2015

# DEVELOPMENTS IN NONPARAMETRIC REGRESSION METHODS WITH APPLICATION TO RAMAN SPECTROSCOPY ANALYSIS

Jing Guo
*University of Kentucky*, jgu232@g.uky.edu

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Jing Guo, Student

Dr. Richard Charnigo, Major Professor

Dr. Steven Browning, Director of Graduate Studies

# DEVELOPMENTS IN NONPARAMETRIC REGRESSION METHODS WITH APPLICATION TO RAMAN SPECTROSCOPY ANALYSIS

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Public Health at
the University of Kentucky

By
Jing Guo
Lexington, Kentucky

| | | |
|---|---|---|
| Co-Directors: | Dr. Richard Charnigo | Professor of Biostatistics |
| and | Dr. Bin Huang | Professor of Biostatistics |

Lexington, Kentucky

2015

ABSTRACT OF DISSERTATION

DEVELOPMENTS IN NONPARAMETRIC REGRESSION METHODS WITH APPLICATION TO RAMAN SPECTROSCOPY ANALYSIS

Raman spectroscopy has been successfully employed in the classification of breast pathologies involving basis spectra for chemical constituents of breast tissue and resulted in high sensitivity (94%) and specificity (96%) (Haka et al, 2005). Motivated by recent developments in nonparametric regression, in this work, we adapt stacking, boosting, and dynamic ensemble learning into a nonparametric regression framework with application to Raman spectroscopy analysis for breast cancer diagnosis. In Chapter 2, we apply compound estimation (Charnigo and Srinivasan, 2011) in Raman spectra analysis to classify normal, benign, and malignant breast tissue. We explore both the spectra profiles and their derivatives to differentiate different types of breast tissue. In Chapters 3-5 of this dissertation, we develop a novel paradigm for incorporating ensemble learning classification methodology into a nonparametric regression framework. Specifically, in Chapter 3 we set up modified stacking framework and combine different classifiers together to make better predictions in nonparametric regression settings. In Chapter 4 we develop a method by incorporating a modified AdaBoost algorithm in nonparametric regression settings to improve classification accuracy. In Chapter 5 we propose a dynamic ensemble integration based on multiple meta-learning strategies for nonparametric regression based classification. In Chapter 6, we revisit the Raman spectroscopy data in Chapter 2, and make improvements based on the developments of the methods from Chapter 3 to Chapter 4. Finally we summarize the major findings and contributions of this work as well as identify opportunities for future research and their public health implications.

KEYWORDS: Nonparametric regression, Stacking, Boosting, Raman Spectroscopy, Cancer

Author's signature: Jing Guo

Date: April 14, 2015

# DEVELOPMENTS IN NONPARAMETRIC REGRESSION METHODS WITH APPLICATION TO RAMAN SPECTROSCOPY ANALYSIS

By
Jing Guo

Co-Director of Dissertation: Richard Charnigo

Co-Director of Dissertation: Bin Huang

Director of Graduate Studies: Steven Browning

Date: April 14, 2015

# ACKNOWLEDGMENTS

I would like to express my deepest appreciations to my advisor Dr. Richard Charnigo. This work would not have been possible without his countless hours of mentoring, discussion, reviewing, and editing. I am grateful for his excellent guidance and encouragement boosting my confidence to explore and develop new ideas. I would also like to express the gratitude for my coadvisor Dr. Bin Huang for providing valuable insights and provoking suggestions.

I would like to acknowledge my committee members: Dr. Cidambi Srinivasan (Department of Statistics), Dr. Thomas Tucker (Department of Epidemiology), and my outside examiner Dr. Janelle A. Molloy (Department of Radiation Medicine). In addition, I share the credit of my work with Dr. Abigail Haka (Weill Cornell Medical College), Dr. Ramachandra Dasari (Massachusetts Institute of Technology), and Dr. Maryann Fitzmaurice (Case Western Reserve University) for contributing the Raman spectra data in this dissertation work.

I would like to thank the College of Public Health at the University of Kentucky, for the financial support and research experience during my graduate study, with special thanks to Dr. Bin Huang and Dr. Heather Bush whom I worked with during my research assistantships. I am also grateful for the teaching experience in Department of Statistics at the University of Kentucky. Thanks also go to the faculty, staff and fellow graduate students who continually offered support, advice, and assistance.

Finally, I am indebted to my parents for their endless love and encouragement, and my husband for his support and great help.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# Chapter 1 Introduction

## 1.1 Nonparametric regression methods

### Background

Parametric estimates of mean curves depend on the chosen parametric model. Although they can be suitable for small sample sizes $n$, and are often easy to interpret, parametric estimates might be too restricted within the assumed parametric structure to fit unexpected features or complex features in the mean curve [35]. Furthermore, it is not always clear how to choose a proper form of a parametric estimate for multivariate $X$, because it is hard to visualize the data for multivariate $X$ [31]. In these cases, nonparametric regression or semiparametric regression may be appropriate when an underlying parametric model cannot be identified.

Suppose the regression model has the form

$$Y_i = \mu(x_i) + \epsilon_i \text{ for } i \in \{1, \ldots, n\}, \tag{1.1}$$

where $x_i \in \mathcal{X}$, $Y_i$ are observed responses, and $\epsilon_i$ are random errors with zero mean. In parametric settings, the functional form of $\mu$ is known, and the parameters or coefficients are to be determined. In nonparametric regression, on the other hand, the functional form of $\mu$ is unspecified. Rather, a data driven technique is used to determine the shape of a curve. The modeler determines the amount of local curvature to be depicted in the curve [31]. The goal of nonparametric regression is to directly estimate the regression function $\mu$ instead of estimating parameters [24]. Most methods of nonparametric regression implicitly assume that $\mu$ is a smooth, continuous function. When using some nonparametric regression methods, decisions must be made regarding polynomial order and bandwidth. Such decisions depend on

the presence of local curvature, desired degree of smoothing, and the minimization of some global error criterion for tuning parameter selection [35].

This section describes four forms of nonparametric regression including kernel estimation, local regression, smoothing splines and compound estimation. Then an illustrative example is presented to show practical applications of these nonparametric regression methods.

## Common nonparametric regression methods

The following description pertains to scalar $X$, for simplicity.

### Kernel Smooth

The main idea of the kernel method is to estimate the mean response at every point $x_0$.

The Nadaraya-Watson kernel estimator[6][55][94] is defined as

$$\hat{\mu}(x_0) = \frac{\sum_i K_h(x_0, x_i) Y_i}{\sum_i K_h(x_0, x_i)}, \tag{1.2}$$

where $K_h(x_0, x_i) = \frac{1}{h} K(\frac{x_0 - x_i}{h})$. The function $K$ is called the kernel, and along with the bandwidth $h$, it controls the weight given to the observations $\{x_i\}$ at each point $x_0$ based on their proximity. Two of the popular compactly supported kernels are the Epanechnikov kernel

$$K(u) = \frac{3}{4}(1 - u^2) 1_{\{|u| \leq 1\}}$$

and the tri-cube kernel

$$K(u) = (1 - |u|^3) 1_{\{|u| \leq 1\}}.$$

The kernel function can also be chosen in another form such as Charnigo and Srinivasan(2011) described [18],

$$K(u) = 1_{|u| \leq 1} \sum_{m=0}^{\infty} a_m u^{2m},$$

with various constraints on the $a_m$. The smoothing parameter $h$, also known as bandwidth, controls the size of the neighbourhood around $x_0$. If the bandwidth decreases with the sample size at an appropriate rate, then the kernel estimators are consistent: that is, $\hat{\mu}(x) \xrightarrow{P} \mu(x)$. The mean squared error is $d_\mu(x,h) = E[\hat{\mu}_h(x) - \mu(x)]^2$. As $n \longrightarrow \infty$, $h \longrightarrow 0$, $nh \longrightarrow \infty$, under certain conditions, we have

$$d_\mu(x,h) \approx (nh)^{-1}\sigma^2 c_K + h^4 d_K^2 [\mu^{''}(x)]^2/4,$$

where $\sigma^2 = Var(\epsilon_i)$, and $c_K = \int K^2(u)du$, $d_K = \int u^2 K(u)du$ [36][35]. Suppose that $x_i$'s are drawn from a distribution with density $g(x)$, then the bias, which has the form $h^2\left(\frac{1}{2}\mu^{''}(x) + \frac{\mu'(x)g'(x)}{g(x)}\right)\int u^2K(u)du + o(h^2)$ is increasing in $h$ whereas the variance $\frac{\sigma^2}{g(x)nh}\int K^2(u)du + o(\frac{1}{nh})$ is decreasing in $h$.

**Rate of convergence** Let AMSE denote the asymptotic MSE. We may write denoting constant terms by $C_1$ and $C_2$, respectively

$$AMSE(n,h) = \frac{1}{nh}C_1 + h^4 C_2.$$

Minimizing the expression with respect to $h$ gives the optimal bandwidth $h_{opt} \sim n^{-1/5}$ with mean square convergence $O(n^{-2/5})$ [36]. In the more general case that $\mu^{(J+1)}(x)$ is continuous, where $J$ is a positive integer, the optimal bandwidth is $O(n^{-\frac{1}{2J+3}})$, the corresponding mean square error is $O(n^{-\frac{2(J+1)}{2J+3}})$, and the convergence rate is $O_p(n^{-\frac{J+1}{2J+3}})$. Since the above optimal bandwidth is unavailable due to its dependence on the unknown function $\mu(x)$, two approaches for choosing appropriate bandwidth including cross validation and the generalized cross validation are commonly used [93].

**Cross validation** Define a leave-one-out estimator:

$$\hat{\mu}_{h,-i}(X_i) = \frac{\sum_{j\neq i} K_h(X_i - X_j)Y_j}{\sum_{j\neq i} K_h(X_i - X_j)},$$

and cross validation function is

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{\mu}_{h,-i}(X_i)\}^2,$$

where $\hat{\mu}_{(-i)}$ denotes the estimator of $\mu$ obtained when $(x_i, y_i)$ is excluded. Then the smoothing parameter can then be chosen by minimizing $CV(h)$.

**Generalized cross validation**

Using cross validation sometimes leads to too small an $h$. Generalized cross validation is one type of penalizing function for $CV(h)$ which aims at an asymptotic cancellation of the bias[35](Craven and Wahba, 1979). In the setting of equidistant $X_i$ on the unit interval, the $GCV$ functions can be written as

$$GCV(h) = CV(h)(1 - n^{-1}h^{-1}K(0))^{-2}.$$

The smoothing parameter chosen by $GCV$ is asymptotically optimal.

The Nadaraya-Watson kernel estimator often suffers from bias, both at the boundaries and in the interior when the $x_i$ are not uniformly distributed due to the asymmetric effect of the kernel in these regions. Moreover, it is not infinitely differentiable if the underlying function $K$ is not.

**Local Regression**

In local regression[48], we can fit straight lines locally instead of constants, which could reduce the problems arising from kernel smoothing, therefore local regression modifies kernel smoothing in such a way that the bias is largely eliminated. In local linear regression, a separate weighted least squares problem is solved at each target point $x_0$,

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_i K_h(x_0, x_i)(y_i - \alpha - x_i\beta)^2. \tag{1.3}$$

The estimate is then $\hat{\mu}(x_0) = \hat{\alpha} + x_0\hat{\beta}$. Again, selection of the bandwidth $h$ is critical.

The LLR estimator behaves better at the boundary of the support of $X_i$. Its

asymptotic distribution can be given by

$$\sqrt{nh}(\hat{\mu}(x) - \mu(x) - \frac{1}{2}\sigma_K^2 h^2 \mu''(x)) \xrightarrow{D} N(0, \frac{C_K \sigma^2(x)}{g(x)}),$$

assuming $x$ is not a boundary point. Local polynomial regression extends the linear model to the polynomial model, which could further correct for bias in regions of high curvature, yet at the expense of increased variability. The local regression smooth is not in general infinitely differentiable, and even when differentiable, may not satisfy $\widehat{\mu'(x_0)} = \frac{d}{dx}\hat{\mu}(x)|_{x=x_0}$.

### Smoothing Splines

A spline that could pass close to the observations $\{x_i, y_i\}$ without passing through all of them is called a smoothing spline [62]. In smoothing splines, the target is to find the function $\mu(x)$ with two continuous derivatives that minimizes the penalized sum of squares

$$\sum_{i=1}^{n}\{y_i - \mu(x_i)\}^2 + \lambda \int \{(\mu''(u))\}^2 du. \tag{1.4}$$

The solution turns out to be a natural cubic spline [29]. Speckman(1985)[78] obtained the optimal rates of convergence for smoothing spline estimators under certain conditions. The tuning parameter $\lambda$ plays a role like the bandwidth of local regression in trading off bias for variability. However, $\lambda$ is not immediately interpretable as the size of a neighbourhood.

### Compound estimator and the self-consistency property

Charnigo and Srinivasan (2011) have proposed a compound estimator that has the self-consistency property in that its derivative estimates the derivatives of the mean response function [18]. This is a desirable property when not only the regression curve itself is the target of interest, but also its derivatives. Self-consistency is important if estimates of multiple derivatives are used to make inferences.

**Compound estimator** The first step in defining the compound estimator is to specify pointwise estimators of $c_{j;a} := \mu^{(j)}(a)/j!$ for $0 \leq j \leq J$ and $a \in I_n$, where $I_n$ is a finite subset of $[-1, 1]$. Let $\tilde{c}_{j;a}$ denote these estimators.

Secondly, define a polynomial $\tilde{\mu}_{J;a}(x) := \sum_{j=0}^{J} \tilde{c}_{j;a}(x-a)^j$ for each $a \in I_n$.

Finally, the compound estimator is defined by

$$\mu^*(x) := \sum_{a \in I_n} W_{a,n}(x)\tilde{\mu}_{J;a}(x),$$

and for $1 \leq j \leq J$, $\mu^{(j)}(x)$ is estimated by

$$\frac{d^j}{dx^j}\mu^*(x) = \sum_{a \in I_n} \sum_{k=0}^{j} \binom{j}{k} \frac{d^k}{dx^k}\tilde{\mu}_{J;a}(x) \frac{d^{j-k}}{dx^{j-k}} W_{a,n}(x),$$

where $W_{a,n}(x)$ can be defined as $\frac{\exp[-\beta_n(x-a)^2]}{\sum_{c \in I_n} \exp[-\beta_n(x-c)^2]}$, in which $\beta_n$ is a nondecreasing sequence of positive real numbers.

The compound estimator $\mu^*(x)$ is self-consistent by construction and it can also achieve an almost optimal convergence rate arbitrarily close to $O_p(n^{-\frac{J+1}{2J+3}})$. Moreover, the derivatives $\frac{d^j}{dx^j}\mu^*(x)$ achieve almost optimal convergence rates arbitrarily close to $O_p(n^{-\frac{J+1-j}{2J+3}})$.

**Self consistency property** An estimator $\widehat{\mu(x)}$ and companion estimators $\widehat{\frac{d^j}{dx^j}\mu(x)}$, $j \in \mathcal{J}$, are self-consistent if $\frac{d^j}{dx^j}\widehat{\mu(x)}$ exists and equals $\widehat{\frac{d^j}{dx^j}\mu(x)}$ for every $j \in \mathcal{J}$, where $\mathcal{J} \subset \mathcal{N}$, and $\mathcal{N}$ denotes the set of natural numbers[18].

A spline smooth in the form of a degree $J$ piecewise polynomial is self-consistent with $\mathcal{J} = \{1, 2, \ldots, J-1\}$. However, the results may not be stable (i.e. may exhibit considerable variability) when approaching the maximum number of differentiations that can be performed on $\widehat{\mu}(x)$[18].

## Application to real dataset

Now consider an illustrative example using the above different nonparametric regression methods. The dataset is from the 2006 National Health and Nutrition Examination Survey(NHANES) study on the serum levels of two variables for 3026 adult women: triglycerides(TRG), and apolipoprotein B (APB). Apolipoprotein B (Apo B) is the dominant protein constituent of LDL. It is known that elevated Apo B is associated with increased risk of vascular disease[3]. The objective is to identify the relationship between serum levels of APB and TRG in order to predict serum APB level by TRG, since lab tests for TRG are more commonly conducted than lab tests for APB in the general population. Since the covariate TRG is not close to uniformly spaced, this example will illustrate the utility of a nearest neighbour approach to bandwidth/tuning parameter specification.

First of all, we can test whether there is a significant difference between the results of non-parametric regression and parametric regression. A linear (or quadratic) relationship between APB and TRG differs significantly from nonparametric regression results if the fitted line (or parabola) lies at least partially outside the confidence bands accompanying the non-parametric regression results. This idea is illustrated in Figure 1.1. Linear regression (or quadratic regression) would not be an appropriate choice to estimate the APB by TRG relationship because the line (parabola)falls partially outside the confidence bands from local regression. Locfit.raw function was employed to perform local regression, in which the nearest neighbour parameter was set to 0.1.

Figure 1.2 shows the estimated mean responses of APB level using nonparametric regression methods including kernel smooth, local regression, smoothing spline and compound estimation. In the kernel regression, Gaussian kernel function was used, and bandwidth was chosen to be 28 by cross validation. In local regression, the degree

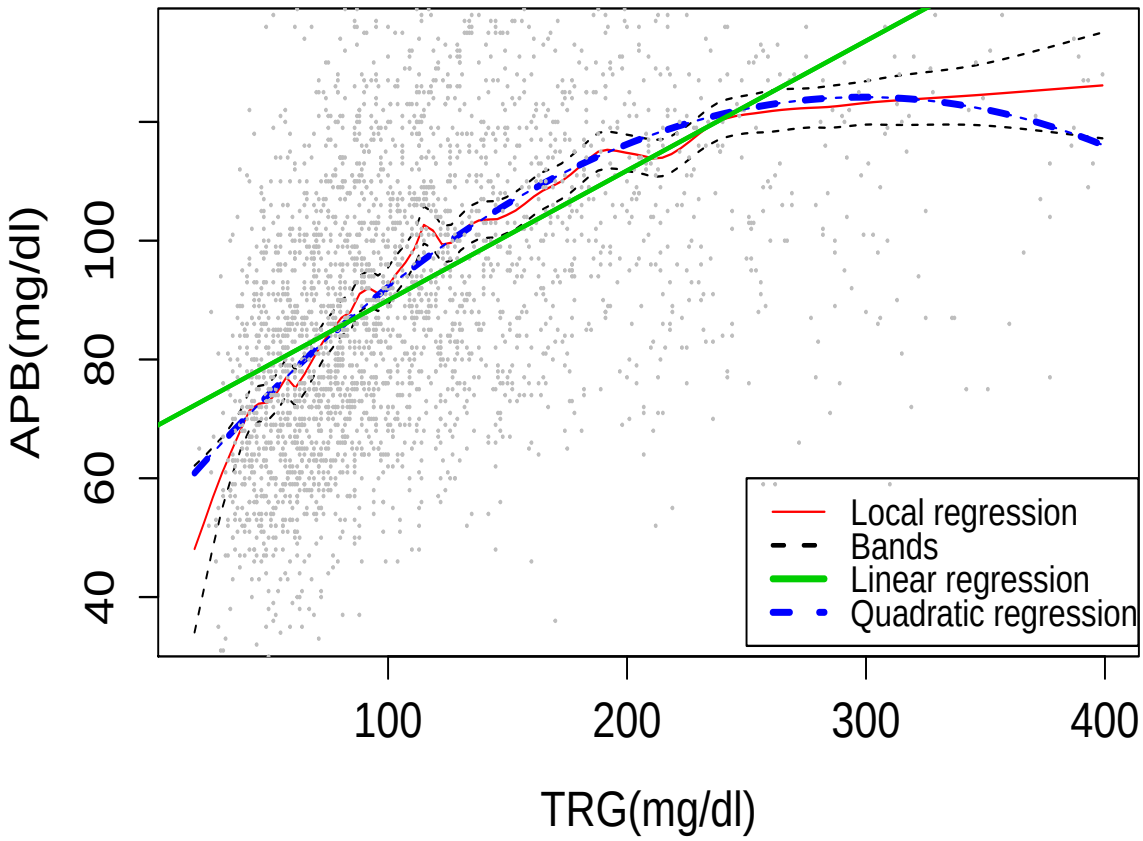


Figure 1.1: Estimated mean APB level using parametric models versus confidence bands from Local regression. Shown are estimated mean responses for APB level by local regression and parametric methods. Parametric results (line and parabola), falling partially outside the confidence bands of local regression, are not satisfactory to demonstrate the relationship.

of local polynomials was set to be 1, and the tricube weight function was used. The smooth. spline function was used to perform spline smoothing, and the smoothing parameter was chosen by generalized cross validation to be 1. The nearest neighbours fraction and convolution parameter with local regression pointwise estimators were selected to be 0.18 and 0.25 by generalized cross validation. Overall the relationship between APB and TRG appears to be approximately a concave function. The fitted smoothing spline is the smoothest curve among the four. Kernel regression and local regression fits appear less smooth, and are noticeably more oscillatory in regions

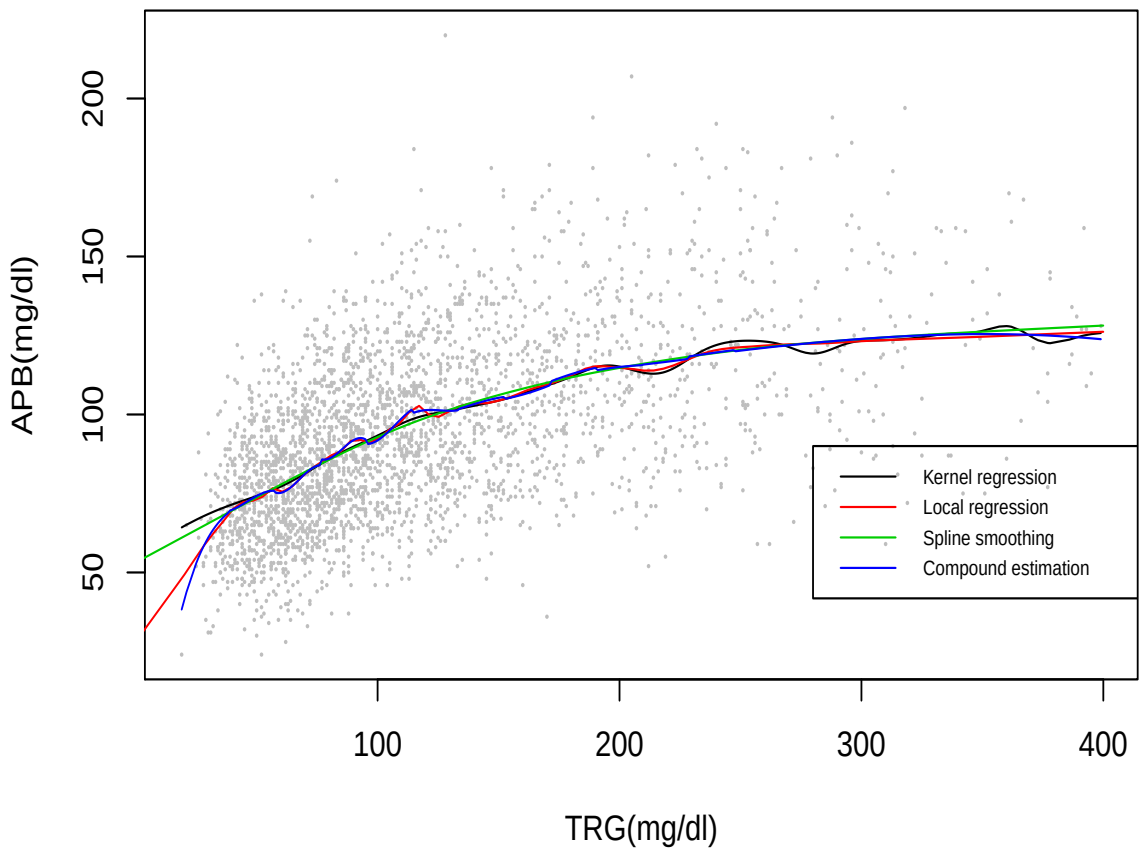


Figure 1.2: Estimated mean responses of APB level using different nonparametric regression methods

where TRG often falls compared with smoothing spline fit in the same regions. The compound estimation fit appears intermediate.

The estimated first derivatives of the mean response using different nonparametric regression methods are shown in Figure 1.3. Locfit.raw function was used to generate the estimated first derivative of the mean response using local regression, since this is not equal to the derivative of the estimated mean response. Due to lack of self-consistency, in local regression, zeros of the estimated first derivatives do not occur at local extrema. For example, there is an estimated peak at which TRG is 120 in local regression. Yet the estimated derivative when TRG is equal to 120 is negative. Regarding kernel smoothing, using a fixed bandwidth instead of a nearest

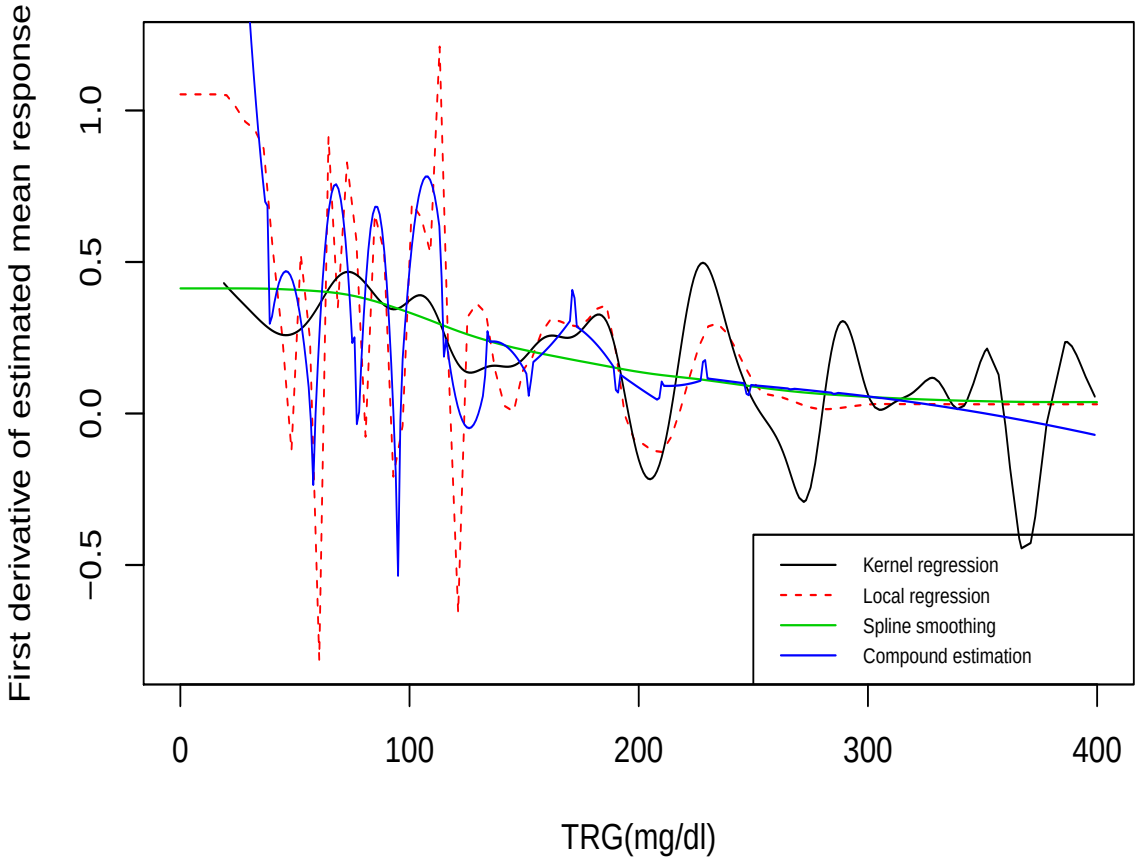


Figure 1.3: Estimated first derivatives of mean responses using different nonparametric regression methods

neighbour approach may have been anticipated to result in too much smoothing for small TRG(high bias, low variance), and in too little smoothing for large TRG(high variance, low bias), although from Figure 1.3, this appears not to have been the case, because the fitted first derivative from kernel smoothing is oscillatory across all TRG values. With compound estimation, the fitted first derivative is rather oscillatory for small TRG but less so for large TRG, exhibiting a similar pattern to local regression but less pronounced. As expected, the fitted first derivative by spline smoothing is very smooth.

Notably, small differences in fitted mean responses translated into larger differences in fitted first derivatives. We emphasize that compound estimation, like kernel

smoothing and spline smoothing, enjoys self-consistency so that derivatives of mean response may be estimated by differentiating estimates of the mean response. One caveat for all four methods in this example is that they have not incorporated the NHANES survey weights.

## 1.2 Pattern Recognition

Pattern recognition has been an active area of research in a wide range of fields, such as diagnosing diseases, recognizing dangerous driving conditions, identifying which customers will be spotting good opportunities on the financial markets, classifying galaxies by shape, and identifying handwritten symbols[64]. Recently nonparametric regression methods have also been applied in pattern recognition to characterize the nanoparticles[15]. Different from traditional statistical methods, which usually emphasize hypothesis testing, in pattern recognition(also known as statistical learning or data mining), the goal is to use available inputs to predict or classify an output[28]. In more traditional parlance, inputs and outputs correspond to independent and dependent variables respectively. This is more exploratory and associated with "past experience or knowledge"(Kennedy 1997)[42].The framework in its simplest form is as follows. Each object(person or other experimental unit) gives rise to certain measurements which together form the input which we call the feature vector $X$. The output will be from among a fixed number of categories or classes, say $1, \ldots, K$. The task is to classify an object to one category on the basis of the observed value $X = x$. Stacking[95] and boosting [68] are two learning algorithms which we plan to apply in nonparametric regression settings.

## Stacking

Stacked generalization, also known as stacking, combines multiple classifiers using another classifier, often referred to as a meta-level classifier, to give improved prediction accuracy by learning the way that their multiple classifiers' outputs correlate with the true class[95]. Generally cross-validation and least squares are used under non-negativity constraints to determine the coefficients in the combination[8]. Stacking can be applied recursively, which generates a hierarchical combiner.

The aim of stacking is combining multiple classifiers generated by different learning algorithms $L_1, \ldots, L_N$ on a single dataset $S$, which consists of examples $s_i = (x_i, y_i)$, i.e., pairs of feature vectors ($x_i$) and their classifications ($y_i$). Its framework can be described as follows [22]: in the first phase, a set of base-level classifiers $C_1, C_2, \ldots, C_N$ is generated, where $C_i = L_i(S)$. In the second phase, a meta-level classifier learns to combine the outputs of the base-level classifiers. To generate a training set for learning of the meta-level classifier, a cross validation procedure is applied [84]. Using meta-level classifier, stacking infers reliable and unreliable base classifiers. Using output probabilities corresponding to each label can improve the performance of stacking [1].The most important issues in stacking are the choice of the features and the algorithm for learning at the meta-level [22].

The idea of stacking is also applicable to a continuous response variable. To illustrate stacking in a non-parametric setting, suppose that a scalar response variable $C$ and a vector of predictor variables $X$ are governed by $C_i = \mu(X_i) + \epsilon_i$ for $1 \leq i \leq n$, where $\mu$ is an unknown mean function and $\epsilon_1, \ldots, \epsilon_n$ are error terms. In this scenario, letting $\widehat{\mu_1}, \ldots, \widehat{\mu_J}$ denote candidate estimators of $\mu$, one might stack the candidate estimators into a final estimator of the form $\sum_{j=1}^{J} \widehat{w_j}\widehat{\mu_j(X)}$, where $\widehat{w_1}, \ldots, \widehat{w_J}$ are chosen to minimize the cross-validated sum of squares $\sum_{i=1}^{n}(C_i - \sum_{j=1} w_j \widehat{\mu_j^{-i}(X_i)})^2$, where superscript $-i$ denotes a calculation excluding observation $i$.

## Boosting

Boosting is one of the most important recent developments in classification methodology [68]. Boosting uses a set of weak learners to create a single strong learner[69]. A weak learner is a classifier which is only slightly correlated with the true classification, whereas a strong learner is a classifier that is arbitrarily well-correlated with the true classification[70].

The idea of boosting is applying a classification algorithm sequentially to reweighted versions of the training data and then taking a weighted majority vote of the sequence of classifiers[27]. For many classification algorithms, this simple strategy results in dramatic improvements in performance. Some well-known boosting algorithms include AdaBoost algorithm (Schapire)[68], functional gradient descent(FGD)(Breiman) and $L_2$ Boosting algorithm[10].

To illustrate boosting, suppose that a binary scalar response variable $C \in \{-1, 1\}$ is to be predicted from a vector $X$ and that sample data $(X_1, C_1), \ldots, (X_n, C_n)$ have been acquired for the purpose of developing a prediction rule. Let $\mu_1(X)$ denote the initial classification rule, define weights $w_i := \frac{\exp(\alpha_1 \mathtt{I}\{C_i \neq \mu_1(X_i)\})}{\sum_{j=1}^{n} \exp(\alpha_1 \mathtt{I}\{C_j \neq \mu_1(X_j)\})}$ for $1 \leq i \leq n$ , where $\alpha_1 := -\text{logit}(\sum_{i=1}^{n} \mathtt{I}\{Ci \neq \mu_1(X_i)\}/n)$ and $\mathtt{I}\{\}$ is an indicator function that equals 1 when its argument is true and 0 otherwise. The sample data are then assigned the weights $w_1, \ldots, w_n$ during the development of a new prediction rule, which we label $\mu_2(X)$. Thus, $\mu_2(X)$ gives greater weight to observations that were misclassified by $\mu_1(X)$. The performance of $\mu_2(X)$ is likewise used to define another prediction rule $\mu_3(X)$ that gives greater weight to observations that were misclassified by both $\mu_1(X)$ and $\mu_2(X)$. This process continues for a specified number of iterations. At the last iteration, observations have differing weights according to how frequently they were misclassified by the various intermediate prediction rules, and one may define an overall prediction rule by appropriately combining the intermediate prediction

rules(e.g., by factoring a majority vote).

## Nonparametric methods applied in pattern recognition

How does pattern recognition relate to nonparametric regression? The basic idea is illustrated as follows. Suppose that there exists a variable $C \in \{-1, 1\}$ representing some characteristic of an object. Suppose that the relationship between two other variable $X$ and $Y$ is governed by $Y_i = \mu_1(x) + \epsilon_i$ for all subjects with $c = 1$ and by $Y_i = \mu_{-1}(x) + \epsilon_i$ for all subjects with $c = -1$. If we need to classify an object as $c = 1$ or $c = -1$, one way to do so is to estimate the mean response function relating $Y$ to $X$ for that object, let this estimate be denoted $\hat{\mu}$. Then we compare $\hat{\mu}$ to $\mu_1(x)$ and $\mu_{-1}(x)$. If $C = 1$ then $\mu(x) = \mu_1(x)$ and if $C = -1$ then $\mu(x) = \mu_{-1}(x)$. If $\hat{\mu}(x)$ is "closer" to $\mu_1(x)$, then the object is classified as $c = 1$, otherwise the object is classified as $c = -1$. However, one must define "close". We do this in one of two ways, as described next.

### $L^1$ method

The $L^1$ distance between two functions $f_1$ and $f_2$ is defined as $\int |f_1(x) - f_2(x)|dx$. Besides having been employed as a tool to estimate regression and density functions nonparametrically[34][92][83] , $L^1$ distance may also be applied in pattern recognition. This idea has previously been employed by Charnigo et al (2007) for nanoparticle characterization[15] and by Charnigo, Hall, and Srinivasan(2011) for identification of a chemical compound[18].

The basic idea is as follows. Consider the example of using Raman spectroscopy to diagnose breast cancer. Let $\widehat{v_i}(x)$ denote the Raman spectrum for person $i$ to be classified, and $v_N(x)$ and $v_A(x)$ denote known normal and abnormal Raman spectra. One can classify person $i$ as normal or abnormal based on the closeness of $\widehat{v_i}(x)$ to each of the known Raman spectra with respect to $L^1$ distance. Explic-

itly, person $i$ is classified as normal if $\int |\widehat{v_i}(x) - v_N(x)|dx < \int |\widehat{v_i}(x) - v_A(x)|dx$ and abnormal if $\int |\widehat{v_i}(x) - v_A(x)|dx < \int |\widehat{v_i}(x) - v_N(x)|dx$. Moreover, derivatives may also be considered, so that person $i$ is classified as normal if $\int |\widehat{v_i}'(x) - v_N'(x)|dx < \int |\widehat{v_i}'(x) - v_A'(x)|dx$ and abnormal if $\int |\widehat{v_i}'(x) - v_A'(x)|dx < \int |\widehat{v_i}'(x) - v_N'(x)|dx$.

**Confidence band method**

Confidence band from nonparametric regression can also be applied in pattern recognition, although references describing the use of confidence bands in machine learning are quite limited [38].

Confidence bands are curves enclosing a model (function) being estimated by regression [38]. They represent the areas where the true model resides with a probability of $1 - \alpha$. Usually a value of 0.05 is used for $\alpha$ so that the bands enclose the true model with a probability of 95%. There exist many approaches to compute confidence bands, e.g. by Monte Carlo[41] or bootstrapping methods[37]. Charnigo, Hall and Srinivasan(2013) have also developed a method to estimate simultaneous confidence bands for both a mean response and its derivatives in nonparametric regression(under review), and they also present the idea of solving a pattern recognition problem through confidence band method[17].

To illustrate, consider again the example of using Raman spectroscopy to diagnose breast cancer. Let $\widehat{v_i}(x)$ denote the Raman spectrum for person $i$ to be classified, and $v_N(x)$ and $v_A(x)$ denote known normal and abnormal Raman spectra. One can fit confidence bands around the Raman spectrum of person $i$ $\widehat{v_i}(x)$. Person $i$ can be classified as normal or abnormal based upon whether $v_N(x)$ or $v_A(x)$ is inside the bands. Explicitly, if $v_N(x)$ is inside the bands, but $v_A(x)$ is outside, person $i$ is classified as normal, whereas if $v_A(x)$ is inside the bands, but $v_N(x)$ is outside, person $i$ is classified as abnormal. While both $v_A(x)$ or $v_N(x)$ may be inside or outside at $\alpha = 0.05$, in general there will exist a choice of $\alpha$ at which the proportion of curve $v_A(x)$ or $v_N(x)$ inside the bands is larger.

## Cancer

Cancer is a growing public health problem. The 2009 age adjusted invasive cancer(all sites) incidence rate in the US is 633.1 per 100,000 and in Kentucky is as high as 696.6 per 100,000[73]. The most frequently diagnosed cancers are prostate cancer, accounting for 31% of new cancers in men in the US, and breast cancer, accounting for 32% of new cancers among females[39]. Lung and colorectal cancers are the third and fourth most commonly diagnosed cancers[39].

The growing number of cancer diagnoses puts enormous pressure on health systems. According to the Medical Expenditure Panel Survey (MEPS), each year $38.4 billion of direct medical services is consumed for cancer-associated care. Another $59.2 billion is spent on concurrent conditions affecting cancer patients. On average, a patient with cancer incurs annual expenses of $9,753 [97]. About one in eight (12.29%) U.S. women will develop invasive breast cancer over the course of her lifetime[77].

Previous epidemiological studies found various risk factors for cancer including lifestyles, genetic factors and environmental factors. For example, smoking increases the risk of developing cancer. According to the National Cancer Institute, smoking causes 30% of all cancer deaths in the U.S. and is responsible for 87% of cases of lung cancer [67]. Not only does smoking affect the lungs, it can cause kidney, pancreatic, cervical, and stomach cancers and acute myeloid leukemia[67]. Breast feeding could protect against breast cancer [40]. Compared with parous women who never breastfed, women who had breastfed for 25 months or more had an estimated relative risk of 0.67 (95% CI, 0.52-0.85)[44]. An increased risk of breast cancer in women with a family history (any relative) of breast cancer has been demonstrated with an

estimated relative risk of RR = 1.9 (95% CI, 1.7-2.0)[59].

## Raman spectroscopy

Raman spectroscopy is one of the most common vibrational spectroscopies for assessing molecular motion and fingerprinting species. It is based on inelastic scattering of a monochromatic excitation source. The routine energy range is usually from 200 - 4000 $cm^{-1}$[47][19].

More specifically, Raman spectroscopy has been defined as a "coherent two-photon process in which a molecule simultaneously absorbs an incident photon and emits a Raman photon, accompanied by its transition from one energy level to another, giving rise to a frequency (i.e., energy) shift of the emitted photon" [32].

Raman spectra can provide detailed quantitative chemical information about a tissue. One particular advantage is that Raman spectroscopy can be used for in vivo measurements. Therefore it has potential to distinguish benign tissue and malignant tissue without many biopsies[32][50]. Compared with fluorescence, Raman spectra provide high information content, yet the signals are often weaker, which made the analysis of Raman spectra data complex[33].

## Raman spectroscopy in cancer diagnosis

Raman spectroscopy has been proposed for early cancer diagnosis by a number of authors [81][75][32][33]. Dr Stone used Raman spectroscopy for early diagnosis of laryngeal malignancy in 2000 [81].

Dr Haka, Dr Feld, and their collaborators applied Raman spectroscopy in breast cancer diagnosis in 2005[32]. The Raman spectra peaks correspond to different molecules. Whereas normal mammary spectra primarily contain peaks associated with lipids, tumor-containing mammary glands show an increase in peaks indicating

proteins and a decrease in those indicating lipids[56]. Their results [32] show that the Raman spectroscopy has better specificity and sensitivity in diagnosing breast cancer than optical tomography, and is less likely to be influenced by the patient's menopausal status and breasts' density than fluorescence spectroscopy. In 2009, Dr Haka and collaborators conducted a prospective study to diagnose normal, benign, and malignant human breast tissues using Raman spectroscopy [33]. This analysis of the prospectively obtained clinical data set showed Raman spectroscopy has a sensitivity of 83% and specificity of 93%, a positive predictive value of 36% and a negative predictive value of 99%.

## 1.4 Scope of Dissertation

To complete chapter 1, we now briefly describe our agenda for the remainder of this dissertation document.

**Chapter 2: Nonparametric regression in Raman spectroscopy**

In Chapter 2, we apply nonparametric regression in Raman spectra analysis to classify normal, benign, and malignant breast tissue. Instead of requiring basis spectra of chemical constituents for breast tissue, our study will explore both the spectra profiles and their derivatives(high frequency information) to differentiate different types of breast tissue. We will employ minimum distance approach and confidence bands approach for classification.

**Chapter 3: Stacking for nonparametric regression**

Motivated by the advancement of stacking in statistical learning, in Chapter 3 we will set up modified stacking framework and propose method to combine different classifiers together to make better prediction in nonparametric regression settings. Unlike existing Stacking approaches assuming that inputs are known and attempting to relate inputs to noise-corrupted or otherwise distorted outputs, our approach will solve the inverse problem, and the object is to infer the inputs from noise-corrupted

or otherwise distorted outputs such as in Raman spectra analysis.

**Chapter 4: Boosting for nonparametric regression**

In Chapter 4 we will develop a method by incorporating boosting into the nonparametric regression to improve classification accuracy. Boosting uses a set of weak learners to create a single strong learner. Different from stacking, in Boosting, examples that are incorrectly predicted by previous classifiers in the series are weighted more heavily than examples that were correctly predicted. Our Boosting methodology will differ from existing approaches, however, so that we can employ it to address inverse problems such as classifying different types of breast tissue in Raman spectra analysis.

**Chapter 5: Dynamic ensemble integration for nonparametric regression**

In Chapter 5 we will propose development of a novel dynamic framework based on multiple meta-learning strategies for classification problem in nonparametric regression settings. We will present promising lines for future work and the potential applications of this novel framework.

**Chapter 6: Nonparametric regression in Raman spectroscopy, revisited**

In Chapter 6, we will revisit the Raman spectroscopy data in Chapter 2, and make improvement based on the developments of the methods from Chapter 3 and Chapter 4. We will compare the method by incorporating stacking with the method incorporating boosting in the nonparametric regression settings. Finally we will summarize the major findings and contributions of this work as well as identify opportunities for future research and their public health implications.

# Chapter 2 Nonparametric Regression Techniques in Pattern Recognition

## 2.1 Background

As a novel and rapidly developing imaging tool in cancer diagnosis, Raman spectroscopy has been successfully employed in the classification of normal, benign, and malignant breast tissue, based on coefficient estimates from a linear combination model involving basis spectra for chemical constituents of breast tissue(Haka et al, 2005)[32]. The Raman spectra peaks correspond to different Raman active biological molecules in tissues. Whereas normal mammary spectra primarily contain peaks associated with lipids, tumor-containing mammary glands show an increase in peaks indicating proteins and a decrease in those indicating lipids[56]. Apart from less invasive feature, Raman spectroscopy has better specificity and sensitivity in diagnosing breast cancer than optical tomography, and is less likely to be influenced by the patient's menopausal status and breasts' density than fluorescence spectroscopy.

Various techniques, such as neural networks and linear regression for classification, hierarchical cluster analysis (HCA), linear discriminant analysis (LDA) for disease differentiation, partial least squares, a regression based technique and hybrid linear analysis are used to analyze the Raman spectra[23]. Nonparametric regression or semiparametric regression has advantages over parametric models when an underlying parametric model cannot be identified. Motivated by study on how the derivatives of Raman spectra might be employed to address a pattern recognition problem in analytic chemistry Charnigo et al 2011[14], we further explore how derivatives of Raman spectra can be used for diagnosing breast cancer. The data we used were similar but not exactly the same as those considered by Haka et al (2005)[33].

We use two approaches for using Raman spectra and their derivatives in the diagnosis of breast cancer: one based on minimum distance, and the other based on

confidence bands.

### 2.2 Previous work and definitions

Our implementations of the two approaches rely on compound estimation as described in Chapter 1(Charnigo and Srinivasan, 2011), generalized $C_p$ criterion as briefly described below (Charnigo, Hall, and Srinivasan, 2011), and confidence bands(as briefly described below)(Charnigo, Hall, and Srinivasan, 2013) procedures coded in the R statistical software language.

## Generalized $C_p$ criterion

Charnigo, Hall, and Srinivasan(2011) have proposed a generalized $C_p$ ($GC_p$) criterion which can be used when selecting tuning parameters in derivative estimation[16] in non-parametric regression estimation settings. In general, assume that model (1.1) from Chapter 1 holds, and $\widehat{\frac{d^q}{dx^q}\mu_\lambda(x_i)}$ has the form $\sum_{m=1}^n l_{m;\lambda}^{(q)}(x)Y_m$ for some specified functions $l_{1;\lambda}^{(q)}(x),\ldots,l_{n;\lambda}^{(q)}(x)$ that do not depend on $Y_1,\ldots,Y_n$. Then $GC_p$ criterion is defined as

$$GC_p(\mathbf{Y},\widehat{\mu_\lambda}) := \sum_{i=1}^n s_i(Y_i^{(q)} - \widehat{\frac{d^q}{dx^q}\mu_\lambda(x_i)})^2 + \sigma^2\sum_{i=1}^n s_i\sum_{m=1}^n(2c_{i,m}l_{m;\lambda}^{(q)}(x_i) - c_{i,m}^2),$$

where $s_1,\ldots,s_n$ are observation weights between 0 and 1, $\mathbf{Y}$ stands for $(Y_1,\ldots,Y_n)^T$. $Y_i^{(q)}$ is called an empirical derivative and defined as $\sum_{m=1}^n c_{i,m}Y_m$, where the $c_{i,m}$ are some specified constants. $\lambda$ denotes a vector containing the tuning parameters[16]. In essence, this $GC_p$ criterion is a residual sum of squares for the fitted derivative of order $q$ plus a penalty term so that $GC_p$ has expected value approximately equal to the target $\sum_{i=1}^n(\frac{d^q}{dx^q}\mu(x_i) - \widehat{\frac{d^q}{dx^q}\mu_\lambda(x_i)})^2$. In this study, compound estimation is used to construct unknown $\mu(x)$ , therefore $\lambda$ contains tuning parameters for compound estimation: $\beta_n$, and nearest neighbour fraction for the pointwise estimators. Also we

have

$$l^{(j)}_{m;\lambda}(x) = \sum_{a\in I_{n;\lambda}} \sum_{k=0}^{j} \binom{j}{k} \frac{d^k}{dx^k} \sum_{p=0}^{J} r_{a;p;m;\lambda}(x-a)^p \frac{d^{j-k}}{dx^{j-k}} W_{a,n;\lambda}(x),$$

Suppose $\tilde{c}_{p;a}$ has the form $\sum_{i=1}^{n} r_{a;p;m;\lambda} Y_i$ for $a \in I_{n,\lambda}$ and $W_{a,n;\lambda}$ can be defined as $\frac{\exp[-\beta_n(x-a)^2]}{\sum_{c\in I_{n,\lambda}} \exp[-\beta_n(x-c)^2]}$, in which $\beta_n$ is a nondecreasing sequence of positive real numbers. Further we take $q = 1$ and choose $Y_i^{(1)}$ to be a difference quotient-like approximation to $\mu^{(1)}(x_i)$:

$$Y_i^{(1)} := \sum_{m=1}^{k} (m^2 / \sum_{s=1}^{k} s^2)(Y_{i+m} - Y_{i-m})/(x_{i+m} - x_{i-m}) 1_{(k+1)\le i\le (n-k)},$$

where $k \approx 0.05n$, and approximate $\sigma^2$ by

$$\hat{\sigma}^2 := \frac{\sum_{i=1}^{n} (Y_i - \sum_{m=1}^{n} l^{(0)}_{m;\lambda_0}(x_i) Y_m)^2}{n - \sum_{m=1}^{n} l^{(0)}_{m;\lambda_0}(x_m)};$$

where $\lambda_0$ contains the most extreme tuning parameters under consideration.

## Simultaneous confidence bands

We employ the approach of Charnigo, Hall and Srinivasan (2013) to construct confidence bands that are simultaneous over different orders of derivatives[17]. They have shown that under certain conditions, for all sufficiently large $n$,

$$\mathbb{P}[\ L_p(x) \le \mu^{(p)}(x) \le U_p(x)\ \forall p \in \{0, 1, \ldots, J\} \text{ and } x \in E\ ] \ge 1 - \alpha_n - \delta_n(h_n), \text{ where}$$

$$L_p(x) := \widehat{\mu^{(p)}}_I(x) - \widehat{M}_p - z_{\alpha_n}(1+\gamma)\widehat{\sigma} D_{p,I}(x) \quad \text{and}$$

$$U_p(x) := \widehat{\mu^{(p)}}_I(x) + \widehat{M}_p + z_{\alpha_n}(1+\gamma)\widehat{\sigma} D_{p,I}(x).$$

Above, $\mu^{(p)}(x) := \frac{d^p}{dx^p}\mu(x)$, $E \subset \mathcal{X}$ is compact, and $\alpha_n$ is a non-increasing sequence in $(0, 1)$. Also $\delta_n(h_n)$ is a positive quantity that turns to 0 as $n \to \infty$,

$$\widehat{\mu^{(p)}}(x;\xi) := \sum_{i=1}^{n} l_{p;i}(x;\xi) Y_i, \quad D_p(x;\xi) := \sqrt{\sum_{i=1}^{n} l_{p;i}(x;\xi)^2},$$

$$\widehat{M_p}(\xi) := \sup_{x \in E} \left| \widehat{\mu^{(p)}}_I(x;\xi) - \widehat{\mu^{(p)}}(x;\xi) \right|,$$

$$\widehat{\sigma}^2 := \frac{1}{2\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} (Y_{2i} - Y_{2i-1})^2.$$

$z_{\alpha_n}$ is not the upper $\alpha_n$ quantile of the standard normal distribution but rather is a number chosen to ensure the desired confidence level and for convenience may be replaced by its upper bound $\sqrt{-2\log\left[\frac{1-(1-\alpha_n)^{1/|\mathbf{G}_n|}}{2(J+1)}\right]}$. $\gamma$ is set to a very small value (such as 0.05), and $\mathbf{G}_n$ denotes a grid on which confidence interval are constructed prior to linear interpolations to acquire confidence bands[17].

## 2.3 Methods

We now describe these two approaches in detail. For ease of exposition the first approach is presented in Sections 1, 2, 3, 4a, and 5a below, while the second is constituted by Sections 1, 2, 3, 4b, and 5b.

Also, we refer to the units of observation as "subjects", although they may be samples of tissue rather than people. To help the reader keep track of notation, we provide a glossary in Table 2.1.

Table 2.1: Glossary of notation

| Notation | Definition |
|---|---|
| $x$ | Raman shift |
| $y(x)$ | Unnormalized Raman spectrum |
| $x_i$ | Value of Raman shift for $i^{th}$ observation |
| $y(x_i)$ | Value of unnormalized Raman spectrum for $i^{th}$ observation |
| $y^*(x)$ | Normalized Raman spectrum |
| $\widehat{\mu^{(j)}}(x)$ | Estimated $j^{th}$ derivative of normalized Raman spectrum |
| $c$ | Index of possible diagnoses |
| $\widehat{\mu_c^{(j)}}(x)$ | Reference curve $j$ for diagnosis $c$ (average estimated $j^{th}$ derivative of Raman spectra for diagnosis $c$) |
| $u$ | Symbol to indicate an unknown diagnosis |
| $\widehat{\mu_u^{(j)}}(x)$ | Estimated $j^{th}$ derivative for Raman spectrum from unknown diagnosis |
| $\widehat{\zeta_{c,u}^{(j)}}$ | $L^1$ distance between $\widehat{\mu_u^{(j)}}(x)$ and $\widehat{\mu_c^{(j)}}(x)$ |
| $\widehat{\mu_{low,u}^{(j)}}(x)$ | Lower confidence band of $j^{th}$ derivative for Raman spectrum from unknown diagnosis |
| $\widehat{\mu_{up,u}^{(j)}}(x)$ | Upper confidence band of $j^{th}$ derivative for Raman spectrum from unknown diagnosis |
| $\widehat{c}_{D,j}$ | Symbol to indicate inference for an unknown diagnosis based on minimum distance to reference curve $j$ |
| $\widehat{\rho^{(j)}}_{c,u}$ | Portion of $j^{th}$ reference curve for diagnosis $c$ inside confidence bands of $j^{th}$ derivative from unknown diagnosis |
| $1_A$ | Indicator function, equals 1 when assertion $A$ is true and 0 otherwise |
| $\widehat{c}_{B,j}$ | Symbol to indicate inference for an unknown diagnosis based on confidence bands containing reference curve $j$ |
| $\widehat{\rho^{(j,k)}}_{c,u}$ | Portion of $j^{th}$ and $k^{th}$ reference curves for diagnosis $c$ inside respective confidence bands |
| $\widehat{c}_{B,j,k}$ | Symbol to indicate inference for an unknown diagnosis based on confidence bands containing reference curves $j$ and $k$ |

The column "Definition" refers to a brief definition for each notation.

**Section 1: Normalize the Raman spectra**

The data from spectra instruments are often influenced by subtle changes in settings or conditions and hence are often contaminated by noise[61]. The point of normalization is to adjust for a range of experimental conditions. Suppose that, for each subject, we observe $(x_i, y(x_i))_{i=1}^n$, where $x_i$ denotes a value of the Raman shift and $y(x_i)$ the corresponding value of the Raman spectrum for that subject. For simplicity we make the (usually realistic) assumption that $x_1, \ldots, x_n$ are common to all subjects (in this study, $x_i = 686 + 2(i-1)$ for $i \in \{1, 2, ..., 548\}$), although this assumption is not essential to the deployment of our methodology. On the other hand, $y(x_1), \ldots, y(x_n)$ will vary from subject to subject, both because there are biochemical variations across different subjects and because different subjects are exposed to various amounts of radiation. The former source of variation interests us as a means for distinguishing subjects with cancer from subjects without cancer, but the latter source is a nuisance that actually impedes making such distinctions. Therefore, we normalize the Raman spectrum for each subject.

The two normalization schemes considered herein are called "RANGE" and "STDEV". The first scheme linearly re-scales each patient's Raman spectrum so that the minimum value is 0 and the maximum value is 1, which is represented symbolically by

$$y^*(x) := \frac{y(x) - \min_{1 \le i \le n} y(x_i)}{\max_{1 \le i \le n} y(x_i) - \min_{1 \le i \le n} y(x_i)}.$$

Above, $y^*(x)$ denotes the normalized Raman spectrum. The second scheme linearly re-scales each patient's Raman spectrum so that the mean value is 0 and the standard deviation is 1, which is represented symbolically by

$$y^*(x) := \frac{y(x) - n^{-1} \sum_{i=1}^n y(x)}{\sqrt{(n-1)^{-1} \sum_{i=1}^n (y(x_i) - n^{-1} \sum_{i=1}^n y(x_i))^2}}.$$

Because our data were already centered at 0 (mean 0) for each person, the second scheme is also equivalent to the "SCALE" method, which represented symbolically

by

$$y^*(x) := \frac{y(x_i)}{\sqrt{\sum_{i=1}^{n}(y(x_i))^2}}.$$

**Section 2: Estimate the derivatives of the normalized Raman spectra**

Since $y(x)$ is only acquired for $x$ in the finite grid $\{x_1, \ldots, x_n\}$, $y^*(x)$ as defined in either normalization scheme is also available only on that same finite grid. Even if we are willing to assume that the stochastic errors associated with the observations of $y^*(x)$ are negligibly small (and especially if we are *not* willing to assume this), we must employ a statistical smoothing method to estimate $y^*(x)$ for $x$ in the continuum of the interval $[x_1, x_n]$. An naive approach such as calculating difference quotients will have a large variance and the random noise will be considerably magnified. Furthermore if the raw data have the spikes, these will be dominant in estimated derivatives using naive difference quotients; statistical smoothing methods will reduce the distortions caused by such spikes. Figure 2.1 illustrates this idea.

The statistical smoothing method that we use for this purpose is called compound estimation. Pioneered and described in detail by Charnigo and Srinivasan (2011), compound estimation has several advantages over its competitors. First, not only is $y^*(x)$ estimated but so are its derivatives (with respect to $x$). We let $\widehat{\mu^{(j)}}(x)$ denote the estimated $j^{th}$ derivative, where $j$ can in principle be any nonnegative integer; however, in practice attention is often directed to $j \leq 2$. The special case $j = 0$ refers to estimation of $y^*(x)$; generic use of the symbol $j$ will include that special case. Second, the estimated derivatives satisfy the "self-consistency" property that $\frac{d^k}{dx^k}\widehat{\mu^{(j-k)}}(x) = \widehat{\mu^{(j)}}(x)$ for any $k \leq j$. Thus, self-consistency implies a sort of interchangeability between the processes of estimation and differentiation. While this seems like a natural requirement for any statistical smoothing method, local regression (Loader, 1999) and some other statistical smoothing methods do not possess this property[48]; for some methods, the derivatives of the estimates may not even

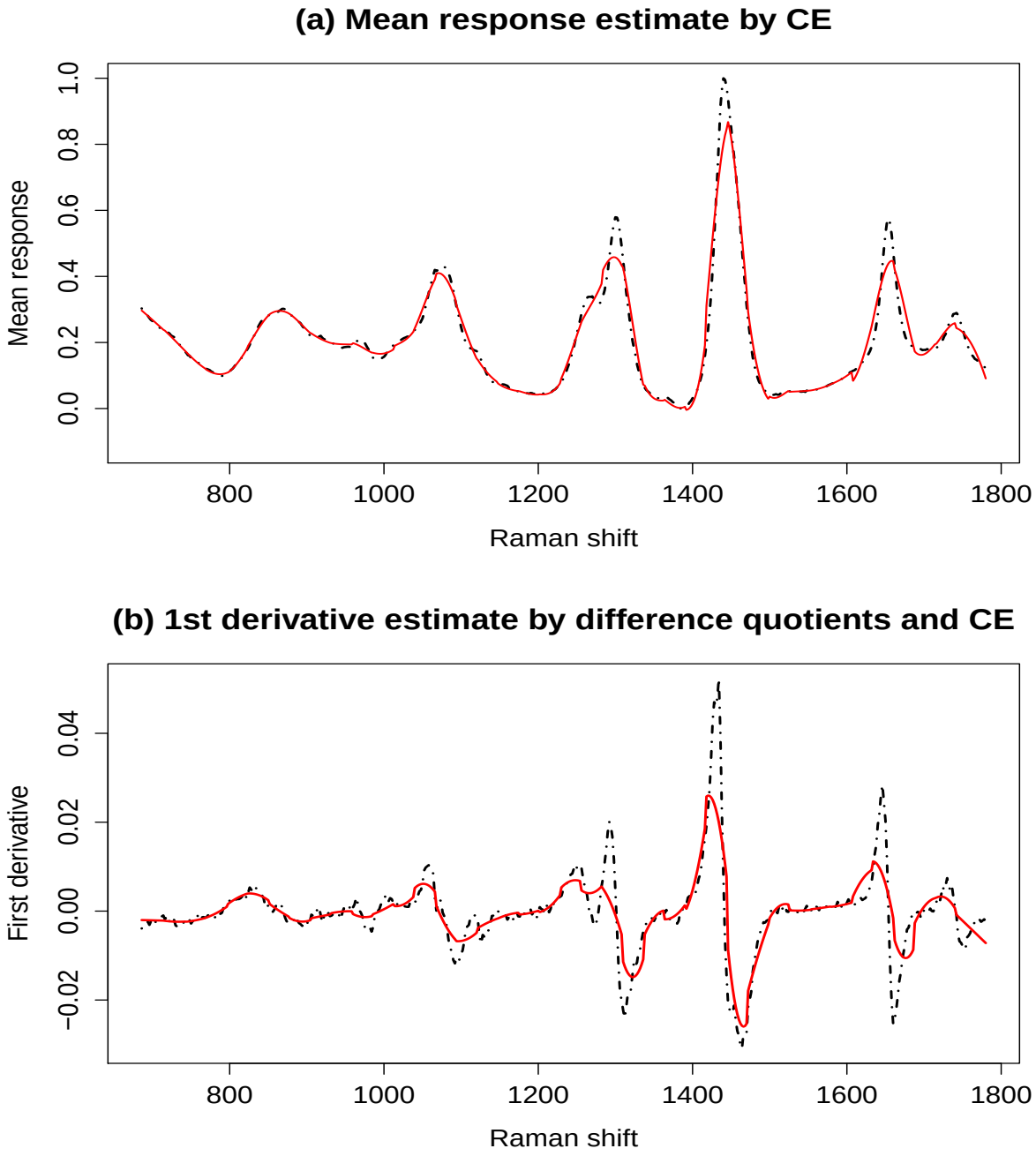


Figure 2.1: Estimated first derivative using difference quotients and compound estimation. Panel a are the raw data for Raman intensity level (black dot dashed line) and estimated mean response by compound estimation(red solid line). Panel b displays first derivative estimate by ordinary difference quotients method(black dot dashed line) as well as compound estimation(red solid line).

exist. Third, compound estimation enjoys near-optimal convergence rates (Stone, 1980; Stone, 1982)[79][80]. Stone has shown that under regularity conditions the optimal convergence rate for a collection of estimating derivative functions that has derivatives of order $J+1$ is $O_p(n^{(j-J-1)/(2J+3)})$ for $j \leq J$[79]. The self-consistent compound estimator and its derivatives achieve a nearly optimal convergence rate of $O_p(n^{(j-J-1))/(2J+3)+\nu})$ for $j \leq J$ as shown by Charnigo and Srinivasan [18]. Roughly speaking, this means that the error in estimating the $j^{th}$ derivative decreases almost as quickly (in relation to an increasing sample size) as is theoretically possible; more precisely, $\hat{\mu}^{(j)}(x) - \mu^{(j)}(x) = O_p(n^{(j-J-1)/(2J+3)+\nu})$ for $j \leq J$. Fourth, numerical studies have demonstrated that compound estimation may recover derivatives substantially more accurately than local regression or splines (Wahba, 1990)[91], even when the sample size is modest.

Like any other statistical smoothing method, compound estimation relies on the selection of tuning parameters to obtain a reasonable compromise between ordinary linear regression (i.e., drawing a straight line through the data) and interpolation (i.e., literally connecting the dots)(Figure 2.2). Therefore, we used the generalized $C_p$ criterion of Charnigo, Hall, and Srinivasan (2011) to select tuning parameters for compound estimation[16]. This criterion was specifically developed to enhance recovery of derivatives and, besides being theoretically justified, compared favorably to several other criteria in simulation studies assessing the accuracy with which derivatives were recovered.

**Section 3: Obtain the reference curves for various diagnoses**

We introduce the symbol $c$ as an index of possible diagnoses. More specifically we identify $c = 1$ with a normal diagnosis, $c = 2$ with cancer, $c = 3$ with fibroadenoma ("FA"), and $c = 4$ with fibrocystic change ("FC"). Reference curve $j$ for diagnosis $c$, denoted $\widehat{\mu_c^{(j)}}(x)$, is defined as the average of $\widehat{\mu^{(j)}}(x)$ over all subjects known to have diagnosis $c$. In a retrospective study, such as the one for which results are presented

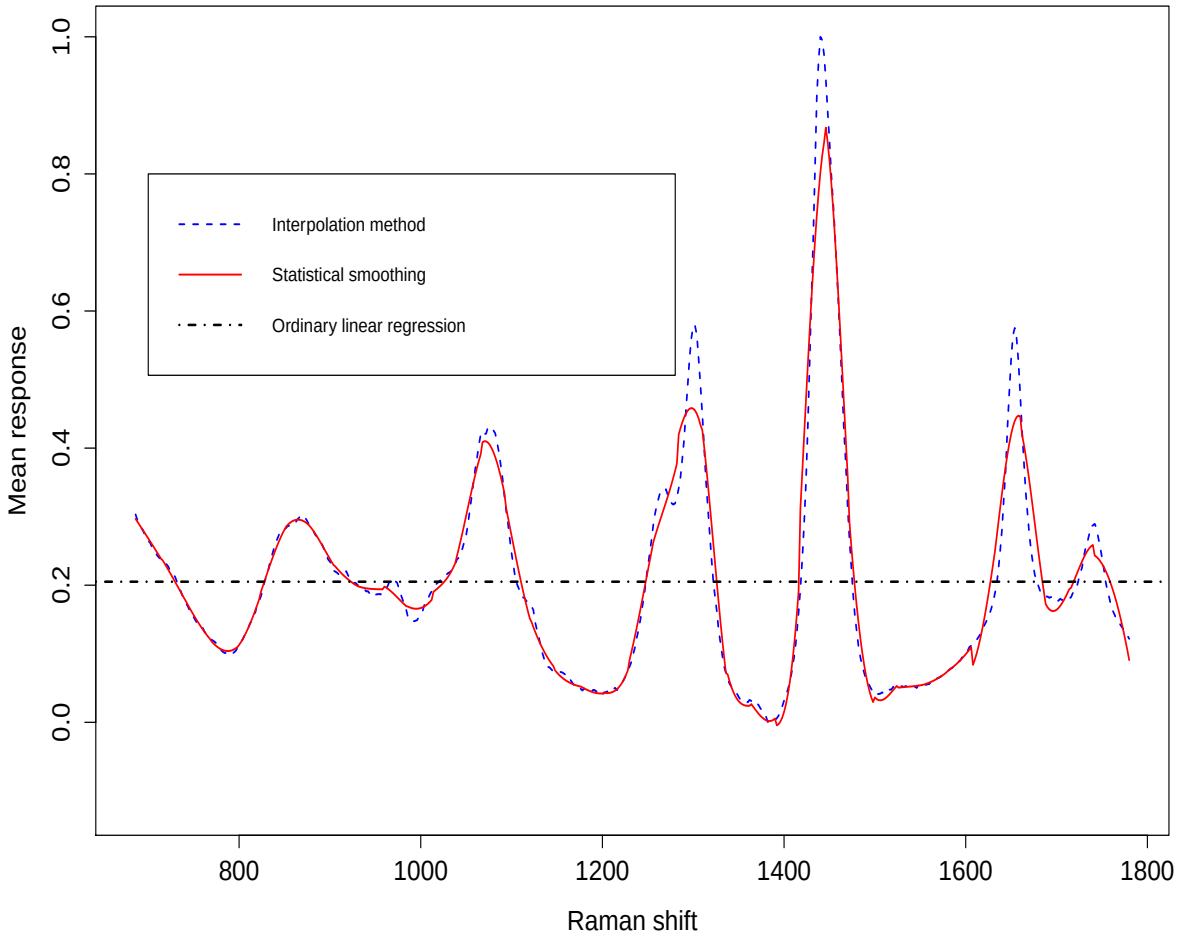


Figure 2.2: Estimated mean response using interpolation, statistical smoothing, and ordinary linear regression. Shown are estimated mean responses for Raman intensity level by interpolation, statistical smoothing, and ordinary linear regression methods. Statistical smoothing method (here we use compound estimation) as shown by red solid line serves as a reasonable compromise between interpolation method(blue dashed line) and ordinary linear regression(black dot dashed line).

herein, all of the subjects' diagnoses are known. However, in a prospective study there would be one group of subjects with known diagnoses and another group of subjects with unknown diagnoses.

**Section 4a: Calculate distances from a subject's estimated derivatives to the reference curves** Consider a specific subject for whom a diagnosis is to be inferred from the estimated derivatives of her Raman spectrum. In a retrospective study this can be any subject, and we temporarily "blind" ourselves as to her actual

diagnosis. (Thus, that subject's data are temporarily excluded from the calculation of any reference curves.) In a prospective study this can be any subject whose diagnosis is unknown. Let the symbol $u$ stand in place of $c$ for this subject, so that $\widehat{\mu_u^{(j)}}(x)$ represents the estimated $j^{th}$ derivative of her Raman spectrum.

For $c \in \{1, 2, 3, 4\}$ and $j \in \{0, 1, 2\}$ we define

$$\widehat{\zeta_{c,u}^{(j)}} := \int_{x_1}^{x_n} \left| \widehat{\mu_u^{(j)}}(x) - \widehat{\mu_c^{(j)}}(x) \right| \, dx,$$

which is the $L^1$ distance between the subject's estimated $j^{th}$ derivative and reference curve $j$ for diagnosis $c$. Up to a multiplicative constant determined by the spacing between successive Raman shifts, which will be irrelevant for our inferential purpose since all the $L^1$ distances $\widehat{\zeta_{c,u}^{(j)}}$ share the same multiplicative constant, which is $|x_{i+1} - x_i|$, we may approximate $\widehat{\zeta_{c,u}^{(j)}}$ by $\sum_{i=1}^{n} \left| \widehat{\mu_u^{(j)}}(x_i) - \widehat{\mu_c^{(j)}}(x_i) \right|$.

**Section 4b: Construct confidence bands for the derivatives of a subject's normalized Raman spectrum**

The first paragraph in Section 4a still applies, but the second paragraph is replaced by the following.

Using the method of Charnigo, Hall, and Srinivasan (2013), we create confidence bands for the $j^{th}$ derivative of the subject's Raman spectrum; these are denoted $\widehat{\mu_{low,u}^{(j)}}(x)$ and $\widehat{\mu_{up,u}^{(j)}}(x)$ respectively. Confidence bands generalize the concept of a confidence interval. While a confidence interval contains a range of values thought to contain an unknown number, confidence bands define an area in two-dimensional space thought to contain an unknown curve, in this case the $j^{th}$ derivative of the subject's Raman spectrum. At this juncture we remind the reader that this derivative is unknown and that $\widehat{\mu_u^{(j)}}(x)$ is only an estimate. Indeed, the confidence bands are constructed below and above this estimate in much the same way that a confidence interval equals an estimate of an unknown number minus and plus some multiple of the standard error.

Figure 2.3 provides an illustration (with $j = 0$). The dashed black curves represent

the confidence bands; the lower dashed curve is $\widehat{\mu^{(j)}_{low,u}}(x)$, and the upper dashed curve is $\widehat{\mu^{(j)}_{up,u}}(x)$. The solid blue curve is $\widehat{\mu^{(j)}_{1}}(x)$ (reference curve for a normal diagnosis), while the dot dashed green curve is $\widehat{\mu^{(j)}_{2}}(x)$ (reference curve for a cancer diagnosis). Further explanation of Figure 2.3 appears in Section 5b below.

**Section 5a: Infer a diagnosis by minimizing distance to a reference curve**

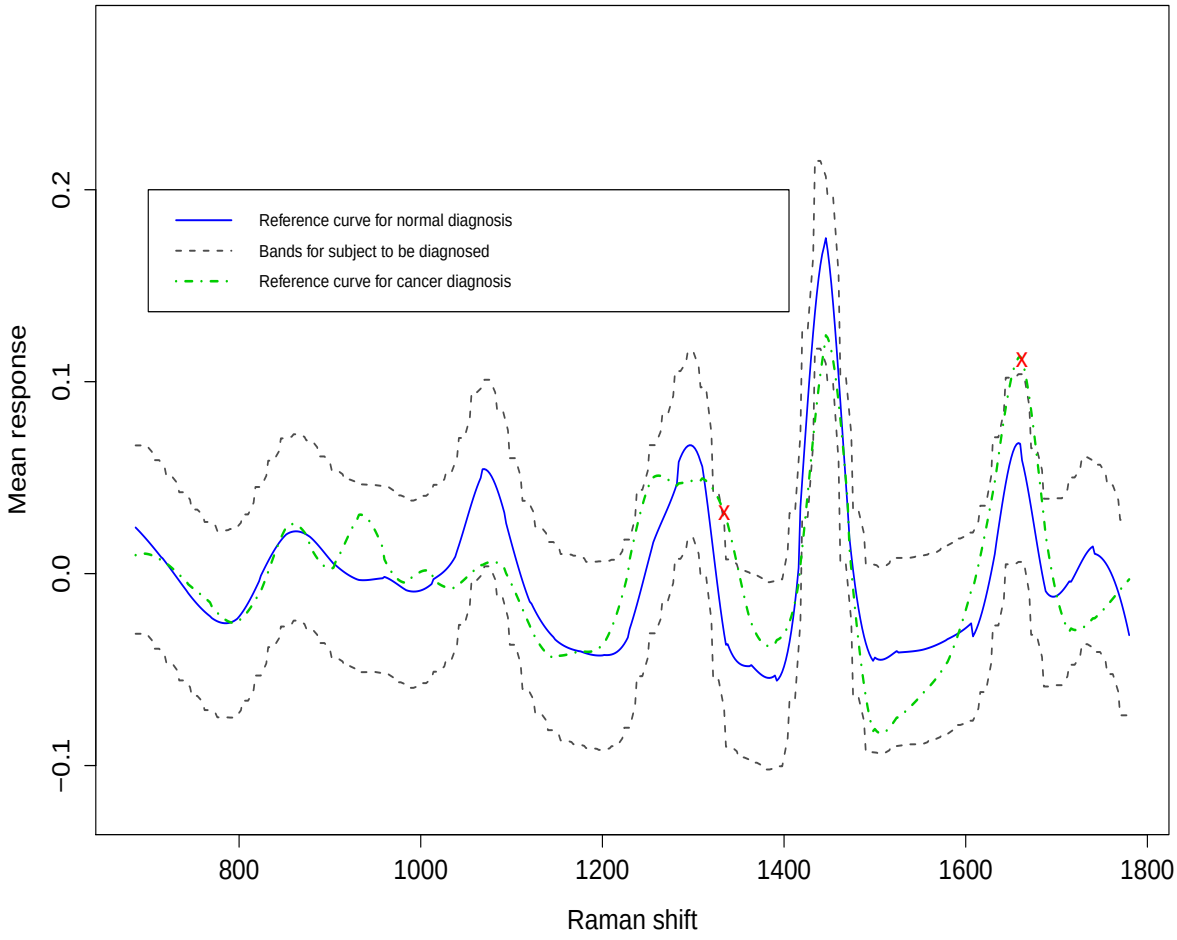


Figure 2.3: Simultaneous confidence bands for mean responses of Raman spectra. Shown are estimated mean response reference curve for a normal diagnosis (blue solid curve), reference curve for a cancer diagnosis(green dot-dashed curve) and a confidence bands of subject to be diagnosed(gray dashed curve). "X" symbols identify two of the locations where the reference curve for cancer diagnosis lies outside the confidence bands for subject to be diagnosed. On the other hand, the reference curve for normal diagnosis lies entirely within the confidence bands.

Having obtained (discrete approximations to) $\widehat{\zeta_{c,u}^{(j)}}$ for $c \in \{1,2,3,4\}$, we infer the unknown diagnosis to be the minimizer

$$\widehat{c}_{D,j} := \arg \min_{c \in \{1,2,3,4\}} \widehat{\zeta_{c,u}^{(j)}}.$$

In words, the unknown diagnosis is guessed to be the one whose $j^{th}$ reference curve is closest (in the sense of $L^1$ distance) to the subject's estimated $j^{th}$ derivative as shown by Figure 2.4. The notation $\widehat{c}_{D,j}$ indicates the dependence of this inference both on the approach used ("D" for minimizing distance) and the derivative considered ("j" for the $j^{th}$ derivative).

**Section 5b: Infer a diagnosis by capturing a reference curve inside confidence bands** Consider again Figure 2.3.

The $j^{th}$ reference curve for a normal diagnosis (blue solid curve) lies entirely within the confidence bands for the $j^{th}$ derivative of the subject's Raman spectrum (gray dashed curves), while the reference curve for a cancer diagnosis (green dot-dashed curve) breaches the confidence bands in multiple locations; two such locations are identified with "X" symbols. This suggests that a normal diagnosis is consistent with the subject's data, while a cancer diagnosis is not.

The situation in Figure 2.3 is rather idealized, however. FA and FC were excluded for simplicity, but in reality we have to consider these possibilities. What if the confidence bands had contained both reference curves, or if the confidence bands had been breached by both reference curves ? One might try to remove either contingency by adjusting the confidence level lower (e.g., from 95% to 90%) in the former instance to obtain narrower confidence bands or higher (e.g., from 95% to 99%) in the latter instance to obtain wider confidence bands. Unfortunately, confidence bands are more complicated than ordinary confidence intervals in that their length is influenced not only by the confidence level but also by an additive adjustment that accounts for their construction over a continuum such as the interval $[x_1, x_n]$. Therefore, we instead proceed as follows.

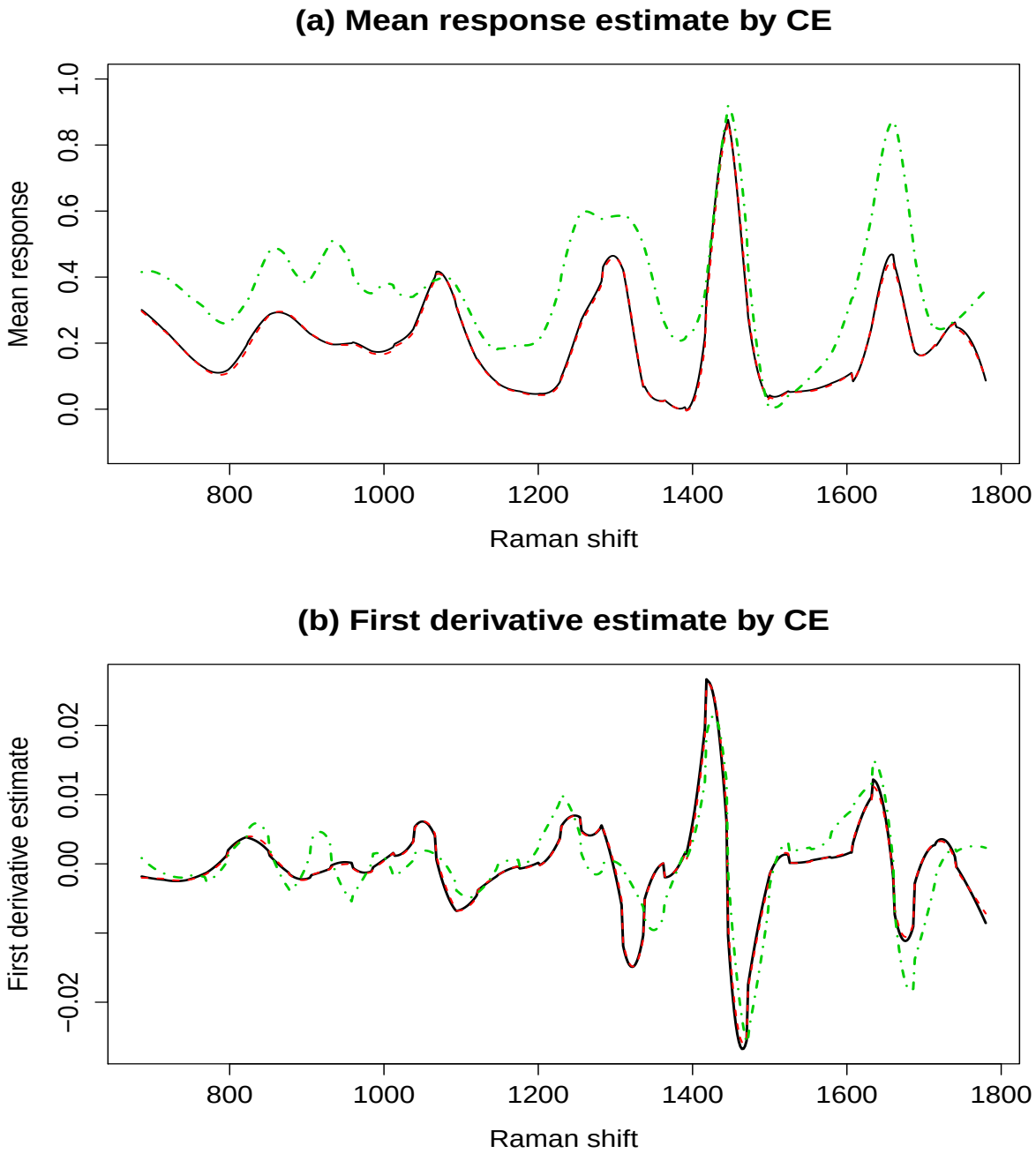


Figure 2.4: Estimated curves for mean responses and first derivatives by CE. Shown are estimated mean responses(Panel a) and first derivatives (Panel b) for Raman intensity level of reference curve for a normal diagnosis (black solid curve), reference curve for a cancer diagnosis(green dot dashed curve) and a curve of subject to be diagnosed(red dashed curve). A normal diagnosis is consistent with the subject's data because the curve of subject is closer to the reference curve to the normal reference curve with respect to both mean responses and first derivatives.

Let

$$\widehat{\rho^{(j)}}_{c,u} := n^{-1} \sum_{i=1}^{n} 1_{\widehat{\mu^{(j)}_{low,u}}(x_i) < \widehat{\mu^{(j)}_{c}}(x_i) < \widehat{\mu^{(j)}_{up,u}}(x_i)},$$

where $1_A$ denotes an indicator function that equals 1 when the assertion $A$ is true and 0 otherwise. In words, this is a discrete approximation to the fraction of the continuum $[x_1, x_n]$ over which the $j^{th}$ reference curve for diagnosis $c$ falls within the confidence bands. We can infer the unknown diagnosis to be the maximizer

$$\widehat{c}_{B,j} := \arg \max_{c \in \{1,2,3,4\}} \widehat{\rho^{(j)}}_{c,u}.$$

Thus, the unknown diagnosis is guessed to be the one whose $j^{th}$ reference curve is most contained in the confidence bands for the $j^{th}$ derivative of the subject's Raman spectrum. The notation $\widehat{c}_{B,j}$ indicates the dependence of this inference both on the approach used ("B" for confidence bands) and the derivative considered. If there is a "tie" (in particular, if multiple reference curves fall entirely inside the confidence bands), then the unknown diagnosis is inferred to be that which is more prevalent among all of the subjects with known diagnoses. Table 2.2 provides a hypothetical example. There is a tie between diagnosis "2" and "4", so we choose "4" due to higher prevalence.

In addition, the confidence bands approach allows multiple derivatives to be

Table 2.2: A hypothetical example showing how to solve a "tie" problem

| Diagnosis | $\widehat{\rho}$ | Prevalence |
|---|---|---|
| 1 | 0.98 | 0.3 |
| 2 | 1 | .2 |
| 3 | 0.84 | 0.1 |
| 4 | 1 | .4 |

considered simultaneously in a way that is not possible with the minimum distance

approach. More specifically, let

$$\widehat{\rho^{(j,k)}}_{c,u} := n^{-1} \sum_{i=1}^{n} 1_{\widehat{\mu^{(j)}_{low,u}}(x_i) < \widehat{\mu^{(j)}_{c}}(x_i) < \widehat{\mu^{(j)}_{up,u}}(x_i) \text{ and } \widehat{\mu^{(k)}_{low,u}}(x_i) < \widehat{\mu^{(k)}_{c}}(x_i) < \widehat{\mu^{(k)}_{up,u}}(x_i)}.$$

This is a discrete approximation to the fraction of the continuum $[x_1, x_n]$ over which *both* the $j^{th}$ and $k^{th}$ reference curves for diagnosis $c$ fall within the respective confidence bands. Again, we can infer the unknown diagnosis to be the maximizer

$$\widehat{c}_{B,j,k} := \arg \max_{c \in \{1,2,3,4\}} \widehat{\rho^{(j,k)}}_{c,u},$$

and we may break a "tie" based on prevalence.

In principle this idea generalizes to any number of derivatives(i.e., consider 0, 1, and 2 together).

### 2.4 Results

To illustrate and evaluate the two approaches described in the Methods section, we applied them to Raman spectrum data originally considered by Haka et al (2005). The data were de-identified Raman spectra from 124 *ex vivo* samples of human breast tissue, which we regard as "subjects" in the language of the Methods section. Among the 124 samples, 47 were normal, 31 were cancerous, 15 exhibited fibroadenoma ("FA"), and 31 exhibited fibrocystic change ("FC"). The use of these data for the present study was cleared by the University of Kentucky's Institutional Review Board. Table 2.3 provides the correct classification rates overall and within each diagnosis, using both approaches. More specifically, the minimum distance approach was applied with $j = 0$, $j = 1$, and $j = 2$, using both the RANGE and the STDEV normalizations. The confidence bands approach was applied with $j = 0$, $j = 1$, and $(j, k) = (0, 1)$, again using both the RANGE and the STDEV normalizations. The intersection in Table 2.3 indicates both the reference curves for mean response and the first derivative are considered simultaneously. The confidence bands approach was not applied with

$j = 2$ because the confidence bands for $2^{nd}$ derivatives would be uninformative due to their extremely large widths.

All normal tissues (47 out of 47) were correctly classified, and most FA tissues (between 11 and 14 out of 15) were also correctly classified. However, cancer and FC tissues proved more difficult. The confidence bands approach with $(j, k) = (0, 1)$ using the STDEV normalization was most successful overall (92 out of 124) and with cancer tissues (22 out of 31), while the minimum distance approach with $j = 2$ using either normalization fared best with FC tissues (14 out of 31).

Table 2.4 displays results from an alternate analysis in which only two diagnoses were permitted (normal and abnormal). In effect this grouped together the cancer, FA, and FC tissues. Several approach and normalization combinations yielded an overall correct classification rate of 114 out of 124, with 10 FC tissues misclassified as normal.

Additional findings appear in Tables 2.5 and 2.6. The former displays results from an alternate analysis in which only three diagnoses were permitted (normal, cancer, and non-cancer abnormal), with the overall best performer being the minimum distance approach with $j = 2$ using the RANGE normalization (94 out of 124). The latter displays results from an alternate analysis in which the normal tissues were set aside, with the overall best performer being the confidence bands approach with $(j, k) = (0, 1)$ using the STDEV normalization (54 out of 77).

## 2.5 Discussion

In this chapter, we have applied compound estimation(Charnigo and Srinivasan, 2011) and simultaneous confidence bands(Charnigo, Hall and Srinivasan, 2013) in the analysis of Raman spectra to classify different types of breast tissues including normal, fibroadenoma(FA), fibrocystic change(FC) and cancer. Different ways of classifications are used involving two normalization schemes and two strategies for choosing

Table 2.3: Raman spectra diagnosis results for four different types of tissues

| Different Methods | | Normal (out of 47) | Cancer (out of 31) | FA (out of 15) | FC (out of 31) | Total Correct (out of 124) |
|---|---|---|---|---|---|---|
| MD1 | $\mu_0$ | 47 (100%) | 16 (51.6%) | 14 (93.3%) | 6 (19.4%) | 83 (66.9%) |
| | $\mu_1$ | 47 (100%) | 19 (61.3%) | 12 (80.0%) | 10 (32.4%) | 88 (71.0%) |
| | $\mu_2$ | 47 (100%) | 15 (48.4%) | 13 (86.7%) | 14 (45.2%) | 89 (71.8%) |
| MD2 | $\mu_0$ | 47 (100%) | 19 (61.3%) | 14 (93.3%) | 7 (22.6%) | 87 (70.2%) |
| | $\mu_1$ | 47 (100%) | 19 (61.3%) | 12 (93.3%) | 8 (22.6%) | 86 (70.2%) |
| | $\mu_2$ | 47 (100%) | 15 (48.4%) | 13 (86.7%) | 14 (45.2%) | 89 (71.8%) |
| Band1 | $\mu_0$ | 47 (100%) | 18 (58.1%) | 13 (86.7%) | 6 (19.4%) | 84 (67.7%) |
| | $\mu_1$ | 47 (100%) | 1 (3.2%) | 11 (73.3%) | 7 (22.6%) | 66 (53.2%) |
| | Intersection | 47 (100%) | 15 (48.4%) | 14 (93.3%) | 6 (19.4%) | 82 (66.1%) |
| Band2 | $\mu_0$ | 47 (100%) | 21 (67.7%) | 13 (86.7%) | 9 (29.0%) | 90 (72.6%) |
| | $\mu_1$ | 47 (100%) | 1 (3.2%) | 13 (86.7%) | 7 (22.6%) | 68 (54.8%) |
| | Intersection | 47 (100%) | 22 (71.0%) | 14 (93.3%) | 9 (29.0%) | 92 (74.2%) |

We use MD and Band as abbreviations for Minimum distance method and Confidence band method respectively. Indices 1 and 2 refer to the first and second normalization methods respectively. Rows labeled $\mu_j$ indicate diagnosis based on derivatives of order $j$. Entries are numbers(and percentages) of subjects correctly diagnosed.

the optimal diagnosis.

One strategy is based on the minimum L1 distance between a patient's curve and a reference curve, while the other is based on maximum proportion of a reference curve falling within the confidence bands surrounding a patient's curve. We consider not only the mean response, but also the derivatives of the Raman intensity level. In addition, by using compound estimation, we can achieve the self consistency property that the derivatives of the estimator estimate the derivatives of the mean response. Simultaneous confidence bands can achieve a specified coverage probability for more than one derivative at a time, which provides a natural way for multiple derivatives

Table 2.4: Raman spectra diagnosis results for normal and abnormal tissues

| Different Methods | | Normal (out of 47) | Abnormal | | | | Total Correct (out of 124) |
|---|---|---|---|---|---|---|---|
| | | | Cancer (out of 31) | FA (out of 15) | FC (out of 31) | Total abnormal (out of 77) | |
| MD1 | $\mu_0$ | 47 (100%) | 31 (100%) | 15 (100%) | 21 (67.7%) | 67 (87.0%) | 114 (91.9%) |
| | $\mu_1$ | 47 (100%) | 31 (100%) | 15 (100%) | 21 (67.7%) | 67 (87.0%) | 114 (91.9%) |
| | $\mu_2$ | 47 (100%) | 25 (80.6%) | 15 (100%) | 19 (61.2%) | 59 (76.6%) | 106 (85.5%) |
| MD2 | $\mu_0$ | 47 (100%) | 31 (100%) | 15 (100%) | 21 (67.7%) | 67 (87.0%) | 114 (91.9%) |
| | $\mu_1$ | 47 (100%) | 31 (100%) | 15 (100%) | 21 (67.7%) | 67 (87.0%) | 114 (91.9%) |
| | $\mu_2$ | 47 (100%) | 29 (93.5%) | 15 (100%) | 20 (64.5%) | 64 (83.1%) | 111 (89.5%) |
| Band1 | $\mu_0$ | 47 (100%) | 31 (100%) | 15 (100%) | 21 (67.7%) | 67 (87.0%) | 114 (91.9%) |
| | $\mu_1$ | 47 (100%) | 27 (87.1%) | 15 (100%) | 18 (58.1%) | 60 (77.9%) | 107 (86.3%) |
| | Intersection | 47 (100%) | 31 (100%) | 15 (100%) | 21 (67.7%) | 67 (87.0%) | 114 (91.9%) |
| Band2 | $\mu_0$ | 47 (100%) | 31 (100%) | 15 (100%) | 21 (67.7%) | 67 (87.0%) | 114 (91.9%) |
| | $\mu_1$ | 47 (100%) | 30 (100%) | 15 (100%) | 19 (61.3%) | 64 (83.1%) | 111 (89.5%) |
| | Intersection | 47 (100%) | 31 (100%) | 15 (100%) | 21 (67.7%) | 67 (87.0%) | 114 (91.9%) |

We use MD and Band as abbreviations for Minimum distance method and Confidence band method respectively. Indices 1 and 2 refer to the first and second normalization methods respectively. Rows labeled $\mu_j$ indicate diagnosis based on derivatives of order $j$. Entries are numbers(and percentages) of subjects correctly diagnosed.

to be used simultaneously in the diagnosis of breast tissue.

Different classification schemes are presented including "normal" vs "abnormal", "normal" vs "cancer" vs "abnormal benign", finer classification of "abnormal" into three types "cancer" vs "FA" vs "FC", and finer classification into four types as Haka et al(2005) showed in their work.

Our results indicate there is only a slight difference among the two normalization schemes "RANGE" and "STDEV". Under the strategy of minimum distance, using derivatives may in some cases yield better results than employing the mean response, while using the mean response (or using both the mean response and the first derivative simultaneously) can yield better results than using the first derivative if applying confidence bands method. Additionally, the derivatives mostly yield better results

Table 2.5: Raman spectra diagnosis results for normal, cancer and abnormal non cancer tissues

| Different Methods | | Normal (out of 47) | Cancer (out of 31) | FA/FC (out of 46) | Total Correct (out of 124) |
|---|---|---|---|---|---|
| MD1 | $\mu_0$ | 47 (100%) | 23 (74.2%) | 6 (13.0%) | 76 (61.3%) |
| | $\mu_1$ | 47 (100%) | 19 (61.3%) | 23 (50%) | 89 (71.8%) |
| | $\mu_2$ | 47 (100%) | 20 (64.5%) | 27 (58.7%) | 94 (75.8%) |
| MD2 | $\mu_0$ | 47 (100%) | 19 (61.3%) | 22 (47.8%) | 88 (71.0%) |
| | $\mu_1$ | 47 (100%) | 19 (61.3%) | 22 (47.8%) | 88 (71.0%) |
| | $\mu_2$ | 47 (100%) | 19 (64.5%) | 26 (56.5%) | 92 (74.2%) |
| Band1 | $\mu_0$ | 47 (100%) | 18 (58.1%) | 11 (23.9%) | 76 (61.3%) |
| | $\mu_1$ | 47 (100%) | 3 (9.7%) | 25 (54.3%) | 75 (60.5%) |
| | Intersection | 47 (100%) | 24 (77.4%) | 10 (66.7%) | 81 (65.3%) |
| Band2 | $\mu_0$ | 47 (100%) | 13 (41.9%) | 25 (54.3%) | 85 (68.5%) |
| | $\mu_1$ | 47 (100%) | 2 (6.5%) | 26 (56.5%) | 75 (60.5%) |
| | Intersection | 47 (100%) | 19 (61.3%) | 22 (54.3%) | 88 (71.0%) |

We use MD and Band as abbreviations for Minimum distance method and Confidence band method respectively. Indices 1 and 2 refer to the first and second normalization methods respectively. Rows labeled $\mu_j$ indicate diagnosis based on derivatives of order $j$. Entries are numbers(and percentages) of subjects correctly diagnosed.

Table 2.6: Raman spectra diagnosis results for cancer, FA and FC tissues

| Different Methods | | Cancer (out of 31) | FA (out of 15) | FC (out of 31) | Total Correct (out of 77) |
|---|---|---|---|---|---|
| MD1 | $\mu_0$ | 16 (51.3%) | 14 (93.3%) | 16 (51.6%) | 46 (59.7%) |
| | $\mu_1$ | 19 (61.3%) | 12 (80.0%) | 20 (64.5%) | 51 (66.2%) |
| | $\mu_2$ | 15 (48.4%) | 13 (86.7%) | 24 (77.4%) | 52 (67.5%) |
| MD2 | $\mu_0$ | 19 (61.3%) | 14 (93.3%) | 17 (54.8%) | 50 (64.9%) |
| | $\mu_1$ | 19 (61.3%) | 12 (80.0%) | 18 (58.1%) | 49 (63.6%) |
| | $\mu_2$ | 15 (48.4%) | 13 (86.7%) | 24 (77.4%) | 52 (67.5%) |
| Band1 | $\mu_0$ | 18 (58.1%) | 13 (86.7%) | 16 (51.6%) | 47 (61.0%) |
| | $\mu_1$ | 2 (6.5%) | 11 (73.3%) | 22 (71.0%) | 35 (45.5%) |
| | Intersection | 15 (48.4%) | 14 (93.3%) | 16 (51.6%) | 45 (58.4%) |
| Band2 | $\mu_0$ | 17 (54.8%) | 13 (86.7%) | 19 (61.3%) | 49 (63.6%) |
| | $\mu_1$ | 2 (6.5%) | 11 (73.3%) | 21 (67.7%) | 34 (44.2%) |
| | Intersection | 22 (71.0%) | 14 (93.3%) | 18 (58.1%) | 54 (70.1%) |

We use MD and Band as abbreviations for Minimum distance method and Confidence band method respectively. Indices 1 and 2 refer to the first and second normalization methods respectively. Rows labeled $\mu_j$ indicate diagnosis based on derivatives of order $j$. Entries are numbers(and percentages) of subjects correctly diagnosed.

with the minimum distance approach than the confidence bands approach. However, neither the minimum distance approach nor the confidence bands approach yielded as high a correct classification rate as the methodology by Haka et al (2005). This can be explained by two possible reasons.

First, the methodology by Haka et al (2005) related the Raman spectra for patients to the Raman spectra for the biochemical constituents of breast tissue, whereas the approaches considered here did not exploit any knowledge of biochemistry. Second, Haka et al (2005) did not normalize the patients' Raman spectra but rather the coefficients by which the patients' Raman spectra were related to the biochemical constituents of breast tissue. The RANGE and STDEV normalizations appear too simple and may have discarded some information that may have been useful for diagnosis. For instance, normal patients' Raman spectra have greater intensity/amplitude than abnormal patients' Raman spectra, but this information is discarded when each patient's Raman spectrum is constrained to lie between 0 and 1 as in the RANGE normalization. Although the normal patients were quite readily distinguished by the shapes of the Raman spectra, a similar principle may have applied in the comparison of cancer patients to benign abnormal patients. In other words, benign abnormal and cancer patients may have had spectra with similar shape but different amplitude, in which case a RANGE normalization would erase much of the distinction.

For further research, one possible improvement could be to enhance the minimum distance approach and the confidence bands approach by either exploiting knowledge of biochemistry or using a more sophisticated normalization. Also we can potentially improve the prediction accuracy by incorporating advanced statistical learning approaches such as "Stacking" or "Boosting". "Stacking" entails constructing ensembles of heterogeneous classifiers to combine individual predictions and outperforms selecting the best classifier[22], while "Boosting" uses a set of weak learners to create a single strong learner[70]. Stacking is historically one of the first ensemble learning

methods (combination of classifier) to improve classification accuracy[95]. It combines several base classifiers by means of a "meta - classifier" and shows excellent performance in many practical studies[46]. In Boosting, examples that are incorrectly predicted by previous classifiers in the series are weighted more heavily than examples that were correctly predicted. Thus Boosting attempts to produce new classifiers that are better able to predict examples for which the current ensemble's performance is poor.

However, existing approaches to Boosting and Stacking assume that inputs are known and attempt to relate inputs to noise-corrupted or otherwise distorted outputs. In Raman spectroscopy and other inverse problems, the inputs themselves are unknown and the object is to infer the inputs from noise-corrupted or otherwise distorted outputs. For instances, an output could be a patient's breast cancer diagnosis while the corresponding output could be the patient's Raman spectrum. Hence, our work in succeeding chapters is not merely to apply existing approaches but to modify them for use in inverse problems.

# Chapter 3 Stacking for Nonparametric Regression

## 3.1 Background

Stacked generalization, also called stacking (Wolpert, 1992), is the most well-known meta-learning method[95]. Breiman translated it into statistical language for the ordinary regression setting in 1993[8]. The idea is to combine predicted classifications from different classifiers into a new data set, and then employ a second-stage learning algorithm based on this new data set to improve the prediction of classification. Although this process as a whole can be iterated to obtain multiple level stacking, we will consider two levels of stacked generalization for now and describe its mathematical algorithm as specified by Ting and Witten(1999)[85].

Figure 3.1 illustrates this stacking framework. It comprises two levels: level-0 (with cross-validation) and level-1. Suppose the input is a dataset $L = (y_n, x_n)$, $n = 1, ..., N$ with $N$ entries, each standing for an observation. In each entry, $y_n$ is the class value while $x_n$ is a vector of the attribute values. Now we randomly divide the dataset $L$ into $J$ partitions as $L_1, ..., L_J$, each with almost equal length. At level-0, in using $J$-partitioned cross-validation, we denote $L_j$ as the $j$th partition and $L^{(-j)} = L - L_j$. They will be used as test and training sets respectively. We choose $K$ learning algorithms as level-0 generalizers so that the level-0 models $M_k^{(-j)}$, for $k = 1, ..., K$, are produced by applying the $k$th algorithm on the data of $L^{(-j)}$. Some examples of level 0 generalizers include linear discriminant analysis, logistic regression, decision trees, neural networks, Naive Bayesian(NB), nearest neighbor, Support Vector Machines and so on[66].

Let $z_{kjn}$ denote the resulting prediction of the model $M_k^{(-j)}$ on $x_{jn}$, the $n$th instance in the test set $L_j$. The output of the level-0 process is the level-1 data ($L'_{CV}$) that is

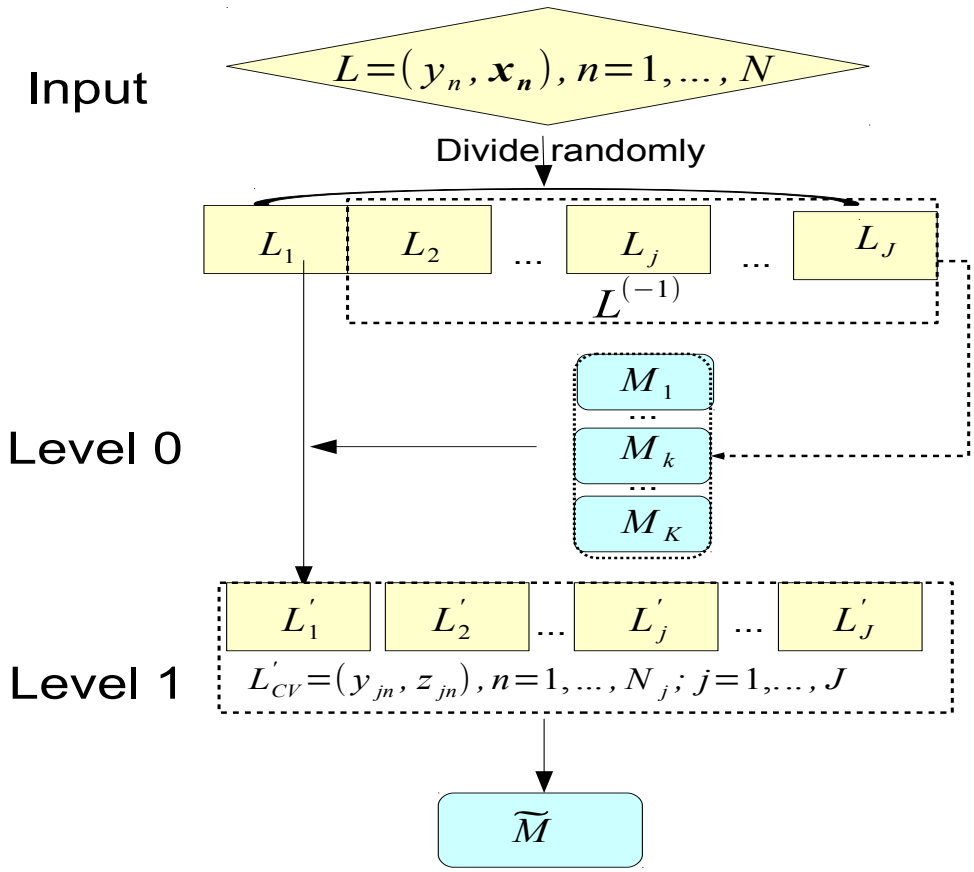


Figure 3.1: Stacking framework illustrating the two levels of model training. In level 0, $L_1$ and $L^{(-1)}$, defined by a random partition of input dataset $L$, are used as an example to show a typical cross validation to obtain level 1 data. In level 1, $z$ is a vector containing the outputs from different models $M_k$ used in level 0. $\tilde{M}$ is the level 1 model built from level 1 data $L_{CV}^{'}$ using a level 1 generalizer.

assembled from the outputs of the above $K$ models:

$$L'_{CV} = (y_{jn}, z_{1jn}, \ldots, z_{Kjn}), n = 1, \ldots, N_j; j = 1, \ldots, J.$$

From level-1 data ($L'_{CV}$) we use some learning algorithm, the level-1 generalizer, to derive a model $\tilde{M}$ for $y$ as a function of $(z_1, ..., z_K)$, the level-1 model. Level-1 generalizer can be the same as level-0 generalizers. Several level-1 generalizers such as NB, multi-response linear regression algorithm, meta decision trees (MDTs) have been used[85][53]. To complete the training process, the ultimate level-0 models $M_k, k = 1, ..., K$ should be constructed using all data in $L$; these will be used to classify any new instance that may arise.

Next, let us consider the classification process using the previous induced models $M_k, k = 1, ..., K$ along with $\tilde{M}$. Given a new observation, models $M_k$ produce a vector $(z_1, ..., z_K)$. The final classification result for that observation could be obtained by using the level-1 model $\tilde{M}$ on the vector $(z_1, ..., z_K)$.

It is recommended that using class probabilities from base classifiers rather than a single class prediction could improve stacking performance[85]. Ensembles of diverse base-level classifiers are known to reduce multicollinearity and yield good performance [52]. Merz (1999) proposes SCANN (Stacking, Correspondence Analysis and Nearest Neighbor) that combines correspondence analysis into stacking to remove correlations between base-level classifiers[53]. StackingC (Seewald, 2002), a variant of stacking, is proposed to perform well in both binary class and multi-class problem[74]. Meta-decision trees (MDTs) were first implemented into stacking by Todorovski and Dzeroski and claimed to perform better than stacking with ordinary decision tress[86].

However, the previous work has been mainly focused on the computing implementation of stacking, and improved accuracy was only justified by reducing classification error rate from a collection of datasets. We, too, will examine our methodology using a collection of datasets. However, we will also provide some theoretical rationale for our methodology.

The rest of this Chapter will be organized as follows. In section 2, we will establish a probabilistic framework for convex combination of nonparametric regression based classifiers employing the minimum distance methods described in Chapter 2 with illustration in Raman spectroscopy data. In section 3, we will theoretically justify the asymptotic properties of a stacked generalizer based on exponentially weighted vote in the two step sequential classifications. In section 4, we will perform simulations using an artificial waveform data set acquired from the UCI Machine Learning Repository to show the performance of stacking under different noise levels[2]. Finally, in section 5, we will summarize the major findings and discuss the implications of stacking in classification problems.

### 3.2 Probability framework for convex combination of base classifiers

In this section, we establish a probability framework for convex combination of base classifiers with the aim to investigate whether combination of base classifiers would perform better than a single classifier alone. Convex combination is a linear combination which requires the coefficients to sum to 1, and that these coefficients are non-negative. In a simple case when there are two classifiers, convex combination $h(x)$ has the form $h(x) = \lambda h_1(x) + (1-\lambda)h_2(x)$, while $1 \geq \lambda \geq 0$, $h_1(x)$, $h_2(x)$ are two classifiers to be combined.

Recall we have generated a group of classifiers employing minimum distance approach in Chapter 2 defined as the minimizer

$$\arg \min_{c \in \{1,2,3,4\}} \sum_{i=1}^{n} \left| \widehat{\mu_u^{(j)}}(x_i) - \widehat{\mu_c^{(j)}}(x_i) \right|$$

for $c \in \{1,2,3,4\}$ and $j \in \{0,1,2\}$ based on $L^1$ distance.

This is similar to the minimizer based on $L^2$ distance:

$$\arg \min_{c \in \{1,2,3,4\}} \sum_{i=1}^{n} \{\widehat{\mu_u^{(j)}}(x_i) - \widehat{\mu_c^{(j)}}(x_i)\}^2$$

for $c \in \{1,2,3,4\}$ and $j \in \{0,1,2\}$.

Since the estimators from compound estimation are linear in the observed responses, we could write $\widehat{\mu_u^{(j)}}(x_i)$ in the form $\widehat{\mu_u^{(j)}}(x) := L^{(j)}(x)Y$, where $Y$ denotes the observed spectrum for a sample of unknown diagnoses $u$. Suppose the reference curves are known, and that the Raman intensity level $Y_i$ given the Raman spectrum truly belongs to $c$ denoted as $Y_i|c$ follows $Y_i|c := \mu_c(x_i) + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$ for $c \in \{1,2,3,4\}$, the classifier that whether a breast sample tissue being classified as invasive carcinoma (for example: $c = 4$) based on minimum $L^2$ distance approach from compound estimation for mean response or derivative can be defined as

$$h_j(Y) = 1_{\forall s \in \{1,2,3\}: \sum 2(L^{(j)}Y)_i(\mu_{si} - \mu_{4i}) < \sum(\mu_{si}^2 - \mu_{4i}^2)},$$

which represents the classifier from compound estimation for mean response or first derivative respectively when $j = 0$ or $j = 1$.

Therefore, given a Raman spectrum truly belongs to cancer $c = 4$, we have

$$\begin{pmatrix} (2(\mu_1 - \mu_4)^T L^{(j)} \\ (2(\mu_2 - \mu_4)^T L^{(j)} \\ (2(\mu_3 - \mu_4)^T L^{(j)} \end{pmatrix} Y \sim$$

$$N\left(\begin{pmatrix} 2(\mu_1 - \mu_4)^T L^{(j)} \\ 2(\mu_2 - \mu_4)^T L^{(j)} \\ 2(\mu_3 - \mu_4)^T L^{(j)} \end{pmatrix} \mu_4, 4\sigma^2 \begin{pmatrix} (\mu_1 - \mu_4)^T L^{(j)} \\ (\mu_2 - \mu_4)^T L^{(j)} \\ (\mu_3 - \mu_4)^T L^{(j)} \end{pmatrix} \begin{pmatrix} (\mu_1 - \mu_4)^T L^{(j)} \\ (\mu_2 - \mu_4)^T L^{(j)} \\ (\mu_3 - \mu_4)^T L^{(j)} \end{pmatrix}^T\right)$$

for $j \in \{0,1\}$.

Suppose we have two classifiers, such as indicator functions $h_0(Y)$ and $h_1(Y)$ defined above based on minimum $L^2$ distance approach from mean response and its derivative. There are two scenarios for convex combination of the two classifiers.

One obvious way is to define $h(Y) = \alpha h_0(Y) + (1-\alpha)h_1(Y)$. In this case, we have

the following conclusions:

$$P(h(Y) = 0|c) = P(h_0(Y) = 0 \cap h_1(Y) = 0|c)$$

$$P(h(Y) = \alpha|c) = P(h_0(Y) = 1 \cap h_1(Y) = 0|c)$$

$$P(h(Y) = 1 - \alpha|c) = P(h_0(Y) = 0 \cap h_1(Y) = 1|c)$$

$$P(h(Y) = 1|c) = P(h_0(Y) = 1 \cap h_1(Y) = 1|c),$$

for $c \in \{1, 2, 3, 4\}$.

Under assumption of $Y_i|c := \mu_c(x_i) + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$ for $c \in \{1, 2, 3, 4\}$, the probability $P(h(Y) = j|c)$ for $j \in \{0, \alpha, 1 - \alpha, 1\}$ is a function of $\sigma^2$, and can be calculated by way of multinomial distribution. Table 3.1 shows probability calculations under different $\sigma^2$ given the Raman spectrum truly belongs to $c$.

When $\sigma^2$ is very small, for instance, $\sigma^2 = 0.0001$, there is no difference for the two classifiers for all pathologies to be classified into; indeed, all classifications are correct. However, as expected, the probability of correct classification by both classifiers will become lower as $\sigma^2$ increases (such as increasing from 0.01 to 0.1). Also, the impacts of $\sigma^2$ on the probabilities are different when classifying different pathologies. For example, it might be easier to make correct classifications for both classifiers when classifying normal than fibroadenoma if $\sigma^2 = 0.01$ ($P(h(Y) = 1|c) = 1$ for normal(c=1), and $P(h(Y) = 1|c) = 0.684$ for fibroadenoma(c=3)). Moreover, we can infer from $P(h(Y) = \alpha|c)$ and $P(h(Y) = 1 - \alpha|c)$ that if $\sigma^2 = 0.1$, the classifier from compound estimation for mean response curve is more likely to make correct classification than the classifier from compound estimation for first derivative. However, an obvious difficulty with the preceding scheme is what decision to make if $h_0(Y)$ and $h_1(Y)$ disagree. The results suggest we may defer to $h_0(Y)$, but then we do not really take into account $h_1(Y)$.

This motivates consideration of a second scenario in which the indicator conditions

Table 3.1: Probability calculation for $P(h(Y) = j|c)$ under different $\sigma^2$

| $P(h(Y) = j\|c)$ | $\sigma^2$ | $j = 0$ | $j = \alpha$ | $j = 1 - \alpha$ | $j = 1$ |
|---|---|---|---|---|---|
| c=1(N) | 0.0001 | 0 | 0 | 0 | 1 |
| | 0.001 | 0 | 0 | 0 | 1 |
| | 0.01 | 0 | 0 | 0 | 1 |
| | 0.1 | 0 | 0.022 | 0 | 0.978 |
| c=2(FC) | 0.0001 | 0 | 0 | 0 | 1 |
| | 0.001 | 0 | 0 | 0 | 1 |
| | 0.01 | 0 | 0 | 0 | 1 |
| | 0.1 | 0.012 | 0.175 | 0.015 | 0.798 |
| c=3(FA) | 0.0001 | 0 | 0 | 0 | 1 |
| | 0.001 | 0 | 0.065 | 0 | 0.935 |
| | 0.01 | 0 | 0.316 | 0 | 0.684 |
| | 0.1 | 0.036 | 0.431 | 0.012 | 0.521 |
| c=4(C) | 0.0001 | 0 | 0 | 0 | 1 |
| | 0.001 | 0 | 0 | 0 | 1 |
| | 0.01 | 0 | 0.097 | 0 | 0.903 |
| | 0.1 | 0.036 | 0.352 | 0.029 | 0.583 |

Entries are probabilities of $P(h(Y) = j|c)$ for $j \in \{0, \alpha, 1-\alpha, 1\}$ and $c \in \{1, 2, 3, 4\}$. $j$ represents the value of convex combination $h(Y)$ of two classifiers $h_0(Y)$ and $h_1(Y)$. If $j = 0$, both the two classifiers make mistake, while both classifier are correct if $j = 1$. When $j = \alpha$, only the classifier $h_0(Y)$ is correct, while if $j = 1 - \alpha$, only $h_1(Y)$ is correct. $c$ denotes the true pathology type, with 1, 2, 3, 4 represent normal, fibrocystic change, fibroadenoma, and invasive carcinoma, respectively .

are convexly combined rather than the indicators themselves:

$$h(Y) = 1_{\forall s\in\{1,2,3\}:\alpha\sum 2(L^{(0)}Y)_i(\mu_{si}-\mu_{4i})+(1-\alpha)\sum 2(L^{(1)}Y)_i(\mu_{si}-\mu_{4i})<\sum(\mu_{si}^2-\mu_{4i}^2)}.$$

Here, we can choose weight $\alpha$ to maximize $P(h(Y) = 1|c)$, and this will not (in general) be equivalent to using $h_0(Y)$.

Figure 3.2 illustrates this idea. It is shown that, for invasive carcinoma, maximization of the probability of making a correct prediction could be achieved when the weight $1 - \alpha$ falls between 0 and 1, more specifically when $1 - \alpha \approx 0.25(\sigma^2 = 0.01)$. However, for normal tissue, choosing $\alpha = 0$ appears best. Overall, it is most likely to make correct classification for normal among all pathology types. The pattern of likelihood of correct classifications for different pathology types differs for different

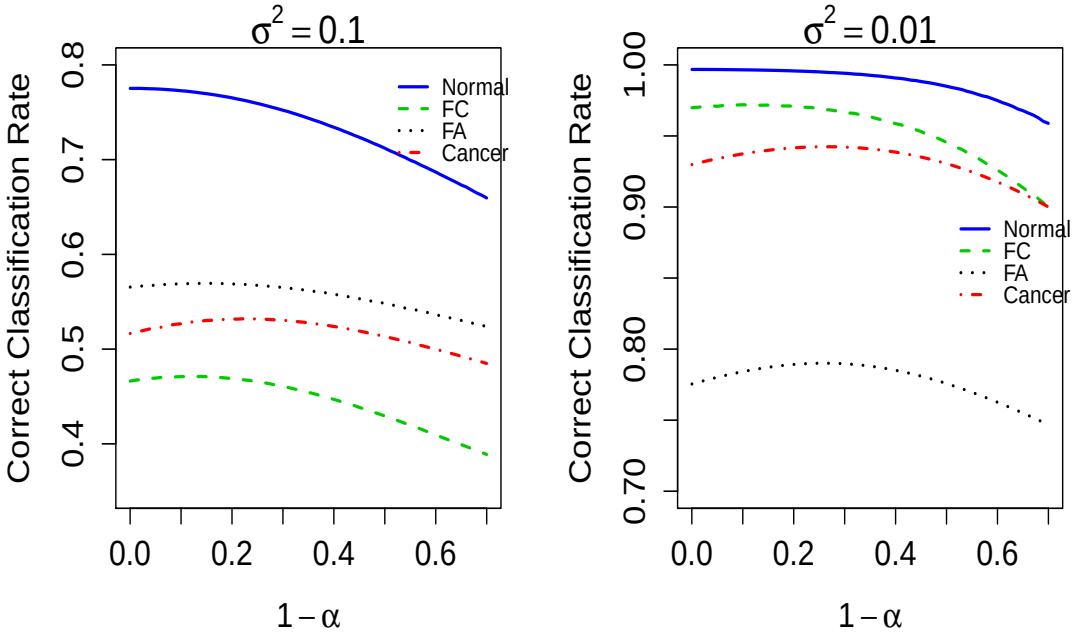


Figure 3.2: Correct classification probability by convex combination $h(Y)$ vs weight $1-\alpha$ in the 2nd scenario

$\sigma^2$. For example, fibroadenoma is least likely to be correctly classified by $h(Y)$ if $\sigma^2 = 0.1$, while it is not this case if $\sigma^2 = 0.01$. Also, as $\sigma^2$ decreases from 0.1 to 0.01, the correct classification probability by convex combination of indicator conditions increases for all pathology types.

To sum up, in the second scenario for defining $h(Y)$ considered above, we can conclude that combination of base classifiers could perform better than a single classifier alone for most pathological types using Raman spectra data. And the weight for $h_0(Y)$ should generally be chosen larger than that for $h_1(Y)$ assuming that $Y_i|c := \mu_c(x_i) + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$ for $c \in \{1, 2, 3, 4\}$.

However, there are some limitations for these convex combinations. First of all, the probability of correct classification depends on $\sigma^2$. Since $\sigma^2$ is an unknown parameter, it is difficult to estimate the probability in practice, and this makes the choice of optimal weight more difficult. Further, the assumption of homoscedasticity

for $\sigma^2$ may not be true in real situations. This may be part of the reason why the classifier based on compound estimation for first derivative generates better results than mean response shown in Chapter 2, while under homoscedasticity assumption, the probability of making correct classification from mean response $h_0(Y)$ is generally higher than from first derivative $h_1(Y)$, and should be weighted more than $h_1(Y)$ in the second scenario. In addition, the preceding assumes independent errors and additive normally distributed errors which may not be realistic. Finally, the true $\mu_c(x)$ was unknown and estimated by the average of the curves for each pathology type, which might contain some error. Therefore we will explore a more flexible method for stacking of base classifiers in section 3.3, in that strong assumptions on the nature of the data are not required.

### 3.3 Performance for ensemble classifiers

Freund, Mansour and Schapire (2004) proposed a weighted average vote of base classifiers for binary classification learning which can be applied to any type of base classifiers[25]. Specifically, suppose the predicted outcome $C \in \{+1, -1\}$, and let H be a fixed class of hypotheses (also called base classifiers), mappings from $Y$ to $C$. In this weighted average algorithm, each base classifier is weighted exponentially with respect to its training error. The weight is defined as $w_{h_i} := \exp(-\eta\hat{\epsilon}_{h_i})$, where $\hat{\epsilon}_{h_i}$ is the training error of base classifier $h_i$; $\eta = \ln(8|H|)m^{1/2-\theta}$, in which $|H|$ is the number of base classifiers, $m$ is the size of training data, and $0 < \theta < \frac{1}{2}$. Let

$$\hat{l}_\eta(Y) = \frac{1}{\eta} \ln \left( \frac{\sum_{h,h(Y)=+1} w(h)}{\sum_{h,h(Y)=-1} w(h)} \right),$$

the weighted average prediction is defined to be $\text{sign}(\hat{l}_\eta(Y))$ if $|\hat{l}_\eta(Y)| > \Delta$.

Moreover, the authors have claimed that overfitting could be avoided by allowing the algorithm to abstain from predicting some instances with output "no prediction", which occurs when $|\hat{l}_\eta(Y)| \leq \Delta$. Through the authors' theoretical justification, the

probability that this classifier makes an error when it does not abstain is at most $2\epsilon + \tilde{O}(1/\sqrt{m})$, where $\epsilon$ is the smallest classification error among the base classifiers, and $\tilde{O}$ indicates an approximation that neglects tuning parameter $\theta$ and an additional reliability parameter $\delta$. Further, the probability that the classification rule will output "no prediction" is upper bounded by $5\epsilon + \tilde{O}(\ln(|H|)/\sqrt{m})$. Note that $2\epsilon$ and $5\epsilon$ are worst case scenarios; correct classification is often better than for any base classifier.

Now we are establishing theoretical results showing a justification of sequentially applying such a weighted average algorithm; suppose there are four categories and we perform classification in 2 steps.

Let $C$ denote different categories to be classified into, and $C \in \{1, 2, 3, 4\}$. In the first step classification, we combine $\{1, 2\}$ as one group denoted as $+1$, and the rest $\{3, 4\}$ as the other group denoted as $-1$. Let $g_1$ denote the first step classification, so we have $g_1 \in \{+1, -1\}$.

In the second step classification, we separate class 1 from class 2 in the first group, and separate class 3 from class 4 in the second group, respectively, based on results from first step classification. Let $g_{2(+1)}$ denote the second step prediction separating class 1 denoted as $+1$ from class 2 denoted as $-1$, so we have $g_{2(+1)} \in \{+1, -1\}$. In the same way, let $g_{2(-1)}$ denote the second step prediction separating class 3 denoted as $+1$ from class 4 denoted as $-1$, so we have $g_{2(-1)} \in \{+1, -1\}$.

Figure 3.3 shows an illustration of multiclass prediction matching $\{g_1, g_{2g_1}\}$ to $C$ for a four class problem.

### Notations and definition of the algorithm

Let $D$ be a fixed but unknown distribution over (Y, C) pairs. Suppose $H$ is a fixed class of hypotheses mapping from $Y$ to $\{g_1, g_{2g_1}\}$, corresponding to $C$. Each hypothesis $h$ comprises two stage estimators $\{h_1(Y), h_{2(h_1)}(Y)\}$. We denote the true error

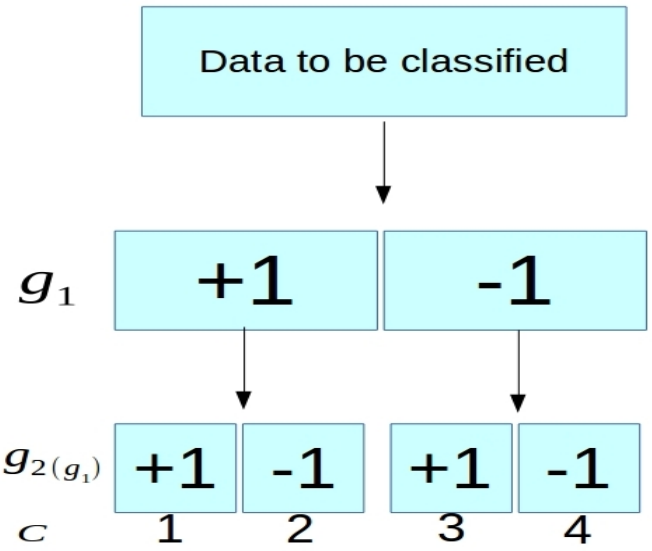


Figure 3.3: Two stage multiclass prediction for a four class problem

of a hypothesis $h$ in the first and second stage classification by $\varepsilon(h_1)$ and $\varepsilon(h_{2(k)})$, respectively. Then we have

$$\varepsilon(h_1) := \Pr_{(Y,C)\sim D}[h_1(Y) \neq g_1]$$

and

$$\varepsilon(h_{2(k)}) := \Pr_{(Y,C)\sim D}[h_{2(h_1)}(Y) \neq g_2 | h_1(Y) = g_1 = k]$$

for $k \in \{+1, -1\}$. Suppose $\zeta(h)$ is the probability of a hypothesis $h$ classifying as $+1$ given a correct classification in the first tep: $\zeta(h) := \Pr_{(Y,C)\sim D}[h_1 = g_1 = +1 | h_1 = g_1]$, the total error of a hypothesis is

$$\begin{aligned} \varepsilon(h) &= \varepsilon(h_1) + (1 - \varepsilon(h_1))\zeta(h)\varepsilon(h_{2(+1)}) \\ &+ (1 - \varepsilon(h_1))(1 - \zeta(h))\varepsilon(h_{2(-1)}). \end{aligned}$$

Let $m$ be the sample size, and $|H|$ be number of base classifiers. The estimated errors for $\varepsilon(h_1)$ and $\varepsilon(h_{2(k)})$ are defined by

$$\hat{\varepsilon}(h_1) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{h_1(Y_i) \neq g_{i1}\}}$$

and

$$\hat{\varepsilon}(h_{2(k)}) = \frac{\sum_{i=1}^{m} 1_{\{h_2(Y_i) \neq g_{i2g_{i1}} \cap h_{i1} = g_{i1} = k\}}}{\sum_{i=1}^{m} 1_{\{h_{i1} = g_{i1} = k\}}}$$

for $k \in \{+1, -1\}$.

For each hypothesis $h$, there are three weights defined as

$$w_1(h) := \exp(-\eta_1 \hat{\varepsilon}(h_1))$$

$$w_{2(+1)}(h) := \exp(-\eta_{2(+1)} \hat{\varepsilon}(h_{2(+1)}))$$

$$w_{2(-1)}(h) := \exp(-\eta_{2(-1)} \hat{\varepsilon}(h_{2(-1)}))$$

The prediction on a new instance is based on a combination of two stage estimators $\{\hat{l}_1(Y), \hat{l}_{2sign(\hat{l}_1(Y))}(Y)\}$, where

$$\hat{l}_1(Y) \;=\; \frac{1}{\eta_1} \ln \left( \frac{\sum_{h,h_1(Y)=+1} w_1(h)}{\sum_{h,h_1(Y)=-1} w_1(h)} \right).$$

and

$$\hat{l}_{2sign(\hat{l}_1(Y))}(Y) \;=\; \frac{1}{\eta_{2sign(\hat{l}_1(Y))}} \ln \left( \frac{\sum_{h,h_{2sign(\hat{l}_1(Y))}(Y)=+1} w_{2sign(\hat{l}_1(Y))}(h)}{\sum_{h,h_{2sign(\hat{l}_1(Y))}(Y)=-1} w_{2sign(\hat{l}_1(Y))}(h)} \right).$$

The true log ratios $l_1(Y)$ and $l_{2sign(l_1(Y))}(Y)$ are defined to be replacing $\hat{\varepsilon}$ with $\varepsilon$. Suppose $\Delta_1$ and $\Delta_{2sign(\hat{l}_1(Y))}$ are the abstention thresholds in the first and second stage estimations. The final prediction is defined to be

$$\hat{p}(Y) = \begin{cases} \{sign(\hat{l}_1(Y)), sign(\hat{l}_{2sign(\hat{l}_1(Y))}(Y))\}, & \text{if } |\hat{l}_1(Y)| > \Delta_1 \quad \text{and } |\hat{l}_{2sign(\hat{l}_1(Y))}(Y)| > \Delta_{2sign(\hat{l}_1(Y))}, \\ \{sign(\hat{l}_1(Y)), 0\}, & \text{if } |\hat{l}_1(Y)| > \Delta_1 \quad \text{and } |\hat{l}_{2sign(\hat{l}_1(Y))}(Y)| \leq \Delta_{2sign(\hat{l}_1(Y))}, \\ 0, & \text{otherwise.} \end{cases}$$

**Analysis of the algorithm**

We will first show how to choose the thresholds: $\Delta_1$, $\Delta_{2(+1)}$, and $\Delta_{2(-1)}$ by presenting the theorems that differences between the estimated log ratios and the true log ratios are small.

Recall Theorem 1 and Theorem 2 by Freund and colleagues:

" Theorem 1: For any distribution D, any instance x, any $\lambda$, $\eta > 0$ and any $s \in \{-1, +1\}$,

$$\Pr_{S \sim D^m}[s(l(x) - \hat{l}(x)) \geq 2\lambda + \frac{\eta}{8m}] \;\leq\; 2e^{-2\lambda^2 m},"$$

While Theorem 1 holds for any fixed instance, Theorem 2 holds with respect to a randomly chosen instance.

" Theorem 2: For any $\delta > 0$ and $\eta > 0$, if we set

$$\Delta = 2\sqrt{\frac{\ln(\sqrt{2}/\delta)}{m}} + \frac{\eta}{8m},$$

then we have

$$\Pr_{(x,y)\sim D}[s(l(x) - \hat{l}(x)) \geq \Delta] \leq \delta$$

with probability at least $1 - \delta$. "

Now we state analogues of Theorem 2 for our sequential stacked classifier.

**Theorem 3.3.1.** *If we set* $\Delta_1 = 2\sqrt{\frac{\ln(\sqrt{2}/\delta_1)}{m}} + \frac{\eta_1}{8m}$ *for any* $\delta_1 > 0$, $\eta_1 > 0$, *then with probability at least* $1 - \delta_1$,

$$\Pr_{(Y,C)\sim D}[|l_1(Y) - \hat{l}_1(Y)| \geq \Delta_1] \leq \delta_1.$$

*Proof.* We apply Theorem 2 of Freund and colleagues with the sample distribution $(x, y) \sim D$ replaced by $(Y, C) \sim D$, the function $l$, $\hat{l}$ set equal to $l_1$ and $\hat{l}_1$ respectively, $\eta$ set to $\eta_1$, $\Delta$ set to $\Delta_1$ and $\delta$ set to $\delta_1$. □

**Theorem 3.3.2.** *Let* $m_{2(+1)} = \sum_{i=1}^{m} 1_{\{sign(\hat{l}_{i1}(Y))=+1\}}$ *and* $m_{2(-1)} = \sum_{i=1}^{m} 1_{\{sign(\hat{l}_{i1}(Y))=-1\}}$, *then we have the following statements for any* $\delta_{2(+1)} = \delta_{2(-1)} = \delta_2 > 0$*:*

*i. If we set* $\Delta_{2(+1)} = 2\sqrt{\frac{\ln(\sqrt{2}/\delta_2)}{m_{2(+1)}}} + \frac{\eta_{2(+1)}}{8m_{2(+1)}}$ *for any* $\eta_{2(+1)} > 0$, *then with probability at least* $1 - (\delta_1 + \delta_2)$,

$$\Pr_{(Y,C)\sim D}[|l_{2(+1)}(Y) - \hat{l}_{2(+1)}(Y)| \geq \Delta_{2(+1)} \bigcap |l_1(Y) - \hat{l}_1(Y)| \geq \Delta_1] \leq \delta_1\delta_2;$$

*ii. Likewise, if we set* $\Delta_{2(-1)} = 2\sqrt{\frac{\ln(\sqrt{2}/\delta_2)}{m_{2(-1)}}} + \frac{\eta_{2(-1)}}{8m_{2(-1)}}$ *for any* $\eta_{2(-1)} > 0$, *then with probability at least* $1 - (\delta_1 + \delta_2)$,

$$\Pr_{(Y,C)\sim D}[|l_{2(-1)}(Y) - \hat{l}_{2(-1)}(Y)| \geq \Delta_{2(-1)} \bigcap |l_1(Y) - \hat{l}_1(Y)| \geq \Delta_1] \leq \delta_1\delta_2.$$

*Proof.* From Theorem 3.3.1, let $\Delta_1 = 2\sqrt{\frac{\ln(\sqrt{2}/\delta_1)}{m}} + \frac{\eta_1}{8m}$ for any $\delta_1 > 0$, $\eta_1 > 0$,

$$\Pr[\Pr_{(Y,C)\sim D}[|l_1(Y) - \hat{l}_1(Y)| \geq \Delta_1] \leq \delta_1] \geq 1 - \delta_1. \tag{3.1}$$

Similarly, in the second step classification, we have

$$\Pr[\Pr_{(Y,C)\sim D}[|l_{2(+1)}(Y) - \hat{l}_{2(+1)}(Y)| \geq \Delta_{2(+1)}] \leq \delta_2] \geq 1 - \delta_2. \tag{3.2}$$

Let A denote $|l_1(Y) - \hat{l}_1(Y)| \geq \Delta_1$, and B denote $|l_{2(+1)}(Y) - \hat{l}_{2(+1)}(Y)| \geq \Delta_{2(+1)}$, 3.1 and 3.2 can be rewritten as

$$\Pr[\Pr_{(Y,C)\sim D}[A] \leq \delta_1] \geq 1 - \delta_1 \tag{3.3}$$

and

$$\Pr[\Pr_{(Y,C)\sim D}[B] \leq \delta_2] \geq 1 - \delta_2. \tag{3.4}$$

Combining 3.3 and 3.4, we have

$$\Pr[\Pr_{(Y,C)\sim D}[A \cap B] \leq \delta_1\delta_2] \geq 1 - (\delta_1 + \delta_2), \tag{3.5}$$

which is equivalent to that with probability at least $1 - (\delta_1 + \delta_2)$,

$$\Pr_{(Y,C)\sim D}[|l_{2(+1)}(Y) - \hat{l}_{2(+1)}(Y)| \geq \Delta_{2(+1)} \bigcap |l_1(Y) - \hat{l}_1(Y)| \geq \Delta_1] \leq \delta_1\delta_2.$$

So far we prove the claim for i. The proof for ii is almost identical. □

Then we begin to show the performance of sequential stacked classifier by recalling Theorem 4 of Freund and colleagues: "Let $H$ be a finite hypothesis class and let $\epsilon$ be the error of the best hypothesis in $H$ with respect to the distribution $D$ over the examples, that is, $\epsilon = \min\{\varepsilon(h) : h \in H\}$. Let $\eta > 0$ and $\Delta \geq 0$ be such that $\Delta\eta \leq 1/2$. Then for any $\gamma \geq \ln(8|H|)/\eta$,

$$\Pr_{(x,y)\sim D}[yl(x) \leq 0] \leq 2(1 + 2|H|e^{-\eta\gamma})(\epsilon + \gamma)$$

and

$$\begin{aligned} \Pr_{(x,y)\sim D}[yl(x) \leq 2\Delta] &\leq (1 + e^{2\Delta\eta})(1 + 2|H|e^{-\eta\gamma})(\epsilon + \gamma) \\ &\leq 4(1 + 2|H|e^{-\eta\gamma})(\epsilon + \gamma)." \end{aligned}$$

With this context, we now state an analogue for our sequential stacked classifier.

**Theorem 3.3.3.** *Let $\epsilon$ be the error of the best hypothesis in $H$. Suppose the best hypothesis $h^*$ meets the conditions below:*

*1.* $\epsilon = \min\{\varepsilon(h) : h \in H\} = \epsilon(h^*)$;

*2.* $\epsilon_1 = \min\{\varepsilon(h_1) : h \in H\} = \epsilon(h_1^*)$;

*3.* $\epsilon_{2(+1)} = \min\{\varepsilon(h_{2(+1)}) : h \in H\} = \epsilon(h_{2(+1)}^*)$, *where*

$$\varepsilon(h_{2(+1)}) := \Pr_{(Y,C)\sim D}[h_{2(+1)}(Y) \neq g_2 | h_1(Y) = g_1 = +1];$$

*4.* $\epsilon_{2(-1)} = \min\{\varepsilon(h_{2(-1)}) : h \in H\} = \epsilon(h_{2(-1)}^*)$, *where*

$$\varepsilon(h_{2(-1)}) := \Pr_{(Y,C)\sim D}[h_{2(-1)}(Y) \neq g_2 | h_1(Y) = g_1 = -1].$$

$\Pr_{(Y,C)\sim D}[\chi]$ *be the probability of total classification error, defined as*

$$\Pr_{(Y,C)\sim D}[\chi] = \Pr_{(Y,C)\sim D}[g_1 l_1(Y) \leq 0 \cup g_{2(g_1)} l_{2(g_1)}(Y) \leq 0].$$

*Let* $\eta_1, \eta_{2(+1)}, \eta_{2(-1)} > 0$ *and* $\Delta_1, \Delta_{2(+1)}, \Delta_{2(-1)} \geq 0$ *be such that* $\Delta_1\eta_1 \leq 1/2$, $\Delta_{2(+1)}\eta_{2(+1)} \leq 1/2$, $\Delta_{2(-1)}\eta_{2(-1)} \leq 1/2$. *Then for any* $\gamma \geq \ln(8|H|)/\min\{\eta_1, \eta_{2(+1)}, \eta_{2(-1)}\}$,

$$\begin{aligned}\Pr_{(Y,C)\sim D}[\chi] &\leq 2(1 + 2|H|e^{-\eta_1\gamma})(\epsilon_1 + \gamma) \\ &+ 2\max\{(1 + 2|H|e^{-\eta_{2(+1)}\gamma})(\epsilon_{2(+1)} + \gamma), (1 + 2|H|e^{-\eta_{2(-1)}\gamma})(\epsilon_{2(-1)} + \gamma)\}.\end{aligned}$$

*Proof.* Define the probability of error in the first step classification as $\Pr_{(Y,C)\sim D}[\chi_1] := \Pr_{(Y,C)\sim D}(g_1 l_1(Y) \leq 0)$. Since $\gamma \geq \ln(8|H|)/\min\{\eta_1, \eta_{2(+1)}, \eta_{2(-1)}\}$, then $\gamma$ satisfies $\gamma \geq \ln(8|H|)/\eta_1$, $\gamma \geq \ln(8|H|)/\eta_{2(+1)}$, $\gamma \geq \ln(8|H|)/\eta_{2(-1)}$. Let $\Pr_{(Y,C)\sim D}(\chi_{2(+1)})$, $\Pr_{(Y,C)\sim D}(\chi_{2(-1)})$ be the conditional probabilities of error in the second step classification given correctly classified in the first step as $+1$ and $-1$ respectively:

$$\Pr_{(Y,C)\sim D}(\chi_{2(+1)}) := \Pr_{(Y,C)\sim D}(g_{2(+1)} l_{2(+1)}(Y) \leq 0 \mid sign(l_1(Y)) = g_1 = +1);$$

$$\Pr_{(Y,C)\sim D}(\chi_{2(-1)}) := \Pr_{(Y,C)\sim D}(g_{2(-1)}l_{2(-1)}(Y) \leq 0 \mid sign(l_1(Y)) = g_1 = -1).$$

By using Theorem 4 from Freund and colleagues, we have

$$\Pr_{(Y,C)\sim D}[\chi_1] \leq 2(1 + 2|H|e^{-\eta_1\gamma})(\epsilon_1 + \gamma),$$

$$\Pr_{(Y,C)\sim D}[\chi_{2(+1)}] \leq 2(1 + 2|H|e^{-\eta_{2(+1)}\gamma})(\epsilon_{2(+1)} + \gamma),$$

$$\Pr_{(Y,C)\sim D}[\chi_{2(-1)}] \leq 2(1 + 2|H|e^{-\eta_{2(-1)}\gamma})(\epsilon_{2(-1)} + \gamma).$$

Let $\zeta$ be the probability of classifying as $+1$ given a correct classification in the first step, $0 < \zeta < 1$, then we have

$$\begin{aligned}
\Pr_{(Y,C)\sim D}[\chi] &= \Pr_{(Y,C)\sim D}[\chi_1] + (1 - \Pr_{(Y,C)\sim D}[\chi_1])\zeta \Pr_{(Y,C)\sim D}[\chi_{2(+1)}] \\
&+ (1 - \Pr_{(Y,C)\sim D}[\chi_1])(1 - \zeta) \Pr_{(Y,C)\sim D}[\chi_{2(-1)}] \\
&\leq 2(1 + 2|H|e^{-\eta_1\gamma})(\epsilon_1 + \gamma) \\
&+ (1 - \Pr_{(Y,C)\sim D}[\chi_1])\max\{\Pr_{(Y,C)\sim D}[\chi_{2(+1)}], \Pr_{(Y,C)\sim D}[\chi_{2(-1)}]\} \\
&\leq 2(1 + 2|H|e^{-\eta_1\gamma})(\epsilon_1 + \gamma) \\
&+ 2\max\{(1 + 2|H|e^{-\eta_{2(+1)}\gamma})(\epsilon_{2(+1)} + \gamma), (1 + 2|H|e^{-\eta_{2(-1)}\gamma})(\epsilon_{2(-1)} + \gamma)\}.
\end{aligned}$$

□

### 3.4 Simulations on waveform data

#### Description of waveform data

We use waveform database (version 2) from UCI machine learning repository for simulations. This dataset, first described by Breiman et al[9], has three classes of waves, each of which is based on random convex combination of two of three base waveforms with noise added to all of the attributes. In the waveform dataset version 1, there are 21 attributes corresponding to positions on a horizontal axis such as in

Figure 3.4, while in the version 2 waveform dataset, there are 40 attributes, with the latter added 19 attributes all noise with mean 0 and variance 1. More specifically, suppose $b_1$, $b_2$, $b_3$ are three base waveform data shown in Figure 3.4, the waveform data of three classes $w_1$, $w_2$, $w_3$ are generated according to the equations below:

$$w_1(x) = \alpha b_1(x) + (1-\alpha)b_2(x) + \epsilon(x);$$

$$w_2(x) = \alpha b_1(x) + (1-\alpha)b_3(x) + \epsilon(x);$$

$$w_3(x) = \alpha b_2(x) + (1-\alpha)b_3(x) + \epsilon(x);$$

where $x = 1, \ldots, 40$, $\alpha$ is a uniform random number between 0 and 1, $\epsilon(x)$ are normally distributed with mean 0 and variance 1.

Figure 3.5 shows the first example of waveform data in each class. In our simulation study, we considered data with different noise levels, $\sigma \in \{1, 1.5\}$, generated by the same seed 12345 using the C code by David Aha in 1988.

This waveform dataset has 5000 observations. By using different proportions of data for training, we also evaluated the performance of stacking with respect to different size of training data.

**Base classifiers**

Now we will briefly describe how to generate the base classifiers. Since there are a total of three waveform classes, in the first step classification, we will group the first and second class together, and separate them from the third class. Then in the second step classification, we classify the first class from the second class. Each base classifier is based on this grouping strategy.

By using compound estimation, we are able to obtain smoothed curves for mean response and first derivative of both base waveforms and different types of waveform data. In compound estimation, there are a total of 10 centering points ($2 + 4 * n, n \in$

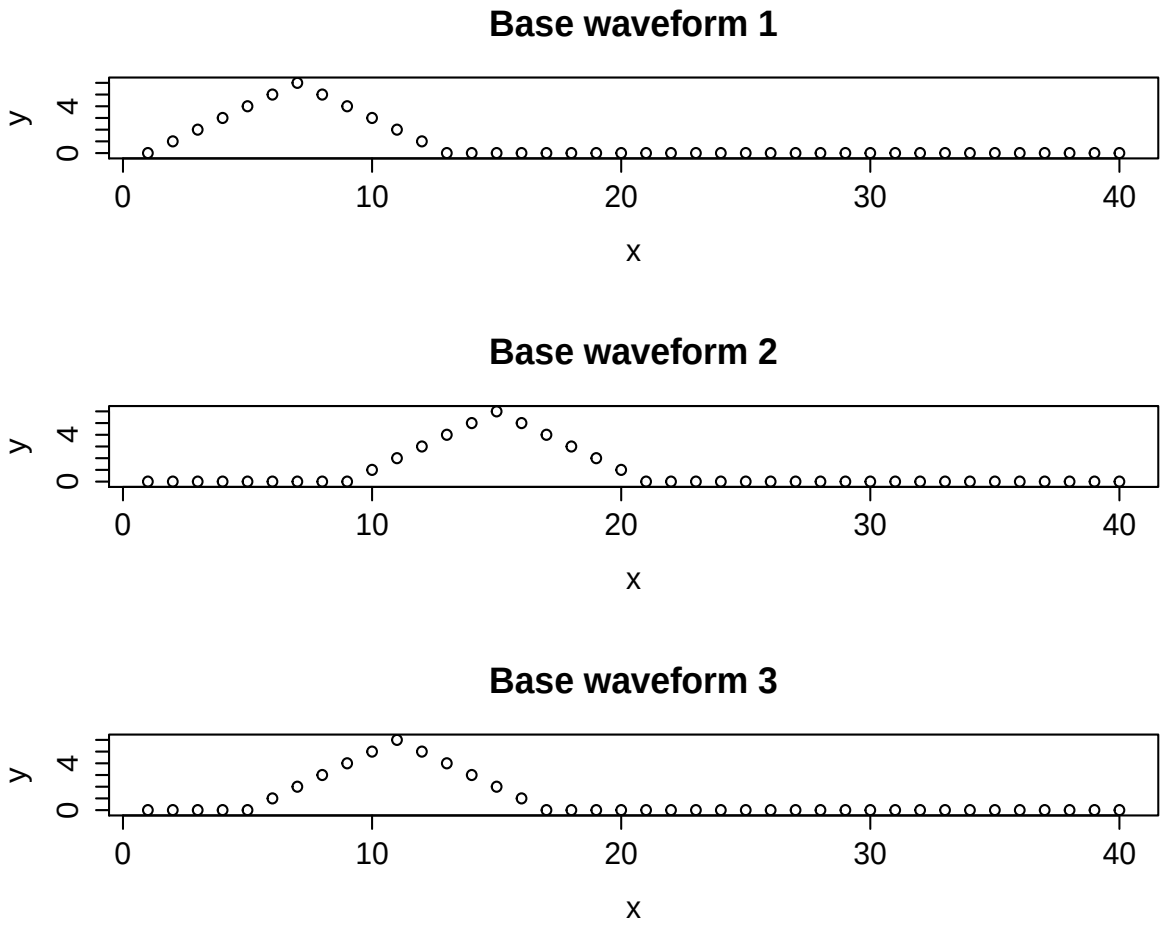


Figure 3.4: Base waveforms

$\{0, \ldots, 9\}$), in which the local polynomials of degree three were estimated. 0.6 and 0.18 were selected as convolution parameter and nearest neighbor fraction, respectively. Then we use least square estimation with nonnegative constraints to estimate convex combination coefficients of a generic waveform in terms of the types in Figure 3.5 based on raw data, compound estimation for mean response and its derivative of base waveform data. Finally, logistic regression, classification tree, and support vector machine were implemented for level 0 classification based on coefficients from previous step.

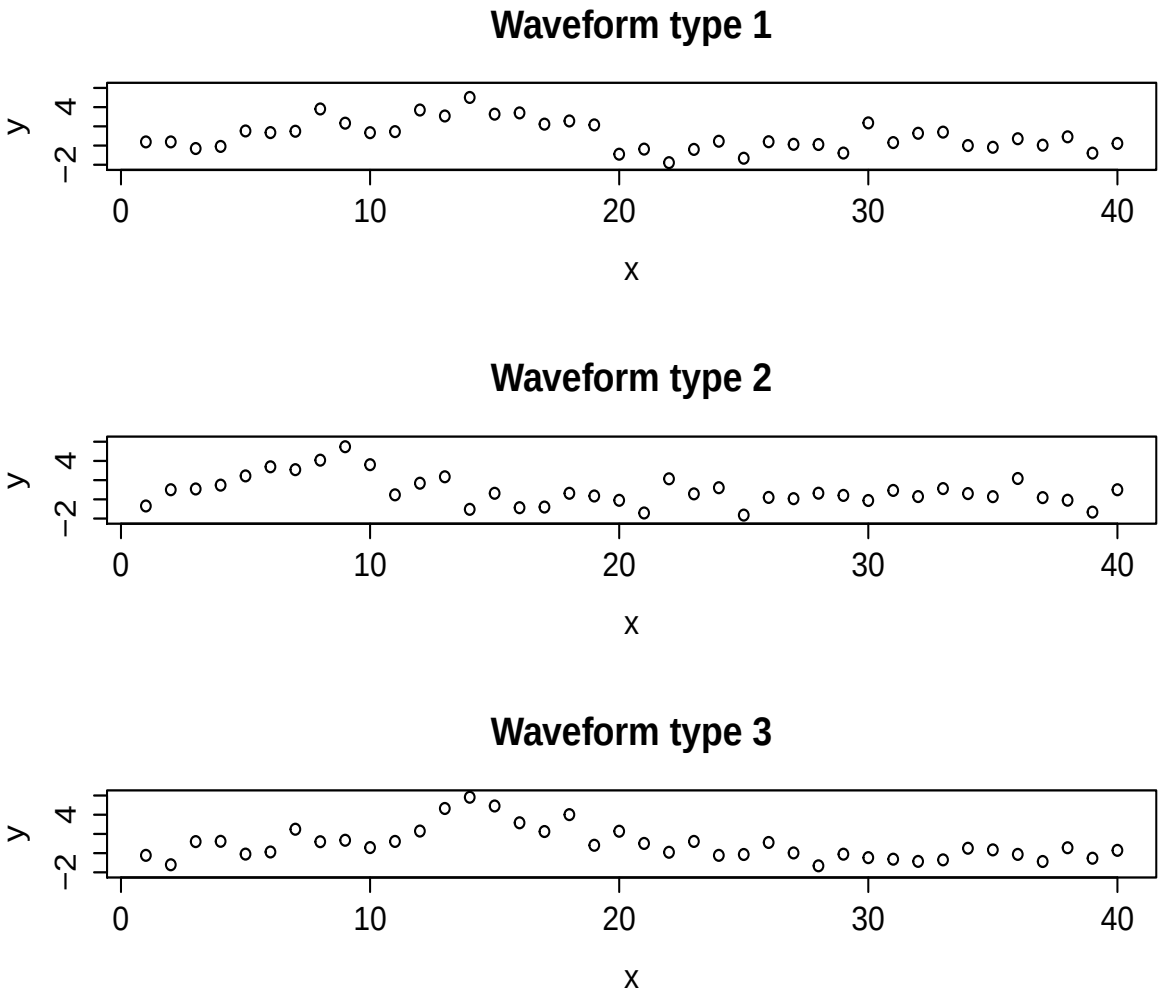


Figure 3.5: Waveform data in three classes

**Stacking results**

Two step sequential classification described in Section 3.3 was used in our simulation study. The results are shown in Table 3.2 and Table 3.3 under the noise levels $\sigma = 1$ and $\sigma = 1.5$. The first step separates the first and second types of waveform from the third type of waveform, while the second step separates the first from the second type of waveform data.

The data were randomly divided into three equal parts, and we used one third,

two thirds for training respectively in order to investigate the impact of different proportions of training data on the results. The base classification testing errors were used to define $\hat{\epsilon}(h)$ quantities for stacking. As expected, with larger training data, the testing error decreases for most level 0 classifiers. It is also shown that stacking performs better than or at least the same as the best base classifiers.

Table 3.2: Testing errors in waveform data simulation study ($\sigma = 1$)

| Different Methods | | First step | | Second step | |
|---|---|---|---|---|---|
| | | Training(1/3) | Training (2/3) | Training(1/3) | Training (2/3) |
| Logistic regression | Raw data | 0.128 | 0.125 | 0.093 | 0.095 |
| | $\mu_0$ | 0.128 | 0.127 | 0.100 | 0.096 |
| | $\mu_1$ | 0.182 | 0.175 | 0.121 | 0.119 |
| Classification tree | Raw data | 0.104 | 0.093 | 0.096 | 0.096 |
| | $\mu_0$ | 0.101 | 0.100 | 0.094 | 0.094 |
| | $\mu_1$ | 0.142 | 0.125 | 0.115 | 0.118 |
| SVM | Raw data | 0.097 | 0.086 | 0.091 | 0.088 |
| | $\mu_0$ | 0.097 | 0.088 | 0.091 | 0.088 |
| | $\mu_1$ | 0.128 | 0.112 | 0.101 | 0.106 |
| Stacking | | 0.095 | 0.085 | 0.087 | 0.086 |

Table 3.3: Testing errors in waveform data simulation study ($\sigma = 1.5$)

| Different Methods | | First step | | Second step | |
|---|---|---|---|---|---|
| | | Training(1/3) | Training (2/3) | Training(1/3) | Training (2/3) |
| Logistic regression | Raw data | 0.163 | 0.163 | 0.149 | 0.136 |
| | $\mu_0$ | 0.164 | 0.167 | 0.150 | 0.152 |
| | $\mu_1$ | 0.234 | 0.219 | 0.202 | 0.180 |
| Classification tree | Raw data | 0.143 | 0.144 | 0.148 | 0.133 |
| | $\mu_0$ | 0.143 | 0.136 | 0.143 | 0.137 |
| | $\mu_1$ | 0.178 | 0.167 | 0.161 | 0.168 |
| SVM | Raw data | 0.144 | 0.137 | 0.141 | 0.143 |
| | $\mu_0$ | 0.138 | 0.132 | 0.144 | 0.133 |
| | $\mu_1$ | 0.181 | 0.160 | 0.156 | 0.155 |
| Stacking | | 0.137 | 0.130 | 0.141 | 0.127 |

Compared with Table 3.2, Table 3.3 demonstrates that the testing errors of base classifiers and stacking increase with greater noise level $\sigma = 1.5$. Classification based

on compound estimation for derivatives does not perform as well as raw data and compound estimation for mean response if using the same classification technique. Also, among the base classification methods, Support Vector Machine (SVM) generally produces less testing errors than logistic regression and classification tree method, particularly when using compound estimation for mean response.

## 3.5 Discussion

In this simulation study, the base waveform data are employed for classification in all base classifiers, which is one reason that these base classifiers perform better than in the previous paper [85]. More specifically, our base classifiers perform better than the best base classifier Naive Bayesian (NB) in the previous paper [85]. Derivative can amplify the fine structure of the data over a short range of values. As such, derivative estimation might be helpful if there is high frequency information to be captured in the functional data for classification, such as in Raman spectroscopy data. However, in our simulation study, there is no high frequency information in the waveform data and classifiers based on derivative will only amplify the noise. In addition, waveforms have points of non-differentiability. Therefore the classifiers based on derivative have greater error rate than based on mean response estimation and raw data. Support Vector Machine (SVM) constructs a hyperplane that maximizes the margin (i.e., distance between the hyperplane and data) while allowing for misclassified training data and can perform well in testing data. By employing the weighting function on the testing errors of base classifiers, stacking performs at least the same as the best base classifier in this simulation study, therefore it has great potential to be applied in real data sets, such as Raman spectroscopy data.

# Chapter 4 Boosting for Nonparametric Regression

## 4.1 Background

Boosting is an advanced machine learning approach based on the idea of updating weak learner to a single strong learner through differential sampling [70]. Specifically, suppose we have a base weak and simple learner, which is slightly better than random guessing on the training data. In the training step, through many rounds of different sampling, many new classifiers will be generated based on the weak learner. At the same time, much more attention will be focused on those "hard examples", which are difficult to correctly classify. Then these generated hypotheses are combined so as to achieve higher prediction accuracy. Many research papers on boosting have focused on theoretical studies as well as its applications such as optical character recognition (OCR) [72][21].

The AdaBoost algorithm, proposed by Freund and Schapire in 1995 [26], has nice theoretically justified properties including "driving the generalization errors down close to 0" and "resistance to overfitting"[71], and it has been widely used in many pattern recognition fields such as face detection [90][69]. The AdaBoost algorithm for a binary outcome $\in \{-1, 1\}$ can be briefly described as follows:

i. Initialize equal weight on each example in the training dataset $D_1(i) = 1/m$ for $i \in \{1, \ldots, m\}$.

ii. Do a loop: for each round $t = 1, \ldots, T$,

a. Select the hypothesis $h_t$ minimizing the weighted error: $\epsilon_t = Pr_{D_t}[h_t(x_i) \neq y_i]$;

b. Choose weight for hypothesis $h_t$: $\alpha_t = 1/2(\ln(\frac{1-\epsilon_t}{\epsilon_t}))$;

c. Update the weight for each example: $D_{t+1}(i) = \frac{D_t(i)exp(-\alpha_t y_i h_t(x))}{Z_t}$, where $Z_t$ is chosen so that $D_{t+1}$ will be a distribution.

iii. Combine these hypotheses: $f(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$.

One caveat about AdaBoost algorithm is its poor performance on the noisy data, which leads to development of other boosting algorithm as "BrownBoost" algorithm, having greater tolerance to noise. There are also some other variants of AdaBoost algorithm including "LogitBoost", "GentleBoost" and "Boosting for multiclass outcomes" [71].

There are two general requirements for a learning algorithm: i. it fits the data well; ii. it is simple[71]. The simplicity can be measured by Vapnik-Chervonenkis (VC)-dimension[89]. When VC dimension is finite, the probability of difference between generalization error and training error (the latter regarded as random) approaches to zero will be high. However, on the other hand, when VC-dimension is infinite, the bound for difference between generalization error and training error is of order $O(1)$.

The theoretical bound on the training error of Adaboost by Freund and Schapire is given as follows[26][71]:

"Theorem 1. Given the notations of AdaBoost algorithm above, let $\gamma_t := \frac{1}{2} - \epsilon_t$, and let $D_1$ be an arbitrary initial distribution over the training set. Then the weighted training error of the combined classifier $H$ with respect to $D_1$ is bounded as

$$\Pr_{i\sim D_1}[H(x_i) \neq y_i] \leq \prod_{t=1}^{T} \sqrt{1-4\gamma_t^2} \leq \exp(-2\sum_{t=1}^{T}\gamma_t^2)."$$

This theorem implies under the weak learner assumption (i.e. $\gamma_t \geq k > 0$), the training error drops exponentially fast as a function of the number of rounds $T$. $\gamma_t$ is the "edge" measuring how much better is the error rate of the $t$-th weak classifier $h_t$ than random guessing rate of 1/2. AdaBoost is called "adaptive boosting" in that it does not need a prior knowledge of $\gamma_t$, but rather adjusts to the error $\epsilon_t$ from each round as it becomes available.

Schapire et al. also provided the margin explanations for Boosting's effectiveness and bounded the generalization error by Theorem 2 for finite base classifier space and Theorem 3 for infinite base classifier space[72].

"Theorem 2. Let $D$ be a distribution over $X \times \{-1, 1\}$, and let $S$ be a sample of $m$

examples chosen independently at random according to $D$. Assume that the base-classifier space $H$ is finite, and let $\delta > 0$. Then with probability at least $1 - \delta$ over the random choice of the training set $S$, every weighted average function $f$ satisfies the following bound for all $\theta > 0$:

$$\Pr_D[yf(x) \leq 0] \leq \Pr_S[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}}\left(\frac{\log m \log |H|}{\theta^2} + \log(1/\delta)\right)^{1/2}\right).$$

"Theorem 3. Let $D$ be a distribution over $X \times \{-1, 1\}$, and let $S$ be a sample of $m$ examples chosen independently at random according to $D$. Suppose that the base-classifier space $H$ has VC-dimension $d$, and let $\delta > 0$. Assume that $m \geq d \geq 1$. Then with probability at least $1 - \delta$ over the random choice of the training set $S$, every weighted average function $f$ satisfies the following bound for all $\theta > 0$:

$$\Pr_D[yf(x) \leq 0] \leq \Pr_S[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}}\left(\frac{d \log^2(m/d)}{\theta^2} + \log(1/\delta)\right)^{1/2}\right).$$

Theorem 2 and Theorem 3 imply the generalization error bound depends on the entire distribution of margins of training examples (as measured by $\theta$), number of training examples ($m$), and "complexity" of weak classifiers ($d$). Previous research showed evidence that improving the margin distribution instead of the minimum margin could produce better ensembles [63] [76]. However, designing an algorithm with each observation's margin as large or larger than that produced by AdaBoost did not result in better performance than AdaBoost, suggesting the large margin distribution theory appears to be insufficient for explaining the performance of ensemble methods [49].

### 4.2 Modified AdaBoost

We have modified AdaBoost algorithm by adding a scaling factor $c$ in the weight function for each example in the training data: $D_{t+1}(i) = \frac{D_t(i)\exp(-c\alpha_t y_i h_t(x))}{Z_t}$, where

$0 \leq c \leq 1$, and $Z_t$ is chosen so that $D_{t+1}$ will be a distribution. The scaling factor $c$ can measure the degree of boosting; when $c = 1$, it turns out to be AdaBoost. In this section, we examine the condition under which $c < 1$ suffices to ensure that the training error tends to 0.

## Analysis of Modified AdaBoost Algorithm

Theorem 4.2.1 states the bounds for training error of the modified AdaBoost algorithm. We follow the basic approach of [26], let $D_1(i)$ $(1 \leq i \leq m)$ denote the initial weight for each example in the training data and $\epsilon_t$ be the weighted error for each round: $\epsilon_t = Pr_{D_t}[h_t(x_i) \neq y_i]$; also define the weight adjustment factor for hypothesis $h_t$: $\alpha_t = 1/2(\ln(\frac{1-\epsilon_t}{\epsilon_t}))$. The weight function for each example in the training data is updated in each round by: $D_{t+1}(i) = \frac{D_t(i)\exp(-c\alpha_t y_i h_t(x))}{Z_t}$, where $0 \leq c \leq 1$, and $Z_t$ is chosen so that $D_{t+1}$ will be a distribution.

**Theorem 4.2.1.** *Let $\gamma_t := \frac{1}{2} - \epsilon_t$, and let $D_1$ be an arbitrary initial distribution over the training set. Then the weighted training error of the combined classifier $H$ with respect to $D_1$ is bounded as*

$$\Pr_{i\sim D_1}[H(x_i) \neq y_i] \leq \prod_{t=1}^{T} 2^{1-c}(1-4\gamma_t^2)^{c/2} \leq 2^{(1-c)T}\exp(-2c\sum_{t=1}^{T}\gamma_t^2).$$

*Proof.* Let

$$F(x) := \sum_{t=1}^{T} \alpha_t h_t(x).$$

According to the definition of $D_{t+1}$ in terms of $D_t$ in the modified AdaBoost algorithm,

$$\begin{aligned} D_{T+1}(i) &= D_1(i) \times \frac{e^{-cy_i\alpha_1 h_1(x_i)}}{Z_1} \times \ldots \times \frac{e^{-cy_i\alpha_T h_T(x_i)}}{Z_T} \\ &= \frac{D_1(i)\exp(-cy_i\sum_{t=1}^{T}\alpha_t h_t(x_i))}{\prod_{t=1}^{T} Z_t} \\ &= \frac{D_1(i)\exp(-cy_i F(x_i))}{\prod_{t=1}^{T} Z_t}. \end{aligned} \tag{4.1}$$

$Z_t$ is the normalization factor, and can be calculated as

$$\begin{aligned}
Z_t &= \sum_{i=1}^{m} D_t(i) e^{-c\alpha_t y_i h_t(x_i)} \\
&= \sum_{i:y_i=h_t(x_i)} D_t(i) e^{-c\alpha_t} + \sum_{i:y_i \neq h_t(x_i)} D_t(i) e^{c\alpha_t} \\
&= e^{-c\alpha_t}(1-\epsilon_t) + e^{c\alpha_t}\epsilon_t \\
&= e^{-c\alpha_t}(\frac{1}{2}+\gamma_t) + e^{c\alpha_t}(\frac{1}{2}-\gamma_t) \qquad (4.2) \\
&= (\frac{1}{2}-\gamma_t)^{c/2}(\frac{1}{2}+\gamma_t)^{c/2}((\frac{1}{2}+\gamma_t)^{1-c} + (\frac{1}{2}-\gamma_t)^{1-c}) \qquad (4.3) \\
&\leq 2(\frac{1}{2}-\gamma_t)^{c/2}(\frac{1}{2}+\gamma_t)^{c/2} \\
&= 2^{1-c}(1-4\gamma_t^2)^{c/2}. \qquad (4.4)
\end{aligned}$$

Equation (4.2) uses $\gamma_t := \frac{1}{2} - \epsilon_t$, and equation (4.3) follows from our choice of $\alpha_t = 1/2(\ln(\frac{1-\epsilon_t}{\epsilon_t}))$. Then the training error is

$$\begin{aligned}
\Pr_{i\sim D_1}[H(x_i) \neq y_i] &= \sum_{i=1}^{m} D_1(i) 1\{H(x_i) \neq y_i\} \\
&\leq \sum_{i=1}^{m} D_1(i) \exp(-c y_i F(x_i)) \\
&= \sum_{i=1}^{m} D_{T+1}(i) \prod_{t=1}^{T} Z_t \qquad (4.5) \\
&= \prod_{t=1}^{T} Z_t \qquad (4.6) \\
&\leq \prod_{t=1}^{T} 2^{1-c}(1-4\gamma_t^2)^{c/2} \qquad (4.7) \\
&\leq 2^{(1-c)T} \exp(-2c\sum_{t=1}^{T}\gamma_t^2). \qquad (4.8)
\end{aligned}$$

Equation (4.5) uses equation (4.1). Substituting equation (4.4) into equation (4.6) gives the first bound of the theorem (4.7). For the second bound (4.8), we apply the approximation $1 + x \leq e^x$. □

**Corollary 4.2.1.** *Assume* $R := \lim_{T\to\infty} \frac{\sum \gamma_t^2}{T}$ *exists and is positive. Then* $c > \frac{\log 2}{\log 2 + 2R}$ *ensures* $\Pr_{i\sim D_1}[H(x_i) \neq y_i] \to 0$ *as* $T \to \infty$.

*Proof.*

$$\begin{aligned}\Pr_{i\sim D_1}[H(x_i)\neq y_i] &\leq 2^{(1-c)T}\exp(-2c\sum_{t=1}^{T}\gamma_t^2)\\ &= \exp((1-c)T\log 2-2c\sum_{t=1}^{T}\gamma_t^2).\end{aligned}$$

Given $R:=\lim_{T\to\infty}\frac{\sum\gamma_t^2}{T}$ and $c>\frac{\log 2}{\log 2+2R}$, $\lim_{T\to\infty}\exp((1-c)T\log 2-2cRT)=0$. □

### 4.3 Simulations on waveform data

We use the same waveform dataset (version 2) as in Chapter 3 from UCI machine learning repository for simulations. In this waveform dataset (version 2), there are three types of waveforms including a total of 5000 observations and 40 attributes for each observation as described in detail in Chapter 3. Since boosting can improve a weak classifier into a strong classifier, we investigate the improvement of minimum distance approach based on compound estimation by boosting under the noise level $\sigma = 1$. Compound estimation is used to obtain smoothed curves for mean response and first derivative of all base waveforms and different types of waveform data (see Chapter 3).

Figure 4.1 presents 10 of the three types of original waveform data, while Figure 4.2 shows the smoothed waveform data by compound estimation of mean response curves. Notably, the distortion caused by presence of spikes in the original data is largely reduced by compound estimation. For boosting, we will only focus on the minimum distance approach based on smoothed mean response curves, because the smoothed first derivatives and second derivatives can not distinguish the three types of waveforms (Figure 4.3 and Figure 4.4) due to the noise level ($\sigma = 1$) and non-differentiability of the original waveform data. Now we will describe our boosting approach and investigate the performance of boosting under different proportions of data randomly allocated to training (2/3 vs 1/3) as well as different scaling factors $c$

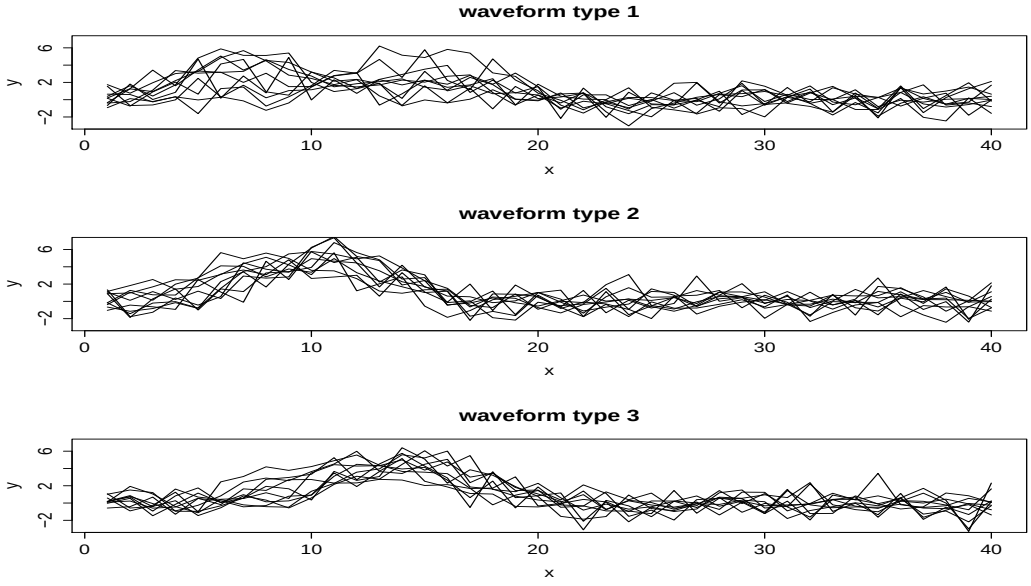


Figure 4.1: Three types of original waveform data

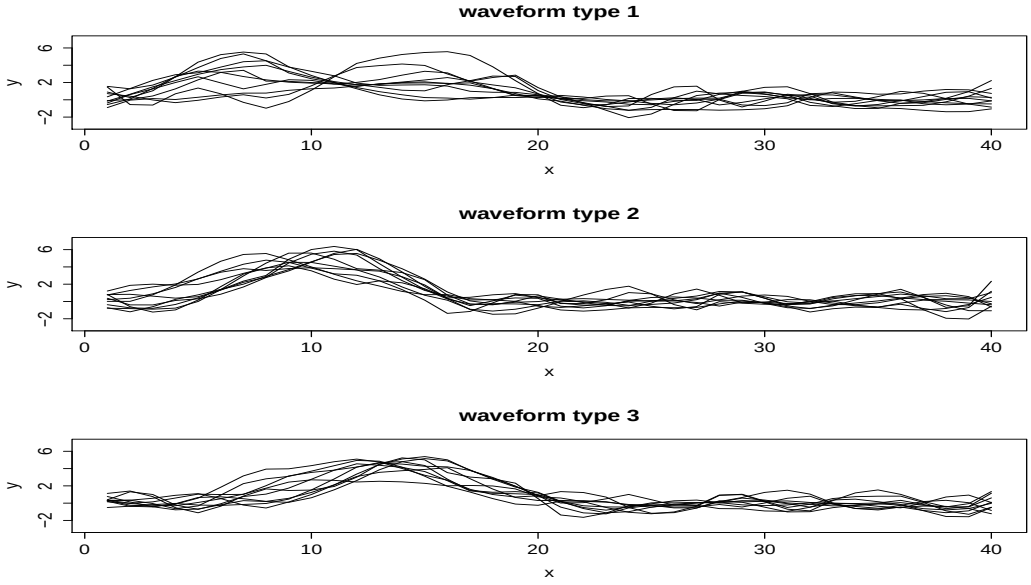


Figure 4.2: Three types of smoothed mean response waveform curves

discussed in Section 4.2.

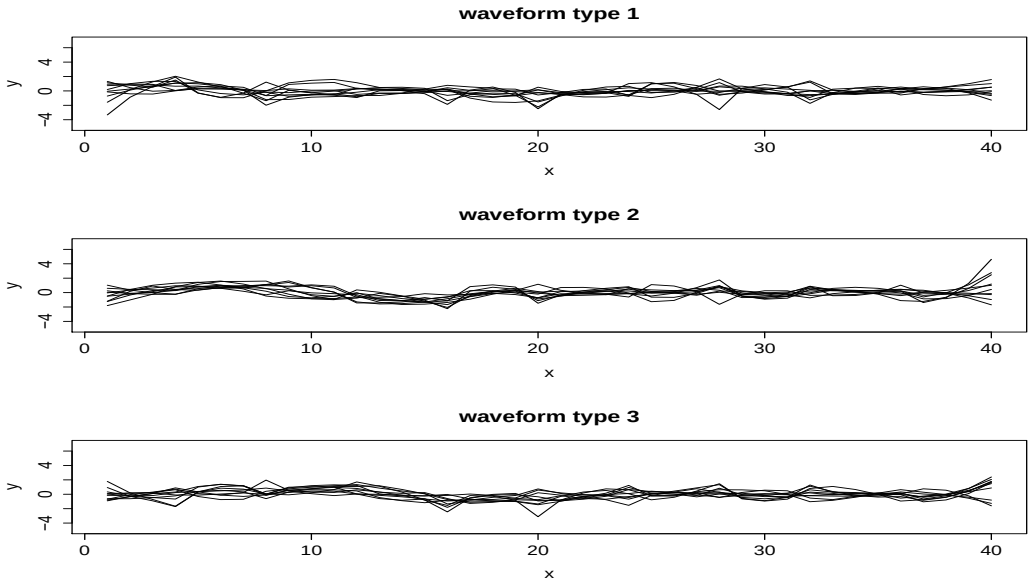


Figure 4.3: Three types of smoothed first derivatives of waveform data

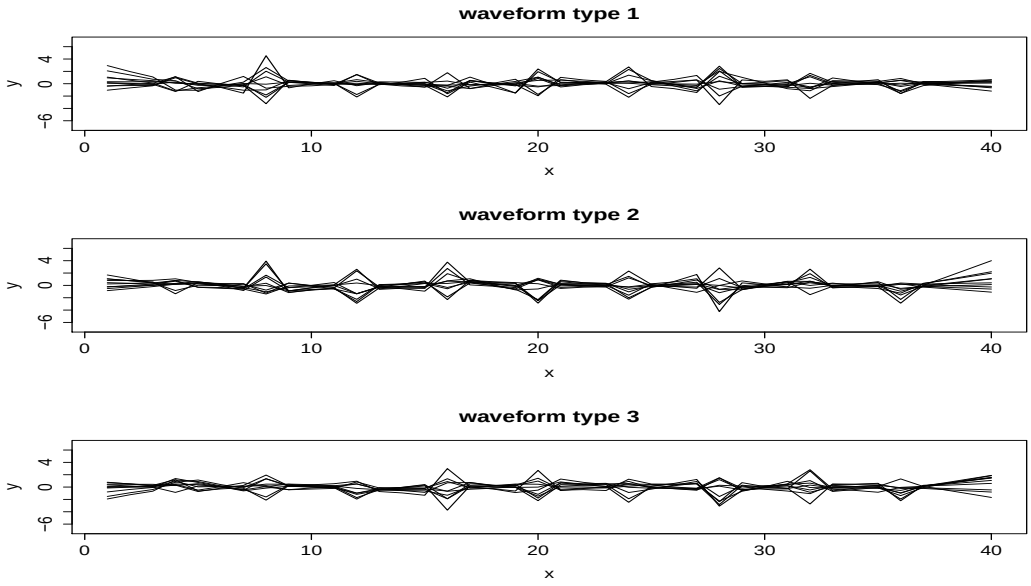


Figure 4.4: Three types of smoothed second derivatives of waveform data

### Boosting based on minimum distance approach

Two step sequential classification was used in our simulation study. The first step separates the first and the second type from the third type of waveform, while the

second step separates the first type from the second type of waveform data.

At the first step, initially in the first round of boosting ($t = 1$ means without boosting), a combined mean response reference curve for waveform type 1 and type 2 is defined as $\widehat{\mu_d}(x)$ over all smoothed waveform data in the training set known to be type 1 or 2, and similarly, a mean response reference curve for waveform type 3 is defined to be $\widehat{\mu_d}(x)$ over all smoothed waveform data in the training set known to be type 3 ($d = 1$ with type 1 or 2, and $d = -1$ with type 3). In the same way, two reference curves will be defined as $\widehat{\mu_d}(x)$ from the smoothed training waveform data separating waveform type 1 from type 2 at the second step ($d = 1$ with type 1, and $d = -1$ with type 2). Then the reference curves are updated by weighted average of $\widehat{\mu}(x)$ known to be type $d$ with the weight equal to the sampling weight in each round of boosting. Specifically, the weight on each observation was initialized to be equal, and then updated in each round according to $D_{t+1}(i) = \frac{D_t(i)exp(-c\alpha_t y_i h_t(x))}{Z_t}$, where $Z_t$ is chosen so that $D_{t+1}$ will be a distribution, $\alpha_t = 1/2(\ln(\frac{1-\epsilon_t}{\epsilon_t}))$, and $\epsilon_t$ is the weighted classification error of the training data defined as $\epsilon_t = Pr_{D_t}[h_t(x_i) \neq y_i]$. The reference curves updated from the training data during each round $t$ will then be used to define the classification algorithm $h_t(x)$ in order to classify the testing data. In addition, we investigate different scaling factors $c \in \{0.25, 0.5, 0.75, 0.85, 0.95, 1\}$. Number of rounds $T$ is chosen to be 15, thus a consistency pattern for the overall classification error is reached. Finally, the prediction for the testing data will be a weighted combination of $h_1(x), h_2(x), \ldots, h_T(x)$ defined as $sign(\sum_{t=1}^{T} \alpha_t h_t(x))$.

Table 4.1 and Table 4.2 present the misclassification error rates from the testing data for these two steps when training proportion is 2/3. Table 4.3 and Table 4.4 report the results when training proportion is 1/3. There are improvements for overall correct classification rates from boosting at both steps.

At the first step, without boosting ($T = 1$) the overall testing classification error is a little bit higher when using 1/3 training data than that using 2/3 training data

Table 4.1: Classification testing error in boosting for the first step with 2/3 training data

| Number of rounds | $T = 1$ | $T = 15$ | | | | | |
|---|---|---|---|---|---|---|---|
| Scaling factor | | c=0.25 | c=0.5 | c=0.75 | c=0.85 | c=0.95 | c=1 |
| Type 1 or Type 2 | 0.2263 | 0.1825 | 0.1825 | 0.1861 | 0.1797 | 0.1734 | 0.1615 |
| Type 3 | 0.0404 | 0.0281 | 0.0298 | 0.0263 | 0.0263 | 0.0316 | 0.0509 |
| Overall | 0.1627 | 0.1297 | 0.1303 | 0.1314 | 0.1273 | 0.1248 | 0.1236 |

Table 4.2: Classification testing error in boosting for the second step with 2/3 training data

| Scaling factor | c=0.25 | | c=0.5 | | c=0.75 | |
|---|---|---|---|---|---|---|
| Number of rounds | $T = 1$ | $T = 15$ | $T = 1$ | $T = 15$ | $T = 1$ | $T = 15$ |
| Type 1 | 0.3441 | 0.0990 | 0.3366 | 0.0702 | 0.3407 | 0.0931 |
| Type 2 | 0.0000 | 0.0976 | 0.0000 | 0.1263 | 0.0000 | 0.1116 |
| Overall | 0.1551 | 0.0982 | 0.1551 | 0.1004 | 0.1558 | 0.1031 |
| Scaling factor | c=0.85 | | c=0.95 | | c=1 | |
| Number of rounds | $T = 1$ | $T = 15$ | $T = 1$ | $T = 15$ | $T = 1$ | $T = 15$ |
| Type 1 | 0.3278 | 0.0637 | 0.3181 | 0.1167 | 0.3035 | 0.1397 |
| Type 2 | 0.0000 | 0.1326 | 0.0000 | 0.0810 | 0.0000 | 0.0499 |
| Overall | 0.1546 | 0.1001 | 0.1534 | 0.0982 | 0.1513 | 0.0947 |

Table 4.3: Classification testing error in boosting for the first step with 1/3 training data

| Number of rounds | $T = 1$ | $T = 15$ | | | | | |
|---|---|---|---|---|---|---|---|
| Scaling factor | | c=0.25 | c=0.5 | c=0.75 | c=0.85 | c=0.95 | c=1 |
| Type 1 or Type 2 | 0.2344 | 0.1920 | 0.1870 | 0.1956 | 0.1847 | 0.1810 | 0.1491 |
| Type 3 | 0.0263 | 0.0298 | 0.0289 | 0.0237 | 0.0219 | 0.0316 | 0.0491 |
| Overall | 0.1632 | 0.1365 | 0.1329 | 0.1368 | 0.1290 | 0.1299 | 0.1149 |

Table 4.4: Classification testing error in boosting for the second step with 1/3 training data

| Scaling factor | c=0.25 | | c=0.5 | | c=0.75 | |
|---|---|---|---|---|---|---|
| Number of rounds | $T = 1$ | $T = 15$ | $T = 1$ | $T = 15$ | $T = 1$ | $T = 15$ |
| Type 1 | 0.3855 | 0.1139 | 0.3742 | 0.0863 | 0.3798 | 0.0912 |
| Type 2 | 0.0031 | 0.0832 | 0.0031 | 0.1104 | 0.0031 | 0.1133 |
| Overall | 0.1755 | 0.0971 | 0.1744 | 0.0993 | 0.1763 | 0.1032 |
| Scaling factor | c=0.85 | | c=0.95 | | c=1 | |
| Number of rounds | $T = 1$ | $T = 15$ | $T = 1$ | $T = 15$ | $T = 1$ | $T = 15$ |
| Type 1 | 0.3641 | 0.1052 | 0.3532 | 0.0906 | 0.3565 | 0.1007 |
| Type 2 | 0.0032 | 0.0945 | 0.0032 | 0.1255 | 0.0030 | 0.1108 |
| Overall | 0.1740 | 0.0996 | 0.1732 | 0.1086 | 0.1667 | 0.1061 |

when separating waveform Type 1 or Type 2 from Type 3. Boosting ($T = 15$) drives the overall testing errors down and improves classification accuracy. Generally, the improvement effect from "boosting" is more noticeable when training proportion is 2/3 than that using 1/3 training data except for the case "c=1". Moreover, the overall testing error is different when choosing different scaling factors "c". Specifically, the overall testing error is relatively larger when c = 0.75, and gets smaller as c = 1. The testing error is not an approximately linear nor a quadratic function of c, suggesting that c may play a complex role in the overall testing error from boosting.

At the second step, the testing errors without boosting are calculated from the testing data that are correctly classified from boosting in the first step corresponding to the same "c". Considering the results from step 1, the classification errors are different without boosting for the step 2. It can be seen that the classification error for type 1 is much greater than that for type 2. Boosting also gives better overall accuracy when separating type 1 from type 2. Again, the different overall classification errors regarding to various values of "c" can be observed at the second step. It is interesting that boosting improves the classification error most noticeably when c = 0.25: from 0.1551 to 0.0982 with 2/3 training data, and from 0.1755 to 0.0971 with 1/3 training data.

## 4.4 Discussion

In this Chapter, we have described a generalization of AdaBoost by introducing a scaling factor "c", so that AdaBoost is a special case of this modified AdaBoost algorithm (c = 1). In addition, we have explored this generalization of AdaBoost sequentially in our simulation study. The theoretical bound on the training error obtained by this modified AdaBoost algorithm shows that given c is large enough, the training error drops exponentially fast as a function of the number of rounds in boosting under the weak learning assumption. The simulation study on the waveform data gave an

example on how boosting would improve minimum distance approach by assigning different weights for the training data to obtain the "reference curves". In addition, from the simulations, we have also seen the effect of boosting varies with respect to different choices of scaling factor in our modified AdaBoost algorithm. Simulation experiments done by combining the second and third type of waveform data together in the first step were considerably not successful (results not shown here). The reason is obvious: by taking an average of the smoothed curves of all of the second and the third types of waveform data, the reference curve for these two types of waveform data would be similar to that for the first type, thus it is difficult to use minimum distance approach for classification of the first type from the second and third types of waveform data in the first step. On the other hand, this problem does not exist if combining the first type and second type of waveform together as one group in the first step as we presented above.

# Chapter 5 Dynamic Ensemble Integration for Nonparametric Regression

## 5.1 Background

Ensemble learning typically include three phases: classifier generation, selection, integration. The goal in the generation phase is to obtain a set of classifiers. These classifiers can be generated by different algorithms (e.g. classification tree[9] and support vector machine[20]) or based on differential sampling of the objects (e.g. bagging[7] and boosting[70]). The purpose in the selection phase is to prune the classifiers in the ensemble to increase diversity among the classifiers and classification accuracy, as well as reduce computational complexity. In the integration phase, the selected classifiers will be combined together. There are many strategies for combination, such as majority voting and bagging. In Chapter 3, we have discussed stacking: combining the classifiers by assigning the weights to the predicted classifiers based on their performance.

In Chapter 5, we will have a brief review on ensemble selection, and call for a dynamic ensemble integration scheme to consist of stacking and boosting together with the aim to further improve the accuracy of prediction.

## 5.2 Ensemble selection

In this section, we will review general ensemble selection framework and then discuss recent advances in dynamic ensemble selection.

Ensemble selection, also called ensemble pruning, is a process to select a subset of the classifiers from the pool of classifiers obtained in the generation phase. The various ensemble selection methods can be classified into the following categories: Search-based, Clustering-based, Ranking-based and other [87]. Drawing some

inspiration from feature selection, the Search-based approaches search for a subset of classifiers by adding or removing classifiers from the candidate subset including forward subset selection, backward subset selection or a combination of both. The clustering-based methods consist of a partitioning step prior to selection step. In the partitioning step, a cluster algorithm is used to discover groups of classifiers that make similar predictions. For example, Giacinto et al. defined a distance metric between two classifiers as the probability that the classifiers don't make coincident errors, and assign classifiers that make few coincident errors to different clusters [30]. Lazarevic and Obradovic used a k-means clustering algorithm to guide the clustering process [45]. By selecting the subset of classifiers from each cluster, we could increase the diversity among the classifiers and thus improve classification accuracy. The Ranking-based selection orders the classifiers based on an evaluation measure such as accuracy given in Partridge and Yates [58].

Regarding to the evaluation measures, it is important to consider both of those that are based on performance and those on diversity to guide the search process and/or to establish the stopping criterion. Because different metrics are appropriate in different learning settings, Caruana et al. experimented with several performance metrics, including accuracy, root-mean-squared-error, mean cross-entropy, lift, precision/recall break-even point, precision/recall F-score, average precision and ROC area [13]. Traditional diversity measures including diversity measures disagreement, double fault, Kohavi-Wolpert variance, inter-rater agreement, generalized diversity and difficulty were used for greedy ensemble selection in [82]. Concurrency, margin distance minimization, Complementariness and Focused Selection Diversity are four diversity measures designed specifically for greedy ensemble selection[57].

Most of previous researches have focused on static ensemble selection: selecting a subset of base classifiers and combining them together for all test samples. However, for different test samples the subsets of different base classifiers may have different

performance, thus a dynamic ensemble selection scheme may be appealing. Ko et al. proposed four dynamic ensemble selection schemes based on K-nearest-oracles (KNORA) to select different subsets of base classifiers for different test samples [43]. The idea is that, for any test data point, it first selects nearest K neighbors in the validation set, figures out which classifiers correctly classify those neighbors in the validation set and uses them as the ensemble for classifying that test sample. Their results suggest that the proposed schemes perform better than the static selection method when using the majority voting rule for combining classifiers.

## 5.3 Static ensemble integration scheme

Before proposing the dynamic ensemble integration scheme (DEIS), we will first introduce the static ensemble integration scheme for three reasons. First, our dynamic ensemble integration scheme is built upon the static ensemble integration scheme. Second, although the idea of dynamic ensemble integration is appealing, static ensemble integration may be less computationally intensive and time consuming than dynamic ensemble integration. Thirdly, DEIS may not always have better prediction accuracy than the static ensemble integration.

Our static ensemble integration calls for a combination of boosting, ensemble selection, and stacking. Specifically, it includes the following steps. The first four steps entail training data to develop the final prediction algorithm, while the last step uses the final algorithm to make predictions on the test data set.

*Step 1: Generation of base classifiers*

In nonparametric regression settings, the base classifiers could be based on the smoothed data or smoothed derivatives from nonparametric regression methods in addition to raw data which may contain noise and spikes. Further, given various dimension reduction techniques such as using linear combinations of basis functions, principal component analysis, the important features or patterns from the raw data

or smoothed data can be captured. Finally, based on the outputs from different dimension reduction techniques, different classification algorithms including classification tree, logistic regression, neural network and support vector machines can be applied. In addition to dimension reduction, we have also proposed two approaches called "minimum distance approach" and "confidence bands approach" in Chapter 2 to be used for pattern recognition problems in non-parametric regression settings. Hence, a lot of base classifiers will be generated with different training errors.

*Step 2: Boosting to improve/upgrade weak classifiers*

For those "weak" classifiers, we could upgrade them using boosting through differential sampling as discussed in Chapter 4. For instance, minimum distance approach could be improved by boosting with classification prediction accuracy increasing from 0.163 to 0.124 from the simulation study on the waveform data. However, this step is not limited to those weak classifiers. Some good base classifiers can also be improved by boosting.

*Step 3: Ensemble selection*

Figure 5.1 gives the static ensemble selection and integration scheme. In this static ensemble selection, only one set of base classifiers is selected for all testing data. There are many ensemble selection procedures as discussed in section 5.2. Here, we will describe two common approaches: search-based compound ensemble selection and clustering based ensemble selection [30].

The search-based compound ensemble selection is essentially a mixture of forward step and backward step selection. It starts by randomly selecting a predefined number of $K$ classifiers. At each iteration, one forward step and one backward step are applied. The forward step selects one classifier from the pool of base classifiers which improves the accuracy of the ensemble the most and results in $K+1$ models in the ensemble. Then the backward step selects the $K$ classifiers among the $K+1$ models with highest accuracy. This process stops when the same set of ensemble is selected

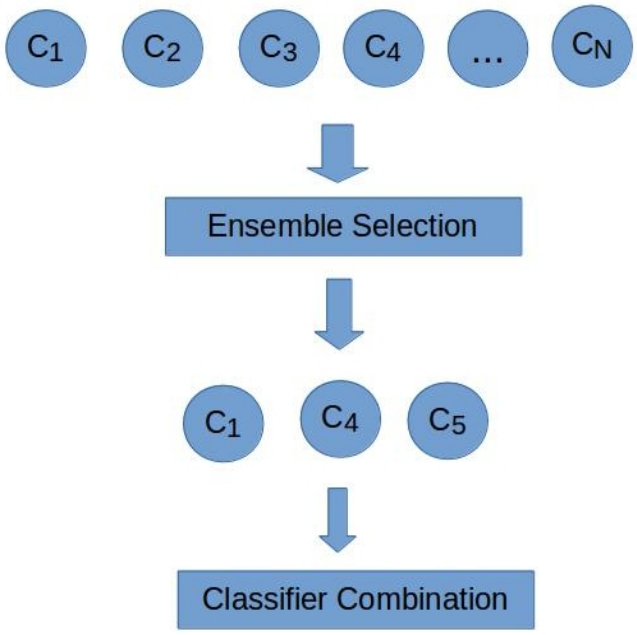


Figure 5.1: Static ensemble selection scheme

again after one iteration.

The clustering based ensemble selection consists of two phases: partitioning and selection. It first partitions the base classifiers according to the probability of making coincidence errors among two classifiers in order to choose diverse classifiers for the pruned ensemble [30]. Let $C$ be the collection of N base classifiers: $C = \{c_1, c_2, c_3, \ldots, c_N\}$, and let $C_i$ be a subset or cluster, so that $C$ is made up of the union of $M$ subsets: $C = \bigcup_{i=1}^{i=M} C_i$ for $1 \leq M \leq N$, where $C_i$ and $C_j$ are mutually exclusive if $i \neq j$,. The probability of making coincident errors for two classifiers is defined by a compound error probability: $Pr(c_i$ fails, $c_j$ fails). The compound error probabilities between any two classifiers within the same cluster should be higher than that of two classifiers belonging to different clusters. Hierarchical agglomerative clustering (HAC) algorithm is implemented to identify the subsets [30]. In addition, the distance between two classifiers is defined as 1 -$Pr(c_i$ fails, $c_j$ fails). In the selection phase, one representative classifier exhibiting the maximum average distance

from all other clusters is chosen from each cluster to create the pruned ensemble. For each clustering result, the performance of the pruned ensemble on a validation set will be evaluated using stacking as the combination method. The final pruned ensemble is the one that achieves the highest classification accuracy.

*Step 4: Stacking for combination of the selected ensemble*

As discussed in Chapter 3, stacking can be used to combine a group of classifiers. In addition to classifier combination, stacking is used to guide ensemble selection in step 3. The selected ensemble is one that achieves the highest classification accuracy through validation set from stacking.

*Step 5: Prediction and classification*

In static ensemble integration, the final combined algorithm for the selected ensemble will be used to classify the testing data or predict a new instance.

## 5.4 Dynamic ensemble integration scheme

Different from static ensemble integration, dynamic ensemble integration will not always select the same ensemble for different testing data. Given a new instance, it chooses the predictors that are expected to make the best combined prediction. Figure 5.2 shows this dynamic ensemble selection scheme. Now we will describe the six steps in the dynamic ensemble integration scheme.

*Step 1: Find similar data for testing data*

Dynamic ensemble integration starts with the testing data to find similar data from the training set. Figure 5.3 illustrates this step: different training data might be selected for different test data. The test data could be either one observation (such as one patient's data) or a group of samples (such as patients' registry data in one place). The standard method for obtaining similar data is the well-known k-nearest neighbors with the Euclidean distance [96]. This method weighs equally all the input variables. Some authors used attribute weighted metrics to find similar data [65] [88].

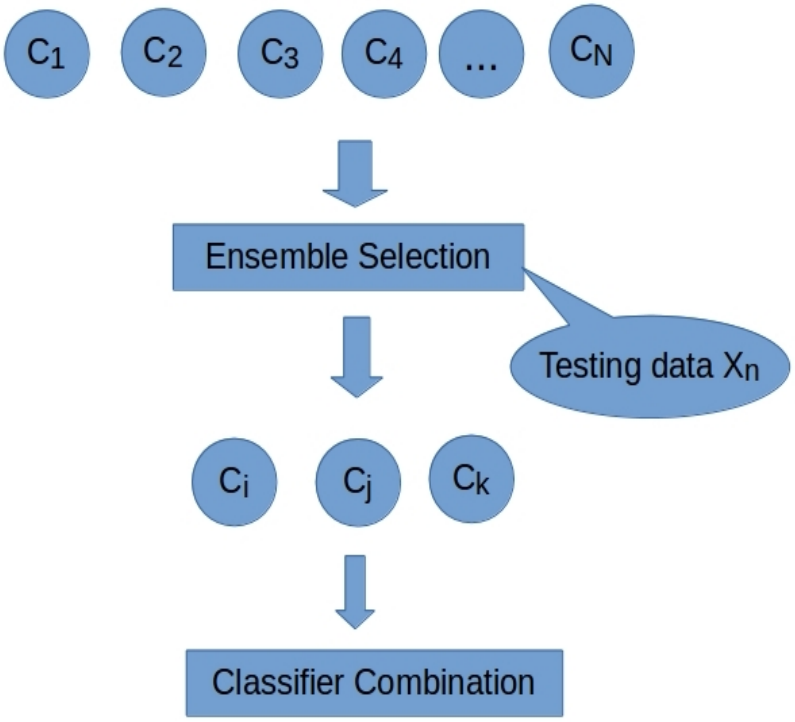


Figure 5.2: Dynamic ensemble selection scheme

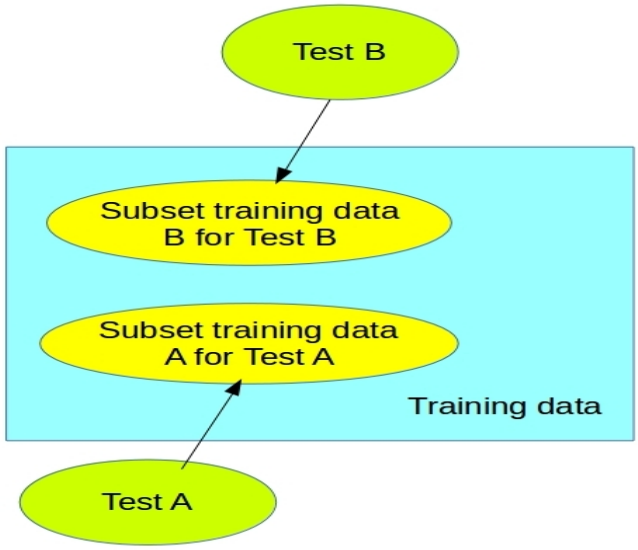


Figure 5.3: Selection of similar data from training set for testing data

It would be compelling to define the similarity measures in nonparametric regression settings.

*Step 2: Generation of base classifiers*

*Step 3: Boosting to improve/upgrade weak classifiers*

*Step 4: Ensemble selection*

*Step 5: Stacking for combination of the selected ensemble*

The above Step 2 to Step 5 are similar to that in static ensemble integration except they are based on the selected data from step 1 rather than all the training data. In this dynamic ensemble integration scheme, the step 4 and 5 are also dynamic in that different testing data may have different subsets of the ensemble used for prediction, and the weights assigned to the classifiers in the ensemble when using stacking for classifier combination may be different as well.

*Step 6: Prediction and classification*

This prediction step is straightforward as in static ensemble integration. For a given input value, we will obtain the prediction for each classifier in the selected ensemble, and then combine the results using stacking to obtain the ensemble prediction.

Our proposed ensemble scheme can be implemented in sequential classifications when there are more than two classes in the output. The idea is similar to that has been investigated in Chapter 3 and Chapter 4. As for future work, evaluation on the performance of dynamic ensemble selection versus static ensemble selection in nonparametric settings could be explored.

## 5.5 Discussion

In this Chapter, we propose development of a novel framework that fuses the ensemble techniques of boosting, stacking, and dynamic integration for classification problems in nonparametric regression settings. The main advantage of ensemble

methods is their well known accuracy and robustness [52]. However, they typically require large training data [12]. Future work needs to be carried out regarding the training sample size as well as the amount of similar data needed to be selected in the dynamic ensemble integration. Also, there is lack of comparative studies on the performance of ensemble selection procedures such as search-based ensemble selection and cluster-based ensemble selection. It would be interesting to explore which procedure could provide better prediction accuracy in static and dynamic ensemble integration schemes.

For future application of ensemble integration, especially dynamic ensemble integration, this ensemble-learning framework could be of practical importance in medical applications in personalized medicine [54], personalized treatment optimization [4], and many cost-effective applications relevant to e-healthcare and web-enabled diagnostics. For example, Electroencephalography (EEG) data analysis can be used for detecting abnormalities in order to diagnose epilepsy [98] or sleep disorders [11]. Similar to Raman spectroscopy data, EEG data can also be analyzed in the same way using stacking, boosting and ensemble integration schemes for classification problem in nonparametric regression settings. If the dynamic ensemble integration methodology is proved to be successful in the application of EEG to detect abnormality, this can facilitate remote medical screening or diagnosis through inexpensive medical devices carried out by patients, and thus perhaps improve the health care of patients with chronic diseases such as epilepsy, and sleep disorders.

# Chapter 6 Nonparametric Regression in Raman Spectroscopy, Revisited

## 6.1 Background

Raman spectroscopy is a spectroscopy technique that can measure the vibrational modes of molecules[60]. It can be done in two ways, ex vivo or in vivo. For ex vivo, the sample tissue only requiring about 1 cubic millimeter sampled tissue volume can be obtained from needle biopsies and Raman spectra data are measured in laboratory setting. For in vivo, the development of deep sub-surface Raman techniques (sub spatially offset Raman spectroscopy (SORS)) provides new opportunities offering a promising way of non-invasive characterization of biological tissues[51]. In either way, Raman spectroscopy can be less invasive than surgical biopsy. Also, it can provide timely information on several different molecules to infer chemical or morphological composition of biological tissue. Thus, Raman spectroscopy has the potential to reduce repeated needle biopsies in clinical cancer diagnosis and a patient's anxiety.

The schematic of the clinical Raman system mainly includes five parts (Figure 6.1): light from an 830-nm diode laser, sample, Raman probe's excitation fiber, wavelength separation device, and detector electronics such as CCD detector. In one previous study in 2005, Raman spectroscopy has been successfully employed in the classification of breast pathologies involving basis spectra for chemical constituents of breast tissue and resulted in high sensitivity (94%) and specificity (96%)[32]. And when the same classification algorithm developed in 2005 was applied in a prospective study in 2009, they obtained sensitivity of 83% and specificity of 93%[33].

In this Chapter, we will revisit the Raman spectroscopy data from Chapter 2, and in sections 6.2 and 6.3 make improvements based on the developments of the methods from Chapter 3 to Chapter 4. We will evaluate the performances of these developments incorporating different ensemble learning methods in a nonparametric

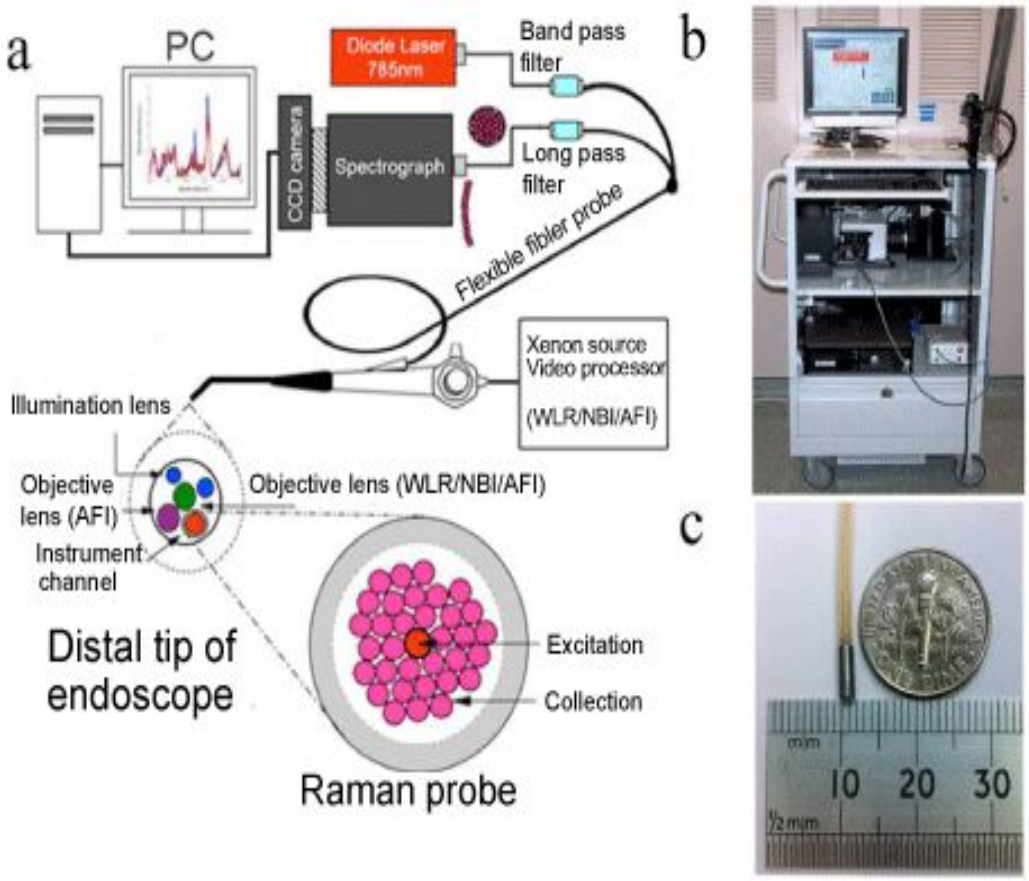


Figure 6.1: (a) Schematic diagram of the clinical Raman spectroscopy system; (b) Photo of the Raman endoscopic system in clinic; (c) Photo of the fiber-optic Raman endoscopic probe. Adapted from "Raman Endoscopy for Objective Diagnosis of Early Cancer in the Gastrointestinal System," by Bergholt MS, et al, 2013, J Gastroint Dig Syst, S1(008)[5].

regression setting. Finally in section 6.4 we will summarize the major findings and contributions of this work as well as identify opportunities for future research and their public health implications.

## 6.2 Combination of Nonparametric Regression Based Classifiers for Breast Tissue Diagnosis from Raman Spectra

In Chapter 2, we have generated different base classifiers based on compound estimation of a mean response function and its derivative, and the simultaneous confidence bands method in conjunction with compound estimation. In this section, we apply

an innovative stacking-type method proposed in Chapter 3 to combine different nonparametric regression based classifiers that rely on basis spectra and/or derivatives of basis spectra with the aim to improve the sensitivity and specificity for breast tissue diagnosis.

The Raman spectra data for the present study are essentially the same as in the previous study in 2005. There are four pathological types including 31 spectra for infiltrating carcinoma, 31 spectra for fibrocystic change, 15 spectra for fibroadenoma, and 47 spectra for normal. Also, a total of 9 basis morphological Raman spectra were used to establish the Raman spectroscopy model including calcium oxalate, calcium hydroxyapatite, cholesterol-like, water, beta-carotene, fat, collagen, cell nucleus, and cell cytoplasm.

## Generating base classifiers

Now we introduce the steps for generating base classifiers.

**Data processing** As we did in Chapter 2, we first performed normalization on the raw spectra data so that each spectrum has minimum value of 0 and maximum value of 1 applying the formula

$$y^*(x) := \frac{y(x) - \min_{1\le i\le n} y(x_i)}{\max_{1\le i\le n} y(x_i) - \min_{1\le i\le n} y(x_i)}.$$

where $y^*(x)$ denotes the normalized Raman spectrum, $x_i$ denotes a value of the Raman shift and $y(x_i)$ is the corresponding value of the Raman spectrum for that observation.

**Smoothing data** Then the compound estimation approach developed by Charnigo and Srinivasan [18] was used to smooth the Raman spectra curves and obtain estimation for mean response and their derivatives. First of all, like most nonparametric regression methods, it can be used to smooth the raw data which contain noise and oscillatory parts. Secondly, this approach allows us not only to use the information

of spectra but also their derivatives which may contain high frequency information while enjoying self-consistency. Thirdly, it can enable us to generate a couple of base classifiers for stacking. In this study, we used generalized Cp criterion proposed by Charnigo et al to select tuning parameters[16], so that the derivatives will be well estimated.

**Raman spectroscopy model** In order to establish the Raman spectroscopy model, first of all, for each spectrum to be classified, linear regression with nonnegative constraints was used to obtain fitted coefficients using 9 basis spectra. Intuitively the fitted coefficients can be viewed as the proportion of contribution of each basis spectrum. Secondly, the 8 fitted coefficients excluding water were normalized so that they sum to 1. Finally we performed two step sequential classifications using binary logistic regression or classification tree based on the fitted coefficients. Specifically, N and FC were initially separated from FA and C based on fitted coefficients for collagen and fat. Further, fibrocystic change and normal tissue were separated using fitted coefficients for collagen and fat again, while cancer and fibroadenoma were separated based only on fat using logistic regression. The decision threshold was chosen to maximize the correct classification rate on the training data in each logistic regression. Although Haka et al(2005) did not do so, we have replaced raw data by smoothed data from compound estimation (including estimated derivatives). Table 6.1 is a summary of 6 base classifiers which will be combined later.

The first classifier is what Haka et al did in the previous paper in 2005. They only selected fitted coefficients for fat and collagen to enter into a logistic regression model. The second and third classifiers are based on compound estimation to smooth the mean response curves and first derivatives, respectively. Also we used backward elimination to select the model (i.e., choose basis spectrum coefficients) when applying logistic regression. The fourth and fifth classifiers are variants of the first and second classifiers respectively, except that we used classification trees to perform prediction

Table 6.1: Review of base classifiers

| Id | Base classifier | Detail |
|---|---|---|
| 1 | Haka's method | Fitted Coefficients(Fat and Collagen); logistic regression |
| 2 | Compound estimation for mean response | Other combination of fit coefficients in logistic regression(backward) |
| 3 | Compound estimation for first derivative | logistic regression(backward) |
| 4 | Variant of Haka's method | Fit coefficients decision tree |
| 5 | Variant of (2) | Fit coefficients decision tree |
| 6 | Compound estimation for first derivative | Fit coefficients decision tree |

All procedures use two step classification: first separate N/FC from C/FA, then separate N from FC, and C from FA, respectively.

instead of logistic regression. And the sixth classifier is based on compound estimation for first derivative using classification tree.

## Stacked generalization

We perform two sequential stacking procedures discussed in Chapter 3 to combine the base classifiers generated above. First of all, the input L dataset is randomly divided into J parts, level 0 is cross validated prediction using K different algorithms, in our case, we performed five fold cross validation using 6 base classifiers so that J=5 and K = 6. In level 1, we combine the K base classifiers based on the output from level 0 to achieve better prediction. We have used exponentially weighted average vote proposed by Freund et al in 2004. The idea is similar to majority vote, except that each base classifier is weighted exponentially with respect to its training error. The weight is defined as an exponential function of $\eta$ and the training error $\hat{\epsilon}_{h_i}$ as follows[25]:

$$w_{h_i} := \exp(-\eta\hat{\epsilon}_{h_i}),$$

where $\hat{\epsilon}_{h_i}$ is the training error of base classifier $h_i$; $\eta = \ln(8|H|)m^{1/2-\theta}$, in which $|H|$ is the number of base classifiers, $m$ is the size of training data, and $0 < \theta < \frac{1}{2}$. In our study, $\theta$ was chosen to be 0.1 so as to obtain the least classification error. Suppose we have binary outcome, the final weighted average prediction is defined to be the outcome maximizing summation of weights of the base classifiers. Symbolically, let

$$\hat{l}_\eta(Y) = \frac{1}{\eta} \ln \left( \frac{\sum_{h,h(Y)=+1} w(h)}{\sum_{h,h(Y)=-1} w(h)} \right),$$

the weighted average prediction is defined to be $\text{sign}(\hat{l}_\eta(Y))$. Although not present in this section, an abstention feature is available as we have discussed in Chapter 3. The abstention feature can be interpreted as "no prediction", and it helps to identify the locations of potential overfitting and allow special actions on these cases. Sometimes sequential classification can perform better than direct classification. This is especially the case when the first step classification can yield better prediction, either because certain errors are forgiven at the first step (e.g. Normal called FC, or vice versa) or because errors incurred at the second step are less serious if the first step proceeded successfully (e.g. FA called cancer, or vice versa). Both of these considerations reflect that some categories may be inherently closer than others. Another advantage of sequential classification is that different classifiers can be applied in sequential steps and the stacking itself can be changed (e.g. variation of tuning parameters). However, if the observation is misclassified in the first step, it will certainly not be correctly classified in the second step, so it will not be entered into the cross validation in the second step classification either. Rather, we will use all other correctly classified observations from the first step classification for training in the second step.

## Results and conclusions

Table 6.2 shows the performance of stacking compared with base classifiers.

Table 6.2: Correctly classification rate of base classifiers and stacking

| Classification | M1 | M2 | M3 | M4 | M5 | M6 | Stacking |
|---|---|---|---|---|---|---|---|
| Step 1 | 71 | 72 | 70 | 69 | 69 | 69 | 73/78 |
| (N FC vs C FA) | 45 | 45 | 45 | 43 | 42 | 44 | 45/46 |
| Step 2 | 45 | 45 | 46 | 42 | 46 | 45 | 46/47 |
| (N vs FC) | 22 | 23 | 22 | 24 | 23 | 23 | 23/26 |
| Step 2 | 11 | 11 | 9 | 12 | 14 | 8 | 11/15 |
| (C vs FA) | 28 | 26 | 26 | 25 | 24 | 24 | 28/30 |

$\theta = 0.1$

In step 1, M2 is the best base model among the base models, In step 2 when separating N from FC, classifier number 5 is the best base model, while when separating cancer from FA, the first classifier performs the best among all base classifiers. Comparing stacking with base classifiers, stacking performs either better than the best base model as shown in the first step or equal to the best model as shown in the second step.

In conclusion, stacking yields at least the same quality of results as the best base classifier in the two step sequential classifications, which is all the more impressive because the best base classifier varies from situation to situation. Thus, overall generalization error is smaller when using stacking. Also, in our Raman spectra classification, we use five fold cross validation which enables us to make better assessment of predictive ability from the training data.

## 6.3 Boosting of Nonparametric Regression Based Classifiers for Breast Tissue Diagnosis from Raman Spectra

In this section, we are going to apply boosting algorithm described in Chapter 4 to the Raman spectroscopy data in nonparametric regression settings. Different from stacking which combines a set of different classifiers into an ensemble classifier, in boosting, the performance of one classifier can be improved through weighting: examples/instances incorrectly predicted by previous classifiers in the series are weighted

more heavily than examples that were correctly predicted. We will first revisit the classifier employing minimum distance approach based on compound estimation for the first derivative in Chapter 2 due to its relatively good performance.

## Boosting of minimum distance approach

Sequential classification has been shown to perform better than direct classification in the four breast pathological types of Raman spectra data. For instance, when using the base classifier of minimum distance approach on the smoothed derivative estimation from compound estimation, the sequential classification has an overall correct classification rate of 103/124 versus 88/124 for the direct classification into four types. Thus, we will also apply sequential boosting to improve the performance of minimum distance approach.

Data normalization and smoothing steps are the same as described in section 6.2. Recall that in Chapter 2, in order to apply the minimum distance approach, a reference curve $j$ of derivative $j$ for diagnosis $c$, denoted $\widehat{\mu_c^{(j)}}(x)$, is defined as the average of $\widehat{\mu^{(j)}}(x)$ over all subjects known to have diagnosis $c$, where $c = 1$ with a normal diagnosis, $c = 2$ with cancer, $c = 3$ with fibroadenoma ("FA"), and $c = 4$ with fibrocystic change ("FC"), and $\widehat{\mu^{(j)}}(x)$ denote the estimated $j^{th}$ derivative. For boosting, as mentioned above, we will focus on the first derivative estimation ($j = 1$) when employing minimum distance approach. There are two ways to implement the sequential classification in the first step with regard to the reference curves. First, we can define two combined reference curves for N/FC and FA/C respectively, while each reference curve is defined as the average of $\widehat{\mu_d^{(j)}}(x)$ over all subjects known to have diagnosis $d$, where $d = 1$ with normal or fibrocystic change ("FC"), and $d = 2$ with fibroadenoma ("FA") or cancer. Another way is to use four reference curves as

we defined in Chapter 2, and output the predicted classification as N/FC if normal reference curve or FC reference curve is closest to the estimated first derivative to be classified or FA/C otherwise based on $L^1$ distance between the subject's estimated $j^{th}$ derivative and reference curve $j$ for diagnosis $d$. We will explore both approaches and discuss the performance of boosting for each approach. For brevity, we will call the first approach "Two reference curves" and the latter one "Four reference curves".

**Two reference curves**

We adapt the AdaBoost algorithm described in Chapter 4 with modifications in the sequential classification process using minimum distance approach.

First, the reference curve defined as $\widehat{\mu_d^{(j)}}(x)$ is updated by weighted average of $\widehat{\mu^{(j)}}(x)$ known to have diagnosis $d$ with the weight equal to the sampling weight in each round of boosting. Specifically, the weight on each observation was initialized to be equal, and then updated in each round according to $D_{t+1}(i) = \frac{D_t(i)exp(-c\alpha_t y_i h_t(x))}{Z_t}$, where $Z_t$ is chosen so that $D_{t+1}$ will be a distribution, $\alpha_t = 1/2(ln(\frac{1-\epsilon_t}{\epsilon_t}))$, and $\epsilon_t$ is the weighted classification error of the leave one out cross validation data defined as $\epsilon_t = Pr_{D_t}[h_t(x_i) \neq y_i]$.

Second, we introduce a scaling factor $c$ in $D_{t+1}(i) = \frac{D_t(i)exp(-c\alpha_t y_i h_t(x))}{Z_t}$ with $0 \leq c \leq 1$ and explore the effect of this scaling factor on the performance of boosting as shown in Figure 6.2 to 6.4 for total classification error, classification error for N/FC, and classification error for FA/C in the first step, respectively. This scaling factor $c$ could control the degree of boosting: the larger $c$ is, the higher the degree of boosting ($c = 0$ representing no boosting at all, and $c = 1$ representing AdaBoost). We have found out the total classification error of boosting with scaling factor 0.75 is better than with other scaling factors. This finding suggests that there may exist an optimal boosting scaling factor leading to the best performance of boosting. Also, at

the beginning of boosting (when number of rounds = 1), the classification error for N/FC is relatively higher than the classification error for FA/C (figure 6.3 and 6.4).

Number of rounds $T$ is another important parameter in AdaBoost. As number of

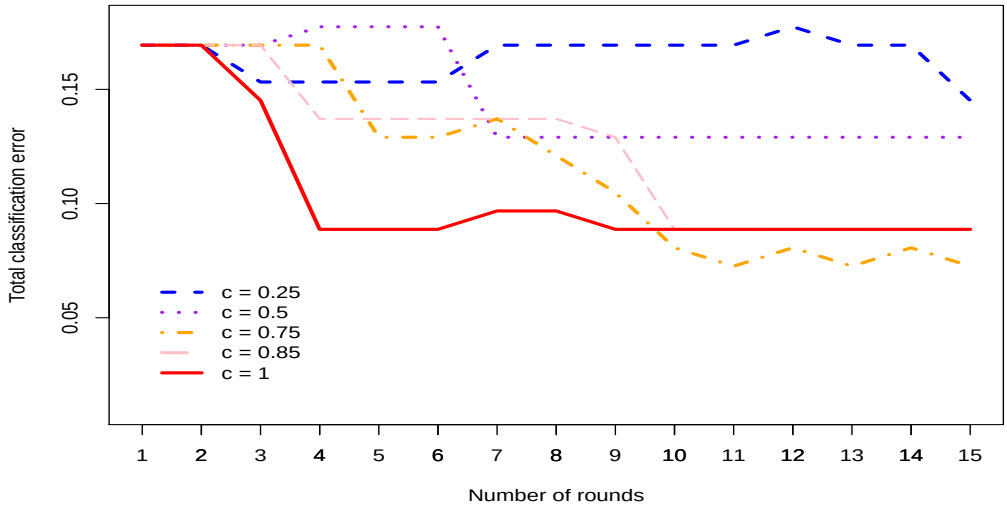


Figure 6.2: Total classification error against number of rounds with respect to scaling factor $c$ in the first step. The blue line represents the scaling factor $c = 0.25$; the green line represents the scaling factor $c = 0.5$; the orange line represents the scaling factor $c = 0.75$; the pink line represents the scaling factor $c = 0.85$; the red line represents the scaling factor $c = 1$.

rounds in boosting increases, the classification error for N/FC goes down, while the classification error for FA/C generally goes up. In practice, we usually stop iterations of boosting long after a consistency pattern for the overall classification error of training data is reached. Figure 6.2 shows early stopping at rounds 4 through 6 may result in poorer performance of boosting than the starting point when $c = 0.5$. Additionally, for all nonzero $c$ considered, there is a trade off between classification error of boosting for N/FC and the classification error for FA/C sometimes appearing as soon as in the first step, although the total classification error of boosting decreases as long as $c \neq 0$.

Table 6.3 shows the boosting performance via number of misclassifications at the two steps with regard to different scaling factor $c$. The second step of boosting

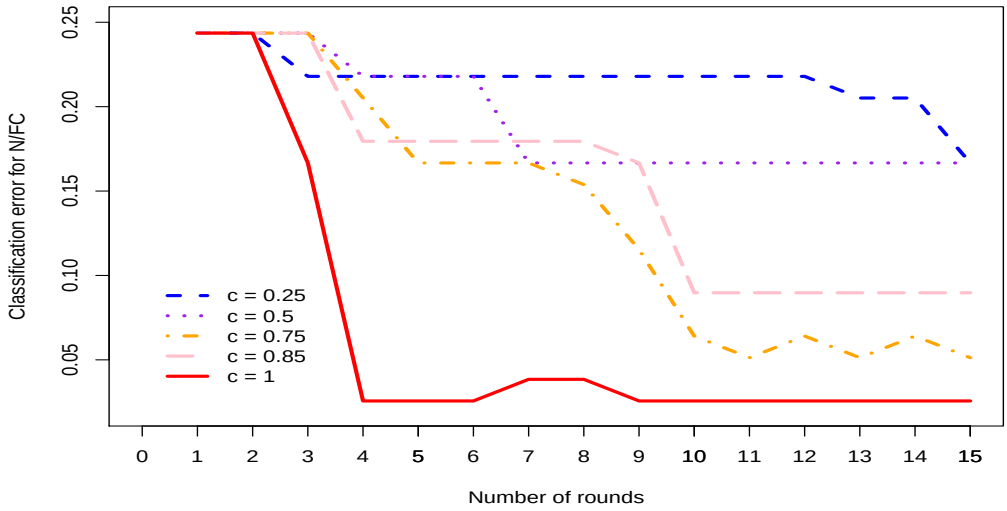


Figure 6.3: Classification error of N/FC against number of rounds with respect to scaling factor $c$ in the first step. The blue line represents the scaling factor $c = 0.25$; the green line represents the scaling factor $c = 0.5$; the orange line represents the scaling factor $c = 0.75$; the pink line represents the scaling factor $c = 0.85$; the red line represents the scaling factor $c = 1$.

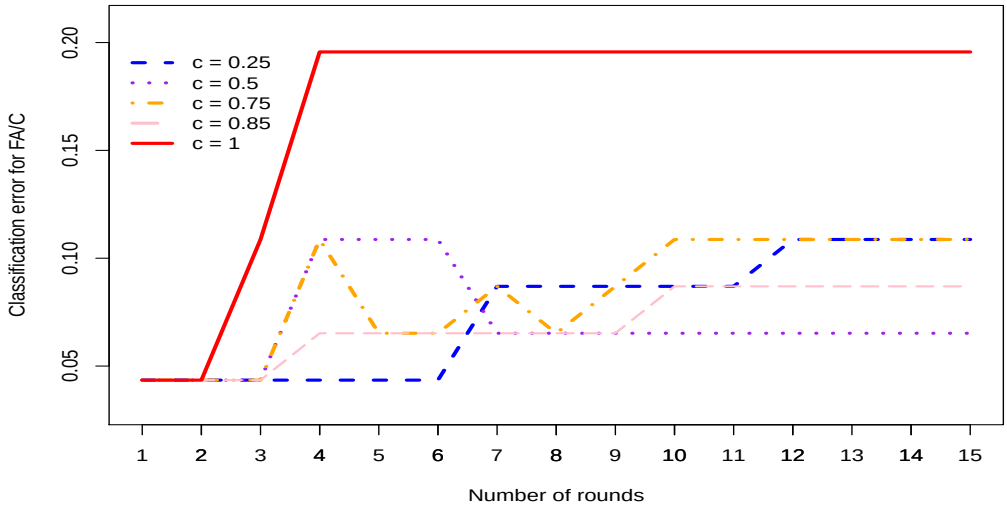


Figure 6.4: Classification error of FA/C against number of rounds with respect to scaling factor $c$ in the first step. The blue line represents the scaling factor $c = 0.25$; the green line represents the scaling factor $c = 0.5$; the orange line represents the scaling factor $c = 0.75$; the pink line represents the scaling factor $c = 0.85$; the red line represents the scaling factor $c = 1$.

separates N and FC, as well as C and FA respectively. As shown in table 6.3, the second step is based on the results from the first step using $c = 0.75$. When $c = 0.75$ in the first step, not only could we achieve better overall prediction in the first step than the other choices of $c$, but also obtain a satisfactory classification for FA/C. The results also indicate that when the sample size becomes small in the second step, there is not much improvement for classification using boosting.

Table 6.3: Number of misclassifications in boosting based on Two reference curves

| Step | Number of rounds | $T = 15$ | | | | | No |
|---|---|---|---|---|---|---|---|
| | Scaling factor | c=0.25 | c=0.5 | c=0.75 | c=0.85 | c=1 | boosting |
| Step 1 | N/FC | 13 | 13 | 4 | 7 | 2 | 19/78 |
| | C/FA | 5 | 3 | 5 | 4 | 9 | 2/46 |
| | Overall | 18 | 16 | 9 | 11 | 11 | 21/124 |
| Step 2 | N | 0 | 0 | 0 | 0 | 0 | 0/47 |
| (N vs FC) | FC | 8 | 7 | 7 | 7 | 10 | 10/27 |
| | N/FC | 8 | 7 | 7 | 7 | 10 | 10/74 |
| Step 2 | FA | 3 | 3 | 3 | 3 | 3 | 3/15 |
| (FA vs C) | C | 5 | 5 | 5 | 5 | 5 | 5/26 |
| | FA/C | 8 | 8 | 8 | 8 | 8 | 8/41 |

### Four reference curves

Boosting using four reference curves in the first step is similar to the approach using two reference curves described above except for the updated weights used to calculate the four reference curves in the first step. The reference curve defined as $\widehat{\mu_c^{(j)}}(x)$ is updated by weighted average of $\widehat{\mu^{(j)}}(x)$ known to have diagnosis $c$ in each round of boosting with the weight a function of the weighted classification error from the previous round for the first step "equivalence class" to which $c$ belongs, denoted $k$. Symbolically, in the Four reference curves approach, let $D_{tk}$ denote the round $t$ weight used to calculate both reference curves in equivalence class $k$; the weights are initialized to be equal. Then $D_{tk}$ will be updated in each round according to $D_{t+1,k}(i) = \frac{D_{tk}(i)exp(-c\alpha_{tk}y_i h_{tk}(x_i))}{Z_{tk}}$, where $Z_{tk}$ is chosen so that $D_{t+1,k}$ will be a distribution, $\alpha_{tk} = 1/2(ln(\frac{1-\epsilon_{tk}}{\epsilon_{tk}}))$, and $\epsilon_{tk} = Pr_{D_{tk}}[h_{tk}(x_i) \neq y_i | y_i = k]$ for $k = 1$ denoting "N/FC" and $k = -1$ denoting "FA/C".

Without boosting, compared with Two reference curves approach, the Four reference curves approach has less total misclassifications (17/124 vs. 24/124) for the first step. Table 6.4 shows the number of misclassifications in boosting based on four reference curves approach. Number of rounds $T$ is chosen to be 15, thus a consistency pattern for the overall classification error is reached. In the first step, the overall number of misclassifications (10/124) is smaller when the scaling factor $c$ was chosen to be 0.5 than the other choices for $c$. This result is comparable to that using two reference curves approach when $c = 0.75$ (9/124). For the second step, the classification is based on the results from the first step using $c = 0.5$. And the number of misclassifications is comparable to that using two reference curves approach. Again, there is no improvement from the boosting when separating "FA" from "C".

Table 6.4: Number of misclassifications in boosting based on Four reference curves

| Step | Number of rounds | $T = 15$ | | | | | No |
|---|---|---|---|---|---|---|---|
| | Scaling factor | c=0.25 | c=0.5 | c=0.75 | c=0.85 | c=1 | boosting |
| Step 1 | N | 0 | 0 | 0 | 0 | 0 | 0/47 |
| | FC | 4 | 5 | 8 | 5 | 7 | 11/31 |
| | FA | 1 | 1 | 2 | 2 | 2 | 0/15 |
| | C | 6 | 4 | 5 | 5 | 6 | 6/31 |
| | N/FC | 4 | 5 | 8 | 5 | 7 | 11/78 |
| | FA/C | 7 | 5 | 7 | 7 | 8 | 6/46 |
| | Overall | 11 | 10 | 15 | 12 | 15 | 17/124 |
| Step 2 (N vs FC) | N | 0 | 0 | 0 | 0 | 0 | 0/47 |
| | FC | 8 | 7 | 7 | 7 | 10 | 10/26 |
| | N/FC | 8 | 7 | 7 | 7 | 10 | 10/73 |
| Step 2 (FA vs C) | FA | 2 | 2 | 2 | 2 | 2 | 2/14 |
| | C | 4 | 4 | 4 | 4 | 4 | 4/27 |
| | FA/C | 6 | 6 | 6 | 6 | 6 | 6/41 |

Figure 6.5 to Figure 6.7 show the total classification error, classification error for N/FC, and classification error for FA/C in the first step against the number of rounds under different scaling factors for the Four reference curves approach. At the beginning without boosting, figure 6.5 shows the total classification error is less than that using Two reference approach. As number of rounds increases, the total misclassification errors of boosting go down, and stabilized for $c = 0.85$ and $c = 0.75$.

Yet, the total classification error goes up from round 14 to 15 when using $c = 0.25$ and $c = 1$. The "optimal scaling factors" are different from that in the Two reference curves approach. Specifically, $c = 0.5$ if using Four reference curves, while $c = 0.75$ if using Two reference curve.

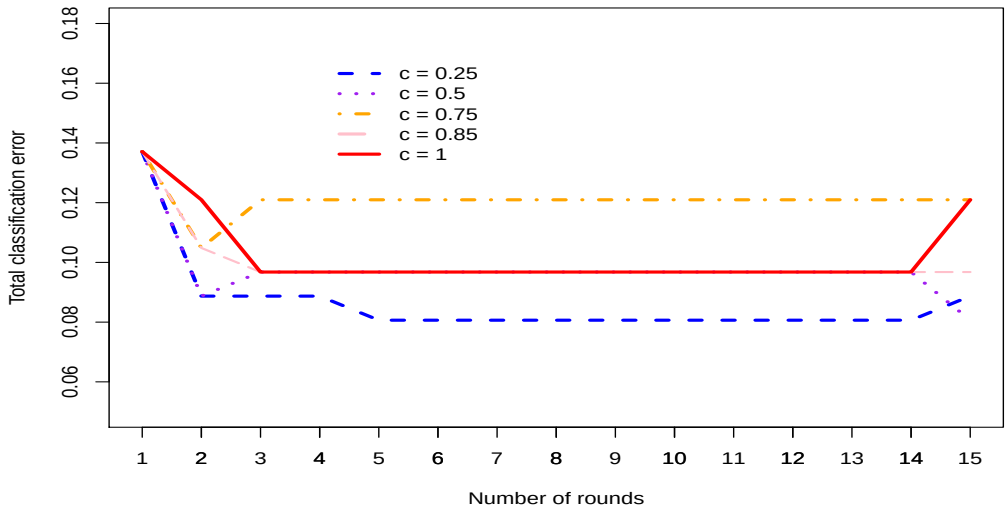


Figure 6.5: Total classification error against number of rounds with respect to scaling factor $c$ in the first step using Four reference curves approach. The blue line represents the scaling factor $c = 0.25$; the green line represents the scaling factor $c = 0.5$; the orange line represents the scaling factor $c = 0.75$; the pink line represents the scaling factor $c = 0.85$; the red line represents the scaling factor $c = 1$.

## Discussion

To summarize, boosting drives down the classification errors in both Two Reference curves (from 21/124 to 9/124) and Four Reference curves (from 17/124 to 10/124) approaches in the first step for classification of Raman spectroscopy data. It suggests that boosting could turn a weak learner into a relatively strong learner through weighted sampling and combination of classifiers from each round. Yet, boosting does not improve the classification rate significantly in Step Two for this particular data set. One reason is that during the second step, the sample size gets smaller. The other

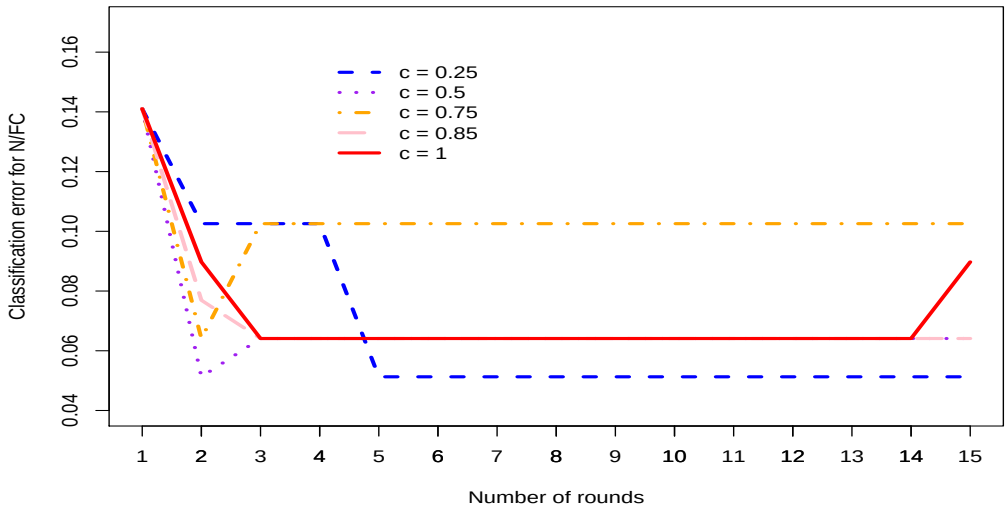


Figure 6.6: Classification error of N/FC against number of rounds with respect to scaling factor $c$ in the first step using Four reference curves approach. The blue line represents the scaling factor $c = 0.25$; the green line represents the scaling factor $c = 0.5$; the orange line represents the scaling factor $c = 0.75$; the pink line represents the scaling factor $c = 0.85$; the red line represents the scaling factor $c = 1$.

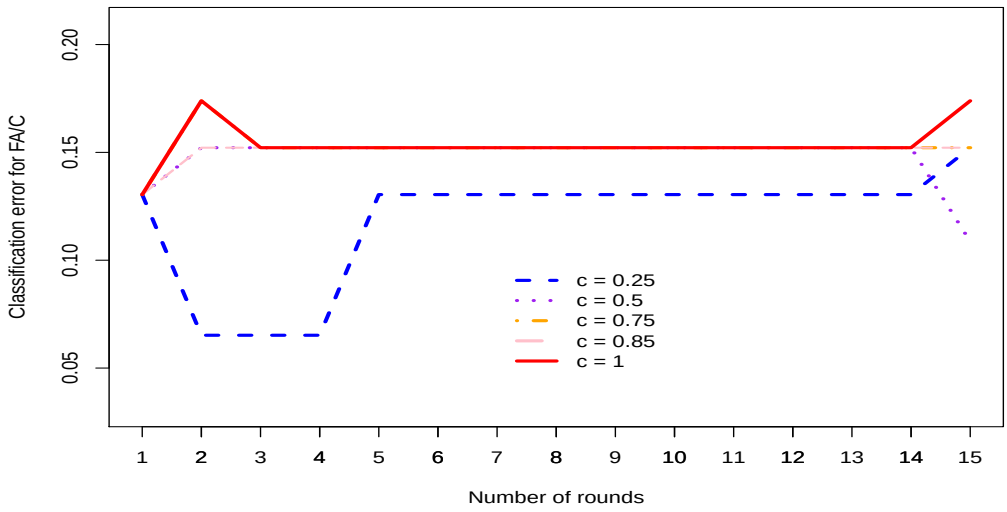


Figure 6.7: Classification error of FA/C against number of rounds with respect to scaling factor $c$ in the first step using Four reference curves approach. The blue line represents the scaling factor $c = 0.25$; the green line represents the scaling factor $c = 0.5$; the orange line represents the scaling factor $c = 0.75$; the pink line represents the scaling factor $c = 0.85$; the red line represents the scaling factor $c = 1$.

reason may be inherent closeness of "N" and "FC", "FA" and "C" makes the second step classifications more difficult, and thus Boosting's effectiveness in improvement of classification accuracy is minimal. Another important finding by modification of AdaBoost is that adding a scaling factor in AdaBoost could potentially improve the classification rate in this real data application. Finally, we have defined weights that can be updated in each round of boosting in order to calculate the "reference curves" and generate different classifiers in different rounds of boosting.

## 6.4 Conclusions

In this dissertation, we developed a novel hybridization of nonparametric smoothing, dimension reduction, and statistical learning techniques with application to Raman Spectroscopy data for breast cancer diagnosis.

This dissertation work is primarily motivated by the practical application of using a Raman spectrum derived from human breast tissue to classify the tissue as normal, cancerous, or abnormal but benign. From an epidemiologic perspective, if such a diagnostic procedure proves to be sufficiently sensitive and specific, then Raman spectroscopy (which is non-invasive) may provide a useful intermediary between mammography and surgical biopsy. For example, if a mammogram is inconclusive but Raman spectroscopy clearly suggests that no cancer is present, then an unnecessary surgical biopsy may be avoided. Given the immense numbers of mammograms that are performed, Raman spectroscopy may thus provide a mechanism to more effectively cope with the large numbers of false positives that occur.

The motivation for our methodology is that, besides spectra profiles, their derivatives may be useful for classifying different types of breast pathologies because they reveal high frequency features of signals. Also, by appealing to ideas from machine learning, diagnoses can be improved and conflicts between different diagnostic methods can be settled.

In Chapter 2, we applied compound estimation to smooth the existing spectrum data in order to remove or lessen stochastic noise and acquire smooth objects for differentiation. Compound estimation is a very recent innovation, whose hallmark is the self-consistency property: the derivatives of the estimated mean response function estimate the derivatives of the mean response function (Charnigo and Srinivasan, 2011)[18]. This interchange of differentiation with estimation means that logical conflicts are avoided, such as an estimated local maximum of a mean response function occurring where the estimated first derivative is nonzero. Importantly, not all smoothing techniques possess the self-consistency property.

Later in Chapter 6, we represented both the compound-estimated Raman spectra profiles and their derivatives as linear combinations of compound-estimated basis Raman spectra and their derivatives, respectively. Although in principle coefficients from these two linear regressions should be the same, coefficient estimates will be slightly different from each other and those of Haka et al. (2005) because of smoothing and whether high-amplitude or high-frequency features are emphasized, per consideration of the spectra or their derivatives respectively. Moreover, coefficient estimates obtained from these two linear regressions may ultimately lead to improved classification.

Stacking is a process to combine multiple classifiers together to make better predictions. For example, we may have one classifier based on the approach proposed by Haka et al using the raw data (2005), and the other two classifiers based on smoothed Raman spectrum data from compound estimation, and the derivatives of smoothed Raman spectrum data. The conflicts from different classifier predictions could be reconciled by assigning them weights based on their performance. The idea, roughly speaking, is that one classifier which makes a mistake can be overridden by the other two classifiers. However, the number of classifiers is not limited to three in stacking. We set up a sequential classification framework for multiple class prediction (repre-

senting four types of breast tissue) and provided a theoretical rationale for two stage sequential classification.

Boosting is a process to improve an individual classifier by iteratively reweighting observations based on where the mistakes were made in the previous step. We implemented it sequentially in two steps and justified a family of reweighting schemes, of which the standard AdaBoost algorithm is a special case in Chapter 4. We also proposed an ensemble integration scheme to consist of stacking and boosting together in Chapter 5.

Methodology developed in this dissertation exhibited some success in improving accuracy of prediction in both numerical simulation and practical application to actual Raman Spectroscopy data for diagnosing breast cancer. For example, by combining classifiers using stacking, the prediction of classification is better than or at least the same as the best individual classifier (the overall correct classification rate 87.1% vs 85.4%, and sensitivity is 90.3% vs 90.3%). Also, there are improvements from boosting in simulation studies based on minimum distance approach. The overall classification errors in the first step decrease from 0.1627 to 0.1236 with 2/3 training data, and from 0.1632 to 0.1149 with 1/3 training data, while in the second step they presented more remarkable improvements: from 0.1551 to 0.0982 with 2/3 training data, and from 0.1755 to 0.0971 with 1/3 training data.

## 6.5 Future research work and applications

Dynamic ensemble integration scheme has the potential to further improve classification accuracy in Raman spectra data analysis if data set is large enough. An immediate future work is needed to define similarity metrics between observations in order to implement dynamic ensemble integration scheme in nonparametric regression settings.

The methodology developed in this dissertation has numerous potential applica-

tions, especially to functional data considered as observations varying over a continuum (e.g. spectrometric curves, brain scans (fMRI), or protein and gene expression profiles). Other scenarios in which an estimated mean response function may indicate how to classify or characterize some person or object may also be examined using some variant of the methodology in this dissertation. For example, glucose tolerance tests or lipoprotein kinetics typically entail acquiring a set of data points from each subject corresponding to time and serum concentration or specific activity, and one may imagine that there is some interest in classifying such a subject with respect to his/her present condition (which is perhaps already known), then with regard to a prognosis for his/her future status. Similar to Raman spectroscopy data, Electroencephalography (EEG) data can also be analyzed in the same way to detect abnormality among those patients with chronic diseases such as epilepsy, and sleep disorders.

It is desirable to use functional data as predictors to guide clinical decision making and personalized treatment. For example, functional data as a predictor (such as gene expression profiles of patients with leukemia or brain scans of patients with schizophrenia) can be used to predict a categorical outcome (such as different types of diagnosis) or a continuous outcome (such as survival time or time to recurrence). Following the idea of dynamic ensemble integration in Chapter 5, we could refine the training data for prediction on a particular observation in order to achieve better prediction and guide personalized clinic decision making.

## Bibliography


[1] M. Awais, F. Yan, K. Mikolajczyk, and J. Kittler. Novel fusion methods for pattern recognition. *Machine Learning and Knowledge Discovery in Databases*, pages 140–155, 2011.

[2] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[3] P. Barter, C. Ballantyne, R. Carmena, M. Cabezas, M. Chapman, P. Couture, J. Graaf, P. Durrington, O. Faergeman, J. Frohlich, et al. Apo b versus cholesterol in estimating cardiovascular risk and in guiding therapy: report of the thirty-person/ten-country panel. *Journal of internal medicine*, 259(3):247–258, 2006.

[4] C. Bennett and T. Doub. Data mining and electronic health records: selecting optimal clinical treatments in practice. *arXiv preprint arXiv:1112.1668*, 2011.

[5] M. S. Bergholt, W. Zheng, K. Y. Ho, K. G. Yeoh, and Z. Huang. Raman endoscopy for objective diagnosis of early cancer in the gastrointestinal system. *J Gastroint. Dig. Syst. S*, 1:008, 2013.

[6] H. Bierens. Kernel estimators of regression functions. In *Advances in Econometrics: Fifth world congress*, volume 1, pages 99–144. Cambridge University Press New York, 1987.

[7] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[8] L. Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

[9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and regression trees. wadsworth & brooks. *Monterey, CA*, 1984.

[10] P. Bühlmann and T. Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

[11] I. G. Campbell. Eeg recording and analysis for sleep research. *Current Protocols in Neuroscience*, pages 10–2, 2009.

[12] R. Caruana, A. Munson, and A. Niculescu-Mizil. Getting the most out of ensemble selection. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 828–833. IEEE, 2006.

[13] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, page 18. ACM, 2004.

[14] R. Charnigo, M. Francoeur, P. Kenkel, M. Pinar Mengüç, B. Hall, and C. Srinivasan. Estimating quantitative features of nanoparticles using multiple derivatives of scattering profiles. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 112(8):1369–1382, 2011.

[15] R. Charnigo, M. Francoeur, M. Mengüç, A. Brock, M. Leichter, and C. Srinivasan. Derivatives of scattering profiles: tools for nanoparticle characterization. *JOSA A*, 24(9):2578–2589, 2007.

[16] R. Charnigo, B. Hall, and C. Srinivasan. A generalized cp criterion for derivative estimation. *Technometrics*, 53(3):238–253, 2011.

[17] R. Charnigo, B. Hall, and C. Srinivasan. Simultaneous confidence bands for derivatives of mean responses. Submitted for publication and available at www.richardcharnigo.net/TechReports, 2013.

[18] R. Charnigo and C. Srinivasan. Self-consistent estimation of mean response functions and their derivatives. *Canadian Journal of Statistics*, 39(2):280–299, 2011.

[19] N. Clothup, L. Daly, and S. Wiberley. Introduction to infrared and raman spectroscopy. *Academic, San Diego*, 1990.

[20] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[21] H. Drucker, R. Schapire, and P. Simard. Boosting performance in neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):705–719, 1993.

[22] S. Džeroski and B. Ženko. Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 54(3):255–273, 2004.

[23] M. B. Fenn, P. Xanthopoulos, G. Pyrgiotakis, S. R. Grobmyer, P. M. Pardalos, and L. L. Hench. Raman spectroscopy for clinical oncology. *Advances in Optical Technologies*, 2011, 2011.

[24] J. Fox. Nonparametric regression. *CRAN R Project. January*, 2002.

[25] Y. Freund, Y. Mansour, and R. E. Schapire. Generalization bounds for averaged classifiers. *Annals of Statistics*, pages 1698–1722, 2004.

[26] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[27] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[28] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2001.

[29] L. Galway. Spline models for observational data. *Technometrics*, 34(1):113–114, 1992.

[30] G. Giacinto, F. Roli, and G. Fumera. Design of effective multiple classifier systems by clustering of classifiers. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 2, pages 160–163. IEEE, 2000.

[31] L. Györfi. *A distribution-free theory of nonparametric regression*. Springer Verlag, 2002.

[32] A. Haka, K. Shafer-Peltier, M. Fitzmaurice, J. Crowe, R. Dasari, and M. Feld. Diagnosing breast cancer by using raman spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35):12371, 2005.

[33] A. Haka, Z. Volynskaya, J. Gardecki, J. Nazemi, R. Shenk, N. Wang, R. Dasari, M. Fitzmaurice, and M. Feld. Diagnosing breast cancer using raman spectroscopy: prospective analysis. *Journal of biomedical optics*, 14:054023, 2009.

[34] P. Hall and M. Wand. Minimizing l1 distance in nonparametric density estimation. *Journal of Multivariate Analysis*, 26(1):59–88, 1988.

[35] W. Hardle. *Applied nonparametric regression*, volume 26. Cambridge Univ Press, 1990.

[36] W. Härdle. *Nonparametric and semiparametric models*. Springer Science & Business Media, 2004.

[37] W. Härdle, Y. Ritov, and S. Song. Partial linear quantile regression and bootstrap confidence bands. Technical report, SFB 649 Discussion Paper 2010-002, 2009.

[38] M. Hillebrand, C. Wöhler, L. Krüger, U. Kreßel, and F. Kummert. Self-learning with confidence bands. In *Proc. of the 20th Workshop Computational Intelligence*, pages 302–313. Citeseer, 2010.

[39] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. Thun. Cancer statistics, 2009. *CA: a cancer journal for clinicians*, 59(4):225–249, 2009.

[40] J. Kelsey, M. Gammon, and E. John. Reproductive factors and breast cancer. *Epidemiologic reviews*, 15(1):36, 1993.

[41] W. Kendall, J. Marin, and C. Robert. Confidence bands for brownian motion and applications to monte carlo simulation. *Statistics and Computing*, 17(1):1–10, 2007.

[42] R. Kennedy, Y. Lee, B. Van Roy, C. Reed, and R. Lippmann. *Solving data mining problems through pattern recognition*. Prentice Hall PTR Indianapolis:, 1997.

[43] A. H. Ko, R. Sabourin, and A. S. Britto Jr. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, 41(5):1718–1731, 2008.

[44] P. M. Layde, L. A. Webster, A. L. Baughman, P. A. Wingo, G. L. Rubin, H. W. Ory, Cancer, S. H. S. Group, et al. The independent associations of parity, age at first full term pregnancy, and duration of breastfeeding with the risk of breast cancer. *Journal of clinical epidemiology*, 42(10):963–973, 1989.

[45] A. Lazarevic and Z. Obradovic. Effective pruning of neural network classifier ensembles. In *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*, volume 2, pages 796–801. IEEE, 2001.

[46] A. Ledezma, R. Aler, A. Sanchis, and D. Borrajo. Ga-stacking: Evolutionary stacked generalization. *Intelligent Data Analysis*, 14(1):89–119, 2010.

[47] V. Lin, N. Colthup, W. Fateley, and J. Grasselli. Handbook of infrared and raman characteristic frequencies of organic molecules. *Recherche*, 67:02, 1991.

[48] C. Loader. *Local regression and likelihood*. Springer Verlag, 1999.

[49] W. Martinez and J. B. Gray. The role of margins in boosting and ensemble performance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(2):124–131, 2014.

[50] P. Matousek and N. Stone. Prospects for the diagnosis of breast cancer by noninvasive probing of calcifications using transmission raman spectroscopy. *Journal of Biomedical Optics*, 12:024008, 2007.

[51] P. Matousek and N. Stone. Recent advances in the development of raman spectroscopy for deep non-invasive medical diagnosis. *Journal of biophotonics*, 6(1):7–19, 2013.

[52] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa. Ensemble approaches for regression: A survey. *ACM Computing Surveys (CSUR)*, 45(1):10, 2012.

[53] C. J. Merz. Using correspondence analysis to combine classifiers. *Machine Learning*, 36(1-2):33–58, 1999.

[54] H. Moon, H. Ahn, R. L. Kodell, S. Baek, C.-J. Lin, and J. J. Chen. Ensemble methods for classification of patients for personalized medicine with high-dimensional data. *Artificial intelligence in medicine*, 41(3):197–207, 2007.

[55] È. A. Nadaraya. On estimating regression. *Teoriya Veroyatnostei i ee Primeneniya*, 9(1):157–159, 1964.

[56] A. Nover, S. Jagtap, W. Anjum, H. Yegingil, W. Shih, W. Shih, and A. Brooks. Modern breast cancer detection: a technological review. *Journal of Biomedical Imaging*, 2009:4–4, 2009.

[57] I. Partalas, G. Tsoumakas, and I. P. Vlahavas. Focused ensemble selection: A diversity-based method for greedy ensemble selection. In *ECAI*, pages 117–121, 2008.

[58] D. Partridge and W. B. Yates. Engineering multiversion neural-net systems. *Neural Computation*, 8(4):869–893, 1996.

[59] P. Pharoah, N. Day, S. Duffy, D. Easton, and B. Ponder. Family history and the risk of breast cancer: A systematic review and meta-analysis. *International Journal of Cancer*, 71(5):800–809, 1997.

[60] C. V. Raman and K. S. Krishnan. A new type of secondary radiation. *Nature*, 121(3048):501–502, 1928.

[61] T. W. Randolph. Scale-based normalization of spectral data. *Cancer Biomarkers*, 2(3):135–144, 2006.

[62] C. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183, 1967.

[63] L. Reyzin and R. E. Schapire. How boosting the margin can also boost classifier complexity. In *Proceedings of the 23rd international conference on Machine learning*, pages 753–760. ACM, 2006.

[64] B. Ripley. *Pattern recognition and neural networks*. Cambridge Univ Pr, 1996.

[65] M. Robnik-Šikonja. Improving random forests. In *Machine Learning: ECML 2004*, pages 359–370. Springer, 2004.

[66] L. Rokach. *Pattern classification using ensemble methods*, volume 75. World Scientific, 2009.

[67] A. Sasco, M. Secretan, and K. Straif. Tobacco smoking and cancer: a brief review of recent epidemiological evidence. *Lung cancer*, 45:S3–S9, 2004.

[68] R. Schapire. A brief introduction to boosting. In *International Joint Conference on Artificial Intelligence*, volume 16, pages 1401–1406. LAWRENCE ERLBAUM ASSOCIATES LTD, 1999.

[69] R. Schapire. Theoretical views of boosting and applications. In *Algorithmic Learning Theory*, pages 13–25. Springer, 1999.

[70] R. E. Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[71] R. E. Schapire and Y. Freund. *Boosting: Foundations and algorithms*. MIT Press, 2012.

[72] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of statistics*, pages 1651–1686, 1998.

[73] SEER. National cancer institute seer*stat software. http://www.seer.cancer.gov/seerstat, Accessed September 6, 2012 2012.

[74] A. K. Seewald. How to make stacking better and faster while also taking care of an unknown weakness. In *Proceedings of the nineteenth international conference on machine learning*, pages 554–561. Morgan Kaufmann Publishers Inc., 2002.

[75] K. Shafer-Peltier, A. Haka, M. Fitzmaurice, J. Crowe, J. Myles, R. Dasari, and M. Feld. Raman microspectroscopic model of human breast tissue: implications

for breast cancer diagnosis in vivo. *Journal of Raman Spectroscopy*, 33(7):552–563, 2002.

[76] C. Shen and H. Li. Boosting through optimization of margin distributions. *Neural Networks, IEEE Transactions on*, 21(4):659–666, 2010.

[77] R. Siegel, D. Naishadham, and A. Jemal. Cancer statistics, 2012. *CA: a cancer journal for clinicians*, 62(1):10–29, 2012.

[78] P. Speckman. Spline smoothing and optimal rates of convergence in nonparametric regression models. *The Annals of Statistics*, pages 970–983, 1985.

[79] C. J. Stone. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360, 1980.

[80] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

[81] N. Stone, P. Stavroulaki, C. Kendall, M. Birchall, and H. Barr. Raman spectroscopy for early detection of laryngeal malignancy: preliminary results. *The Laryngoscope*, 110(10):1756–1763, 2000.

[82] E. K. Tang, P. N. Suganthan, and X. Yao. An analysis of diversity measures. *Machine Learning*, 65(1):247–271, 2006.

[83] Q. Tang and J. Wang. L 1-estimation for varying coefficient models. *Statistics*, 39(5):389–404, 2005.

[84] K. Ting and I. Witten. Issues in stacked generalization. *Arxiv preprint arXiv:1105.5466*, 2011.

[85] K. M. Ting and I. H. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:271–289, 1999.

[86] L. Todorovski and S. Džeroski. Combining classifiers with meta decision trees. *Mach. Learn.*, 50(3):223–249, Mar. 2003.

[87] G. Tsoumakas, I. Partalas, and I. Vlahavas. A taxonomy and short review of ensemble selection. In *ECAI 2008, workshop on supervised and unsupervised ensemble methods and their applications*, 2008.

[88] A. Tsymbal, M. Pechenizkiy, and P. Cunningham. Dynamic integration with random forests. In *Machine Learning: ECML 2006*, pages 801–808. Springer, 2006.

[89] V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 2. Wiley New York, 1998.

[90] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[91] G. Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

[92] F. Wang and D. Scott. The l 1 method for robust nonparametric regression. *Journal of the American Statistical Association*, pages 65–76, 1994.

[93] L. Wasserman. *All of nonparametric statistics*. Springer-Verlag New York Inc, 2006.

[94] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[95] D. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[96] K. Woods, W. P. Kegelmeyer Jr, and K. Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 19(4), 1997.

[97] K. Yabroff, E. Lamont, A. Mariotto, J. Warren, M. Topor, A. Meekins, and M. Brown. Cost of care for elderly cancer patients in the united states. *Journal of the National Cancer Institute*, 100(9):630–641, 2008.

[98] P. Zwoliński, M. Roszkowski, J. Żygierewicz, S. Haufe, G. Nolte, and P. J. Durka. Open database of epileptic eeg with mri and postoperational assessment of focia real world verification for the eeg inverse solutions. *Neuroinformatics*, 8(4):285–299, 2010.

## Vita

**Jing Guo**

### Educational Background

- Bachelor of Medicine in Preventive Medicine with honors,
  School of Public Health, Shandong University, Jinan, China, 2009.
- Bachelor of Science in Information science and computation with honors,
  School of Mathematics, Shandong University, Jinan, China, 2009.

### Professional Experience

- Teaching Assistant, 08/2014 - 12/2014. Department of Statistics, University of Kentucky, Lexington, KY, USA.
- Research Assistant, 01/2012 - 08/2014. Kentucky Cancer Registry, Lexington, KY, USA.
- Research Assistant, 08/2010 - 12/2011. School of Public Health, University of Kentucky, Lexington, KY, USA.
- Intern, 02/2009-04/2009. Shandong Center for disease control and prevention, Jinan, China.

### Awards

- Presentation Award 2nd Place at 9th CCTS Conference, March 27 2014

- Student Travel Award to ENAR 2014 from Graduate School at University of Kentucky, Spring 2014
- Excellent Graduate Award (3%) in Shandong Province, 2009
- Academic Scholarship for 4 consecutive years (top 5%), 2005-2008

**Publications**

- Kristin L Long, Bin Huang, **Jing Guo**, Cortney Lee, David Sloan, Shaun McKenzie. "Concerning Trends in Appalachian Patients with Thyroid." American Surgeon. (2014) 80(6): 620-3.
- Bin Huang, **Jing Guo**, Richard Charnigo. "Statistical Methods for Population-Based Cancer Survival in Registry Data." Journal of Biometrics and Biostatistics (2014) 5: e129.
- Katherine E. Campbell, Bin Huang, **Jing Guo**, Timothy W. Mullett, Jeremiah T. Martin, B. Mark Evers, Shaun P. McKenzie. "Tu1547 Esophageal and Gastroesophageal Adenocarcinoma in Young Patients: A Call for a Continued Aggressive Approach to Both Diagnosis and Treatment. " Gastroenterology. (2013) 144(5):S-1124-S-1125.
- Heather M. Bush, Stacey Pagorek, Janice Kuperstein, **Jing Guo**, Katie N. Ballert, and Leslie J. Crofford. "The Association of Chronic Back Pain and Stress Urinary Incontinence: A Cross-sectional Study." Journal of Womens Health Physical Therapy 37, no. 1 (2013): 11-18.