi}(s_{git}),\ g = 1, ..., k,\ i = 1, ..., n_g,\ t = 1, ..., T_{gi}$, in a discretized fashion with

$\cup_{g,i}\{s_{gi1}, ..., s_{giT_{gi}}\} = \{s_1 < s_2 < ... < s_T\} \in \mathcal{S}$. Then discretized $fSNR$ is estimated for each $s_t,\ t = 1, ..., T$.

This approach can track the data better in the pointwise aspect, however, it ignores the within-curve dependence, such as temporal or spatial nature. Especially when each functional sample is not smooth enough due to big noise or being collected over irregular grids under moderate sample size, estimated $fSNR$ might have unrealistic jumps or rapid oscillation within a short interval that leads to hard interpretation. It is therefore plausible to assume smoothness of $fSNR$ and obtain it by estimating mean and deterministic standard deviation functions via smoothing. For example, a natural approach is to regress $y_{gi}(s_t)$ on $s_t,\ t = 1, ..., T$ non-parametrically using kernel or spline smoothing.

The nonparametric regression allows to estimate smooth $fSNR$ through regularization and replication by borrowing strength from nearby observations within as well as between functions. Via one of the smoothing techniques, such as cubic B-splines, smoothing splines (Wahba, 1990) and local polynomial smoothing (Wand and Jones, 1995), $MSB(s)$ and $MSW(s)$ can be replaced by $MSB^s(s)$ and $MSW^s(s)$, where $MSB^s(s) = \sum_{g=1}^{k} n_g\{\hat{\mu}_{g\cdot}(s) - \hat{\mu}_{\cdot\cdot}(s)\}^2/(k-1)$ and $MSW^s(s)$ is the smoothed mean square deviation curve within groups. Here $\hat{\mu}_{g\cdot}(s)$ and $\hat{\mu}_{\cdot\cdot}$ are smoothed group and overall mean functions and $MSW^s(s)$ is a smoothed regression line fitted from square of residuals $r_{gi}^2(s_t) = \{y_{gi}(s_t) - \hat{\mu}_{g\cdot}(s_t)\}^2,\ g = 1, .., k,\ i = 1, ..., n_g,\ t = 1, ..., T_{gi}$. Then we choose an equally spaced grid of m points in $\mathcal{S}$ to calculate the ratio. The use of absolute value of $r_{gi}(s_t)$ to fit the smoothed marginal error curve and replacing denominator by square of it is another possible approach, but experimental studies show that it underestimates the scale of deviation. Details about various nonparametric smoothing techniques can be found in Zhang and Liang (2014, section 2.4). Note that we use a unified modeling approach that estimates group effect and reflects inherent smooth structure simultaneously. It is different from two-step approach in Shen and Faraway (2004) and Zhang and Liang (2014) where reconstructed individual curves via smoothing are used to get the estimated mean functions. However, under irregular data frame two-step approach may lead unreliable reconstruction especially when fitting missing parts. Morris (2015) reviews cases with evidence of benefits of unified modeling.

Among various smoothing methods, we employ the natural cubic splines with equally spaced L interior knots in the rest of this paper. This method is not only easy to be implemented but relieves edge effect by adding constraints beyond the boundary knots (Hastie *et al.*, section 5.2.1, 2009) so that result may stable even when sample size is not large enough. The optimal number of knots is selected via Bayesian information criterion (BIC) which is empirically proven to perform well under irregularly sampled functional structure (Rice and Wu, 2001). Other model selection techniques, Akaike information criterion (AIC) or cross-validation, can be another possibility.

To estimate proposed global measures, two types of functional F-test statistics can be extended and used. (Shen and Faraway, 2004, Cuevas *et al.*, 2004, and Zhang and Liang, 2014) Those statistics are originally proposed for regular structure and developed according to how mean squared functions are integrated over under regular structure. Firstly, we define $\mathcal{F}$ as the integration of F-statistic $function$ over the interval,

$$\mathcal{F} = \frac{1}{|\mathcal{S}|} \int \frac{MSB(s)}{MSW(s)}\, ds \approx \frac{1}{T} \sum_{t=1}^{T} F(s_t), \tag{2.11}$$

where $MSB(s)$ and $MSW(s)$ in section 2.3. We can also define smoothed version by using $F_s(s) = MSB^s(s)/MSW^s(s)$, and approximate the integration as $\sum_{t=1}^{m} F_s(s_t)/m$.

The second type, say $\mathcal{F}^*$, is defined as the ratio of two respectively integrated mean sums-of-squares,

$$\mathcal{F}^* = \frac{MSB_{func}}{MSW_{func}} = \frac{\int MSB(s)\ ds/|\mathcal{S}|}{\int MSW(s)\ ds/|\mathcal{S}|} \approx \frac{\sum_{t=1}^{T} MSB(s_t)/T}{\sum_{t=1}^{T} MSW(s_t)/T}, \tag{2.12}$$

similarly we can define smoothed version and approximates it as $\sum_{t=1}^{m}\ MSB^s(s_t)/ \sum_{t=1}^{m} MSW^s(s_t)$.

Then the global group-effect size can be estimated by extending Wirtzfeld *et al.* (2013),

$$\hat{G}_{fSNR} = \sqrt{(k-1)(\mathcal{F}-1)/N}. \tag{2.13}$$

Let $\hat{G}_{fSNR}$ be zero when $\mathcal{F}$ is less than one. If $\mathcal{F} \geq 1$, under regular structure, $\hat{G}_{fSNR} \xrightarrow{p} G_{fSNR}$ via continuous mapping theorem and dominated convergence theorem under certain conditions. The bias correction term $(k-1)/N$ goes to zero as N increases.

Analogously the G^*_{fSNR} can be estimated by replacing $\mathcal{F}$ with $\mathcal{F}^*$. Under The value of using functional F-statistics in estimating effect size hinges on its simple computation. Analogously the G_{fSNR} can be estimated by replacing $\mathcal{F}^*$ with $\mathcal{F}$. The value of using functional F-statistics in estimating effect size hinges on its simple computation.

2.2.4 Large sample approximation and bootstrap testing

Next we consider hypothesis testing based on $fSNR$ statistics for global hypotheses of the form:

$$H_0 : G_{fSNR} = 0 \qquad \text{versus} \qquad H_A : G_{fSNR} > 0$$

or

$$H_0 : G^*_{fSNR} = 0 \qquad \text{versus} \qquad H_A : G^*_{fSNR} > 0.$$

For balanced functional sampling structures, the test via global measure G_{fSNR} is equivalent to the GPF test proposed by Zhang and Liang (2014). Specifically, under condition A and null hypothesis, $\mathcal{F} \overset{d}{=} (k-1)^{-1}\sum_{r=1}^{m} \lambda_r^{\omega} A_r,\ A_r \overset{i.i.d.}{\sim} \chi^2_{k-1}$, where $\gamma_w(s,t) = \gamma(s,t)/\sqrt{\gamma(s,s)\gamma(t,t)}$, and λ_r^w are the decreasing-ordered eigenvalues of $\gamma_w(s,t)$, with associated eigenfunctions $\phi_r^w(s)$. All conditions are reported in the online Appendix. Similarly test with G^*_{fSNR} is corresponding to the F-type test developed by Shen and Faraway (2004) and Cuevas *et al.* (2004). Under condition B and null hypothesis, $\mathcal{F}^* \overset{d}{=} (k-1)^{-1}\sum_{r=1}^{m} \lambda_r A_r,\ Ar \overset{i.i.d.}{\sim} \chi^2_{k-1}$, where λ_r denoted in Section 2.1. Hereinafter we will call hypothesis inference testing null effect of G_{fSNR} and G^*_{fSNR} as $\mathcal{F}$-test and $\mathcal{F}^*$-test, respectively.

Under irregular structures or small sample sizes the aforementioned asymptotic null distributions are not valid. In such cases we rely on bootstrap resampling methods both for global testing and for construction of pointwise confidence intervals of $fSNR(s)$. The latter application is useful for visualization and detecting subintervals that achieve the most effective separation, as illustrated in Section 4 below. For regular functional data in which all curves span the same domain, Cuevas et al. (2006) considered both a generic nonparametric bootstrap and Gaussian parametric bootstrap, finding no distinct advantage for the parametric bootstrap.

We extend the application of the nonparametric functional bootstrap under sampling framework assumption described in section 2.1. The nonparametric bootstrap method is able to yield consistent result under irregular structure with independence assumption between stochastic process and random interval. In practice, medical or biological data are often collected with repetition from distinct subjects or clusters which have different characteristics. Accordingly, functional data may have correlation structure between observed curves from the same subject. In this case, all multiple curves from the same subject should be resampled together when implementing nonparametric bootstrap method, so that correlation between replicates is preserved.

2.3 Power Analysis

In order to perform power analysis for the fANOVA tests via $\mathcal{F}$ and $\mathcal{F}^*$, we obtain approximate power functions under local alternatives along with a simplifying lower bound. As will be demonstrated in a simulation study, these can be used for planning purposes with moderate to large samples. For smaller samples, we provide a simulation-based power analysis to complement the large sample approximations.

We first show the functional dependence of the asymptotic power on the limiting behavior of the effect size measures G_{fSNR} and G^*_{fSNR} of Section 2, and then derive asymptotic lower bounds that simplify calculations. We also investigate the agreement between approximation based and simulation based estimation of power for moderate sample sizes, and demonstrate sample size analysis for target effect sizes.

2.3.1 Asymptotic power approximation

We first obtain the approximate power functions for local alternatives as the overall sample size N increases. Thus we consider sequences of alternatives of the form,

$$H_{1N}: \ \mu_{Ng}(s) = \mu_0(s) + \beta_{Ng}(s), \quad g = 1, \ldots, k, \tag{2.14}$$

where, as N increases,

$$\beta_{Ng}(s) = N^{-1/2}\eta_g(s) \sim a_g n_g^{-1/2}\eta_g(s), \qquad (2.15)$$

where $\lim_{N\to\infty} n_g/N = a_g \in (0,1)$, and the functions $\eta_g(s)$ are non-zero and square-integrable for $g = 1, 2, \ldots, k$ with $\sum_{g=1}^k a_g\eta_g(s) = 0$. Under these conditions the effect sizes decrease at the same $N^{-1/2}$ rate:

$$G_{fSNR} \sim N^{-1/2}G_0 \quad \text{and} \quad G^*_{fSNR} \sim N^{-1/2}G^*_0 \qquad (2.16)$$

where

$$G_0^2 = \sum_{g=1}^k a_g \int_{\mathcal{S}} \frac{\eta_g^2(s)}{\sigma^2(s)}ds \quad \text{and} \quad (G_0^*)^2 = \frac{\sum_{g=1}^k a_g \int_{\mathcal{S}} \eta_g^2(s)ds}{\int_{\mathcal{S}} \sigma^2(s)ds}.$$

Under condition A, the power of $\mathcal{F}$-test under regular design can be written as,

$$P(\mathcal{F}_0 + 2(k-1)^{-1}\delta_\lambda Z \geq \mathcal{F}_0(\alpha) - (k-1)^{-1}|\mathcal{S}|G_0^2) + o(1), \qquad (2.17)$$

where $Z \sim N(0,1)$, $\mathcal{F}_0$ and $\mathcal{F}_0(\alpha)$ denote null distributions of $\mathcal{F}$ presented in Section 2.3 and its $(1-\alpha)$ quantile, respectively. $\delta_\lambda^2 = \sum_{r=1}^m \lambda_r\delta_r^2$, where $\delta_r^2 = || \int_{\mathcal{S}}(\mathbf{I}_{k-1}, \mathbf{0}),$ $\mathbf{U}^T\mathbf{h}(s)\phi_r(s)ds||^2$, $\mathbf{h}(s) = [\sqrt{a_1}\eta_1(s), ..., \sqrt{a_k}\eta_k(s)]^T/\sigma(s)$ and the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{I}_k - \mathbf{b}\mathbf{b}^T$ with $\mathbf{b} = [\sqrt{a_1}, ..., \sqrt{a_k}]$.

Along similar lines, assuming Condition B of the appendix, Zhang (2011) derived an asymptotic power approximation for $\mathcal{F}^*$. Now for the data after subtracting grand mean function, we improve the approximation slightly by modifying the proof to obtain the local asymptotic power approximation:

$$P(\mathcal{F}_0^* + 2(k-1)^{-1}\delta_\lambda^* Z \geq \mathcal{F}_0^*(\alpha) - (k-1)^{-1}(G_0^*)^2) + o(1), \qquad (2.18)$$

where $Z \sim N(0,1)$, $\mathcal{F}_0^*$ and $\mathcal{F}_0^*(\alpha)$ denote null distribution of $\mathcal{F}^*$ presented in Section 2.3 and its $(1-\alpha)$ quantile, respectively. $\delta_\lambda^{*2} = \{tr(\gamma)\}^{-2} \sum_{r=1}^{m} \lambda_r^* \delta_r^{*2}$, where $\delta_r^{*2} = || \int_{\mathcal{S}} \mathbf{\Omega}^{1/2} \mathbf{d}(s)\ \phi_r^*(s) ds||^2$, with $\mathbf{d}(s) = [\eta_1(s), ..., \eta_k(s)]^T$ and $\mathbf{\Omega} = \text{diag}(a_1, ..., a_k)$. Further details are in the appendix.

To further simplify the power analysis we obtain lower bounds for the local asymptotic power functions of (2.17) and (2.18). These lower bounds can be used to calculate the minimum sample size to achieve a target level of statistical power as a function of effect size based on the approximations:

Proposition 2.1. *The asymptotic lower bounds of the power of $\mathcal{F}$- and $\mathcal{F}^*$- test can be approximated under local alternative by,*

$$Power(\mathcal{F}|H_{1N}) \geq P(\mathcal{F}_0 + W \geq \mathcal{F}_0(\alpha)) + o(1), \tag{2.19}$$

$$Power(\mathcal{F}^*|H_{1N}) \geq P(\mathcal{F}_0^* + W^* \geq \mathcal{F}_0^*(\alpha)) + o(1), \tag{2.20}$$

where $Power(\mathcal{F}|H_{1N})$ and $Power(\mathcal{F}^*|H_{1N})$ denote powers of $\mathcal{F}$- and $\mathcal{F}^*$- test, respectively, under H_{1N}, $\mathcal{F}_0, \mathcal{F}_0(\alpha), \mathcal{F}_0^*$ and $\mathcal{F}_0^*(\alpha)$ denoted in (2.17), (2.18). *$W \sim N(\xi,\ 4(k-1)^{-1}|\mathcal{S}|\ \xi)$ where $\xi = (k-1)^{-1}|\mathcal{S}|\ G_0^2$, and $W^* \sim N(\xi^*,\ 4(k-1)^{-1}\xi^*)$ where $\xi^* = (k-1)^{-1}(G_0^*)^2$.*

Remark. The Welch-Satterthwaite χ^2-approximation can be applied to approximate null distributions, and it helps to conduct sample size estimation in a simpler way. $\mathcal{F}_0$ can be approximated to R, where $R \sim \theta\chi_d^2$, with $\theta = \frac{tr(\gamma_w^{\otimes 2})}{(k-1)|\mathcal{S}|}$ and $d = \frac{(k-1)|\mathcal{S}|^2}{tr(\gamma_w^{\otimes 2})}$, where $\gamma_w^{\otimes 2} = \int_{\mathcal{S}} \gamma_w(s,u)\gamma_w(u,t)du$. Similarly, $\mathcal{F}_0^*$ can be approximated to R^*, where $R^* \sim \theta^*\chi_{d^*}^2$, with $\theta^* = \frac{tr(\gamma^{\otimes 2})}{(k-1)tr(\gamma)}$ and $d = \frac{(k-1)tr(\gamma)^2}{tr(\gamma^{\otimes 2})}$, where $\gamma^{\otimes 2} = \int_{\mathcal{S}} \gamma(s,u)\gamma(u,t)du$. Zhang and Liang (2014) provide the formulas to determine corresponding parameters.

2.3.2 Asymptotic versus simulated power on moderate sample size

Before adopting approximated lower bounds of power in sample size determination, we need examine the accuracy of power approximations under finite sample size. The performance is investigated by comparing approximated and empirical powers via simulation studies under three different scenarios; (i) the model with stationery process, (ii) the model with cyclic marginal error having the minimum variance close to zero, and (iii) the model with heteroscedastic error process, where $\sigma(s)$ proportional to $exp(s)$. Throughout the examples, we fix $k = 3$ and specify three cases of $\mathbf{n} = [n_1, n_2, n_3]$ as $n_g = 20, 50,$ and $100,\ g = 1, 2, 3$, representing small, moderate and large sample size in balanced design. We simulated 1500 sets of discrete response curves over equally spaced grid points $s_t \in [0, 1], t = 1, ..., 80$, to calculate empirical sizes and p-values for $\mathcal{F}$- and $\mathcal{F}^*$-test, with type I error fixed at $\alpha = 0.05$. Two more scenarios and corresponding results are reported in online supplementary material.

Simulation 1 (stationary process)

We generate discrete functional samples from exponentially correlated process,

$$y_{gi}(s_t) = \mu_0(s_t) + \beta_g(s_t) + \epsilon_{gi}(s_t),\ \text{where } \epsilon_{gi}(j, k) = \sigma_e^2 \cdot exp(-|j - k|/d),$$

$$\beta_1(s) = -\delta,\ \beta_2(s) = 0,\ \beta_3(s) = \delta,\ \text{with } \delta > 0,\ g = 1, 2, 3,\ i = 1, ..., n.$$

We specify $\mu_0(s)$ using 3 degrees of freedom B-spline basis functions. Note that the parameter d determines the dependency structure within a curve. Functional samples with values of $0.1, 0.4$ and 0.9 of d implying low, moderate, and high spatial correlation within curve, respectively, are generated. Here δ controls the deviation between mean curves. We set $\sigma_e = 1$ and sequence of δ is applied to study the power under different effect sizes.

Simulation 2 (non-stationary process: cyclic marginal deterministic variance)

Now we consider the model with fluctuating marginal variance function under parallel mean functions as in simulation 1. The discretized response curves under exponentially correlated process with $d = 0.4$ and the following marginal deterministic variance function is simulated; $\sigma_e(s) = cos(8\pi s) + 1.005$. The period and amplitude are 4 and 1. The minimum marginal variance at each cycle is 0.005, that is very close to 0.

Simulation 3 (non-stationary process: heteroscedastic model)

We consider the model in Simulation 1, but with exponentially extreme deterministic marginal variance rather than constant σ_e over s. Specifically, we fix $d = 0.4$ and $\sigma_e^2(s)$ is set to be proportional to $exp(2.4s)$, $s \in [0, 1]$. It leads the heteroscedastic error that has the range of variance as [1,11].

Simulation 4 (non-parallel mean functions)

The stationary functional data samples with group mean functions having one point of intersection are generated,

$$\beta_1(s) = -\varphi(s - 0.5),\ \beta_2(s) = 0,\ \beta_3(s) = \varphi(s - 0.5),\ \text{for } \varphi > 0.$$

The exponentially correlated functional process with $d = 0.4$ and constant σ_e over s as in Simulation 1 is considered and simulated.

Simulation 5 (non-stationary process)

The model simulated in Zhang and Liang (2014, section 3) will be used with little modification to generate discretized functional curves:

$$y_{gi}(s_t) = \mu_0(s_t) + \beta_g(s_t) + \epsilon_{gi}(s_t), \quad \epsilon_{gi}(s) = \mathbf{b}_{gi}^T \mathbf{\Psi}(s), \ i = 1, ..., n, \ g = 1, 2, 3,$$

$$\mathbf{b}_{gi} = [b_{gi1}, b_{gi2}, ..., b_{giq}]^T, \ b_{gir} \overset{d}{=} \sqrt{\lambda_r} z_{gir}, \ r = 1, ..., q,$$

where z_{gir} are *i.i.d.* standard normal random variables. The common covariance function is $\gamma(j,k) = \sum_{r=1}^{q} \lambda_r \psi_r(j) \psi_r(k)$ with the orthonormal basis vector $\mathbf{\Psi}(s) = [\psi_1(s), ..., \psi_q(s)]^T$ and the q decreasing-ordered variance components λ_r. In our example, we set $\mu_0(s) = 1 + 2.3s + 3.4s^2 + 1.5s^3$, $\beta_g(s) = -\delta, 0$ and δ, with $\delta > 0$, $\lambda_r = a\eta^r$, $r = 1, ..., q$, with $\eta = 0.1$, 0.5 or 0.9, representing high, moderate and low within-curve correlation, and $q = 7$. For each value of η, we set $a = 9.5, 1.02$ or 0.21, respectively, to make the overall average of marginal variance to be equal to 1. The orthonormal basis functions are set as $\psi_1(s) = 1$, $\psi_{2r}(s) = \sqrt{2}sin(2\pi rs)$, $\psi_{2r+1}(s) = \sqrt{2}cos(2\pi rs)$, $s \in [0, 1]$, $r = 1, ..., 3$.

Figure 2.1 reports approximated and empirical sizes and powers of fANOVA tests under various scenarios. For Simulation1 and 2, only $\mathcal{F}$-test results are displayed on (a), because both tests give similar results. The lower panels in (b) show the results from two types of test on Simulation 3 to compare performance between them. We present specific powers according to different scales of effect size for both types of test in Table 2.1-2.5, but interpret them with Figure 2.1 for easier comparison.

First of all, we see that the agreement between approximation and empirical estimation is quite good even under small sample size for all scenarios. Second, we find the interesting fact that strong within-curve correlation leads less statistical power.

Apparently, strong dependency structure, larger d in top panels of (a), shows smaller power and it can be inferred that the amount of information at each grid decreases as within-curve correlation becomes stronger. Indeed it results in lack of power. Thus the data with strong within-curve correlation needs larger sample size to achieve the same level of power as we will see in section 3.3. Third, the agreement is rather good for the model with weak within-curve correlation, smaller d in top panels of (a). The rich information at each grid leads less deviations in power estimation. Fourth, slight difference between $\mathcal{F}^*$ and $\mathcal{F}^*$-test is found in Simulation 3. It demonstrate that $\mathcal{F}$-test is powerful over $\mathcal{F}^*$-test under extreme marginal error behaviors. Although difference is likely to be larger than discrepancies between two tests from other simulations, it is not a huge difference as seen in Table 2.3. Lastly, the comparison between Simulation1 and 3 shows that the test under stable marginal error model is more powerful compared to the test from unstable fluctuating $\sigma_e(s)$. For last two simulations, Table 2.4 shows that the shape of mean functions does not affect on the accuracy of asymptotic power. Also we can infer from Table 2.5 that the agreement is not affected by covariance function as well.

2.3.3 Sample size determination

The good accuracy of asymptotic power implies that lower bounds of power in (2.19) and (2.20) can be reasonably used to estimate sample size in practice. Note that two expressions are based on local alternative, thus we slightly modify formulas for practical use under $H_1 : \mu_g(s) = \mu_0(s) + \beta_g(s)$, where $\beta_g(s),\ g = 1, ..., k$, are non-zero and square-integrable functions with $\sum_g n_g \beta_g(s) = 0$. Specifically $\xi = (k-1)^{-1} N \cdot G_0^2$ for (2.19), and $\xi^* = (k-1)^{-1} \|\sigma(\cdot)\|^2 N \cdot (G_0^*)^2$ for (2.20). Now the power is a function of sample size, effect size and significance level as a general case. Provided that working covariance function $\gamma(s,t)$ is assumed and certain level of effect size, say it G, is set, then a Monte Carlo procedure for $\mathcal{F}$-test is implemented as follows: (i) Consider a sequence of sample size $\{N_1 < N_2 < ... < N_m\}$ and specify the sequence of distributions of W for given G and N_j. Let denote it as $W_{G,N_j},\ j = 1, ..., m$. (ii) Generate a large sample of $\mathcal{F}_0$ and W_{G,N_j}, and compute the empirical lower bound of power from empirical distributions. (iii) Obtain the sequence of lower bounds of power as a function of N and choose the sample size that achieves the desired power. A Monte Carlo procedure for $\mathcal{F}^*$-test can be implemented in a similar way. As noted in Zhang and Liang (2014), the Welch-Satterthwaite χ^2-approximation can be applied to approximate $\mathcal{F}_0$. However, we prefer not to use this approximated distribution in this paper, because our experiment shows that this approach yields less accurate result than generating Monte Carlo samples from original $\mathcal{F}_0$.

To illustrate the sample size approximation, we consider the models in Simulation 1 with $d = 0.1,\ 0.4$ and 0.9. We derive the power of $\mathcal{F}$-test as a function of sample size under global effect size $G_{fSNR} = 0.2,\ 0.4$ and 0.8, representing small, medium and large effect size, with type I error at $\alpha = 0.05$. Here sample size means the number of observations in each group for balanced experiment. The powers from $\mathcal{F}$-test are presented and $\mathcal{F}^*$-test gives almost the same result.

Figure 2.2 displays the power curves from three global effect sizes under three magnitudes of dependency. First of all, it can be seen that the power of $\mathcal{F}$-test achieves 0.8 or more even with moderate sample size, around 20, under medium effect size. Secondly, we can see the effect of within-curve dependency in sample size estimation. Obviously, the covariance structure with strong within-curve correlation needs more sample to achieve the same level of power compared to others. It is corresponding to what we found in Figure 2.1 (a).

2.4 Real Data Analysis

2.4.1 Analysis of Mouse and Rat Mammary Tumor Data

We now return to the quantitative ultrasound study. The experiment was conducted with two types of mammary tumors, 13 induced 4T1 tumors on mice and 8 induced MAT tumors on rats. The features of tumor tissues, such as length, height and volume vary across subjects. As mentioned in the introduction, the tumor in each animal is invasively scanned by 5 different transducers from three systems (Siemens, Ultrasonix and VisualSonics) which cover different range of frequency bandwidths. Two transducers, 9L4 and 18L6, from Siemens, L14-5 from Ultrasonix, and MS200 from VisualSonics cover frequencies around 3-13.5 MHz, meanwhile MS400 from VisualSonics covers higher frequencies greater than 13.5 MHz. Different from Wirtzfeld *et al.* (2015), we use the subset of data composed of subjects having large tumor (greater than $70mm^3$) in the analysis. The data pertain to 5 4T1 and 6 MAT large tumors. Large tumor enables transducer to scan the target without much being affected by surrounding normal tissues, so that noise error has been substantially reduced (Wirtzfeld *et al.*, 2015). From here on, we distinguish 55 combinations of animals and transducers by defining variable called as ‘setup’. For each setup, there are 4 or 5 multiple functional records by shifting scan lines within each tumor. The frequency dependent backscatter (BSC) functions were calculated in decibel scale (dB) for each scan based on the collected ultrasound radio frequency signals using a reference phantom technique. More details can be found in Wirtzfeld *et al.* (2015).

The aims of this experiment are as follows: Firstly we want to analyze how well BSC records can separate two tumor types by measuring effect size beyond significance test. Secondly, we are interested in finding the frequencies which achieve the most sufficient precision to distinguish two tumors. Lastly, inter-transducer variation in

BSC is of interest.

Prior to proceeding, note that the repeated measures for each 'setup' lead correlation structure between multiple curves. The mixed ANOVA result with 'setup' as random effect and tumor type as fixed effect is presented in Figure 2.3. They are smoothed from pointwise result via natural cubic splines for comprehensible visualization. The smoothed classic 1-way ANOVA result is shown as well for comparison. We see that the magnitude of marginal error in 1-way ANOVA mostly includes both subject and noise random errors.

The collected BSC curves from Ultrasonix L14-5 are summarize in Figure 2.4. The mean values and standard errors at each grid are presented in (a). Both tumors seem to have nearly constant standard deviation over frequencies. The pointwise and smoothed $fSNR$ are illustrated in (b). This indicates that higher frequencies seem more effective in distinguishing tumors than lower frequencies do.

To make a formal inference, now we apply the $fSNR$ analysis. As the transducers all have different bandwidths, 3 frequency ranges are selected to carry out subsequent analyses. The lower frequency range (3-8.5 MHz) includes data from Ultrasonix and Siemens transducers, middle frequency (8.5-13.5 MHz) includes two VisualSonics transducers, and the higher frequency range (13.5-21.9 MHz) includes one transducer MS400, from VisualSonics. Table 2.6 displays estimated global measures and bootstrapped p-values over each range. We use 1500 non-parametric bootstrapped samples to perform $fSNR$ analysis hereafter. As discussed in section 2.4, all scans for a given animal and transducer combination were sampled together to preserve correlation between replicates from the same 'setup'

We see in Table 2.6 that two types of proposed global measures give almost the same significant tumor effect size with small p-value. The estimated measures at each range suggest that higher frequencies are more effective in separating two different tumors. Figure 2.5 presents the estimate of smoothed $fSNR$ using natural cubic splines with six interior knots selected from BIC, and its 90% pointwise confidence intervals via bootstrapping.(Efron and Tibshirani, 1993). It can yield valuable information about which interval can distinguish two tumors with the most sufficient precision. It demonstrates a trend in increasing separation between MAT and 4T1, which is in agreement with what we found in Table 2.6 and this trend can be explained by the inverse relationship between frequency and wavelength. Higher frequency with short wavelength might collect more information when penetrating a tissue rather than lower frequency with long wavelength can do. Also higher frequencies have relatively wide widths of confidence interval due to small number of curves collected over there.

As a next step, we can compare the efficacy of transducers by comparing estimates of effect size. Table 2.7 presents that transducers covering frequencies less than 13.5 MHz have similar significant effect size around 0.8-0.9 with small p-values except Siemens 9L4. The different behavior in Siemens 9L4 seems to be due to influential observations from 4T1 tumor. For relatively higher frequency range, VisualSonics MS400 apparently shows significantly greater separation with larger effect size, which is corresponding to our finding through Table 2.6 and Figure 2.5.

The last goal is to examine consistency across systems. For this purpose, we use lower and middle frequency ranges that include at least two transducers, and calculate the global measures to investigate the existence of transducer effect for each tumor type. Specifically, Ultrasonix and Siemens are compared over 3-8.5 MHz and VisaulSonics are compared over 8.5-13.5 MHz. Table 2.8 shows that all estimated measures are less than 0.3 with bootstrapped p-values greater than 0.2, thus the claim of consistency across systems is statistically supported. A key observation is that the magnitudes of transducer effect size are much less than those of tumor effect size.

2.4.2 Canadian Weather Data

We analyze the Canadian weather data to illustrate the usefulness of our methodology. The data are the daily temperature and precipitation records of 35 weather stations over a year, 365 days, among which 15 in Atlantic, 12 in Continental, 5 in Pacific and 3 in Arctic. The weather information from each station is collected every day with no missing. The dataset is available through R-package `'fda'`. Various functional data analysis methods were already applied by many authors, including statistical inferences to test significant region effect on temperature. Ramsay and Silverman (2005) characterized the typical temperature pattern and investigated when regional temperature effect is substantial by examining F-ratio *function*. However, although pointwise F-ratio can be used to infer an effect size, the sample size or the number of groups should be known in order to be interpreted. Accordingly, it is hard to compare two statistics in general if they are computed from different designs. Also they did not discuss whether the change of regional effect over a year is statistically significant. Zhang and Liang (2007) assessed the significant differences in temperature between climate zones and investigated its pattern over seasons. However, the change in the magnitude of region effect over seasons was just inferred by comparing the magnitude of p-values, not from precise statistical inference. Here our goals are to quantify region effect on temperature and precipitation by applying proposed $fSNR$ analysis, and see when substantial difference between regions is observed. We will also study if the change of effect size over time is significant via bootstrapped confidence intervals of $fSNR$. Additionally we will see which variable is more affected by geographical factor, among temperature and precipitation.

Figure 2.6 presents the estimated $fSNR$ and its bootstrapped 90% confidence intervals over a year based on 1500 bootstrapped samples. Natural cubic splines with

9 interior knot are used to smooth group mean and marginal error curves. It is seen that the difference of temperature between regions is larger than the difference of precipitation with larger estimated effect sizes over the whole year ($[a, b] = [1, 365]$). Although fANOVA or other proposed test result can give the same conclusion of rejection, $fSNR$ enables to present distinct behaviors over time. Specifically, we observe that magnitudes of region effect are different over seasons. In terms of temperature, the difference between climate zones is less during the summer (June, July and August or $[a, b] = [152, 243]$) than the difference during the spring (March, April and May or $[a, b] = [60, 151]$) or the autumn (September, October and November or $[a, b] = [244, 334]$), which is in agreement with what Ramsay and Silverman (2005) observed from F-ratio $function$. However, we can make further conclusion that seasonal change of region effect in temperature is not statistically significant. The straight horizontal line can be drawn over a year within bootstrapped confidence intervals at around between 1.5 and 2. From (b), it is seen that the region effect on precipitation is large during the winter and early spring. Different from temperature data, we can conclude that this seasonal change is statistically significant.

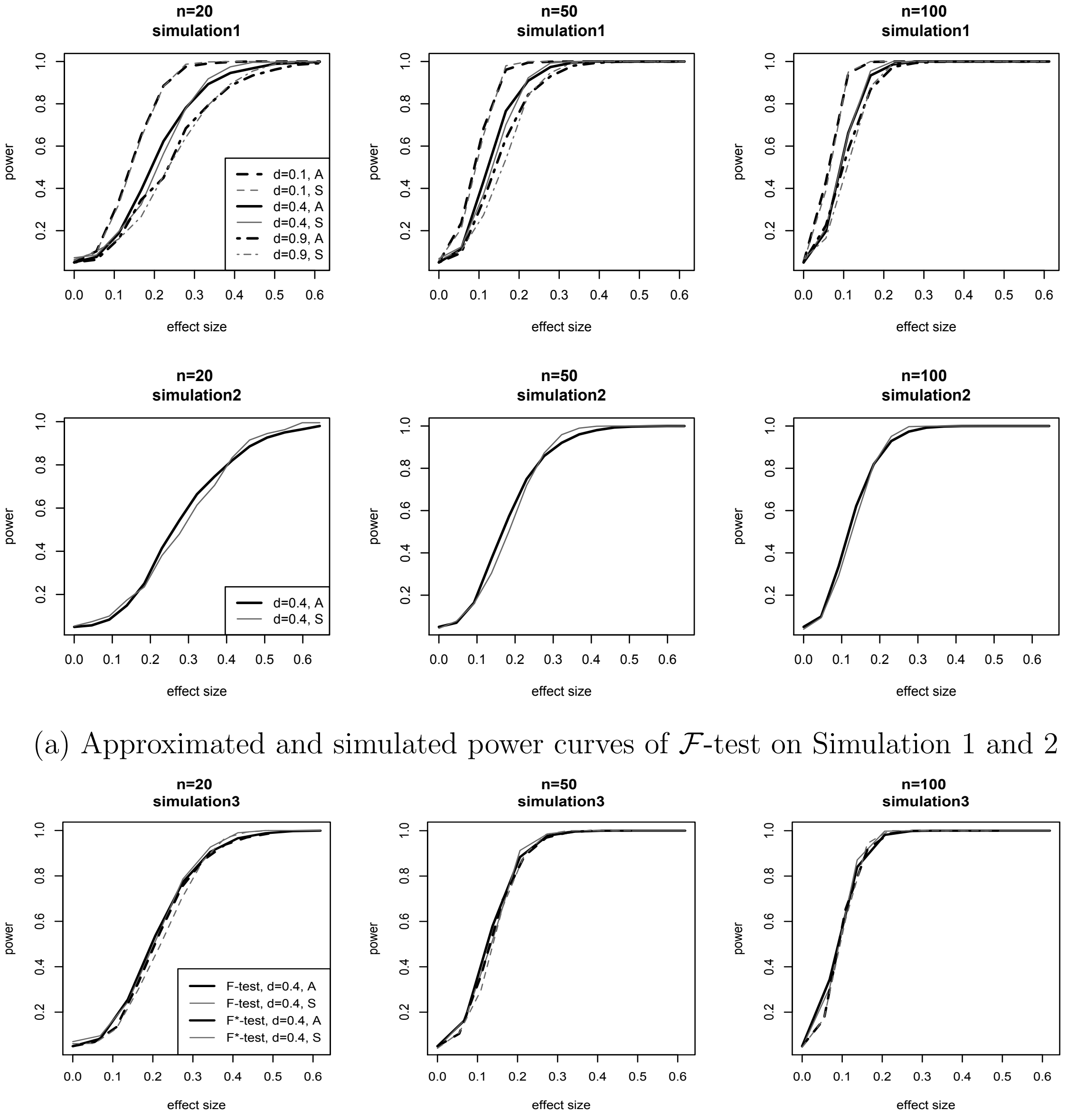
Table 2.6 shows the estimated global measures and bootstrapped p-values. Firstly, we can see the consistency of two measures for each time period. Secondly, as expected, estimated regional effect sizes for temperature data are around twice larger than those for precipitation data for the whole year as well as during each season. Lastly, the magnitudes of the regional differences in temperature and precipitation during the summer are less than the magnitudes during the spring and autumn. Again, all associated p-values are very close to 0, but the proposed measure provides additional information about an effects size and enables to compare to each other.

2.5 Discussion

The advantages of $fSNR$ analysis are as follows: The effect size of the variable of interest is simply computed through F-ratio $function$ or functional F-statistics. Via visualization of local information and its corresponding confidence intervals, the most informative domain of the function can be found. The asymptotic lower bound of power can be used as an handy tool for sample size estimation in planning purpose.

In future works, we aim to provide asymptotic null distribution of functional F-statistics under irregular sampling framework. To do this, central limit theorem in Hilbert space for i.i.d. stochastic precess with random interval should be demonstrated. Another goal is to extend $fSNR$ analysis to 2-dimensional data to quantify the effect size retaining inherent spatial smoothness, to visualize the local effect size in 2-dimensional space and to derive confidence region. It will enable to implement statistical inferences for ultrasound image data for tumor margin assessment as well as spatial data.

2.6 Figures and Tables



(a) Approximated and simulated power curves of $\mathcal{F}$-test on Simulation 1 and 2

(b) Approximated and simulated power curves of $\mathcal{F}$ and $\mathcal{F}^*$-test on Simulation 3

Figure 2.1: Approximated ("A") and simulated ("S") power curves as a function of effect size on each scenario.

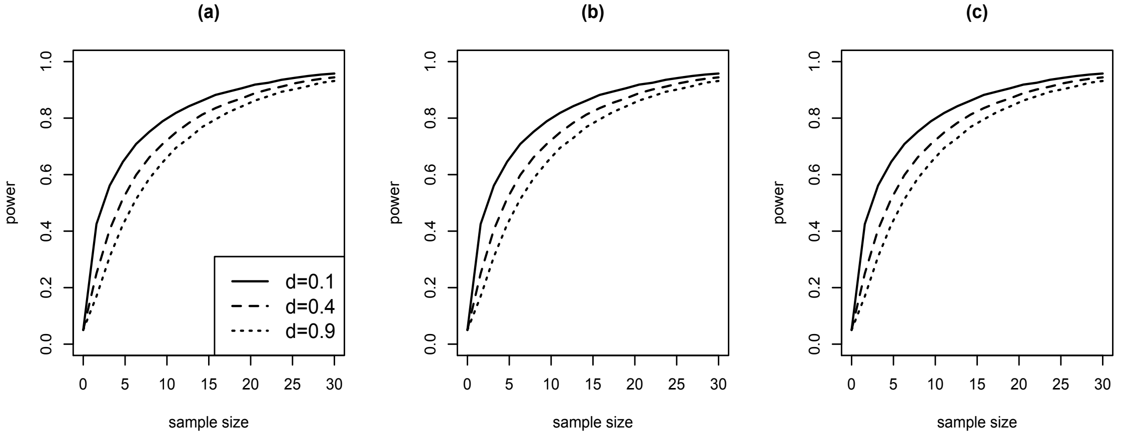


Figure 2.2: Power functions under $d = 0.1,\ 0.4$ and 0.9 for (a) $G_{fSNR} = 0.2$, (b) $G_{fSNR} = 0.4$ and (c) $G_{fSNR} = 0.8$

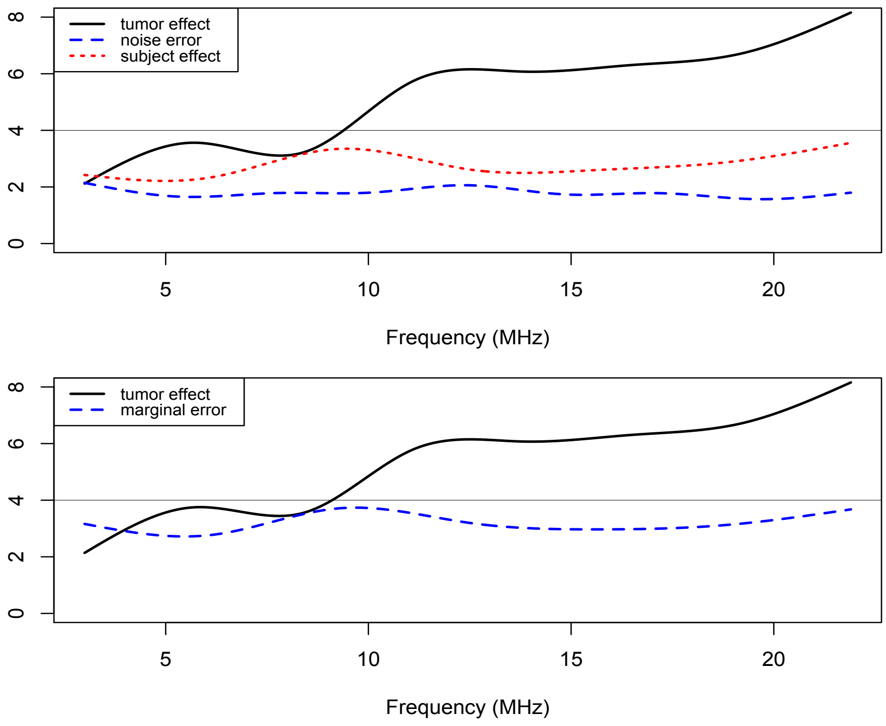


Figure 2.3: (Top) Pointwise estimate of the difference in magnitude of the BSC estimates between two tumors, noise error, and random effect for ‘setup’ from pointwise mixed ANOVA. (Bottom) pointwise tumor effect and marginal error from pointwise 1-way ANOVA

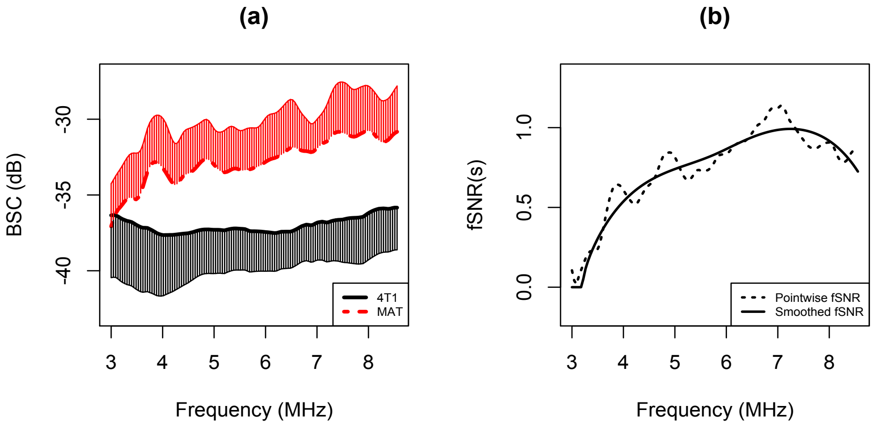


Figure 2.4: Transducer L14-5. (a) The average and SD curves of 4T1 and MAT, (b) Pointwise and smoothed $fSNR$

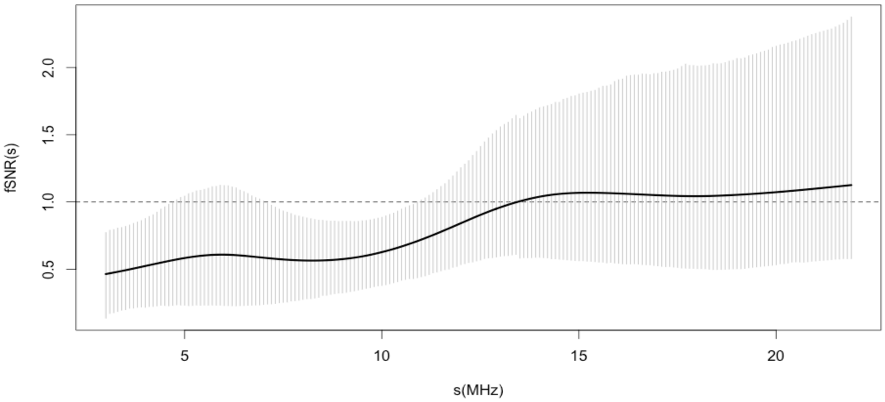


Figure 2.5: The smoothed $fSNR$ and bootstrapped 90 % confidence intervals

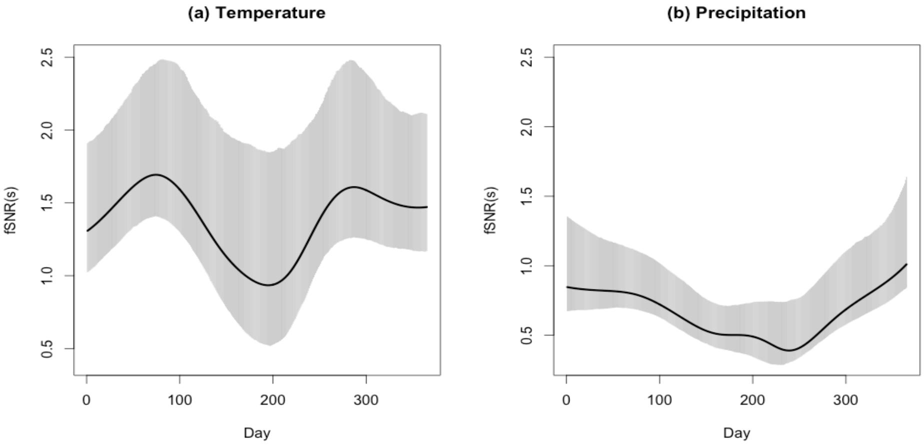


Figure 2.6: The smoothed $fSNR$ and bootstrapped 90% confidence intervals of (a) Temperature and (b) Precipitation

Table 2.1: Simulation 1: approximated and empirical sizes and powers of the $\mathcal{F}$- and $\mathcal{F}^*$-test under stationary process model ($G = G_{fSNR}$, $G^* = G^*_{fSNR}$)

n_g	d	Approximated power/ Empirical power			
		$G = G^* = 0$ ($\delta = 0$)		$G = G^* = 0.11$ ($\delta = 0.14$)	
		GPF	F	GPF	F
20	0.1	0.05/ 0.06	0.05/ 0.05	0.32/ 0.32	0.32/ 0.32
	0.4	0.05/ 0.06	0.05/ 0.06	0.18/ 0.17	0.18/ 0.15
	0.9	0.05/ 0.07	0.05/ 0.06	0.14/ 0.15	0.15/ 0.14
50	0.1	0.05/ 0.05	0.05/ 0.05	0.71/ 0.68	0.71/ 0.68
	0.4	0.05/ 0.06	0.05/ 0.06	0.44/ 0.36	0.43/ 0.36
	0.9	0.05/ 0.05	0.05/ 0.04	0.34/ 0.29	0.34/ 0.28
100	0.1	0.05/ 0.06	0.05/ 0.06	0.94/ 0.95	0.93/ 0.95
	0.4	0.05/ 0.06	0.05/ 0.06	0.69/ 0.66	0.70/ 0.65
	0.9	0.05/ 0.07	0.05/ 0.07	0.59/ 0.51	0.57/ 0.50

n_g	d	Approximated power/ Empirical power			
		$G = G^* = 0.19$ ($\delta = 0.23$)		$G = G^* = 0.26$ ($\delta = 0.32$)	
		GPF	F	GPF	F
20	0.1	0.76/ 0.72	0.74/ 0.72	0.95/ 0.97	0.95/ 0.97
	0.4	0.47/ 0.44	0.48/0.41	0.73/ 0.68	0.72/ 0.66
	0.9	0.37/ 0.34	0.36/ 0.32	0.62/ 0.55	0.62/ 0.52
50	0.1	0.98/ 0.99	0.98/ 0.99	1.00/ 1.00	1.00/ 1.00
	0.4	0.81/ 0.81	0.81/ 0.80	0.96/ 0.99	0.96/ 0.98
	0.9	0.71/ 0.65	0.70/ 0.64	0.90/ 0.92	0.90/ 0.92
100	0.1	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
	0.4	0.95/ 0.99	0.96/ 0.98	1.00/ 1.00	1.00/ 1.00
	0.9	0.91/ 0.95	0.91/ 0.94	0.99/ 1.00	0.99/ 1.00

Table 2.2: Simulation 2: approximated and empirical sizes and powers of $\mathcal{F}$- and $\mathcal{F}^*$-test under the model with cyclic marginal error function

n_i	Approximated power/ Empirical power							
	$G = G^* = 0$ ($\delta = 0$)		$G = G^* = .18$ ($\delta = 13$)		$G = G^* = .28$ ($\delta = 19$)		$G = G^* = .41$ ($\delta = 29$)	
	GPF	F	GPF	F	GPF	F	GPF	F
20	0.05/ 0.05	0.05/ 0.05	0.25/ 0.24	0.26/ 0.22	0.54/ 0.48	0.53/ 0.46	0.82/ 0.83	0.83/ 0.81
50	0.05/ 0.04	0.05/ 0.04	0.57/ 0.50	0.59/ 0.49	086/ 0.87	0.86/ 0.87	0.98/ 1.00	0.98/ 1.00
100	0.05/ 0.04	0.05/ 0.04	0.82/ 0.82	0.82/ 0.82	0.97/ 1.00	0.97/ 1.00	1.00/ 1.00	1.00/ 1.00

Table 2.3: Simulation 3: approximated and empirical sizes and powers of $\mathcal{F}$- and $\mathcal{F}^*$-test under heteroscedisticity model

n_g	Approximated power/ Empirical power			
	$G = G^* = 0$ ($\delta = 0$)		$G = .14, G^* = .11$ ($\delta = .27$)	
	GPF	F	GPF	F
20	0.05/ 0.05	0.05/ 0.05	0.26/ 0.24	0.14/ 0.14
50	0.05/ 0.07	0.05/ 0.06	0.59/ 0.53	0.37/ 0.31
100	0.05/ 0.05	0.05/ 0.05	0.85/ 0.84	0.66/ 0.60

n_g	Approximated power/ Empirical power			
	$G = .21, G^* = .16$ ($\delta = .41$)		$G = .25, G^* = .20$ ($\delta = .50$)	
	GPF	F	GPF	F
20	0.55/ 0.51	0.34/ 0.28	0.72/ 0.67	0.49/ 0.40
50	0.88/ 0.90	0.69/ 0.64	0.95/ 0.97	0.85/ 0.87
100	0.98/ 1.00	0.92/ 0.95	1.00/ 1.00	0.98/ 1.00

Table 2.4: Simulation 4: approximated and empirical sizes and powers of $\mathcal{F}$- and $\mathcal{F}^{*}$-test under the model with non-parallel mean functions

n_i	Approximated power/ Empirical power							
	$G = G^{*} = 0$ ($\varphi = 0$)		$G = G^{*} = .11$ ($\varphi = .45$)		$G = G^{*} = .24$ ($\varphi = 1$)		$G = G^{*} = .22$ ($\varphi = .91$)	
	GPF	F	GPF	F	GPF	F	GPF	F
20	0.05/ 0.07	0.05/ 0.06	0.13/ 0.15	0.13/ 0.14	0.38/ 0.38	0.39/ 0.36	0.73/0.74	0.74/ 0.71
50	0.05/ 0.05	0.05/ 0.05	0.38/ 0.36	0.38/ 0.34	0.85/ 0.87	0.86/ 0.86	0.99/ 1.00	0.99/ 0.99
100	0.05/ 0.06	0.05/ 0.06	0.75/ 0.74	0.75/ 0.73	0.99/ 1.00	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00

Table 2.5: Simulation 5: approximated and empirical sizes and powers of $\mathcal{F}$- and $\mathcal{F}^{*}$-test under non-stationary process model

n_g	η	Approximated power/ Empirical power			
		$G = G^{*} = 0$ ($\delta = 0$)		$G = G^{*} = 0.12$ ($\delta = 0.15$)	
		GPF	F	GPF	F
20	0.1	0.05/ 0.06	0.05/ 0.05	0.13/ 0.11	0.12/ 0.10
	0.5	0.05/ 0.07	0.05/ 0.05	0.20/ 0.20	0.20/ 0.18
	0.9	0.05/ 0.06	0.05/ 0.04	0.29/ 0.28	0.28/ 0.26
50	0.1	0.05/ 0.05	0.05/ 0.05	0.31/ 0.27	0.31/ 0.26
	0.5	0.05/ 0.06	0.05/ 0.06	0.48/ 0.42	0.49/ 0.41
	0.9	0.05/ 0.05	0.05/ 0.04	0.68/ 0.62	0.66/ 0.62
100	0.1	0.05/ 0.05	0.05/ 0.05	0.53/ 0.50	0.55/ 0.49
	0.5	0.05/ 0.05	0.05/ 0.04	0.73/ 0.70	0.74/ 0.69
	0.9	0.05/ 0.05	0.05/ 0.05	0.92/ 0.95	0.93/ 0.95

n_g	η	Approximated power/ Empirical power			
		$G = G^{*} = 0.18$ ($\delta = 0.22$)		$G = G^{*} = 0.24$ ($\delta = 0.29$)	
		GPF	F	GPF	F
20	0.1	0.27/ 0.26	0.28/ 0.24	0.47/ 0.38	0.47/ 0.35
	0.5	0.44/ 0.38	0.43/ 0.35	0.64/ 0.60	0.63/ 0.57
	0.9	0.63/ 0.58	0.63/ 0.57	0.87/ 0.87	0.86/ 0.86
50	0.1	0.61/ 0.55	0.60/ 0.53	0.79/ 0.79	0.80/ 0.79
	0.5	0.77/ 0.78	0.78/ 0.78	0.93/ 0.96	0.93/ 0.95
	0.9	0.94/ 0.97	0.95/ 0.97	1.00/ 1.00	1.00/ 1.00
100	0.1	0.83/ 0.86	0.84/ 0.85	0.95/ 0.98	0.95/ 0.98
	0.5	0.95/ 0.97	0.94/ 0.96	0.99/ 1.00	0.99/ 1.00
	0.9	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00

Table 2.6: Estimates of tumor effect size (bootstrapped p-values in parentheses)

	Low (3-8.5 MHz)	Middle (8.5-13.5 MHz)	High (13.5-21.9 MHz)
$\hat{G}_{fSNR}$	0.56 ($< .001$)	0.99 ($< .001$)	1.06 (0.005)
$\hat{G}^*_{fSNR}$	0.55 ($< .001$)	0.99 ($< .001$)	1.06 (0.008)

Table 2.7: Tumor type effect size for each transducer

Transducer Bandwidth	$\hat{G}_{fSNR}$ / $\hat{G}^*_{fSNR}$	p-value
Ultrasonix L14-5 (3-8.5 MHz)	0.78/ 0.73	$0.005^*/0.008^*$
Siemens 9L4 (3-10.8 MHz)	0.17/ 0.13	0.49/0.57
Siemens 18L6 (3-10.8 MHz)	0.99/ 0.91	$0.004^*/0.007^*$
VisualSonics MS200 (8.5-13.5 MHz)	0.87/ 0.84	$0.02^*/0.02^*$
VisualSonics MS400 (8.5-21.9 MHz)	1.14/ 1.12	$0.004^*/0.004^*$

Table 2.8: Transducer effect size for each tumor type and frequency range

Tumor type	Frequency range	$\hat{G}_{fSNR}$ / $\hat{G}^*_{fSNR}$	p-value
4T1	3 -8.5 MHz	0.31/ 0.31	0.34/0.34
	8.5 -13.5 MHz	0.27/ 0.26	0.16/0.17
MAT	3 -8.5 MHz	0.28 /0.28	0.25/0.24
	8.5-13.5 MHz	0.21/ 0.18	0.44/ 0.50

Table 2.9: The regional effect size for the Canadian daily temperature and precipitation data (with bootstrap p-values in parentheses)

		$[a, b]$				
		Whole year $[1, 365]$	Spring $[60, 151]$	Summer $[152, 243]$	Autumn $[244, 334]$	Winter $[1, 59] \cup [335, 365]$
TMP.	G_{fSNR}	1.36 (< .001)	1.47 (< .001)	1.01 (< .001)	1.48 (< .001)	1.42 (< .001)
	G^*_{fSNR}	1.41 (< .001)	1.5 (< .001)	0.99 (.001)	1.51 (< .001)	1.42 (< .001)
PRC.	G_{fSNR}	0.76 (< .001)	0.81 (< .001)	0.56 (< .001)	0.7 (< .001)	0.92 (< .001)
	G^*_{fSNR}	0.73 (< .001)	0.78 (< .001)	0.53 (< .001)	0.67 (< .001)	0.89 (< .001)

2.7 Technical conditions, proof and numerical algorithm

2.7.1 Condition A

(A.1) $\mu_0(s)$ and $\beta_i(s), i = 1, ..., k \in L^2(\mathcal{S})$ and $tr(\gamma) < \infty$.

(A.2) The marginal error process $\epsilon_i(s), i = 1, ..., k$ are i.i.d.

(A.3) As $n \to \infty$, the k sample sizes satisfy $n_i/N \to a_i \in (0, 1),\ i = 1, ..., k$.

(A.4) The marginal error process $\epsilon_1(s)$ satisfies $E\|\epsilon_1\|^4 = E[\int_{\mathcal{S}} \epsilon_1^2(s)ds]^2 < \infty$.

(A.5) For any $s \in \mathcal{S}$, $\gamma(s, s) > 0$. In addition, the maximum variance $m = max_{s\in\mathcal{S}}$ $\gamma(s, s) < \infty$

(A.6) The expectation $E[\epsilon_1^2(s)\epsilon_1^2(t)]$ is uniformly bounded. That is, for any $(s, t) \in \mathcal{S}^2$, we have $E[\epsilon_1^2(s)\epsilon_1^2(t)] < C < \infty$, where C is some constant independent of (s, t).

2.7.2 Condition B

(B.1) $\delta_r^2 \neq 0$ for at least one $r \in \{1, ..., m\}$

2.7.3 Missing data frame work for irregular functional data (2.5)

Let $\bar{y}(s) = \frac{\sum_{i=1}^n y_i^c(s)\mathbb{1}_{[L_i,U_i]}(s)/n}{\sum_{i=1}^n \mathbb{1}_{[L_i,U_i]}(s)/n} \triangleq \frac{W1}{W2}$. For each $s \in \mathcal{S}$,

$$W_1 \xrightarrow{p} E[(y_i^c(s)\mathbb{1}_{[L_i,U_i]}(s)] = E[y_i^c(s)]P(s \in [L_i, U_i]),$$

by law of large numbers and independence assumption of $y^c(s)$ and I. Similarly

$$W_2 \xrightarrow{p} P(s \in [L_i, U_i]).$$

By continuous mapping theorem, $W_1/W_2 \xrightarrow{p} \mu(s)$. Next, let

$$\hat{\sigma}^2(s) = \frac{\sum_{i=1}^n (y_i^c(s) - \bar{y}(s))^2 \mathbb{1}_{[L_i,U_i]}(s)/n}{\sum_{i=1}^n \mathbb{1}_{[L_i,U_i]}(s)/n - 1/n} \triangleq \frac{W_1'}{W_2'}.$$

For each $s \in \mathcal{S}$,

$$\begin{aligned} W_1' &\xrightarrow{p} E[y_i^c(s) - \bar{y}(s)]^2 P(s \in [L_i, U_i]) \\ &= E[y_i^c(s) - \mu(s)]^2 P(s \in [L_i, U_i]) + E[\bar{y}(s) - \mu(s)]^2 P(s \in [L_i, U_i]) \\ &= \sigma^2(s) P(s \in [L_i, U_i]) + \sigma^2(s) P(s \in [L_i, U_i])/n \\ &= \sigma^2(s) P(s \in [L_i, U_i]) + o(1), \end{aligned}$$

$$W_2' \xrightarrow{p} P(s \in [L_i, U_i])$$

by continuous mapping theorem, $W_1'/W_2' \xrightarrow{p} \sigma^2(s)$

2.7.4 Power of $\mathcal{F}_s^*$- test (2.18)

We follow the method of proof given in Zhang (2011) and Zhang and Liang (2014) with modification. While Zhang (2011) restricted $0 \leq \tau < 1$ for further two steps of approximation for the local alternative,$H_{1n}^c : \mathbf{C}\boldsymbol{\beta}(s) - \mathbf{c}(s) = n^{-\tau/2}\mathbf{d}(s)$, we simplified proof by just using two asymptotic distributions for $MSW \sim AN[tr(\gamma), 2tr(\gamma^{\otimes 2})/(N - k - 1)]$ and $MSB \overset{d}{=} \sum_{r=1}^m \lambda_r A_r/(k-1) + \sum_{r=1}^m \lambda_r^{1/2} \delta_r z_{qr}/(k-1) + \delta^2/(k-1)$, where $A_r \sim \chi^2_{k-1}$. It improves the accuracy of the resulting approximation even for moderate sample size. The accuracy is examined by empirical study.

2.7.5 Lower bounds of powers (2.19) and (2.20)

For (19), combine (17) and upper bound $\delta_\lambda^2 \leq (|S|G_0)^2$. For (20), combine (18) and upper bound $\delta_\lambda^{*2} \leq (G_0^*)^2$.

2.7.6 Algorithms for power estimation

In practice it is not possible to observe functional samples continuously. Rather, the functional data will be collected at a finite number of points in $\mathcal{S}$. Here are algorithms to estimate the powers for discretized observed functions which have a common fine grid $s_t,\ t = 1, ..., T$ over $\mathcal{S}$. We assume the same covariance structure over groups as ANOVA assumption. Let $\mathbf{Y_g}$ be the $T \times n_g$ matrix with each sampled $T \times 1$ vector at each column. And let $\mathbf{Y} = [\mathbf{Y_1}, \mathbf{Y_2}, ..., \mathbf{Y_k}]^\mathbf{T}$ be the $N \times T$ pooled matrix.

- **Estimation of power of $\mathcal{F}$-test** (2.17)

  1. Find $T \times T$ pooled sample correlation matrix on the basis of data $N \times T$ data matrix $\mathbf{Y}$.

     (a) Calculate $T \times T$ pooled sample covariance matrix using centered data matrix $\mathbf{Y}^* = [\mathbf{Y}_1^*, ..., \mathbf{Y}_k^*]^T$, where $\mathbf{Y}_i^*$ represents the data matrix of group g subtracted by its sample row mean vector to make centered data matrix from at each group level.

     $$\hat{\Gamma} = (N-k)^{-1} \sum_{g=1}^{k} \sum_{i=1}^{n_g} y_{gi}^* {y_{gi}^*}^T,$$

     where y_{gi}^* is i^{th} column in $\mathbf{Y}_g^*$, that is $T \times 1$ centered vector. Then

(b) Estimate pooled sample correlation matrix,

$$\hat{\Gamma}_{\omega}(s_t, s_l) = \hat{\Gamma}(s_t, s_l)/\sqrt{\Gamma(s_t, s_t)\Gamma(s_l, s_l)},\ t, l = 1, ..., T.$$

2. Estimate positive eigen values λ_r^{ω}, orthonormal eigen functions $\phi_r^{\omega}(s)$, $r = 1, 2, ..., m$ of $\gamma_w(s, t)$, $\mathbf{U}$, and $\mathbf{h}(s)$.

   (a) Find eigenvalues and eigenvectors of $\hat{\Gamma}_{\omega}(s, t)$ using singular value decomposition (SVD). Let ρ_1^{ω},...,ρ_T^{ω} be all the decreasing-ordered eigenvalues and $\mathbf{u}_1^{\omega}$,...,$\mathbf{u}_T^{\omega}$ be the corresponding normalized eignevectors. Define $w = L/T$ where L is the length of interval τ. Then eigenvalue $\hat{\lambda}_r^{\omega} = w\rho_r$ and the discrete approximation of eigenfunction $\Phi_r^{\omega} = w^{-1/2}\mathbf{u}_r^{\omega}$.

   (b) Estimate m from scree plot by finding an elbow point close to zero.

   (c) Find $\mathbf{U}$ from the following SVD,

$$\mathbf{I}_k - \mathbf{b}\mathbf{b}^T = \mathbf{U}\begin{pmatrix} \mathbf{I}_{k-1} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}\mathbf{U}^T,$$

   where $\mathbf{b}$ defined in (2.17).

   (d) Obtain $k \times T$ matrix $\mathbf{H} = [\mathbf{H}_1, ..., \mathbf{H}_k]^T$, where $T \times 1$ vector $\mathbf{H}_i$ is the the discrete approximation of the i^{th} element of $\mathbf{h}(s)$, $\mathbf{H}_i(s_t) = \sqrt{n_i}(\bar{y}_{i\cdot}(s_t) - \bar{y}_{\cdot\cdot}(s_t))/\sqrt{N\hat{\Gamma}(s_t, s_t)}$

3. Estimate δ_r^2, $\hat{\delta}_r^2 = \|(\mathbf{I}_{k-1}, \mathbf{0})\mathbf{U}^T\mathbf{H}\mathbf{u}_r^{\omega}/T\|^2$, $r = 1, 2, ..., m$, and compute $\hat{\delta}_{\lambda}^2 = \sum_{r=1}^{\hat{m}} \hat{\lambda}^{\omega}\hat{\delta}_r^2$

4. Compute $\hat{\delta}^2 = \hat{G}_{fSNR}$

5. Generate a large sample of $\hat{T}_0^* = (k-1)^{-1} \sum_{r=1}^{\hat{m}} \hat{\lambda}_r A_r,\ A_r \overset{i.i.d.}{\sim} \chi^2_{k-1}$ and Z from $N(0,1)$, and find $\hat{T}_0^*(\alpha)$. Compute the empirical power of a test.

- **Estimation of power of $\mathcal{F}^*$-test** (2.18)

1. Obtain the centered data matrix $\mathbf{Y}^* = [\mathbf{Y}_1^*, ..., \mathbf{Y}_k^*]^T$

2. Find pooled pointwise sample standard deviation $\hat{\sigma}(s_t),\ t = 1, ..., T$, by computing sample standard deviation at each grid s_t.

3. Estimate positive eigenvalues λ_r, orthonormal eigenfunctions $\phi_r(s),\ r = 1, 2, ..., m$, of $\gamma(s,t)$, and $\mathbf{d}(s)$.

   (a) Estimate eigenvalues and eigenfunctions of $\gamma(s,t)$ using the SVD $\mathbf{UDV^T}$ of $\mathbf{Y}^*$. Let ρ_1^*,...,ρ_T^* be all the decreasing-ordered eigenvalues and $\mathbf{v}_1$,...,$\mathbf{v}_T$ be the corresponding normalized eignevectors of $\mathbf{Y}^*$. Then eigenvalue $\hat{\lambda}_r = (N-k)^{-1} w \rho_r^2$ and the discrete approximation of eigenfunction $\Phi_r = w^{-1/2}\mathbf{v}_r$.

   (b) Estimate m from scree plot by finding an elbow point close to zero.

   (c) Obtain $(k-1) \times T$ matrix $\hat{\mathbf{D}} = [\mathbf{D}_1, ..., \mathbf{D}_{k-1}]^T$, where $T \times 1$ vector $\mathbf{D}_i$ is the the discrete approximation of the i^{th} element of $\mathbf{d}(s)$, $\mathbf{D}_i(s_t) = (\bar{y}_{i\cdot}(s_t) - \bar{y}_{k\cdot}(s_t))$

4. Estimate $\delta_r^2,\ \ \hat{\delta}_r^2 = \|(\Omega^{1/2}\mathbf{D}\mathbf{v}_r/T\|^2,\ r = 1, 2, ..., m$, where $\Omega^{1/2}$ =diag($\sqrt{n_1/N}$, ..., $\sqrt{n_k/N}$)

5. Compute $\hat{\delta^{*2}} = \hat{G}^*_{fSNR} \cdot \|\hat{\sigma}(\cdot)\|^2$

6. Generate a large sample of $\hat{T}^* = (k-1)^{-1}\sum_{r=1}^{\hat{m}} \hat{\lambda}_r A_r,\ A_r \overset{i.i.d.}{\sim} \chi^2_{k-1}$ and Z_{qr} from $N(0,1)$ and find $\hat{T}^*(\alpha)$.

Remark. Our empirical finding is that the choice of m, the number of positive eigenvalues in correlation or covariance matrix, does not make a critical impact on power estimation. We suggest to find an elbow point via scree plots and choosing m around them does not make a big difference.

Chapter 3

Robust Probabilistic Classification for Irregularly Sampled Functional Data

3.1 Introduction

A typical diagnosis process in medical data analysis is the following: Collect the measures relevant to particular medical condition as well as the outcome for each patient as training data. Train the classification tool to predict the outcomes of new inputs based on recorded information. The clinical expert makes a diagnosis by combining prediction result with other clinical factors, such as medical history or other related measures, of a patient. As another diagnosis problem, the classification tool can be utilized during surgery to find an area with abnormal tissues or to access tumor margin, for example with biomedical imaging techniques. The classifier can provide a prior information before the image is examined by readers. It can play a pre-screening role to help clinical readers to examine suspicious area in a careful manner. For both examples, reporting a degree of certainty in prediction is more informative than just reporting a single predicted outcome. The subject or area with higher chance of being malignant apparently needs careful inspection compared to one assigned to malignant group but with diagnostic probability close to 0.5. This approach is called probabilistic classification that provides a degree of certainty for the classification result. (Hastie *et al.*, 2009)

In this paper, we build a probabilistic classifier for a functional data, which provides class prediction and probability distribution over a set of classes robust to outlying observations. This work is motivated by the experiment in quantitative ultrasound which collects backscatter coefficients (BSC) curves from two groups of animals with different types of mammary tumor, MAT and 4T1 tumors. The BSC is one example of quantitative ultrasound measurement which offers potential for safe noninvasive diagnosis and derived in the form of functions depending on the microstructural property of target region. Park and Simpson (2017+) studied statistically distinct behaviors in BSC functions over different types of tumor, and an immediate question we may have correspondingly is, whether the functional data classification method can diagnose the future observations into the correct classes. Especially providing stable and informative posterior probabilities to be assigned to each class is of interest in terms of diagnostic purpose. However motivating data has a challenging structure as introduced in Chapter 1, and this type of irregular data should be considered to build classifier.

The extensive tools are developed for classification of functional data, where one assigns a group membership to a new functional form input. Existing methods can be categorized into three subgroups based on their underlying approaches; Regression-based, Density-based, Algorithm-based. Regression approach employs generalized linear models using functional curves as predictors with roughness penalties for regularization. (James, 2002; Muller, 2005; Muller and Stadtmuller, 2005; Goldsmith *et al.*, 2011) Other two approaches perform classification mostly based on dimension reduction to low rank space or via functional principal component (FPC) analysis exploiting a data-driven eigenbasis to represent high-dimensional data on finite dimensional feature space. (Hall, Posit and Presell, 2001; James 2001). Despite different approaches with large number of methods, the robustness to outliers is not seriously considered in our knowledge. Indeed FPC is not robust against outliers because it involves second order moments. Although robust versions of FPCA are proposed (Bali et al., 2011; Boente and Barrera, 2014), they are developed on regular structure. Also most articles assess the results in terms of label prediction accuracy based on misclassification rate, whereby all misclassified observations receive equal weights in evaluation. However, it is fair to impose more penalties in being very confident but incorrect prediction. Furthermore classifiers are commonly built under i.i.d. functions that can be restrictive in medical or biological application where each subject is repeatedly measured in general.

To complement this weakness and to build robustness, Zhu *et al.* (2012) proposed a robust classification method based on the functional mixed model framework in the wavelet space by allowing potentially heavier tails for particular wavelet coefficients. It thereby can handle multiple correlated functions with robust model to outlying functions. However, it does not fit to our data because wavelet approach is more suited for high-dimensional and regular functional samples under stationary periodic characteristic.

The main contribution of this paper is to develop a robust probabilistic classifier based on semiparametric mixed effects model with robust tuning parameter and Bayes rule. Shi, Weiss & Taylor (1996) and Rice & Wu (2000) suggest to model individual curves as spline functions with random coefficients so that covariance function is estimated by approximating the random effects. The key of our method is to impose heavy-tail distribution assumption with robustness parameter ν on random coefficients. This approach enables to fit a robust model for unequally and sparsely collected samples with very flexible spatial covariance structure. Then, given the density of each class, classify according to the largest conditional probability of the class label following Bayes rule. We gain a degree of certainty statements from posterior probabilities. It can be extended to multi-level data and used in cluster-level classification by adding cluster-specific random effect terms. Although it seemingly uses high-dimensional raw information, we indeed regularize the functional curves by the truncation inherent in the knot selection on spline functions. The value of our method hinges on its flexibility and computational efficiency. The spline based mixed-model enables to fit the model under various covariance structures not necessarily stationary.

Another goal is to conduct comparative studies on functional classifiers by examining robustness in terms of label prediction as well as class probability aspect. We aim to compare our proposed method with FPC based method, specifically spline based FPC technique proposed by James, Hastie and Sugar (2000), generalized functional linear model of Muller (2005) and functional linear discriminant tool by James and Hastie (2001). Note that three methods are constructed in the L^2 sense or Gaussian assumption. We implement simulation studies and classify real data to evaluate their performances. The finding is that our robust method is likely to be less certain on false prediction with posterior probabilities close to .5 instead of extreme values.

The paper is organized as follows. We introduce robust semiparametrc mixed effects model in Section 2 with model selection issue. In Section 3 we present (i) simulation studies for comparative studies under different scenarios; (ii) real data analysis returning to two ultrasound quantitative datasets and speech recognition data with artificial contamination.The paper concludes with discussion in Section 4 and shows in detail how to build our proposed classifier in `R`

3.2 Methodology

3.2.1 Robust Semiparametric Mixed Effects Model

A nonparametric linear mixed-effects model (Rice and Wu, 2001) takes the form

$$\begin{gathered} Y_i(s) = \sum_{k=1}^{p} \beta_k \bar{B}_{ik}(s) + \sum_{l=1}^{q} \gamma_{il} B_{il}(s) + \epsilon_i(s), \;\; s \in S \\ \boldsymbol{\gamma}_i \sim N_q(0, \Gamma), \;\; \epsilon_i(s) \overset{i.i.d.}{\sim} N(0, \lambda), \end{gathered} \tag{3.1}$$

where $Y_i(s)$ is the value for the ith curve at s, $\{\bar{B}_{ik}(\cdot)\}$ and $\{B_{il}(\cdot)\}$, $k = 1, ..., p,\ l = 1, ..., q$, are basis spline functions on S, $\boldsymbol{\gamma}_i = (\gamma_{i1}, ..., \gamma_{iq})^T$ is q-dimensional random vector, $\epsilon_i(s)$ is noise random variable, both under normal assumption. The covariance structure is nonparametrically modeled through random coefficient γ_{il}.

We propose the robust semiparametric mixed effects model (RSMM) by imposing heavy-tailed distribution assumption on the errors and random effects, especially the t-distribution with the degrees of freedom (df) ν for both. Thus, $\boldsymbol{\gamma}_i \sim t_q(0, \Gamma, \nu),\ \epsilon_i(s) \overset{i.i.d.}{\sim} t(0, \lambda, \nu)$ in (3.1). In practice functional curves are observed over finite number of grid points and let $\mathbf{Y}_i$ be the n_i-dimensional vector of observed response of ith curve collected over s_{it}, $i = 1, ..., N,\ t = 1, ..., n_i$. The responses are not necessarily regularly sampled over S. The model then can be expressed as,

$$\mathbf{Y}_i = \bar{\mathbf{B}}_i \boldsymbol{\beta} + \mathbf{B}_i \boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \quad i = 1, ..., N \tag{3.2}$$

where $\bar{\mathbf{B}}_i$ and $\mathbf{B}_i$ are corresponding $n_i \times p$ and $n_i \times q$ spline basis matrix evaluated at corresponding grid points and $\boldsymbol{\epsilon}$ are independent with distribution $t_{n_i}(0, \lambda\mathbf{I}, \nu)$ where $\mathbf{I}$ is an identity matrix. Following Pinheiro *et al.* (2001), the resulting marginal density is multivariate t-distribution by using a gamma-normal hierarchical structure. That is,

$$t_n(\bar{\mathbf{B}}\boldsymbol{\beta},\ \mathbf{B}_i\Gamma\mathbf{B}_i^T+\lambda\mathbf{I},\ \nu), \tag{3.3}$$

given $\boldsymbol{\gamma}_i \mid \tau \sim N_q(0,\Gamma/\tau)$, $\boldsymbol{\epsilon}_i \mid \tau \sim N_n(0,\lambda\mathbf{I}/\tau)$ and $\tau \sim \text{Gamma}(\nu/2,\nu/2)$, under conditional independence of $\boldsymbol{\gamma}_i$ and $\boldsymbol{\epsilon}_i$ given τ.

The model can be extend to multi-level data, where between curve correlation is modeled through cluster-level random coefficients. Let $\mathbf{Y}_{ij}$ be the n_{ij}-dimensional observed vector of the jth repetition for ith subject, $i=1,...,N,\ j=1,...,m_i$, over n_{ij} grid points on S. The multi-level RSMM can be written as follows,

$$\mathbf{Y}_{ij}=\bar{\mathbf{B}}_{ij}\boldsymbol{\beta}+\tilde{\mathbf{B}}_{ij}\boldsymbol{\delta}_i+\mathbf{B}_{ij}\boldsymbol{\gamma}_{ij}+\boldsymbol{\epsilon}_i,\quad i=1,...,N,\ j=1,...,m_i, \tag{3.4}$$

where $\tilde{\mathbf{B}}_i$ is $n_i\times r$ spline basis matrix to approximate between curve structure with other terms play the same role in (3.2). Similarly its marginal distribution is multivariate t distribution under conditional mutual independence among $\boldsymbol{\delta}_i,\boldsymbol{\gamma}_{ij}$ and $\boldsymbol{\epsilon}_i$ given τ.

3.2.2 Classification Procedure

The model-based Bayes classification rule is to classify a new observation according to the largest conditional probability of the class label by computing the likelihood. Our method uses the density of each group derived from RSMM. We estimate the robust class probability for a new data object $\mathbf{Y}$,

$$P(class=g|\mathbf{Y})=\frac{f_g(\mathbf{Y})\pi_g}{\sum_j f_j(\mathbf{Y})\pi_j}, \tag{3.5}$$

where the density of the gth class follows multivariate-t distribution in (3.3) and π_g denotes prior probability. We call ν the robustness tuning parameter and use the same level of fixed robustness over groups in this paper. Then $\boldsymbol{\theta}_g = (\boldsymbol{\beta}_g, \boldsymbol{\Gamma}_g, \lambda_g)$, $g = 1, ..., G$, are estimated and used via MLE to compute the likelihood of new input. However it can be generalized as $\nu_g,\ g = 1, ..., G$. The effect on the function of factors such as length and weight of tumors in motivating example can be considered by adding terms to fixed mean function. For valid comparison of likelihoods, the same basis functions are used over groups to approximate random trajectories.

This approach can be extended to classify clusters based on repeated measures. Similarly the class probability for multiple correlated curves can be estimated via the marginal distribution in cluster level.

3.2.3 Model Selection

For practical application; (i) the number of basis functions to fit RSMM, and (ii) robustness tuning parameter ν should be determined. First, Rice and Wu (2001) suggest cross-validated log likelihood, AIC and BIC for model selection and empirically show that last two give similar results to those obtained by cross-validation with faster compute. For classification problem, another way is via cross-validated error. However, note that the unstructured Γ has $q(q+1)/2$ different parameters for q-dimensional spline basis and large q may lead poor prediction error due to over-fitting as well as unstable parameter estimates with local maxima. Thus we empirically suggest to set 6 as the maximum number of basis functions, which is usually enough to achieve good performance on examples. To handle this, James (2001) proposed reduced rank mixed effects framework to stabilize the estimation especially under very sparse curve data with individuals with few measurements. In our computation, we use the B-spline basis and equally spaced knots. Also we recommend to use the same basis functions for $\bar{B}(\cdot)$ and $B(\cdot)$, which empirically leads better performance compared to using different splines for each.

Secondly, we suggest to fix ν a priori at some reasonable value to determine the level of robustness instead of estimating from training set. The final classification appears to be relatively robust to any reasonable choice of ν and our empirical studies in Section 3 show that using $\nu = 3$ even achieves good performance even under Gaussian model. Not only to avoid computational burden in estimating ν but also to maximize the power of robustness, we use fixed conservative ν throughout the applications.

3.3 Numerical Studies

We evaluate the performance of our proposed robust classifier, and compare it with other competitors via simulation study and real data analysis. The sensitivity analysis on tuning parameter is implemented.

3.3.1 Simulation Studies

The simulation study is divided into two parts: in the first part, we examine the robustness of functional classifiers by generating data with different types of outliers, grid-level (local), curve-level (global) and in-between outliers. Our interest lies on investigating the behaviors of posterior probabilities, and see if which approach is likely to be less confident on false prediction. Furthermore Gaussian data is generated in order to consider non-contaminated data; in the second part, we closely mimic the motivating example by generating heavy-tailed and irregular but individually dense data.

We consider two groups and let the population for each group consist of trajectories of the process $Y_g(s) = \mu_g(s) + e(s),\ g = 1, 2$, where $\mu_g(s)$ are group mean curves, and

$e(s)$ is random process under heavy-tailed structure with mean 0. In our example, 50 curves for each group are discretized over 101 equally spaced grid points on $[0, 1]$ by adding exponentially correlated process, $cov(\mathbf{Y}_g(s_j), \mathbf{Y}_g(s_k)) = \sigma_e^2 \cdot exp(-|j-k|/d)$ and independent noise at each grid, both under t distribution with 3 df. In our examples, data are generated with parameters for covariance function as $\sigma_e^2 = 1$ and $d = 0.01, 0.2$ or 1, representing low, moderate or high within-curve dependency under mean-shift model, $\mu_1 = 0,\ \mu_2 = \delta$, with $\delta = 0.5, 1.5$ or 2.5, respectively. We vary δ for each corresponding d, because the classifier trained by weak dependency samples achieves better performance with rich information at each grid. Other types of group means, such as smoothed functions via cubic B-splines or mean functions having some points of intersection, give similar result and the shapes of mean functions do not have significant effect on performance. The Figure 3.1 displays different types of heavy-tail behaviors according to the degrees of magnitude of dependency. The weak dependency leads local outlying behavior otherwise strong dependency apparently shows global abnormality. The non-contaminated data is generated under the same structure with $d = 0.2$ but with Gaussian assumption.

For the second part, we artificially make sparse 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. Among complete individual curves, only corresponding functional values on assigned interval are left and the other two parts are missing. We generate the complete data with 150 equally spaced grids, in other words 50 measurements remain in each curve.

Our simulation results are based on 100 replications. For each replication, the number of basis functions are selected via BIC based on 100 training curves with restriction in maximum number of basis functions as 7. We assume RSMM with the same set of basis functions for fixed and random effect terms with $\nu = 3$. The parameters of the model are estimated for each group and the performance is evaluated on an independent 100 curves of test set containing 50 from each group.

Competitors

We compare our method (RSMM) with four competitors denoted as GSMM, FPC, GFLM and FLDA. All competitors are based on Bayes classification rule that can provide conditional probability for each group, second order moments based classifier, and applicable to irregularly and sparse data.

- GSMM represents the model-based Bayes classifier under Gaussian Semiparametric Mixed effects Model in (3.1).
- The Functional Principal Component analysis (FPC) is a key dimension reduction tool employing truncated Karhunen and (Loeve) basis expansions. Hall, Poskitt and Presell (2001) performed classification based on resulting coefficient through either kernel density estimation or quadratic discriminant analysis (QDA). In our study, we extend this approach by applying FPC technique proposed by James, Hastie and Sugar (2000), which can be applicable to sparse data, and perform QDA on obtained coefficients. Internally mixed effects model with splines is assumed for dimension reduction like our underlying model, but it uses reduced rank framework with normally distributed random coefficients. We choose the same number of B-spline basis selected in RSMM model for fair comparison.
- The Generalized Functional Linear Model (GFLM) of Muller (2005) uses regression model with univariate response variable and FPC scores estimated in the sparse situation. It combines non-parametric local linear smoother to estimate

mean and covariance surface and Gaussian distribution assumption in order to estimate predicted FPC scores.

- Functional Linear Discriminant Analysis (FLDA) for irregularly sample curves is introduced by James and Hastie (2001). It performs LDA by projecting functional curves into subspace where the between-class covariance relative to within-curve covariance is maximized. This rank reduce approach also assumes normality to fit the model.

Evaluation

We evaluate the classification results based on two measures, test error and logarithm loss (LogLoss). The test error is prediction error computed on test set, and LogLoss is the negative log-likelihood of the Bernoulli model which provides punishments for being confident about false classification. This loss function often used as an evaluation metric in kaggle competitions (`https://www.kaggle.com/wiki/LogarithmicLoss`),

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

i indicates individual curves in the data, $N = \sum_{g=1}^{G} n_g$, where n_g denotes the number of curves in group g, y_{ig} is 1 if observation i is the member of class g, and 0 otherwise. p_{ig} denotes $p(\text{class} = g|\mathbf{Y})$. The worst possible case is when a single observation is predicted as definitely false with the estimated probability 0, but it is actually true. It adds infinite to the error score and makes every other result pointless. LogLoss quantifies the accuracy based on membership prediction as well as degree of certainty. In practice, upper and lower bound of probabilities are used in calculation to avoid infinity result under extreme probabilities. Specifically we set `eps` as `1-e15` and replace p_{ig} by `min(max(`p_{ig}`,eps), 1-eps)` as kaggle suggested.

Result

Table 3.1 reports the test error and LogLoss under the data inhering moderate within-curve dependency ($d = 0.2$). Obviously GSMM shows bad performance, especially very large LogLoss. To address the behavior of posterior probabilities for large LogLoss, results of RSMM and GSMM from one sample set are illustrated in Figure 3.2. Gaussian assumed classifier provides unrealistically extreme individual probabilities, most of them close to 0 and 1, while heavy-tailed distribution assumed method appears to present more realistic diagnostic probabilities. The misclassified observations in RSMM have relative large posterior probabilities, which is informative to clinician when making a diagnosis.

Figure 3.3 displays result from different types of functional outliers. The GSMM fails in all situations with extreme probabilities, thus the result is not included for visual comparison. The top panels show the results from local outlying data and apparently GFLM falls behind other three. It can be inferred that grid-level outliers make local smoother technique ineffective. The middle and bottom panels illustrate results of moderate and strong dependency and GFLM and FLDA shows slightly larger LogLoss in general compared to RSMM and FPC methods. In terms of test error, all give comparable results, however FPC seems a bit unstable under heavy tailed data with larger variations. For normally distributed data, we can see in Table 3.2 that all competitors and even our method with conservative ν yield very similar result. To sum up, RSMM generally outperforms others under various types of outliers and at the same time, shows good performance under non-contaminated situation.

As the second part, we examine classification performance on data generated sparsely over intervals but densely over grids. Figure 3.4 illustrates the results obtained from weak dependency with $\delta = 1.5$ and it can be seen that RSMM method outperforms others on both measures. The FLDA that originally targets on discriminant analysis for sparse data yields very similar test errors, but ours provide more accurate and informative posterior probabilities with less LogLoss.

3.3.2 Sensitivity analysis on tuning parameters

Figure 3.5 shows the average test errors and LogLoss in the 50 replications on three different scenarios using different robust tuning parameter; exponentially correlated structure with Cauchy, t distribution with 5 df or Gaussian distribution assumption, representing strong, weak or non-contamination, respectively. The results from Cauchy (top panel) and t distribution (middle panel) display increasing trend but comparable performances among small values of ν. For Gaussian data, there is no significant trend or differences among ν. It implies that the use of fixed robust tuning parameter as 3 works reasonably good regardless of the magnitude of contamination. Indeed it makes the classifier efficient by saving computational cost on estimating ν data as well as by fully exploiting the robustness of the model. It also enables to classify the contaminated new input in a robust manner although training set is non-contaminated.

3.4 Data Analysis

We apply our robust probabilistic classifier as well as competitors to two quantitative ultrasound dataset. Both are irregularly collected but dense in each curve. The performance is evaluated through cross validated errors and LogLoss.

3.4.1 Phantom Data

We consider phantom data, namely A1A2. Specifically backscatter coefficient (BSC) curves in A1A2 dataset are scanned from two types of phantom, A1 and A2, in 9 laboratories with respective transducers covering different frequency ranges. Two phantoms are embedded in different size of glass beads, thus they are physically and acoustically distinct. Figure 3.6 presents BSC curves and there is seemingly well-separated pattern in A1A2. Each curve represents averaged BSC curves from each laboratory so noise is considerably diminished. Also the data seemingly do not have suspicious outlying behaviors due to its experimental design scanning target cell in glass bead directly.

We fit RSMM for each group as training step using cubic B-splines with one knot for fixed and random parts with $\nu = 3$ via BIC criteria. Here we use modified GFLM with spline based FPC approach, because original nonparametric modeling turns out to yield unstable result when only one (or a few) curve is observed over certain frequencies. We compute leave-one-out cross validated (l-o-o cv) error and LogLoss.

Table 3.3 displays similar error rates for all classifiers on A1A2 data, but huge differences are found in LogLoss. It is surprise, because we expect similar performances over methods for this clean data. Our finding is that a certain curve in A2 is closely placed near group A1 and all classifiers misclassify this specific one on cross-validation step. It implies the situation when new observation has relatively heavy-tailed noise compared to collected information on training step. Under this plausible situation, our classifier misclassifies that curve but with weak confidence, while others provide very strong degree of certainty for false prediction. The result shows that our proposed classifier is even robust to outlying new observations.

3.4.2 The Mouse and Rat Mammary Tumor Data

We now build a robust probabilistic classifier for the mouse and rat mammary tumor data. As described in introduction, this experiment scans the target tumor in animal invasively using different kinds of transducers and each transducer covers its own frequency range. They scan the same tumor 4 or 5 times, thus multiple functional curves are expected to be correlated. In detail, BSC curves are derived from 13 mice with 4T1 tumor and 8 rats with MAT tumor by scanning them 5 different transducers. From here on, we distinguish 105 different combinations of animals and transducers with new label called as 'setup'. Each setup forms hierarchical structure with 4 or 5 multiple observations being operated by shifting scan lines.

Under this irregular and hierarchical structure, two kinds of classifier can be considered; A single-curve based classification and a multiple-curve based classification that combines information across correlated curves. To do this, we will consider RSMM imposing random coefficients on 'setup' as follows,

$$\mathbf{Y}_{gij} = \bar{\mathbf{B}}_{gij}\boldsymbol{\beta} + \tilde{\mathbf{B}}_{gij}\boldsymbol{\delta}_{gi} + \boldsymbol{\epsilon}_i, \quad g = 1, 2\ , i = 1, ..., n_g,\ j = 1, ..., m_{gi}, \tag{3.6}$$

where $\mathbf{Y}_{gij}$ denotes observed discretized functional sample for jth repetition of ith setup in group g, and δ_{gi} is t distributed random coefficients that approximate both between and within curve covariance structure. It is a simple version of (3.4). Then we build a probabilistic classifier based on marginal likelihood of each 'setup'.

We evaluate the classification results using 10-fold cross validated misclassification error and LogLoss, and compare our single-curve based (RSMM) and multiple-curve based (m-RSMM) classifiers with FPCA, GFLM and FLDA. Table 3.4 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 shows improved result in class prediction by making use of further information between curves. The LogLoss are quite similar for all approaches, that means their degree of certainty on false classification is not confident.

3.4.3 Phoneme Data

We illustrate our proposed method with example in speech recognition. The phoneme data has 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 no missing. The dataset is available at `http://statweb.stanford.edu/∼tibs/ElemStatLearn/` and details are found in Hastie *et al.* (2015). In this paper, we artificially make data as irregular form by discarding half of information for all curves. Specifically we consider two balanced domains [1,128] and [129, 256], then randomly assign each curve to one of two domains to leave only corresponding measures. Figure 3.7 shows a sample of 10 log-periodograms in each phoneme class from artificially contaminated data. Indeed it forms irregular structure but has dense and rich information for each frequency s. Overall curves seem to have heavy-tailed behavior at each s or in covariance function with similar patterns in simulated t-distributed data in Figure 3.1. Also some groups, for example “iy” or “ao”, show distinct trends on different domains, that implies the potential role of basis functions to unify two trends from different frequency domains. To evaluate the classification results, we select 1,000 training samples for learning step and use remaining 3,509 samples to calculate test error and LogLoss. We compare the performance of our methods with FPCA, modified FGLM as in phantom data analysis and FLDA under both original and contaminated datasets.

Table 3.5 displays the average test errors and LogLoss in 50 replications. For each repetition, the number of basis functions for RSMM was selected via BIC under randomly assigned training and the same number splines are used in FPCA and FLDA for fair comparison. We firstly see that FLDA inhering reduced rank model fails in this dense data set. Next our method and two FPC based methods give similar performances for whole dataset, however, apparently RSMM outperforms under contaminated data. Figure 3.8 reveals the robustness and stability of our method with smaller test error and LogLoss as well as less variations under irregular structure. We see the bad performance of FLDA and instability of FPC based methods. It demonstrates the use of RSMM on general irregular functional data.

3.5 Discussion

When each curve is collected over an extremely fine grid, every kth grid points from original set can be used in analysis to relieve computational load. It does not mean the significant loss of information due to autocorrelation within a curve. The mammary tumor data set empirically give the similar result for the subset of data with $k = 2$ or 4. However, we should be cautious when local atypical behavior is detected over specific area.

Our proposed method has the advantage in its flexibility by employing unstructured scale matrix of random coefficients. However, as discussed in Section 2.3, it can be unstable when each curve is very sparse with just a few measurement. The reduced rank mixed effect model approach (James *et al.*, 2000) which has reduced number of parameters for scale matrix can be an alternative.

3.6 Implementation in `R`

Statistical package `R` has generic functions which fit the liner mixed effects model using heavy-tailed distribution for robust estimation. The `heavyLme()` in package `'heavy'` enables to fit RSMM under unstructured scale matrix of random coefficients and fixed df for t-distribution. This function is based on description in Pinheiro *et al.* (2001). Below we illustrate the implementation of RSMM using subset of mammary tumor data. The m-RSMM can be simply extended as we will see below. The nested-model is not considered in our examples. The other relevant predictors can be added in fixed-effect term.

3.6.1 Training step

The data have functional response variable y, corresponding frequency x, group variable 'type' indicating type of tumor, label 'id' with distinct name for each individual curve, 'setup' representing each subject (cluster) repeatedly scanned multiple times. Given degree and the number of knots K, B-spline basis functions are obtained based on equally spaced knots.

```
library(splines); library(heavy)
tumor.gr1<-tumor[tumor$type="4T1", ]; tumor.gr2<-tumor[tumor$type="MAT",
];
knots<-quantile(range(x), seq(0,1,length=K+2))[-c(1,K+2)]
basis<-bs(tumor$x, degree =degree, knots=knots)
basis.gr1<-basis[tumor$type="4T1", ]; basis.gr2<-basis[tumor$type="MAT",
]
```

The fit of model (3.2) with robust tuning parameter ν for each group is given in:

```
fit.gr1<-heavyLme(y~basis.gr1, random= ~basis.gr1, groups= ~id,
      family=Student(df=nu),control = heavy.control(fix.shape = TRUE),
      data=tumor.gr1)
fit.gr2<-heavyLme(y~basis.gr2, random= ~basis.gr2, groups= ~id,
      family=Student(df=nu),control = heavy.control(fix.shape = TRUE),
      data=tumor.gr2)
```

For m-RSMM (3.6), simply replace 'id' by 'setup'. The number of basis functions can be selected via BIC or cross-validated error.

3.6.2 Prediction step

We compute the likelihood of new input based on estimated scale matrix components and estimates of fixed effects. The B-spline basis functions for new input can be obtained from the same degree and K used above. The unstructured scale matrix of random coefficients are estimated as:

```
Gamma.gr1<-cbind(1,new.basis)%*%fit.gr1$theta %*% t(cbind(1,new.basis))
Gamma.gr2<-cbind(1,new.basis)%*%fit.gr1$theta %*% t(cbind(1,new.basis))
```

The marginal scale is obtained by adding independent heavy-tailed grid level noise:

```
Psi.gr1 <-Gamma.gr1 + diag(fit.gr1$scale,nrow(new.basis))
Psi.gr2 <-Gamma.gr2 + diag(fit.gr2$scale,nrow(new.basis))
```

The posterior probability of a new input to be assigned to type '4T1' is computed as:

```
library(mvtnorm)
```

```
loglik.gr1<-dmvt(new, delta=cbind(1,new.basis) %*%
        fit.gr1$coefficients, sigma=Psi.gr1, df=nu, log=TRUE,
        type="shifted")
loglik.gr2<- dmvt(new, delta=cbind(1,new.basis) %*%
        fit.gr2$coefficients, sigma=Psi.gr2, df=nu, log=TRUE,
        type="shifted")
pprob.gr1<-1/(1+exp(loglik.gr1-loglik.gr2))
```

3.7 Figures and Tables

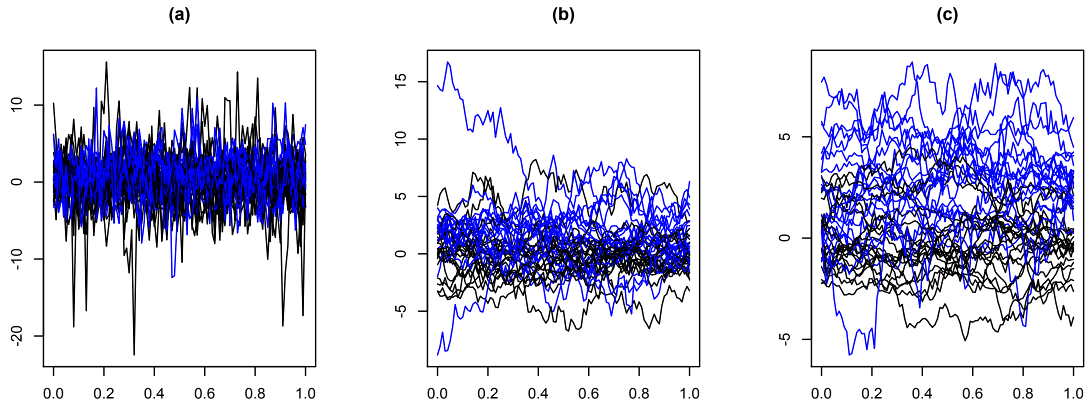


Figure 3.1: Illustration of heavy-tailed sample curves under different magnitudes of within-curve dependency. (a) weak (b) moderate (c) strong.

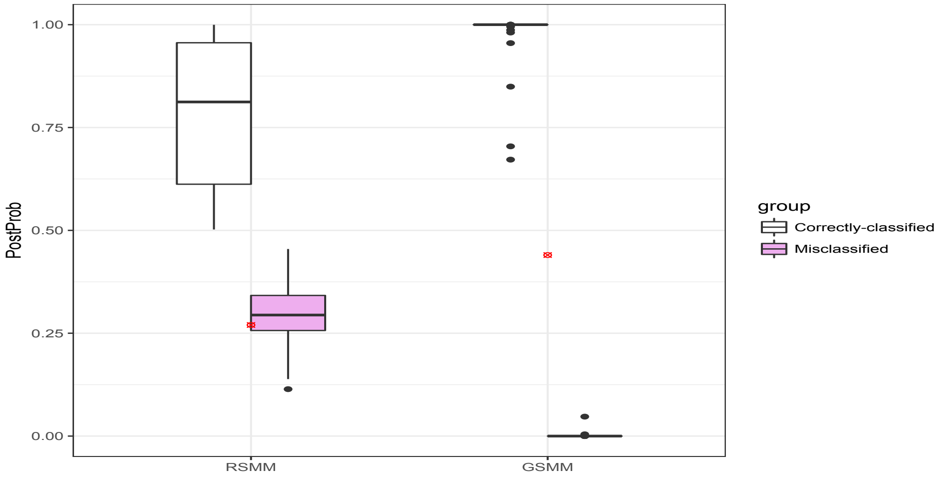


Figure 3.2: Posterior probabilities from RSMM and GSMM. Red dots denote misclassifiction rates

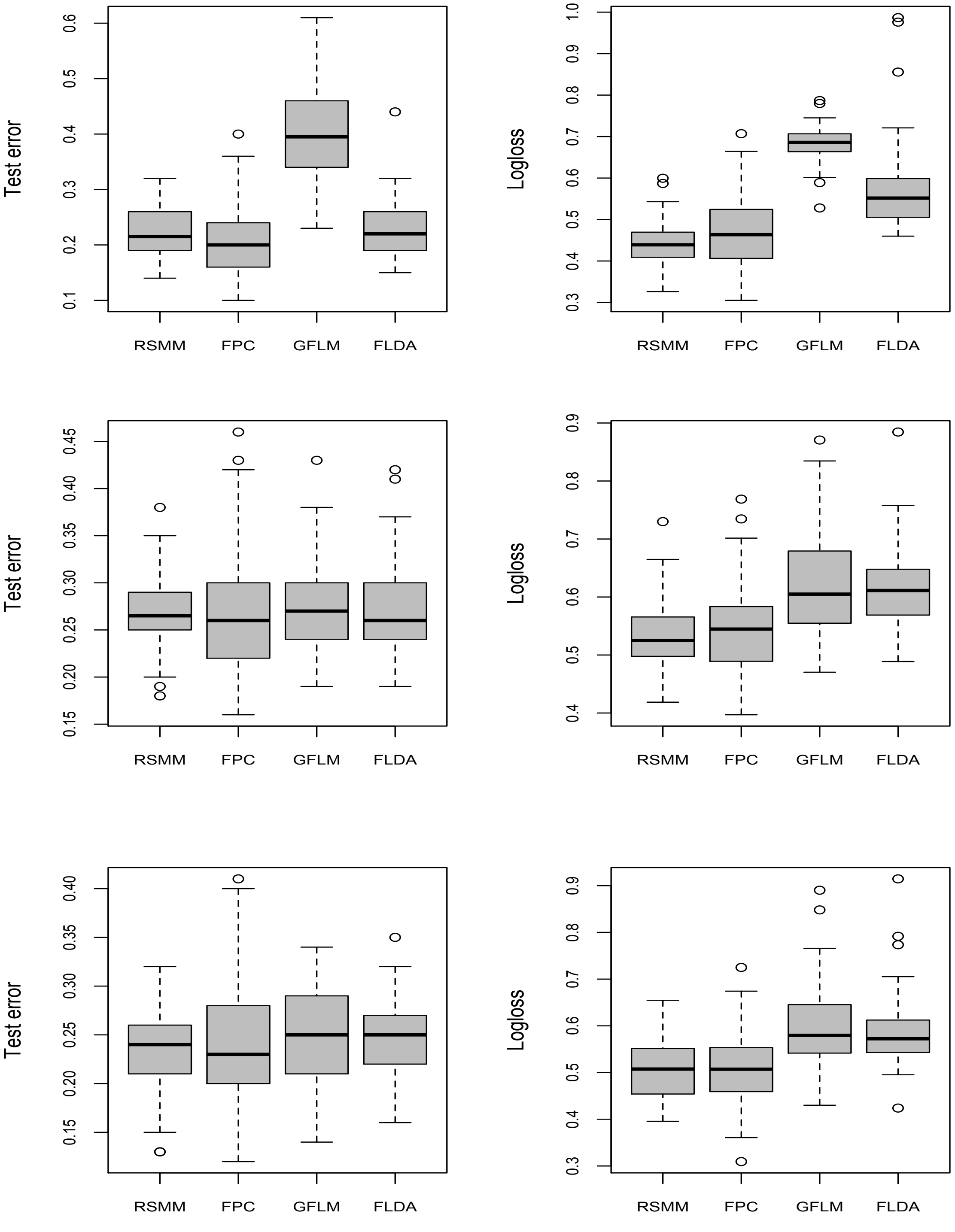


Figure 3.3: Test error (left) and LogLoss (right) under weak (top), moderate (middle), strong (bottom) within-curve dependency structure

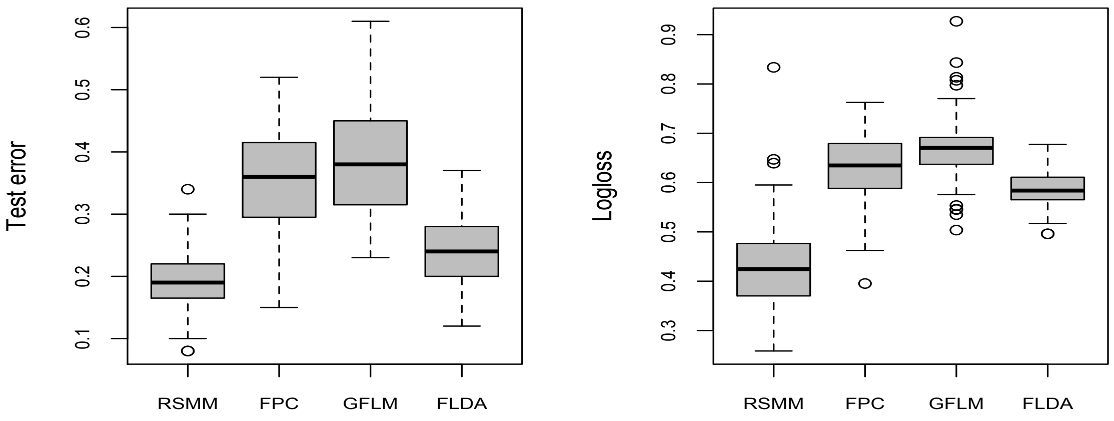


Figure 3.4: Test error and LogLoss under sparse data

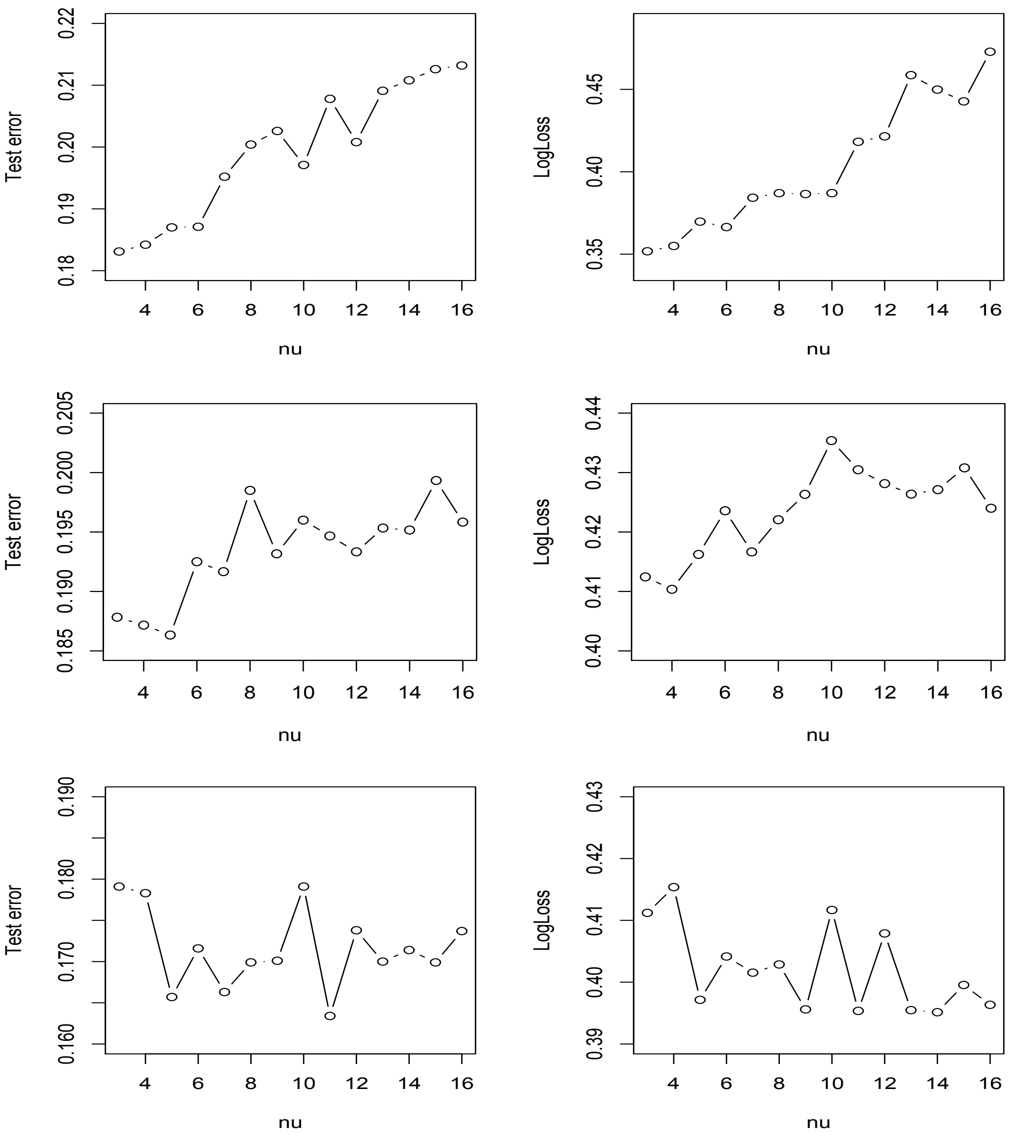


Figure 3.5: Test error and LogLoss under Cauchy (top), t distribution with 5 df (middle) and Gaussian simulated data with different robust tuning parameter ν

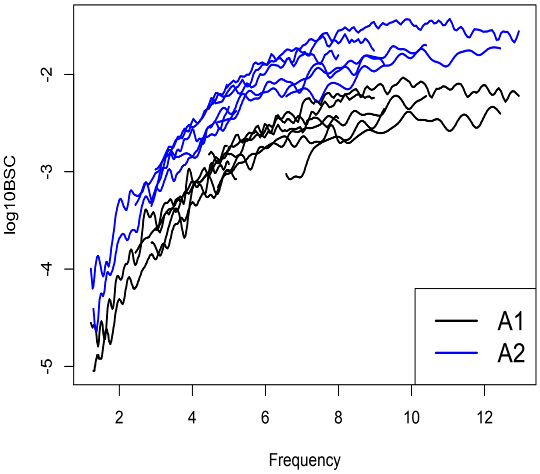


Figure 3.6: Phantom data. A1A2.

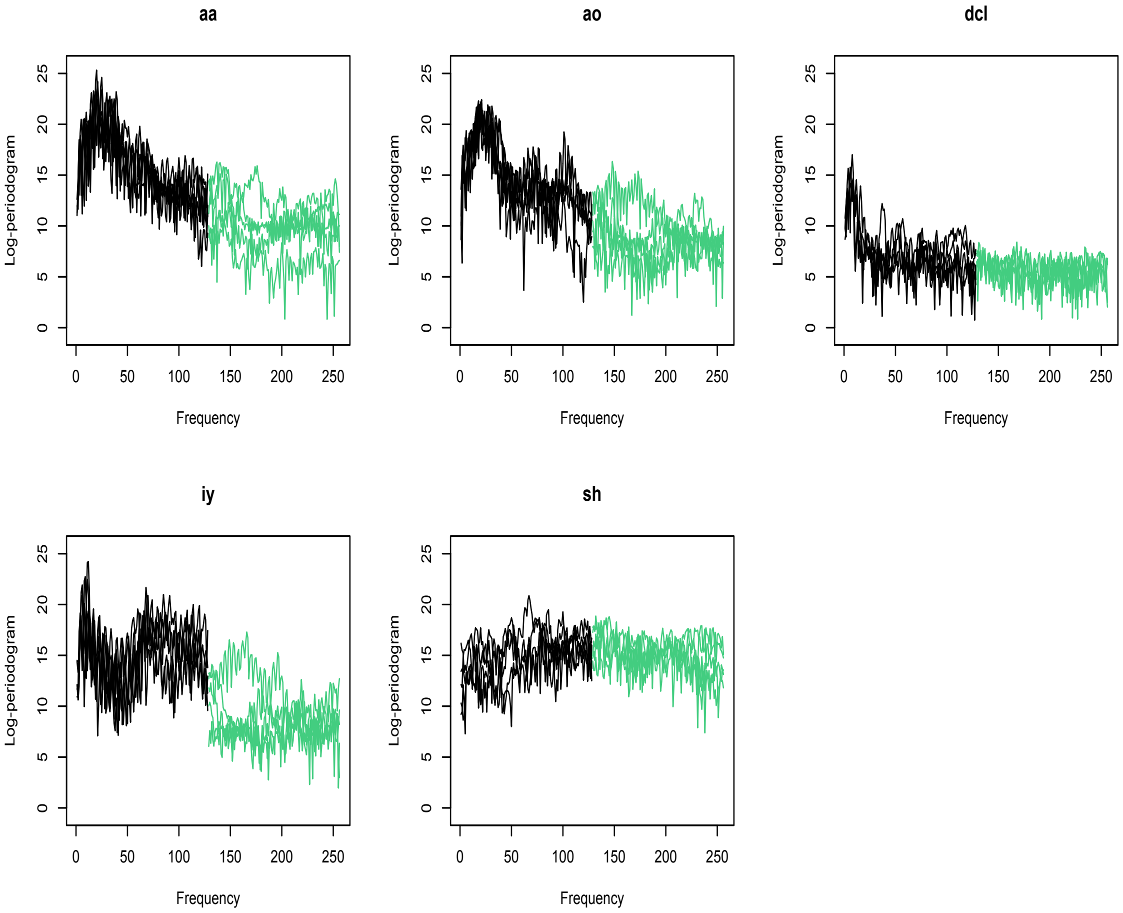


Figure 3.7: A sample of 10 contaminated log-periodograms within each phoneme class. 5 black curves on frequency [1 : 128] and 5 green curves on frequency [129 : 256]

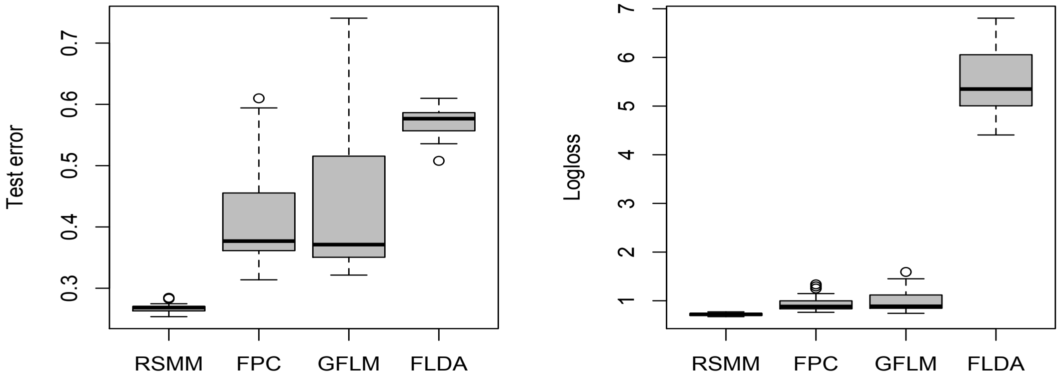


Figure 3.8: Test error and LogLoss in contaminated phoneme example from 50 replications.

Table 3.1: Test error and LogLoss under moderate within-curve dependency

	RSMM	GSMM	FPCA	GFLM	FLDA
error rate	0.33	0.48	0.39	0.23	0.25
LogLoss	0.60	**7.81**	0.66	0.59	0.60

Table 3.2: Test error and LogLoss under Gaussian data

	RSMM	FPCA	GFLM	FLDA
error rate	0.16	0.13	0.16	0.16
LogLoss	0.39	0.30	0.37	0.36

Table 3.3: leave-one-out cv error and LogLoss for phantom A1A2 data

	RSMM	GSMM	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

Table 3.4: 10-fold cross validated error and LogLoss for mammary data

	RSMM	m-RSMM	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 3.5: Mean test errors and LogLoss for contaminated (original) phoneme example

	RSMM	FPCA	GFLM	FLDA
error rate	0.27 (0.10)	0.41 (0.12)	0.43 (0.11)	0.57 (0.57)
LogLoss	0.72 (0.25)	0.94 (0.28)	1.00 (0.28)	5.47 (3.96)

Chapter 4

Functional Central Limit Theorem for Unbalanced Data

While much research on functional data has focused attention on the balanced case, where all functions are observed over the same range, our collaborating research on quantitative ultrasound data analysis across multiple imaging systems (Wirtzfeld *et al.*, 2015) needs to analyze the irregular functional data collected over varying frequency intervals. In the balance case, Zhang and Liang (2014) proposed the GPF test for one-way ANOVA problem for functional data via integrating the usual pointwise F-test. In particular, let $y_{g1}(s), ..., y_{gn_g}(s),\ g = 1, ..., k$ denote k groups of random functions defined over $\mathcal{S}$ and assume that

$$y_{g1}(s), ..., y_{gn_g}(s) \overset{i.i.d.}{\sim} SP(\mu_g, \gamma),\ g = 1, ..., k,$$

where $SP(\mu, \gamma)$ denotes a stochastic process with mean function $\mu(s)$ and covariance function $\gamma(s, t),\ s, t \in \mathcal{S}$. To test the equality of the k mean functions, Zhang and Liang (2014) integrates pointwise F-test over the range of interest and use it as test statistic,

$$T_n = \int_{\mathcal{S}} \frac{SSR_n(s)/(k-1)}{SSE_n(s)/(N-k)},$$

where $N = \sum_{g=1}^{k} n_g,\ SSR_n(s) = \sum_{g=1}^{k} n_g[\bar{y}_{g\cdot}(s) - \bar{y}_{\cdot\cdot}(s)]^2$ and $SSE_n(s) = \sum_{g=1}^{k} \sum_{i=1}^{n_g}$ $[y_{gi}(s) - \bar{y}_{g\cdot}(s)]^2$ with $\bar{y}_{g\cdot} = n^{-1} \sum_{i=1}^{n_g} y_{gi}$ and $\bar{y}_{\cdot\cdot} = n^{-1} \sum_{g=1}^{k} \sum_{i=1}^{n_g} y_{gi}$ representing group and grand mean functions, respectively. They derived the asymptotic null distribution of proposed test statistic that enables to perform fANOVA test with simple calculation without resampling methods. Under certain conditions and the null hypothesis of no group-effect, they showed,

$$T_0 \overset{d}{=} (k-1)^{-1} \sum_{r=1}^{\infty} \lambda_r A_r, \ A_r \overset{i.i.d.}{\sim} \chi^2_{k-1},$$

where $\lambda_r, \ r = 1, ..., \infty$ are the decreasing-ordered eigen values of scaled covariance function $\gamma_w(s,t) = \gamma(s,t)/\sqrt{\gamma(s,s)\gamma(t,t)}$. Also asymptotic power was presented to show the root-n consistency of GPF test.

Here functional central limit theorem (CLT) plays an important role to develop asymptotic behaviors of test statistic. However, note that they consider regular structure where each $s \in \mathcal{S}$ has the same amount of information. We will extend to irregular functional data via missing data formulation. To this end, we will need the following result which provides a CLT for i.i.d. stochastic processes in a separable Hilbert space $\mathbb{H}$ for the complete data.

Theorem 4.1. *Let $X_1, ..., X_n \overset{i.i.d.}{\sim} SP(\mu, \gamma)$ with $\mu(s) \in L^2(\mathcal{S})$ and $tr(\gamma) < \infty$. Define $z_n(s) = n^{-1/2} \sum_{i=1}^{n} \{X_i(s) - \mu(s)\}$, Then,*

$$z_n(s) \xrightarrow{d} GP(0, \gamma),$$

where $GP(0,\gamma)$ denotes gaussian process with zero mean and covariance function γ.

The proof and conditions in a separable Hilbert space $\mathbb{H}$ can be found in Van der Vaart and Wellner (1996) and Hsing and Eubank (2015).

In order to extend this result, we employ missing data framework first introduced in Chapter 2 and sketch the theoretical foundation for irregularly sampled curves.

Specifically we aim to extend functional CLT to unbalanced data under missing data framework described in section 2.2.1,

Corollary 4.1. *Let* $y_i^c, ..., y_n^c \overset{i.i.d.}{\sim} SP(\mu, \gamma)$ *with* $\mu(s) \in L^2(\mathcal{S})$ *and* $tr(\gamma) < \infty$*. Define*

$$z_n(s) = n^{-1/2} \sum_{i=1}^{n} \{y_i^c(s) \mathbb{1}_{[L_i, U_i]}(s) - b(s)\mu(s)\}. \tag{4.1}$$

Under $L_i \overset{i.i.d.}{\sim} F_L$*,* $U_i \overset{i.i.d.}{\sim} F_U$ *with* $P([L_i, U_i] \subset \mathcal{S}) = 1$*,* $\inf_{s \in \mathcal{S}} P(s \in [L_i, U_i]) > 0$ *and* $y_i^c(s) \perp [Li, Ui]$*,*

$$z_n(s) \xrightarrow{d} GP(0, \xi), \tag{4.2}$$

where $b(s) := P(s \in [L_i, U_i])$*, and* ξ *is defined below.*

Here ξ is a covariance function of limiting distribution. Let define the expected product function $E(\mathbb{1}_{[L_i,U_i]}(s) * \mathbb{1}_{[L_i,U_i]}(t)) = c(s,t)$, for $s < t$,

$$c(s,t) := P(L_i \leq s \leq U_i \text{ and } L_i \leq t \leq U_i) = P(L_i \leq s \text{ and } t \leq U_i).$$

Then the covariance function can be written as,

$$\xi(s,t) = c(s,t)\gamma(s,t) + \mu(s)\mu(t)\{c(s,t) - b(s)b(t)\},$$

and the marginal variance if given by,

$$\xi(s,s) = b(s)\sigma^2(s) + \mu^2(s)\{b(s)(1 - b(s))\}.$$

The proposed missing data device of section 2.2.1 reduces the summation of irregularly collected samples to a *modified* i.i.d. summation of random functions defined over common $\mathcal{S}$. This extension will enable to find explicit form of null distribution of fANOVA test-statistic under irregular structure, and to derive asymptotic power function not limited to regular case. By doing so, we will be able to make large sample inferences and also provide theoretical backing for bootstrap method that improves accuracy in small sample sizes.

In future studies, we aim to extend the analysis for test statistics with asymptotic random expressions for irregularly sampled functional data, and to derive asymptotic power analogous to the fANOVA studies of Zhang and Liang (2014). Furthermore the result provides a direction for asymptotic analysis of the robust functional classification method of Chapter 3.

References

[1] Bali, J. L., Boente, G., Tyler, D. E. & Wang, J.-L. (2011). Robust functional principal components: a projection-pursuit approach. *The Annals of Statistics.* 39, 28522882.

[2] Boente, G., Salibian-Barrera, M. (2015). S-estimators for functional principal component. *Journal of the American Statistical Association (Theory and Methods)* Vol. 110, No. 511.

[3] Cuevas, A., Febrero, M., and Fraiman, R. (2004). An ANOVA test for functional data.*Computational Statistics & Data Analysis.* **47**, 111–122.

[4] Cuevas, A., Febrero, M., and Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis.* **51**, 1063–1074.

[5] Edwards, L., Muller, K., Wolfinger, R., Qaqish, B., Schabenberger, O. (2008). An R^2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine* **29**, 6137–6157.

[6] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman & Hall.

[7] Goldsmith, J., Bobb, CM., Crainiceanu, CM., Caffo, BS. & Reich, DS. (2011).

Penalized functional regression. *Journal of Computational and Graphical Statistics*. 20(4): 830851.

[8] Hall, P., Poskitt, DS., & Presnell B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics*. bf 43, 19.

[9] Hastie, T., Buja, A., and Tibshirani, R. Penalized Discriminant Analysis. (1995). *The Annals of Statistics*. 23, 73–102.

[10] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.

[11] Hsing T. and Eubank R. (2015) *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics.

[12] James, GM., Hastie, TJ. and Sugar, C. (2000). Principal component models for sparse functional data. *Biometrika*. 87, 587602.

[13] James, GM., Hastie, TJ. (2001) Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society Series B*. 63, 533550.

[14] Laha, RG., Rohatgi, V. K. (1979). *Probability Theory*, Wiley, New York.

[15] Morris J. S. (2015). Functional regression. *Annual Review of Statistics and its Applications*.2:32159

[16] Mller, H. and Stadtmller, U. Generalized functional linear models (2005). *The Annals of Statistics*. 33, 774805.

[17] Mller, H. (2005). Functional modeling and classification of longitudinal data. *Scandinavian Journal of Statistics* 32, 223240.

[18] Park, Y., Simpson, DG (2017). Effect Size and Power Analysis for Functional ANOVA. *manuscript*

[19] Pinheiro, PC., Liu, C., and Wu, YN. (2001), Efficient Algorithm for RobustEstimation in Linear Mixed-Effects Models Using the MultivariatetDistrib-ution, *Journal of Computational and Graphical Statistics.* 10, 249276.

[20] Ramsay, J. O., and Silverman, B. W. (2005). *Functional Data Analysis*, New York: Springer.

[21] Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics.* **57**, 253–259.

[22] Shen, Q., and Faraway, J. (2004). An F test for linear models with functional responses. *Statistica Sinica.* **14**, 1239–1257.

[23] Van der Vaart, AW. and Wellner, JA (1996). *Weak Convergence and Empirical Processes*, New York: Springer.

[24] Wirtzfeld, L. A., Nam, K., Labyed, Y., Ghoshal, G., Haak, A., Sen-Gupta, E. et al. (2013). Techniques and evaluation from a cross-platform imaging comparison of quantitative ultrasound parameters in an in vivo rodent fibroadenoma model. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **60**, 1386–1400.

[25] Wirtzfeld, L. A., Ghoshal, G., Rosado-Mendez, I. M., Nam, K., Kumar, V., Park, Y. et al. (2015). Quantitative ultrasound comparison of MAT and 4T1 mammary

tumors in mice and rats across multiple imaging systems. *Journal of Ultrasound in Medicine* **34**, 1373–1383.

[26] Yao, F., Muller, H., and Wang, J. L. (2005a). Functional linear regression analysis for longitudinal data. *The Annals of Statistics.* **33**, 2873–2903.

[27] Zhang, J. and Chen, J. (2007). Statistical inferences for functional data. *The Annals of Statistics.* **35**, No. 3, 1052–1079.

[28] Zhang, J. (2011). Statistical inferences for linear models with functional responses. *Statistica Sinica.* **21**, 1431–1451.

[29] Zhang, J. and Liang X. (2014). One-way ANOVA for functional data via globalizing the pointwise F-test. *Scandinavian Journal of Statistics.* **41**, 51–71.

[30] Zhu, HX., Brown, PJ & Morris, JS. (2012). Robust classification of functional and quantitative image data using functional mixed models. *Biometrics.* 68, 12601268.