RICE UNIVERSITY

Functional Data Analysis on Spectroscopic Data

by

Lu Wang

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Dennis D. Cox
Dr. Dennis D. Cox, Chair
Professor of Statistics

David W. Scott
Dr. David W. Scott
Noah Harding Professor of Statistics

Yin Zhang
Dr. Yin Zhang
Professor of Computational and Applied Mathematics.

HOUSTON, TEXAS
DECEMBER 2015
Rice University

Functional Data Analysis on Spectroscopic Data

by

Lu Wang

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Approved, Thesis Committee:

Dr. Dennis D. Cox, Chair
Professor of Statistics

Dr. David W. Scott
Noah Harding Professor of Statistics

Dr. Yin Zhang
Professor of Computational and Applied Mathematics.

Houston, Texas
January 2016
Abstract

Functional Data Analysis on Spectroscopic Data

by

Lu Wang

Functional Data Analysis (FDA) is a collection of statistical methods on analyzing dataset measured along a continuum. It extends the capabilities of conventional statistical techniques, studies the variability of data when the observations are curves instead of points. This thesis addresses problems in two aspects of functional data. The first part is calculation correction factors to remove variabilities using a log additive model with ANOVA estimates on repeated measurements. The second part is about peak estimation algorithm in functional data. Also, there is a side project to deal with classification on unbalanced data using corrupted data technique.

The first project is motivated by data from experiment on medical spectroscopic devices. The goal is to develop statistical models to fit repeated measurements under different devices and groups, and calibrate measurement when the device is used on patients. The calibration factors are calculated both along wavelength and among devices and other factors. An experiment was designed to study the sources of variation in the measurements recorded. Here we present a log additive statistical model that incorporates the sources of variability we identified. Based on
this model, we estimated correction factors from the experimental data needed to eliminate the inter-device variability and other sources of variation. These correction factors are intended to improve the accuracy and repeatability of such devices when making future measurements on patient tissue.

The second part is focused on the kind of data with important information around peak area, so peak locating is a critical step before further analysis. One common problem is that the peak location and height are distorted or shifted in measured data, therefore, we propose a local quadratic fitting model to deal with this kind of problem and further investigated the optimal bandwidth for local polynomial regression, and derived a plug-in bandwidth selection algorithm for estimating peak location and height simultaneously.

The third part is motivated by classification on highly unbalanced patient measurements, when the prediction power is weakened due to the small amount of diseased case. When we use the training set to learn the classifier that predict well on test set data, we could extend the training set with artificially created data, i.e. corrupt the data with noises from a known distribution. Simulation studies and comparisons are presented.
ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest appreciation to my advisor, Dr. Dennis D. Cox, who has been very supportive and encouraging for my study. He guided me with his wisdom and experience, inspired me in the field of statistics, lead me the way through my learning and research experiences.

I would also like to thank Dr. David Scott and Dr. Yin Zhang, who have shared valuable insights in the relevance of the study with me and contributed and extended their valuable assistance in the preparation of this thesis.

Last but not least, many thanks to Dr. Neely Atkinson, Dan Serachitopol, Dr. Jong Soo Lee, Dr. Jose-Miguel Yamal and all other project mates, who helped me and gave suggestions on different aspects of my research.

This dissertation is supported by National Cancer Institute grant PO1-CA82710 (Optical Technologies and Molecular Imaging for Cervical Neoplasia).
I lovingly dedicate this thesis to my parents and my husband, who are always being so supportive to me. A special thanks to my lovely daughter, without whom this thesis could be done sooner.
Abstract

Acknowledgements

1 Introduction

1.1 Functional Data ............................................. 3
1.2 Popular Functional Data Analysis Methods ..................... 6

2 Log Additive Model on Repeated Measurements .................. 9

2.1 Motivating Data Set ......................................... 9
2.2 Heuristic Fitting Approach .................................. 17
2.3 Log Additive Model ......................................... 21
2.4 Discussion and Future Works .................................. 29

3 Local Polynomial Regression for Peak Finding Algorithm ........... 32

3.1 Background .................................................. 33
3.2 Dataset and Peak Problems .................................... 35
3.3 Literature Review on Kernel Methods ............................ 40
3.4 Local Quadratic Fitting Algorithm ............................. 50
3.5 Bandwidth Selection for Local Polynomial Estimates of Peak Location and Height ................................. 56
3.6 Simulation Study ........................................... 65
3.7 Real Data Application ....................................... 80
3.8 Discussion and conclusion .................................... 85
4 Classification with Corrupted Data

4.1 Literature Review ........................................... 89
4.2 Proposed Data Corruption Method ......................... 91
4.3 Corruption on Real Data Studies ............................ 92
4.4 Discussion and Conclusion ................................. 95
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>An example of functional data</td>
<td>4</td>
</tr>
<tr>
<td>2.1</td>
<td>MDC3 device</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>MDC3 Point Probe</td>
<td>11</td>
</tr>
<tr>
<td>2.3</td>
<td>MDC3 imaging</td>
<td>11</td>
</tr>
<tr>
<td>2.4</td>
<td>Log of light intensity measurement</td>
<td>15</td>
</tr>
<tr>
<td>2.5</td>
<td>Gold standard of the devices</td>
<td>16</td>
</tr>
<tr>
<td>2.6</td>
<td>Heuristic approach correction factors</td>
<td>19</td>
</tr>
<tr>
<td>2.7</td>
<td>Log10 of the corrected measurements. Plotted in color of red, green, blue and black for device 1, 2, 3 and 4, but overlapped and only shown black color in the figure.</td>
<td>20</td>
</tr>
<tr>
<td>2.8</td>
<td>Log Additive Model for Filter set 1</td>
<td>24</td>
</tr>
<tr>
<td>2.9</td>
<td>Log Additive Model for Filter set 2</td>
<td>25</td>
</tr>
<tr>
<td>2.10</td>
<td>Log Additive Model for Filter set 3</td>
<td>26</td>
</tr>
<tr>
<td>2.11</td>
<td>Log Additive Model for Filter set 4</td>
<td>27</td>
</tr>
<tr>
<td>2.12</td>
<td>$AC_{ijk} = \exp(lAC_{ijk})$ estimates for filter set 1. Red lines separate the sessions. 5 colors indicate sets of 5 measurements. (Missing values are left blank)</td>
<td>29</td>
</tr>
<tr>
<td>3.1</td>
<td>FastEEM3 device</td>
<td>34</td>
</tr>
<tr>
<td>3.2</td>
<td>FastEEM3 device measuring standard</td>
<td>35</td>
</tr>
<tr>
<td>3.3</td>
<td>Repeatability raw measurements</td>
<td>36</td>
</tr>
<tr>
<td>3.4</td>
<td>Rhodamine measurement excitation 330nm</td>
<td>37</td>
</tr>
<tr>
<td>3.5</td>
<td>Rhodamine peak heights for excitation 330nm</td>
<td>39</td>
</tr>
</tbody>
</table>
3.6 Epanechnikov kernel ........................................... 47
3.7 Boxplot of 3 key features in excitation 360nm ................. 54
3.8 Boxplot of 3 key features by devices .......................... 55
3.9 Asymptotic and empirical bandwidth for location estimator .......................... 66
3.10 Asymptotic and empirical bandwidth for height estimator .................. 68
3.11 3 sets of simulated data ........................................ 70
3.12 MSE for location estimates ....................................... 72
3.13 MSE for height estimates ........................................ 73
3.14 Asymptotic optimal and empirical best bandwidths .................. 75
3.15 Asymptotic and empirical MSE ................................... 77
3.16 Asymptotic and plug-in bandwidth ................................ 79
3.17 Asymptotic and plug-in MSE ..................................... 80
3.18 Rhodamine measurements ....................................... 81
3.19 Rhodamine measurement peak estimates with local quadratic models ............. 83
3.20 Rhodamine measurement peak location and height estimates .................. 84
Functional data, which refer to the data that can be represented as continuous functions, are often observed as a finite number of discrete values contain information of intrinsically smooth curves, and usually there are replications of the samples to make statistical inference possible with a population of curves. The concept of functional data analysis was first brought up by J.O. Ramsay in 1982 [40], and a more in depth analysis methodology paper in 1986 [41]. From then on, treating observations as functions are getting more and more popular among theoretical and applied statisticians.

When representing discrete values to curves and treating each curve as a single observation, analysis on a group of curves arises many complications. Therefore, popular multivariate data analysis approaches are used on functional data, such as principle component analysis, factor analysis. Also, functional linear models, functional ANOVA, analysis on the derivatives of functions, using generalized spline functions or basis functions to represent functional data, smoothing methods using penalty terms or kernels .etc are very important estimation approaches.

In this chapter, I will first give the definition of functional data, followed by some
examples and list of popular functional data analysis techniques. The second part of this chapter is the literature review of some major methods and techniques used in related areas.

In chapter 2, a log additive model is motivated by light intensity data from an experiment on consistency between devices and measurement standards. The device is developed for cervical cancer and pre-cancer status detection. We analyze the repeated measurements from the ideally identical devices, formulate a log additive model with ANOVA coefficient estimates and therefore derived correction factors by device and standard, which will be used to remove systematic differences among them and as calibrations in real time patient diagnoses.

In chapter 3, I propose a local polynomial regression model to estimate peak location and height. The focus of this study is on analyzing the peak area of the curves, which is usually called peak locating or bump hunting according to previous studies. In real data samples, it is often the case that peak area contains more important information of interest than other areas along the curve. To estimate the peak, bandwidth in local polynomial regression plays an important role as a smoothing parameter, the asymptotic optimal bandwidth for estimating peak location and height are derived and compared with empirical results. Based on the asymptotic bandwidth equation, we also proposed a data drive plug-in bandwidth selection algorithm. Simulation and real data application are presented, followed by conclusion and discussion.

In chapter 4, I will focus on patient tissue measurement data with small proportion of diseased case. We propose a method used in regression type of problem, to generate data by adding noise to independent variables with a certain distribution
and regularize the estimates. For example, applying logistic regression on corrupted data set is asymptotically equivalent to add a ridge-like penalty and therefore improve the classification results. Simulation results and a summary of current work are presented.

1.1 Functional Data

1.1.1 Definition

Functional data refers to data that are functions of a continuous variable, which are usually observed as repeated measurements with a complex but relatively smooth pattern [42]. Therefore, the intrinsic structure of the data can be represented as a function. Here, we denote the collection of observations by

$$Y = (Y_1, ..., Y_n) \quad (1.1)$$

Which are stochastically independent with a common distribution. For each discretized data vector

$$Y_i = (Y_i(t_1), ..., Y_i(t_m)), i = 1, ..., n \quad (1.2)$$

The $t_i's$ are not necessarily time points, they could be continuum such as wavelength units, location in space and ages. Often we observe these group of discrete noisy data and assume it’s from a continuous function $m_i(t_j)$ over the continuum $t$ with some random noise $\epsilon_{ij}$:

$$Y_i(t_j) = m_i(t_j) + \epsilon_{ij}, i = 1, ..., n, j = 1, ..., m \quad (1.3)$$

So functional data could be seen as multivariate data, but with ordering on the dimensions. Therefore, functional data analysis methods could making use of derivative
information to enhance the power of the models.

### 1.1.2 Examples of Functional Data

This type of data is often seen in weather data that measuring temperatures over the year; human growth curve that measuring heights of young people over years; music sound signals and electric signals; optical tracking device or spectral measurements etc. The figure below 1.1 are the experimental measurements from a medical spectroscopic device that is used to detect cervical cancer lesions. The x-axis is the emission wavelength in nm, corresponds to the $t'_i$s in equation (1.4). Along the given range of wavelength, the white light reflectance is measured repeatedly for 443 times, as shown in the figure with different colors for 443 curves.

![Figure 1.1: Experimental measurements of a spectroscopic device - for different combinations of factors, white light reflectance is measured across wavelength 360-840nm repeatedly, result in total of 443 curves in the figure.](image)

Functional data analysis is to treat each data function(a curve) as a single entity
instead of a vector of points observed, and then analyze the properties of a group of functions. That is to say, for the example showed in figure 1.1, we can get more information by analyzing the bunch of curves as a group of functions instead of analysis on a single curve or points at a single wavelength. Assuming there exists functions based on the data, functional data analysis expands the finite data dimension to infinity. Therefore, we can evaluate the function at any point in the domain. This means, for the example above, for each curve, we observed 481 intensity values along wavelength 360-840nm for each observation. By fitting a single function on the observed data points, we can evaluate the intensity value at any wavelength in the range. Also, applying to all curves, we get a group of 443 functions.

1.1.3 Estimate the Functional Data

Since functional data is intrinsically with infinite dimensions, functional data analysis methods should not strongly depend on the finite dimensional approximation, it should let the estimation approach the appropriate limiting analogue for true functional observations as the order of approximation \( m \to \infty \) (i.e. Grid Refinement Invariance Principle).

There are two ways to devise such a method, direct and indirect.

• Direct way: Estimate the true functional data, then find (in many cases, interpolate) the finite dimensional approximation.

• Indirect way: Estimate for the finite dimensional data, then show it approaches to a limit when \( m \to \infty \).

Some challenges of estimating the functional data include:

1. Estimation from noisy data with observations on discrete points.
2. Measure of variation in estimates.


### 1.2 Popular Functional Data Analysis Methods

Theory and methods of functional data analysis have been systematically studied by Ramsay, Silverman and Ferraty etc. [39, 38, 18], I will briefly summarize some classical methods below.

**Functional principle Component Analysis and Regression** Principle component analysis (PCA) is used to find the dominant modes of variation in data and how many of these modes are needed to achieve certain level of variation explained from the original data. The functional PCA is defined as the multivariate PCA with integration over the continuum \( t \) instead of summation over the discrete points.

**ANOVA and Functional ANOVA** ANOVA (Analysis of Variance) is a popular method used in experimental data based on completely randomized design, to analyze the differences between group means and their associated procedures. The simplest ANOVA model is a single factor study [25], stated below:

\[
Y_{ij} = \mu_i + \epsilon_{ij} \tag{1.4}
\]

where

- \( Y_{ij} \) is response variable for \( i^{th} \) factor, \( j^{th} \) observation.
- \( \mu_i \) are parameters.
- \( \epsilon_{ij} \) are independent \( N(0, \sigma^2) \)

Therefore, we have \( E[Y_{ij}] = \mu_i \), and \( \sigma^2[Y_{ij}] = \sigma^2[\epsilon_{ij}] = \sigma^2 \).
This ANOVA model could be expressed in matrix form as a linear model:

\[ Y = X\beta + \epsilon \]  

(1.5)

Assume this equation in (1.4) has 3 levels each with 2 observations. Then

\[
Y = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix} \]

(1.6)

ANOVA estimates are the least square estimates in the linear model above and Functional ANOVA are ANOVA when the variables are functions instead of scalars.

**Basis Function and Spline Function Estimations** A typical way of getting functional data is to use nonparametric smoothing methods to represent each observed curve as a functional object, set aside the original data and use the estimated curve instead [26]. One common way is to express the structure by a basis function expansion approximation, say \( Y_i(t) = \sum_{j=1}^{m} \eta_{ij}B_j(t) \).

\( B_j(t) \) is the known basis, with its coefficient for observation \( i \) as \( \eta_{ij} \). There are many ways of choosing the basis functions, such as polynomials, fourier series, powers of \( t \) etc. Then the problem of estimating the functional data becomes multivariate estimation the coefficients \( \eta_{ij} \).

Like the plot in Figure 1.1, it is often the case that the curves are non periodic but complicated. polynomial basis with order \( j \) 4 might not capture the information.
Therefore, a very popular and classical basis functions are first introduced in book by Boor [10]. It is intrinsically a list of polynomials joined along the continuum at a set of points called "knots", so the roughness could be varied along the curve, and it is determined by the number of basis functions and the degree of polynomial used. The coefficient estimate is the relative weight of each basis at the fixed knots, which gives optimal fit of the data.

**Smoothing Penalties**  The basis function above gives a good approximation to functional data, but fitting it by least squares leads to discontinuous control of the degree of smoothing [42]. Therefore, roughness penalty/regularization methods are introduced to overcome such problems. They often represented as

\[
\arg\min_{\{m(t)\}} \sum_{i=1}^{n} [y_i - m(t_i)]^2 + \lambda \int R^2(t)dt
\]

with \( R(t) \) measures roughness of \( m(t) \), \( \lambda \) is the smoothing parameter.

**Other areas and methods**  There are also extensively well studied topics on data with functional response variables, registration/crace alignment problems etc. We won’t be able to go into details about each of these parts here. In this thesis, topics on ANOVA estimates and local polynomial regression as a smoothing technique will be reviewed in more details in later chapters.
Log Additive Model on Repeated Measurements

Currently, we are investigating the spectroscopic device to make in vivo spectroscopy tissue measurement of detect pre-cancer and cancer liaisons in cervix. One problem with this kind of device is that the measurement performance is not stable among devices and other factors. To study the repeatability of the device and assess the non-biological variabilities, our collaborators at Vancouver made a group of experiment measurements using different combinations of devices and standards, under 4 different filter settings. The goal is to learn repeatedly measured data and develop a ”correction factor” model to remove the inter-device variability, and apply the correction factor on real patient measurements later.

2.1 Motivating Data Set

2.1.1 Multispectral Digital Colposcope Device

The repeated experimental measurement data is from the new generation of the optical cervical cancer detection device called ”MDC3” (Multispectral Digital Colposcope), details of the device can be found in [59, 61, 62]. A sample picture is showed
below in Figure 2.1.

Figure 2.1: MDC3 main body parts.

The key part of the device is a point probe, which is a seven-fiber optic cable, i.e. a light-receiving central fiber surrounded by a ring of six illumination fibers. It excites light with a certain wavelength range and collects the reflectance or emission wavelength accordingly [19], which will be sent to the computer for further process and diagnoses.
Figure 2.2: MDC3 key part: point probe, used to excite light and get emission wavelength back into the device.

Figure 2.3: MDC3 structure and optical imaging paths.
Within the MDC3 device, a Xenon arc lamp is installed in the bottom part of the device as the light source. Light is sent to a power meter used to measure the total energy of illumination light delivered to the sample. We use this power meter measurement to normalize the collected data later on. In the upper part of the device as shown in 2.3, light from the lamp goes through an "excitation filter" (at the bottom of Figure 2.3) to give excitation light in a certain wavelength shined on the cervix tissue (here we are making measurements on a standard tile instead of real tissue in the experiment). Then the reflected light is collected and goes through an "emission filter" (on top part of Figure 2.3) back to the device.

The corresponding "excitation filter" and "emission filter" are called "filter sets". We have four different filter sets listed below in Table 2.1. For the white reflectance filter set, a broadband filter is used. For the fluorescence measurements, the optical filter produce narrow band excitation ("Excitation range" in Table 2.1) light to certain range and measured by a spectrometer. The filters are embedded in a rotating wheel, attached with opening and closing shutters. Therefore, one measurement for all four spectra only takes about one second.

Table 2.1: Wavelength ranges for each filter set.

<table>
<thead>
<tr>
<th>Filter set</th>
<th>Description</th>
<th>Excitation range (nm)</th>
<th>Detection range (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter set 1</td>
<td>White reflectance</td>
<td>No filter</td>
<td>410-700</td>
</tr>
<tr>
<td>Filter set 2</td>
<td>Blue fluorescence</td>
<td>455-465</td>
<td>495-550</td>
</tr>
<tr>
<td>Filter set 3</td>
<td>Violet fluorescence</td>
<td>410-420</td>
<td>440-560</td>
</tr>
<tr>
<td>Filter set 4</td>
<td>UV fluorescence</td>
<td>340-360</td>
<td>420-580</td>
</tr>
</tbody>
</table>

For the measurement standards, we are using 4 squared tiles ordered in two batches
to identical specifications. In clinical use, a standard is measured before patient tissue measurement to make standardization and quality assurance.

### 2.1.2 Design of Experiment

The experiment takes 4 ideally identical MDC3 devices, each of them are measured using 4 different filter sets. Within each combination of device and filter set, 4 measuring standards (2 inches squared tiles doped with inorganic blue and green fluorophores) used instead of real tissue sample. This is done so to preclude the biological variability and only study the non-biological variances from the device performance. The goal is to identify the systematic differences among 4 devices and 4 standards.

Due to the nature limit of fluorescence wavelength, different filter sets have different effective wavelength range. In general, we only detect the light intensity measurement in 410 – 700 nm. For each filter set, we use the effective range listed in the last column in Table 2.1.

The measurements were made in a dark room since previous study showed that room light could affect the measurement results. Therefore, for this set of experiment, we call it DoD (Days of Darkness) data. Within each filter set setting, exactly the same measurement steps are followed. There are 4 ideally identical devices used in the experiment, within each device, 4 different standards are applied to take the measurement. Within each combination of device and standard, 3 sessions of experiment were made along the day, within each session, 5 sets of 5 repeated measurement were make. Therefore, we have $4 \times 4 \times 3 \times 5 \times 5 = 1200$ light intensity curves measured for each filter set. After the quality control check, some bad measurements were removed, resulting in 1118 measurements.
After the initial data processes, including background noise reduction, wavelength calibration from pixels to wavelength in nm unit, the quality control selections and effective wavelength range truncation, the dataset is stored as a 4 dimensional data cube, with dimension 4 x 300 x 1118, correspond to 4 filter sets, 300 wavelengths along 401 to 700 nm and 1118 repeated measurements.

2.1.3 Problems and Possible Solution

Currently, the 4 devices have very different performance in measuring the light intensities. This could be easily shown as the log10 values of light intensity measurement plotted below, one color is for one device. Other than the variations of the measurements among devices, different standards show group effects on the measurements and there also exist amplitude fluctuations.
Figure 2.4: Log10 of the light intensity measurements for 4 different filter sets. The color red, green, blue and black correspond to device 1, 2, 3 and 4.

To solve these problems, we developed "correction factors" to calibrate the standard measurements to a predefined gold standard for each filter set. The log10 of the predefined gold standards for 4 filter sets are plotted below.
Figure 2.5: Log10 of the gold standard curves for 4 different filter sets.

Filter set 1 (white reflectance) is a special case since its gold standard is the mean measurement of device 3 standard 0; For filter set 2 to 4, we used NRC(National Research Council of Canada) measurement with excitation wavelength 450,420 and 350 nm, spline interpolated curves as the gold standards.

The details of two approaches of developing the correction factors will be explained in the following sections.
2.2 Heuristic Fitting Approach

We believe the variations among the repeatability experiment measurements could be categorized into variation due to device, variation due to measurement standards and the light intensity amplitude differences among filter sets. Therefore, the heuristic correction factor fitting approach is a stepwise algorithm that calibrate the measurements to gold standards.

2.2.1 Heuristic Correction Algorithm

The algorithm is summarized as below:

1. Power normalization step: divide each measurement curve by its mean intensity value within the given range (refer to the range given in previous section). This step is to reduce the amplitude fluctuation among measured curves.

2. Correction factor step: calculate separately for each combination of device and standard, along the emission wavelength. The resulting correction factors are plotted below.
Figure 2.6: Correction factors of the heuristic approach.
2.2.2 Heuristic Correction Results

After the algorithm described in above, the light intensity measurements become

![Graphs showing corrected measurements](image)

Figure 2.7: Log10 of the corrected measurements. Plotted in color of red, green, blue and black for device 1, 2, 3 and 4, but overlapped and only shown black color in the figure.

The plots are still plotted in different colors with different devices. Because they overlap with each very well, we can only see the last device (black color) in the plot. Heuristic correction algorithm gives a very good results of the light intensity measurements, it corrected the variations among devices and standards. Though it's a good fit in general, we can still see a bit green color (device 2) at the tail area of filter set 3 and filter set 4, this is due to the large variation on the two end of the wavelength ranges. Another concern is that, the correction factors show very different pattern among devices, similar but still different patterns among standards within in each device, along the wavelength. Therefore, fitting a correction model as a function of wavelength is proposed in next section.
2.3 Log Additive Model

2.3.1 The Proposed Model

Inspired by the stepwise correction algorithm above, we propose an additive model with ANOVA estimates on the log10 scale of light intensity measurement. For each filter set, we build a model:

\[ y_{ijk}(t) = \log Y_{ijk}(t) = \log GS(t) + lCFD_i(t) + lCFS_j(t) + lAC_{ijk} + \epsilon_{ijk}(t) \quad (2.1) \]

where

- \(Y_{ijk}(t)\) = measured intensity at wavelength \(t\).
- \(i, j = 1, 2, 3, 4\) for 4 devices and 4 standards respectively.
- \(k = 1, ..., N_{ij}\) refer to the replicates nested within device and standard.
- \(N_{ij}\) = number of repeated measurements for device \(i\) standard \(j\). Note that this is an unbalanced design due to quality control.
- \(t\) = wavelength in manometer (range of \(t\) varies for different filter sets).
- \(lCFD_i(t)\) = log of correction factor based on device \(i\).
- \(lCFS_j(t)\) = log of correction factor based on standard \(j\).
- \(GS(t)\) = Gold Standard of given filter set measurement. This is pre-defined.
- \(lAC_{ijk}\) = log of amplitude correction factor.
- \(\epsilon_{ijk}(t)\) = unexplained variables, with mean 0, variance \(\sigma^2\)

This model is assuming a linear response for the device and it is wavelength dependent with no other terms. The effects of session and set within each session will be discussed later when analyzing amplitude correction factor. Amplitude fluctuation is a multiplicative factor and is assumed to be constant across emission wavelength.
We also take log10 of the data to make the multiplicative relation to additive effect. It reduces the dynamic range and makes the variance more uniform also.

Based on this model, all light intensity measurements could be calibrated to a gold standard with correction factors by device (CFD), correction factors by standard (CFS) and amplitude correction(AC). To estimate these factors, analysis of variance(ANOVA) is used. We will give the details in next section.

### 2.3.2 An alternative to the Gold Standard

In many cases, a gold standard doesn’t exit or not easy to obtain. Therefore, we propose to substitute the $GS(t)$ term to the grand mean $\mu(t)$, which is the average value of the log10 of all measurements at each wavelength [54]. A simple way to estimate each factor is to use ANOVA estimation introduced in Chapter 1. So even if we want to adjust the correction factors to the gold standard later, we just need to subtract off the difference between grand mean measurement and gold standard. Therefore, at each wavelength $t$ in the given range,

$$y_{ijk} = \log Y_{ijk}(t) = \mu(t) + lCFD_i(t) + lCFS_j(t) + lAC_{ijk} + \epsilon_{ijk}(t) \quad (2.2)$$

$$\mu(t) = \frac{1}{N_{ij}} \sum_{ijk} \log Y_{ijk}(t)$$

The ANOVA estimates are the least square means which minimize

$$\sum_{i,j,k,t} [\log Y_{ijk}(t) - \log \mu(t) - CFD_i(t) - CFS_j(t) - AC_{ijk}]^2 \quad (2.3)$$

subject to $\sum_{i=1}^4 CFD_i(t) = 0, \sum_{j=1}^4 CFS_j(t) = 0$. 
Let $\hat{\mu}(t)$ be the average of the log of all measurements at each wavelength, then

$$l\hat{C}FD_i(t) = \frac{\sum_{j=1}^{4} \sum_{k=1}^{N_{ij}} y_{ijk}(t)}{\sum_{j=1}^{4} N_{ij}} - \hat{\mu}(t),$$

(2.4)

$$l\hat{C}FS_j(t) = \frac{\sum_{i=1}^{4} \sum_{k=1}^{N_{ij}} y_{ijk}(t)}{\sum_{i=1}^{4} N_{ij}} - \hat{\mu}(t),$$

$$\hat{\epsilon}_{ijk}^{(1)}(t) = y_{ijk}(t) - \hat{\mu}(t) - l\hat{C}FD_i(t) - l\hat{C}FS_j(t)$$

Therefore, with $n_t$ being the number of emission wavelengths,

$$l\hat{A}C_{ijk} = n_t^{(-1)} \sum_t \hat{\epsilon}_{ijk}^{(1)}(t),$$

(2.5)

$$\epsilon_{ijk}(t) = \hat{\epsilon}_{ijk}^{(1)}(t) - l\hat{A}C_{ijk}$$

These estimates give the factors due to device, standard and aptitude fluctuation precisely. $AC_{ij}$ is taking into account of the interaction effects between device and standard, and also the average of random errors across wavelength. Therefore, the corrected spectra is given by

$$\log Y_{ijk}^{CF}(t) = \log Y_{ijk}(t) - l\hat{C}FD_i(t) - l\hat{C}FS_j(t) - l\hat{A}C_{ijk}$$

(2.6)

### 2.3.3 Results

We present the data and correction factors estimates by each filter set. Starting with filter set 1, the log10 scaled data is plotted followed by residual plot and measurement plots after applying the correction factors. The lower 4 plots in Figure 2.8 are the gold standard (grand mean in this case), $CFD(t)$, $CFS(t)$ and $AC$ estimates respectively.
Figure 2.8: Fitted model, residual checking, correction factors and correction results.
Figure 2.9: Fitted model, residual checking, correction factors and correction results.
Figure 2.10: Fitted model, residual checking, correction factors and correction results.
Figure 2.11: Fitted model, residual checking, correction factors and correction results.
The device and standard factors explained most of the variations in the measurements and also show some overall patterns. $CFD_i(t)$ oscillate along the emission wavelength for all filter sets, it is likely due to the slight differences in lamp spectra and the spectral transmission of the components, such as the interference coatings on the hot mirror, excitation-emission filters, anti-reflection coatings on the projection and collimation lens elements. However, $CFD_2(t)$ has a different trend, after checking with the engineers and clinicians, we found this is due to the backward position of one excitation projection lens in device 2.

For standard correction factors, due to the batch to batch difference (ordered standard 1 and 2 in one batch, 3 and 4 in the other), we see the first batch standard tiles have higher levels of fluorophores, i.e. higher values in $CFS_1(t)$ and $CFS_2(t)$.

In Table 2.2, we give the average across wavelength of the percentage of variances explained by each factor for each filter set. This is showing that the correction factors are effectively removing variations.

**Table 2.2: Percentage of variance explained by each factor.**

<table>
<thead>
<tr>
<th>% of variation</th>
<th>Filter set 1</th>
<th>Filter set 2</th>
<th>Filter set 3</th>
<th>Filter set 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFD</td>
<td>96.53%</td>
<td>34.69%</td>
<td>48.64%</td>
<td>70.40%</td>
</tr>
<tr>
<td>CFS</td>
<td>0.27%</td>
<td>64.50%</td>
<td>50.63%</td>
<td>27.28%</td>
</tr>
<tr>
<td>AC</td>
<td>3.15%</td>
<td>1.23%</td>
<td>1.10%</td>
<td>2.11%</td>
</tr>
<tr>
<td>Unexplained</td>
<td>0.07%</td>
<td>0.29%</td>
<td>0.10%</td>
<td>0.41%</td>
</tr>
</tbody>
</table>

It’s worth to mention that the amplitude correction factors are very similar across all filter sets, but show random patterns across sessions of the day and sets of measurements. In Figure 2.12, we showed a few device and standard combinations to
show some typical patterns. We believe these fluctuations are due to the fluctuation in arc lamp output and lamp position over time. Therefore, these fluctuations suggest that these estimates can’t be used to make any consistent corrections to patient measurements.

Figure 2.12: $AC_{ijk} = \exp(lAC_{ijk})$ estimates for filter set 1. Red lines separate the sessions. 5 colors indicate sets of 5 measurements. (Missing values are left blank)

### 2.4 Discussion and Future Works

#### 2.4.1 Implementation of the Model

The log additive model above is designed to compute the correction factors by device, standards and adjust amplitude fluctuation among each others. In real time patient measurement, the device will be used on a standard right before measuring the cervix tissue, calibrate the output according to our correction factors and on site measurement. Our log additive model provides a unified way to estimate the correction factors along with accuracy assessments, which are confidence intervals (For this part, please refer to [54]). Currently, we have been using the device correction factors
on the new patient measurements made this year, specifically, multiply the patient measurements by $10^{-lCFD}$ to obtain the device corrected spectrum.

Another important application of this model is quality control. We recommend a procedure as follows: After taking the new measurement, take log10 transform, subtract device and standard correction factors and grand mean, if the resulting value is outside of a certain range, we label it as an outlier and suggest repeat the measurement. This confidence interval could be calculated as six-sigma [27] interval for $AC$ estimates.

One problem here is, when the device is used in real time, only one standard corresponds to one device, so how do we incorporate the device and standard iteration effects need to be addressed. Also, necessity of updating the correction factors according to the onsite standard measurement need to be studied with the data from a new experiment. If it’s necessary to update the factors involving time effects, what method to use is the next question.

2.4.2 Discussion on Model Fitting

There are several potential problems with the model and its estimations. One is that the residual plots show a slightly bigger variance at the two end and some valley area. Since the variance of measurement differs along wavelength t, estimation of variance as a function of t for each device could be interesting to explore also. With a good estimation of variance matrix, a more reliable inference on correction factors could be given also, for example, a wavelength dependent weight to refine the model fitting with a two-stage algorithm.
The next step is to implement smoothing function on $CFD_i(t)$ and $CFS_j(t)$. B-spline function would be a good choice here. However, the number of knots and knots location of B-spline functions could be a problem due to the shape of the correction factors. One thought is to use a backward fitting algorithm. Starting with knots on all wavelength units, at each step, drop one knot that gives the best predictability comparing with dropping any other knots. However, this approach might be only valid with constant variance across wavelength, which is not the case in our correction factors. Our correction factors tend to have bigger variance at the two end area and valley between two peaks, so the variance is a function with wavelength $t$.

Another approach we could try is, use the very raw measurement of light intensities, without any processes. That is to say, use the measured curve along pixel numbers instead of the wavelength in nm unit, and leave the background noise in, then apply the smoothing techniques over it first. This might give a more accurate estimates instead of using the processed data we are using now.
Chapter 3

Local Polynomial Regression for Peak Finding

Algorithm

In previous chapter, I studied different sources of variabilities of the standard measurements and built a log additive model to estimate the corresponding correction factors, this is one major problem with the biomedical applications of optical spectroscopy. In this chapter, I will be focused on another aspect, the accuracy of the curve peak estimates.

Peak finding/seeking local maxima is a common problem in spectroscopy and other areas. In our device development experiment, the measured spectra are functional data with variations due to different devices, measurement conditions, excitation wavelengths and some operational effects. The observed curve peak locations may be shifted by contamination fluorescence. To estimate the true spectral peak location and height, we studied performance of local polynomial regression with different bandwidth, derived asymptotic optimal bandwidths for both location and height estimates, and also proposed a plug-in bandwidth selection and peak estimation method. Simulation study and real data application are presented also.
3.1 Background

Every year, there are about 13,500 cases of invasive carcinoma of the cervix diagnosed in the U.S. Cervical cancer is highly curable if detected and treated early. However, for the normal cervical cells, it usually takes several years to change into cancer cells, and pre-cancer or early cancer status usually show no symptoms or signs, but can be detected by having regular Pap tests. If the Pap test is abnormal, one of the most popular further diagnosis options is using a colposcopy spectroscopy device.

We are analyzing the spectroscopic device that can detect cervical cancer automatically. In clinical, it can be used to replace biopsies and facilitate diagnosis and treatment in a single visit. There has been studies that confirm the optical device can potentially diagnose tissue condition based on the molecular and morphologic changes associated with pre-cancer or cancer status. This device is a small, portable and inexpensive optical sensors bounded with computers to automate analysis of signals. The details of the device can be found here [30].

The data analyzed is from an experiment made on the device version called “FastEEM3”. Below is the picture of the whole device and the wand (a point probe) used to excite and collect light intensity data.
Figure 3.1: FastEEM3 device and its key part to excite light and collect reflectance.

Unlike the experiment measuring some standard tiles in previous chapter, the experiment here was conducted by making measurement repeatedly on a cuvette contains Rhodamine solution (in Figure 3.2). This time, we only use one Rhodamine reference standard, but with the point probe in different conditions (dry probe, tipped wet probe and submerge the probe in ultra filtered de-ionized water during measurement).
Figure 3.2: For both dry and wet measurements, Rhodamine standard is placed in the standards try in (a) with an exposed cuvette. For submerged measurements, the standard is placed in the microspheres holder, shown in (b) with exposed cuvette, and (c) with concealed cuvette.

3.2 Dataset and Peak Problems

The measurements are made repeatedly under each point probe condition (dry, wet and submerged), with 3 time period along the day (marked as set 1,2,3). Within each set, 5 blocks of 5 repeats were measured one after the other. The data is unbalanced due to 1 more block of measurements made on wet point probe in set 3. Therefore, the data is a set of $5 \times 5 \times 3 \times 3 + 5 = 230$ Rhodamine solution reflectance intensity curves. This set of data is measured under different excitation wavelengths, which
are 330, 360, 390, 420, 450 and 480 nm.

Figure 3.3: Repeated light intensity measurements are plotted under each excitation wavelength of 330, 360, 390, 420, 450 and 480 nm respectively. X-axis is the emission wavelength in nm unit. Y-axis is the light intensity values.
Figure 3.4: Rhodamine emission light intensity measurement. The left peaks of the curves are contaminations, which could affect the true peak measurement curve on the right.

In theory, Rhodamine has a single fixed emission wavelength, shows as a single peak in light intensity plot. While our plots show there are obviously contaminations peaks before and after the true Rhodamine peaks. These contaminations could shift the true Rhodamine peaks in the measurement (This could be easily understand by the plot above 3.4). Therefore, we summarized the peak location and peak height quantities in the table below.
Table 3.1: Peak key quantities summary: Mean peak height, peak location and standard deviation of peak location within each excitation wavelength.

<table>
<thead>
<tr>
<th>Excitation wavelength</th>
<th>Peak height</th>
<th>Peak location</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>330nm</td>
<td>7855.3</td>
<td>579.69</td>
<td>0.7161</td>
</tr>
<tr>
<td>360nm</td>
<td>22894</td>
<td>579.76</td>
<td>0.48879</td>
</tr>
<tr>
<td>390nm</td>
<td>4315.1</td>
<td>579.67</td>
<td>0.94916</td>
</tr>
<tr>
<td>420nm</td>
<td>36141</td>
<td>579.68</td>
<td>0.49756</td>
</tr>
<tr>
<td>450nm</td>
<td>22985</td>
<td>579.67</td>
<td>0.6224</td>
</tr>
<tr>
<td>480 nm</td>
<td>32309</td>
<td>579.6</td>
<td>0.51988</td>
</tr>
</tbody>
</table>

In terms of the peak locations, the mean values are very close among excitation wavelength, but the standard deviation are relatively large. For example, excitation 330nm has standard deviation of 0.7161, this indicates a peak location shift in range of around 3-4 nm. Also, the table above shows an obvious shift of peak height among different excitation wavelength. A close look at it for excitation 330nm is plotted below. (In theory, the height across devices for fully processed data should be the same.)
Figure 3.5: This is a plot of peak height values for excitation 330nm. Vertical lines separated the data as the 3 sets in the measurement. 5 blocks of 5 measurements effects could be seen in some conditions.

The spread of peaks depend on quality of devices and other factors that we might not be able to change for now. Our goal is to identify both the curve peak location and height accurately. From Chapter 2, we know the systematic variations by devices could be removed, however, with random noise on infinite support exists, as number of points on the curve increase, peak height (maximum point) goes to infinity, peak location converges but could be very slow.

In next section, I will explain why we use nonparametric regression on peak estimation and review some existing kernel methods, then we will focus on study of the local polynomial regression. Specifically we derived the asymptotic optimal band-widths for peak location and height estimates separately, and gives a data-driven
practical bandwidth selection algorithm. But before that, let’s first define the peak location and height.

The curve peak is at point $\theta$ and is defined as

$$m'(\theta) = 0, m''(\theta) < 0$$

Peak height $m(\theta)$ is the maximum value of the curve and estimated peak location $\theta_n$ of $m_n(x)$ satisfies

$$m'_n(\theta_n) = 0, m''_n(\theta_n) < 0$$

By the Taylor series of the first order derivative, we have

$$0 = m'(\hat{\theta}) \approx (\hat{\theta} - \theta)m''(\hat{\theta})$$

(3.1)

According to Parzen 1962 [36], based on Equation 3.1, our sample peak location estimate is asymptotically Normal. Details of the limit distributions are described in Muller 1985 [32] which we will discuss later in this chapter.

### 3.3 Literature Review on Kernel Methods

A lot of literature has been expanded in peak finding/bump hunting algorithms, such as [20, 60, 23], these algorithms are focused on identifying the true signal peaks when the data show multiple peaks and some of them are noises, and applied in flow cytometry data, mass spectrometry data etc.

However, our data of interest is a group of single peaked curves, so regression
model is considered. Parametric regression methods fit the data with predetermined family of functions, which are overly-restrictive for some data, especially when the data is highly variable and with different source of noises. Also, higher orders \( p > 4 \) could lead to unstable numerical results. Since we are focused on single peak curve estimation, and our primary goal is to estimate the peak location and height accurately for a given noisy data, nonparametric regression methods with some smoothing techniques around peak area are considered. In general, they could be categorized into 3 groups: kernel methods, spline methods and wavelets. Here we focus on kernel methods, especially using local polynomial regressions. Before we go any further, let’s see some notations first:

Assuming \((X_i, Y_i)\) are ordered pair data, and the model is

\[
Y_i = m(X_i) + \sigma(X_i)\epsilon_i
\]

with \( \epsilon_i \sim N(0,1) \), \( \sigma^2(X_i) \) is the variance of \( Y_i \) at data point \( X_i \), \( X_i \) is from a design distribution \( f() \). Here we assume \( X_i \) is uniformly distributed and homoskedastic (i.e. \( \sigma^2(X_i) = \sigma^2 \) for all \( X_i \)), but these assumption could be relaxed later. \( X \) and \( \epsilon \) are independent.

Let the conditional mean

\[
m(x) = E(Y|X = x)
\]

with \( m(x) \) is not restricted to any specific parametric family, such as a polynomial function. We usually assume the variance is constant \( Var(Y|X) = \sigma^2 \) and denote \( \nu^{th} \) derivative of \( m(x) \) as \( m^{(\nu)}(x) \). Below is a brief summary of some classical nonparamet-
ric regression methods for univariate case, the multivariate case could be generalized easily.

Kernel estimators are biased, when the mean function $m(x)$ is continuous and sufficiently smooth, they are consistent. These estimators pick up features $x$ in the neighborhood of data point $x_0$ to estimate $m(x_0)$ by assigning a positive weight by kernel function $K()$. Below is a list of well known kernel estimators.

### 3.3.1 Nadaraya-Watson kernel regression estimator

In 1964, Nadaraya [35] and Watson [57] simultaneously but independently proposed a method to overcome the problems with parametric regression. They propose to take weighted average around data points $x$ with the weight function be a symmetric probability density function, called kernel function.

$$
\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K_h(x_i - x) y_i}{\sum_{i=1}^{n} K_h(x_i - x)}
$$

(3.3)

Assume bandwidth $h > 0$, $K_h(x) = K(x/h)/h$. When bandwidth $h$ increases, the smoothness of the estimator increases. The uniform convergence of this estimator and its derivatives is given in [45].

### 3.3.2 Priestley-Chao kernel estimator

In 1972, Priestley and Chao [37] discussed non-parametric function fitting and proposed another consistent kernel estimator as below:

$$
\hat{m}_{PC}(x) = \sum_{i=1}^{n} (x_i - x_{i-1}) K_h(x_i - x) y_i
$$

(3.4)
3.3.3 Gasser-Muller kernel estimator

In 1984, Gasser and Muller [21] proposed a similar but more refined estimator.

\[
\hat{m}_{GM}(x) = \sum_{i=1}^{n} y_i \int_{s_{i-1}}^{s_i} K_h(x_i - u) du
\]

With \( s_i = \frac{x_i + x_{i+1}}{2}, i = 1, 2, ..., n - 2, x_0 = -\infty, x_{n+1} = \infty. \)

This estimator was proposed for equally spaced \( X_i \), but it can be used for non equally spaced design, details of its asymptotic properties and random design can be found in [33, 29, 5].

3.3.4 Local polynomial regression estimator

The three estimators above only give constant fit over certain interval, i.e. they could all be represented in the form of

\[
\hat{m}(x) = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}
\]

with \( w_i = K_h(x_i - x) \) for Nadaraya-Watson estimator and \( w_i = \int_{s_{i-1}}^{s_i} K_h(u - x) du \) for Gasser-Muller estimator. Therefore, they may not be sufficient and accurately estimating the mean function. So a more accurate estimator is needed. To outperform the local constant estimation, a local linear fit is proposed by Stone in 1977 [48], and later generalized it to local polynomial fit and studied the asymptotic properties in Stone 1980 and 1982 [49, 50], also the multivariate case by Ruppert and Wand in 1994 [44]. Based on local polynomial regression concept, several fitting algorithms are proposed, such as a robustness fitting algorithm by Cleveland in 1979 [6], locally weighted regression (loess) by Cleveland and Delvin in 1988 [7].
Local polynomial regression is modeling mean function $m(x_0)$ locally by a polynomial model using Taylor series expansion, and the whole curve $\hat{m}(x)$ is estimated by running the local regression with $x_0$ varies along the estimation domain. i.e. assuming there exists $(p + 1)^{th}$ derivative of the regression function $m(x)$ at point $x_0$, the Taylor expansion for $x$ in the neighborhood of $x_0$ is given by

$$m(x) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x_0)}{j!}(x - x_0)^j = \sum_{j=0}^{p} \beta_j(x - x_0)^j$$

(3.6)

$\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ is solved by minimizing the weighted least squares

$$\sum_{i=1}^{n} [Y_i - \sum_{j=0}^{p}(x_i - x_0)^jK_h(x_i - x_0)]^2K_h(x_i - x_0)$$

(3.7)

where $K()$ is the kernel function assigning weights to each data point and $h$ is the bandwidth.

Local Polynomial regression outperforms the other kernel estimators in many aspects. In Fan 1993 [12], they showed that local linear model optimizes linear minimax risk, and later generalized this property to local polynomial case ($p > 1$) and derivative estimates ($\nu > 0$) in Fan 1997 [13]. Another advantage of local polynomial regression is that it has good performance at the boundary points while some nonparametric estimators have different behaviors at the boundaries compared with those on the interior. Ruppert and Wand 1994 [44] showed the convergence rates are the same at the boundaries and interior points, but in practice, the conditional variance is larger at the boundaries. However, we are primarily interested in the peak area, which is usually on the interior, so we won’t discuss further about the boundary effects in this thesis.
Denote the solution to above formula as $\hat{\beta}_i$ with $j = 0, ..., p$, from Taylor expansion in (3.1), we have

$$\hat{m}_\nu(x_0) = \nu! \hat{\beta}_\nu$$

as the estimators for $m^{(\nu)}(x_0), \nu = 0, ..., p$. In matrix form, this is to find $\beta$ that minimize

$$(y - X\beta)^T W (y - X\beta) \tag{3.8}$$

With $W = \text{diag}(K_h(X_i = x_0))$, $y = (Y_1, ..., Y_n)^T$, $\beta = (\beta_0, ..., \beta_p)^T$ and

$$X = \begin{pmatrix}
1 & (X_i - x_0) & \cdots & (X_i - x_0)^p \\
\vdots & \vdots & \ddots & \vdots \\
1 & (X_n - x_0) & \cdots & (X_n - x_0)^p
\end{pmatrix}$$

Then the weighted least square estimate $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_p)^T$ is given by

$$\hat{\beta} = (X^T W X)^{-1} X^T W y \tag{3.9}$$

So the local polynomial regression estimator for mean function $m(x)$ is the intercept parameter estimate

$$\hat{m}_{LP}(x) = \hat{\beta}_0 \text{ estimated in the neighborhood of data point } x \tag{3.10}$$

It is straight forward to derive the conditional mean and variance of $\hat{\beta}$ as

$$E(\hat{\beta}|X) = (X^T W X)^{-1} X^T W m = \beta + (X^T W X)^{-1} X^T W r \tag{3.11}$$

$$\text{Var}(\hat{\beta}|X) = (X^T W X)^{-1}(X^T \Sigma X)(X^T W X)^{-1} \tag{3.12}$$
With vector of unknown functions $\mathbf{m} = (m(X_1), ..., m(X_n))^T$, polynomial coefficients $\beta = (m(x_0), ..., m^{(p)}(x_0)/p!)^T$, residuals of the local polynomial estimation $\mathbf{r} = \mathbf{m} - \mathbf{X}\beta$ and $\Sigma = diag(K^2_n(X_i - x_0)\sigma^2(X_i))$.

The quality of the estimates are depend on the choices of kernel function $K$, the bandwidth $h$ and order of polynomial function $p$ [1].

**Choice of Kernel Function $K()$** In this thesis, we assume $K$ is a symmetric unimodal probability density function with bounded support, and there exists certain number of moments. According to [1], most of the popular kernel functions can be represented in the form of

$$K(x, q) = \left[2^{2q+1}B(q+1, q+1)\right]^{-1}(1 - x^2)^qI_{|x|<1} \quad (3.13)$$

With $B(a, b)$ being the beta function. For $q = 0, 1, 2, 3$, these kernel functions are uniform, Epanechnikov, biweight nad triweight kernels. The Epanechnikov kernel [11] $K(x) = \frac{3}{4}(1 - x^2)_+$ is optimal in terms of asymptotic mean square error, therefore it is widely used in literature. In some cases, Wand and Jones 1994 [53] pointed out that the discontinuous $1^{st}$ derivative of Epanechnikov kernel is undesirable, so a standard normal kernel $K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ is used also.
However, it is worth to mention that the difference between these kernel functions is small, it doesn’t have critical impact on the fitting results. Therefore, the choice of kernel functions is not discussed further. In this thesis, we use Epanechnikov kernel most of the time.

**Figure 3.6: Epanechnikov kernel**

Choice of polynomial order  \( p \)  The order of polynomial regression function is important in estimation results. There is a trade-off between bias and variance: For larger \( p \), the polynomial gives a precise fitting with small bias but large variance; For small \( p \), the fitting could be overly smoothed with a large bias and small variance. This trade-off relation is not linear: the asymptotic variance of \( \hat{m}(x) \) is the same for \( p = 0 \) and \( p = 1 \) while asymptotic bias decreases when \( p \) increases; The asymptotic
variance only increases when $p = 1$ goes to $p = 2$. So Fan and Gijbels 1995 [16] suggested consider odd order polynomial regressions only and $p - \nu$ odd cases later [17].

In Fan and Gijbels 1995a [15], they proposed an adaptive method to choose polynomial order $p$ differently for different data points. The algorithm is robust to bandwidth, i.e. when the bandwidth is large, a higher order polynomial is chosen; When the bandwidth is small, a lower order polynomial is chosen to make the estimate more stable. In practice, their algorithm chooses $p = 1$ most of the time except the high curvature shaped area. In Cleveland and Loader 1996 [8], they proposed similar idea of modeling data with different polynomial degree for different data and called it “polynomial mixing”, using cross validation to choose mixing degree and $C_p$ to choose bandwidth at each data point $x$.

The choice of polynomial order $p$ is very important, but to a limited extent. Higher order polynomials (usually for $p > 4$) could lead to highly unstable, over-fitted results. This is also another reason of why we study the bandwidth selection instead of polynomial degree selection in this thesis.

**Choice of bandwidth $h$** The choice of bandwidth $h$ is critical for local polynomial regression estimation since it controls the complexity of the model. When $h = 0$, all kernel estimates are equivalent to interpolating data points, which could seen as the most complex model. When $h$ increases, the fitted model is getting more and more smooth, with the variance reduces and bias increases. And when $h \to \infty$, local polynomial regression is just to fit one polynomial model globally, which is the simplest case, with peaks heights been under estimated and valleys being over estimated.
Over the past decades, a lot of effort has been put on to find the data-driven/data-dependent bandwidth for estimate the mean function or its derivatives by minimizing the mean squared error (MSE) or other similar loss function, such as residual squares criterion (RSC in [17]).

Based on the optimization criterion, the bandwidth selection approaches could be classified into two categories: global bandwidth and variable bandwidth. Global bandwidth gives a constant best bandwidth for the whole data domain, it usually minimizes mean integrated squared error (MISE), with unknown terms estimated by plug-in methods or cross-validation methods. In 1992, Fan and Gijbels [14] proposed a leave-one-out cross-validation approach to estimate bandwidth \( h \), and later in 2002, Xia and Li [58] added in a weight function to reduce boundary effects. Another approach is to minimize the asymptotic MSE instead of conditional MISE. In 1995, Ruppert, Sheather and Wand [43] proposed three plug-in estimators: rule of thumb \( \hat{h}_{ROT} \), direct plug-in \( \hat{h}_{DPI} \) and solve the equation \( \hat{h}_{STE} \). \( \hat{h}_{ROT} \) is the simplest, it estimates the mean function by dividing the data range into intervals and fit quartic (\( p = 4 \)) regression functions, it is an inconsistent estimators. \( \hat{h}_{DPI} \) uses \( \hat{h}_{ROT} \) results as a pilot estimates and plug-in the optimal bandwidth formula. \( \hat{h}_{STE} \) is the most complicated one, by solving a system of equations derived with estimates from above two bandwidths. \( \hat{h}_{DPI} \) and \( \hat{h}_{STE} \) are converged to optimal MISE bandwidth and all three estimators perform well in simulation study, with the later two have very small differences in the results.

In order to capture complicated shapes of the curves, more flexible model could be achieved using a variable bandwidth, which means the optimal bandwidth depends
on data point $x$. It uses smaller bandwidths at the area with complicated shapes and larger bandwidths when the curve is smoother or more linear. According to Fan and Gijbels 1996 [17], they proposed a variable bandwidth selection procedure as split up the data domain into subintervals $I_k$, and find bandwidth $\hat{h}_k$ minimizes integrated residual squared criterion (IRSC) within each subinterval, resulting in a bandwidth step function, and then smooth it by taking the average locally using length of $I_k$, then use the resulting bandwidth function $\hat{h}(x)$ to fit polynomial functions of order $p$ for each data point $x$ and lead to a smooth estimated curve. Usually, splitting up the subintervals to $[n/10logn]$ pieces gives satisfactory results. They also proposed a two-stage refined bandwidth selection procedure to improve the convergence rate in their book [17] chapter 4.

Global and variable bandwidths are computed and compared with their asymptotic and finite sample properties in Fan and Gijbels 1996 [17]. However, since asymptotically odd-order polynomial fits are superior to even case, their book only considered estimates when $p - \nu$ is odd, but due to the nature of our application on peak location and height estimates, we consider cases with any polynomial order ($p$ from 1 to 4) and any order of derivative estimates ($\hat{m}_v(x_0) up to \nu = p + 2$)

### 3.4 Local Quadratic Fitting Algorithm

Back to our Rhodamine measurement data showed in Figure 3.4, the curve around true Rhodamine peak area shows a quadratic shape, therefore, it’s a good start to assume the peak area curve is symmetric around the peak location, and fit a quadratic model below:

$$y_{ijk} = C_{ij} f(x_{ijk}) + \epsilon_{ijk}, \epsilon_{ij} \sim N(0, 1) \quad (3.14)$$
where

\[ i \sim \text{each excitation wavelength}, \]

\[ j \sim \text{repeated measurements} \]

\[ k \sim \text{pixels 1:1024} \]

Assuming the true Rhodamine curve is

\[ f(x_{ijk}) = -(x_{ijk} - x_{0ij})^2 + \frac{a_{ij}}{c_{ij}} \]  

(3.15)

where

\[ a_{ij} = \text{peak heights} \]

\[ c_{ij} = \text{precision/ sharpness of the curves} \]

\[ x_{0ij} = \text{peak locations} \]

Therefore, we established an algorithm as fitting the local quadratic model iteratively, with the fitting window shifted according to previous step of peak location estimates. Which is:

1. Initial step:
   - Use global maximum as window center.
   - Fit quadratic model with window width = 15.
     (window range: center +/- 7 pixels)

2. Moving fitting window:
   - Use estimated peak location as the new center.
   - Adjust window width (Increase for noisy data).
<table>
<thead>
<tr>
<th>Test</th>
<th>MANOVA</th>
<th>Height\textsubscript{ANOVA}</th>
<th>Location\textsubscript{ANOVA}</th>
<th>Width\textsubscript{ANOVA}</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-values</td>
<td>0</td>
<td>0</td>
<td>0.774</td>
<td>0.5536</td>
</tr>
<tr>
<td>Permutation Test</td>
<td>Height</td>
<td>Location</td>
<td>Width</td>
<td></td>
</tr>
<tr>
<td>p-values</td>
<td>0</td>
<td>0.767</td>
<td>0.549</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: MANOVA and permutation test result comparison.

3. Local quadratic model fitting in the new window.

4. Repeat above two steps until estimated peak locations converge.

### 3.4.1 Model Assessment

In engineers point of view, it is often of interests to evaluate the peak using 3 key features: peak height, peak location and half width of the peak. Therefore, we assess the model in these 3 key quantities using Multivariate ANOVA and permutation test.

The simulation study is designed with a true curve function

$$f(x_{ij}) = -(x_{ij} - x_0)^2 + h_{ij} \quad (3.16)$$

We generate the observations as

$$y_{ij} = C_{ij} \times f(x_{ij}) + \epsilon_{ij} \quad (3.17)$$

Add random noise with 5 different levels on $C_{ij}$

Apply moving local quadratic fitting algorithm

Analysis on peak height, location, width vs. 5 groups

As the way we designed this simulation, peak heights are different across 5 groups,
while peak locations and width are the same. We can see the ANOVA Wilk’s test table and permutation test have agreed result. So we see the assumptions on ANOVA test are valid, the type I error is good, therefore we can trust the ANOVA on device data later.

### 3.4.2 Model Application on Repeatability Experiment

The structure of Rhodamine intensity measurements are:

- **3 Devices:** BCCA, MDA, LBJ
- **6 excitation Wavelengths:** 330:30:480
- **3 Measurement conditions:** Wet, Dry, Submerged
- **3 Measurement Sessions:** Morning, Noon, Evening
- **5 Blocks of 5 measurements within each group:** (Probe is removed and then reinserted between blocks)

After applying on the moving window local quadratic fitting algorithm, we got estimates of peak locations, heights and half peak widths estimates for all the curves.

The next step is to apply Multivariate ANOVA on the estimates across the possible factors listed below:

- peak height vs. devices
- peak locations vs. excitation wavelengths
- peak half width vs. measurement conditions
- measurement sessions

The conclusion for each key feature estimates are:

- **ANOVA for peak height:** significantly different under all scenarios.
- **ANOVA for peak location:** same across some interaction terms (WL*Sess, WL*Sess*Device).
• ANOVA for peak width: same across almost all interaction terms.

Here is boxplots for 3 key features from excitation wavelength = 360 nm.

![Boxplots for 3 key features](image)

**Figure 3.7:** Boxplot of 3 key features in excitation 360nm. Results are similar across different excitation wavelengths.

From the above plot, we can see reproducibility is poor for height. Now let's have a close look at how height varies within each device. (break down plots (1,2) and (1,3) in 3.7)
Figure 3.8: For excitation wavelength = 360nm, BCCA vs. MDA vs. LBJ.

The results show that the peak locations are repeatable, but peak heights are highly non-reproducible. Therefore, use of peak heights on calibrating patient data need to be re-considered.

### 3.4.3 Discussion

This is a simple and fast way to estimate peak location and height by smoothing with local polynomial concept. It takes only 2-3 steps to get the peak location estimates converged. This algorithm only does 2-3 local regressions to get the fitted results,
it is much faster than the traditional local polynomial regression which does n times local regressions along the estimation domain, n is the number of points on the estimation grid).

Here, the window width is chosen by experience, which is critical in estimation accuracy and efficiency of local polynomial regression. Therefore, we focused on the study of window width/bandwidth for the next part. Also, there are some other aspects we could work on in the future, such as modeling with linear contamination term, registration problem in functional data analysis.

3.5 Bandwidth Selection for Local Polynomial Estimates of Peak Location and Height

In previous part, we reviewed some popular and well established nonparametric methods for estimating a regression function. Now we focus on explore the details of bandwidth effects in local polynomial regression, a generalized version of the local quadratic fitting in previous section and its application to peak location and height estimates.

3.5.1 Peak Location and Height Estimators

Though various bandwidth selection techniques have been expanded in literature, almost all of them are focused on the curve estimation instead of a particular point. Our primary interest of using the local polynomial fitting is to identify and estimate the peak location and height accurately and separately with different bandwidths. In 1985, Muller [32] discussed about location and size estimates at zeros and extrema. He
studied the rate of convergence and asymptotic distributions of the empirical extrema location and size (namely the peak location and height) estimates using Gesser-Muller kernel estimator, and proved that the joint distribution of location and size is asymptotically Normal and uncorrelated. In 1989, Muller [34] also show that consistent estimates of the optimal bandwidths for his peak location and height estimators exist and proposed a practical method lead to the consistent and asymptotically efficient estimates.

We borrow the definition of peak location estimates from Muller 1985 [32], where the peak estimates using bandwidth $h_1$ in domain range $I$ (they use $I = (0,1)$) is given by

$$\hat{\theta}(h_1) = \inf \{ x \in I : \hat{m}(x, h_1) = \max_x \hat{m}(x, h_1) \}$$ (3.18)

They also claimed and proved that $\hat{\theta}(h_1) = \hat{\theta}_1(h_1)$ with

$$\hat{\theta}_1(h_1) = \inf \{ x \in I : \hat{m}_1(x, h_1) = 0 \}$$ (3.19)

In practice, we use $\hat{\theta}_1(h_1)$ as location estimator. In order to derive the asymptotic properties, Muller 1985 [32] Theorem 3.1 implies that we could use the relationship below directly.

$$\hat{\theta} - \theta \approx \frac{m'(\theta) - \hat{m}_1(\theta)}{m''(\theta)}$$ (3.20)

For peak height estimates, we define it as the maximum of the curve estimates instead of the estimate at the estimated peak location, using different bandwidth $h_2$.

$$\hat{m}(\theta, h_2) = \sup \{ x \in I : \hat{m}(x, h_2) \}$$ (3.21)
Our work is motivated by Muller’s approach, but due to the advantages of local polynomial regression mentioned above, we use local polynomial instead of Gesser-Muller kernel estimator. So first we derive the asymptotic MSE of local polynomial regression based on our setting. Throughout this thesis, our regression model only considers equally spaced fixed design case, this assumption could be relaxed with the design distribution \( f(x) \) added in Bias terms. Further, we assume the true function is a single peak curve, and variance \( \sigma^2 \) is constant cross the data range. In Fan and Gijbels 1996 [17], Theorem 3.1 gives the asymptotic bias and variance for \( \hat{m}_p(x) \), so we derive the mean function and its 1st derivative estimates as follows:

Denote estimate of the mean function \( m(x) \) as \( \hat{m}(x) = \hat{\beta}_0 \) and estimate of 1st derivative \( m'(x) \) as \( \hat{m}_1(x) = \hat{\beta}_1 \) in Equation (3.9). Assume equidistant fixed design, \( m^{(p+1)}(x) \) is continuous in the neighborhood of peak area. When \( h \to 0 \), we have \( nh \to \infty \). Then the asymptotic conditional variance and bias of \( \hat{m}(x) \) are:

\[
Var[\hat{m}(x)|X] = e_1^T S^{-1} S^* S^{-1} e_1 \frac{\sigma^2}{nh} + o_p\left(\frac{1}{nh}\right) \quad (3.22)
\]

\[
Bias[\hat{m}(x)|X] = e_1^T S^{-1} c_p m^{(p+1)}(x) h^{p+1} + o_p(h^{p+1}), \text{ If } p \text{ is odd} \quad (3.23)
\]

\[
= e_1^T S^{-1} \tilde{c}_p m^{(p+2)}(x) h^{p+2} + o_p(h^{p+2}), \text{ If } p \text{ is even} \quad (3.24)
\]

Similarly, the asymptotic condition variance and bias of \( \hat{m}_1(x) \) are:

\[
Var[\hat{m}_1(x)|X] = e_2^T S^{-1} S^* S^{-1} e_2 \frac{\sigma^2}{nh^3} + o_p\left(\frac{1}{nh^3}\right) \quad (3.25)
\]
\[ \text{Bias}[\hat{m}_1(x)|X] = e_T S^{-1} \tilde{c}_p \frac{m^{(p+2)}(x)}{(p+2)!} h^{p+1} + o_p(h^{p+1}), \text{ If } p \text{ is odd} \quad (3.26) \]
\[ = e_T S^{-1} \tilde{c}_p \frac{m^{(p+1)}(x)}{(p+1)!} h^p + o_p(h^p), \text{ If } p \text{ is even} \quad (3.27) \]

With moments of kernel \(K\) and \(K^2\): \(\mu_j = \int u^j K(u) du\) and \(\nu_j = \int u^j K^2(u) du\). Matrices \(S = (\mu_{j+l})_{0 \leq j,l \leq p}\) and \(S^* = (\nu_{j+l})_{0 \leq j,l \leq p}\). Vectors \(c_p = (\mu_{p+1}, ..., \mu_{2p+1})^T\), \(\tilde{c}_p = (\mu_{p+2}, ..., \mu_{2p+2})^T\) and unit vector \(e_\nu = (0, ..., 0, 1, 0, ... 0)^T\) with 1 on the \(\nu^{th}\) entry.

To provide further insight to how the local polynomial regression assigns weights to each data points and make it easier to compare with other local constant kernel methods, we will use equivalent kernel introduced in Fand Gijbels 1996 [17] from now on.

\[ K_\nu^*(t) = e_T \nu+1 S^{-1} (1, t, ..., t^p)^T K(t) = K(t) \sum_{l=0}^p S^{\nu l} t^l \quad (3.28) \]

With \(S^{-1} = (S^{jl})_{0 \leq j,l \leq p}\). The equivalent kernel satisfies

\[ \int u^q K_\nu^*(u) du = \delta_{\nu,q} = 1, \text{ for } \nu = q, 0 \leq \nu, q \leq p \quad (3.29) \]
\[ = 0, \text{ for } \nu \neq q, 0 \leq \nu, q \leq p \quad (3.30) \]

The equivalent kernel \(K_\nu^*\) is a kernel of order \((\nu, p+1)\) (definition of kernel order can be found in Gasser, Muller and Mammitzsch 1985 [22]). Therefore, the corresponding asymptotic expressions for Equation (3.22) - (3.26) are as follows:

\[ \text{Var}[\hat{m}(x)|X] = V_0^* \frac{\sigma^2}{nh} + o_p\left(\frac{1}{nh}\right) \quad (3.31) \]
\begin{align*}
\text{Bias}[\hat{m}(x)|X] &= B_{0,p+1}^* m^{(p+1)}(x) \frac{h_{p+1}}{(p+1)!} + o_p(h_{p+1}), \text{ If } p \text{ is odd} \quad (3.32) \\
&= B_{0,p+2}^* m^{(p+2)}(x) \frac{h_{p+2}}{(p+2)!} + o_p(h_{p+2}), \text{ If } p \text{ is even} \quad (3.33)
\end{align*}

Asymptotic variance and bias of first derivative estimates $\hat{m}_1(x)$ are:

\begin{align*}
\text{Var}[\hat{m}_1(x)|X] &= V_1^* \frac{\sigma^2}{nh^3} + o_p \left( \frac{1}{nh^3} \right) \quad (3.34)
\end{align*}

\begin{align*}
\text{Bias}[\hat{m}_1(x)|X] &= B_{1,p+2}^* m^{(p+2)}(x) \frac{h_{p+1}}{(p+2)!} + o_p(h_{p+1}), \text{ If } p \text{ is odd} \quad (3.35) \\
&= B_{1,p+1}^* m^{(p+1)}(x) \frac{h^p}{(p+1)!} + o_p(h^p), \text{ If } p \text{ is even} \quad (3.36)
\end{align*}

With $V_j^* = \int K_j^*(u) du$ and $B_{j,l}^* = \int u^j K_j^*(u) du$. Note that the bias term is zero up to polynomial order $p$.

Therefore, by Equation 3.20, we have the asymptotic properties for peak location estimates $\hat{\theta}$ as follows:

\begin{align*}
\text{Var}[\hat{\theta}|X] &= \frac{1}{m''(\theta)^2} V_1^* \frac{\sigma^2}{nh^3} + o_p \left( \frac{1}{nh^3} \right) \quad (3.37)
\end{align*}

\begin{align*}
\text{Bias}[\hat{\theta}|X] &= \frac{1}{m''(\theta)(p+2)!} B_{1,p+2}^* m^{(p+2)}(\theta) h_{p+1} + o_p(h_{p+1}), \text{ If } p \text{ is odd} \quad (3.38) \\
&= \frac{1}{m''(\theta)(p+1)!} B_{1,p+1}^* m^{(p+1)}(\theta) h^p + o_p(h^p), \text{ If } p \text{ is even} \quad (3.39)
\end{align*}
3.5.2 Optimal Bandwidths for Location and Height Estimators

A theoretical choice of bandwidth for estimating \( m_\nu(\theta) \) is the \( h \) that minimizes the asymptotic MSE

\[
\text{AMSE}[\hat{m}_\nu(x)|X] = \text{Bias}[\hat{m}_\nu(x)|X] + \text{Var}[\hat{m}_\nu(x)|X]
\]

Since location and height estimates have different asymptotic MSE forms, we proposed to use different bandwidth \( h_1 \) and \( h_2 \) for the two estimates separately. So from Equation (3.31) to (3.35), we can get the optimal bandwidth \( h_1^* \) for location estimates \( \hat{\theta} \) as

\[
h_1^* = \left( \frac{3V_1^* \sigma^2 (p + 2)!^2}{2(p + 1)B_{1,p+2}^2 m_1^{(p+2)}(\theta)} \right)^{\frac{1}{2p+5}} n^{-\frac{1}{2p+5}}, \text{ If } p \text{ is odd}
\]

\[
h_1^* = \left( \frac{3V_1^* \sigma^2 (p + 1)!^2}{2pB_{1,p+1}^2 m_1^{(p+1)}(\theta)} \right)^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}, \text{ If } p \text{ is even}
\]

Similarly, for peak height estimates \( \hat{m}(\theta) \), the corresponding optimal bandwidth \( h_2^* \) is

\[
h_2^* = \left( \frac{V_0^* \sigma^2 (p + 1)!^2}{2(p + 1)B_{0,p+1}^2 m_0^{(p+1)}(\theta)} \right)^{\frac{1}{2p+5}} n^{-\frac{1}{2p+5}}, \text{ If } p \text{ is odd}
\]

\[
h_2^* = \left( \frac{V_0^* \sigma^2 (p + 2)!^2}{2(p + 2)B_{0,p+2}^2 m_0^{(p+2)}(\theta)} \right)^{\frac{1}{2p+7}} n^{-\frac{1}{2p+7}}, \text{ If } p \text{ is even}
\]

These two bandwidths are optimal in terms of minimizing the asymptotic MSES of the peak location and height estimates respectively. The reason why we use such estimators are explained in details in Muller 1985 [32]. In Section 3.6.1, we will use simulation examples to see how the peak location and height estimates varies with bandwidth and other properties.
3.5.3 A Practical Bandwidth Selection Algorithm

Local polynomial regression with data-driven bandwidth selector has been explored quite well in literature, and discussion has shown that plug-in methods tend to outperform the cross-validation methods most of the time. Various plug-in methods are using some kind of pilot estimates on the unknown quantities in optimal bandwidth equation, one popular branch is breaking the data into sub-intervals, fit one polynomial model in each interval.

Our focus is on peak estimates instead of the whole curve estimate, we propose a practical data-driven bandwidth selector by minimizing the MSE introduced in previous section. Based on the optimal bandwidths in Equations 3.41 - 3.43 is proposed as below:

1. Locate peak area: location of maximum point +/- 10% of data range; Estimate $\sigma^2$.
2. Within this peak neighborhood, fit a quartic model to estimate mean function and find its maximum location $\hat{\theta}_0$ with $(a = 1 \text{ or } 2)$.
3. Fit $(p + a)$ polynomial model to get $(p + a)^{th}$ derivative estimates $\hat{m}^{(p+a)}(\hat{\theta}_0)$.
4. Plug in the above estimates in optimal bandwidth functions 3.41 - 3.43 to get optimal bandwidth estimates $\hat{h}_1$ and $\hat{h}_2$ respectively.
5. Compute peak location $\hat{\theta}(\hat{h}_1)$ and height $\hat{m}(\theta, \hat{h}_2)$ using $\hat{h}_1$ and $\hat{h}_2$.

This algorithm is using quartic fitting in the neighborhood of peak to get pilot estimates of the derivatives needed in the optimal bandwidth functions, and then plug-in the estimates to get estimated optimal bandwidths, and use them on estimation
of peak location and height respectively. The estimates is easy to apply and data driven. Performance and results of this algorithm will be discussed in next section.

3.5.4 Limit Distributions of the Peak Location and Height Estimators

To derive the limiting distribution and take away the \( n \) term in bandwidth, a special bandwidth \( s \) satisfies \( h = sn^{-\frac{1}{2(k+p+1)}} \) is introduced, i.e. use \( h_1 = sn^{-\frac{1}{2k+3}} \) \( h_2 = tn^{-\frac{1}{2k+5}} \), and our optimal bandwidths for peak location and height estimators \( h^*_1 \) and \( h^*_2 \) are in similar forms of the optimal special bandwidths \( s^* \) and \( t^* \) in Muller 1989 [34] respectively. So all limiting distributions of their estimators based on the optimal bandwidths could be applied to our estimators directly. We have the optimal special bandwidth \( s^* \) for peak location estimate as

\[
s^* = \left( \frac{3V_1^*\sigma^2(p+2)!^2}{2(p+1)B_{1,p+2}^2m(p+2)^2(\theta)} \right)^{\frac{1}{2p+3}}, \quad \text{If } p \text{ is odd, i.e. } k = p+1 \quad (3.45)
\]

\[
= \left( \frac{3V_1^*\sigma^2(p+1)!^2}{2pB_{1,p+1}^2m(p+1)^2(\theta)} \right)^{\frac{1}{2p+3}}, \quad \text{If } p \text{ is even, i.e. } k = p \quad (3.46)
\]

Satisfies

\[
h_1 = s^*n^{-\frac{1}{2k+3}} = s^*n^{-\frac{1}{2k+5}}, \quad \text{If } p \text{ is odd} \quad (3.47)
\]

\[
= s^*n^{-\frac{1}{2k+3}}, \quad \text{If } p \text{ is even} \quad (3.48)
\]

Similarly, the optimal special bandwidth \( t^* \) for peak height estimate is

\[
t^* = \left( \frac{V_0^*\sigma^2(p+1)!^2}{2(p+1)B_{0,p+1}^2m(p+1)^2(\theta)} \right)^{\frac{1}{2p+3}}, \quad \text{If } p \text{ is odd, i.e. } k = p+1 \quad (3.49)
\]

\[
= \left( \frac{V_0^*\sigma^2(p+2)!^2}{2(p+2)B_{0,p+2}^2m(p+2)^2(\theta)} \right)^{\frac{1}{2p+3}}, \quad \text{If } p \text{ is even, i.e. } k = p+2 \quad (3.50)
\]
Satisfies
\[ h_2 = t^* n^{-\frac{1}{2p+1}} = t^* n^{-\frac{1}{2p+3}}, \text{ If } p \text{ is odd} \] (3.51)
\[ = t^* n^{-\frac{1}{2p+3}}, \text{ If } p \text{ is even} \] (3.52)

Note: \( k \) is the kernel order and \( p \) is the polynomial degree.

Using the optimal special bandwidths, we claim the peak location estimate \( \hat{\theta}(s) \) follows an asymptotically Normal distribution

\[
\begin{align*}
\frac{n}{\sqrt{p+1}} (\hat{\theta}(s) - \theta) & \xrightarrow{D} N\left( \frac{B^*_1 v^{(p+2)}(\theta) s^{p+1}}{m''(\theta)(p+2)!}, \frac{V^*_1 \sigma^2}{m''(\theta)^2 s^3} \right), \text{ If } p \text{ is odd} \quad (3.53) \\
\frac{n}{\sqrt{p+3}} (\hat{\theta}(s) - \theta) & \xrightarrow{D} N\left( \frac{B^*_1 v^{(p+1)}(\theta) s^p}{m''(\theta)(p+1)!}, \frac{V^*_1 \sigma^2}{m''(\theta)^2 s^3} \right), \text{ If } p \text{ is even} \quad (3.54)
\end{align*}
\]

\[
\begin{align*}
\frac{n}{\sqrt{p+1}} (\hat{m}(\theta, t) - m(\theta)) & \xrightarrow{D} N\left( \frac{B^*_{0,p+1} v^{(p+1)}(x) t^{p+1}}{(p+1)!}, \frac{V^*_0 \sigma^2}{t} \right), \text{ If } p \text{ is odd} \quad (3.55) \\
\frac{n}{\sqrt{p+3}} (\hat{m}(\theta, t) - m(\theta)) & \xrightarrow{D} N\left( \frac{B^*_{0,p+2} v^{(p+2)}(x) t^{p+2}}{(p+2)!}, \frac{V^*_0 \sigma^2}{t} \right), \text{ If } p \text{ is even} \quad (3.56)
\end{align*}
\]

Muller 1989 [34] also proved that the joint distribution of
\[
\begin{pmatrix}
\frac{n}{\sqrt{p+1}} (\hat{\theta}(s) - \theta) \\
\frac{n}{\sqrt{p+1}} (\hat{m}(\theta, t) - m(\theta))
\end{pmatrix}
\]
with \( k = p + a \) is asymptotically Normal with covariance equals to zero.

We are using similar idea of estimating the peak location as in Muller 1985, which is defined in Equation 3.18, and he claims that it is the same as \( \hat{\theta}_1(s) \) defined in Equation 3.19, we will investigate this in the first part of Simulation section later and proof later.
For the peak height estimate, Muller used the mean function estimate at the estimated location $\hat{\theta}(s)$ with special bandwidth $t$, which is denoted as $\hat{m}(\hat{\theta}(s), t)$, but this may not be the actual maximum point of the mean function. Therefore, in our method, we estimate the peak height as the maximum value of curve estimate using special bandwidth $t$, denote as $\hat{m}(\theta, t)$, discussion and comparison the two approaches will be discussed in the next section.

### 3.6 Simulation Study

In this section, I performed two set of simulation studies to compare how our estimator performs versus Muller’s estimator and investigate how our estimator behaves when sample size goes to infinity under three different scenarios.

#### 3.6.1 Peak Estimator Comparison

In previous section, we discussed the theoretical identification of two kind of estimators, and we denote them as

- Location estimator a: $\hat{\theta}(s)$ (3.18) - maximum point of curve estimation.
- Location estimator b: $\hat{\theta}_1(s)$ (3.19) - cross zero point of 1st derivative estimate.
- Height estimator a: $\hat{m}(\theta, t)$ as the maximum value of curve estimate.
- Height estimator b: $\hat{m}(\hat{\theta}_1(s), t)$ as the curve estimate at $\hat{\theta}_1(s)$.

Theoretically estimator a and b gives the same results, in this section, we will explore how they actually are for the same simulated data set, and also compared the actual bandwidths that gives the best estimators. We use grid search to find
these "best" bandwidth that gives the minimum MSE of location and height estimates respectively for the two approaches listed above. We compared how the MSE of estimator, actual best bandwidth (bandwidth that gives the smallest MSE) and estimated values go with the number of points increases from 30 to 10000. The plot results are for simulation of 200 curves from $m_1(x) = \frac{x(1-x)^5}{B(2,6)}$ with Gaussian noise $(0, 0.2^2)$, the true peak location is at $1/6$, true peak height at around 2.81.

**Figure 3.9:** Log-log scaled plot to check asymptotic performance of actual best bandwidth of location estimator, for simulated data from Beta(2,6). Top 3 plots are for location estimator a; Bottom 3 plots are for location estimator b.

The above plots are color coded for different polynomial degree $p$. Top rows are results for location estimator a, bottom rows are for estimator b. For asymptotic
behavior comparison purpose, each plot has x-axis being the number of points on a curve (from 30 to 10000). In first column, empirical MSE is defined as $MSE(\hat{\theta}(h_1)) = \frac{1}{n} \sum_{i=1}^{n} (\theta_i(h_1) - \theta)^2$ and is plotted with 95% confidence interval. Asymptotic MSE for each polynomial degree $p$ is plotted as dotted lines. For both cases, MSEs are strictly decreasing in an almost linear trend when the number of points increases as expected. However, asymptotic MSE doesn’t always within the confident band of empirical MSE. This could possibly due to small sample size, Monte Carlo sampling error or boundary effect which will be explained later in this section.

For the best bandwidth (the value calculated through grid search that gives the actual best estimate), the trend is decreasing as number of points on the curve increases. For this simulation, we expect boundary effects on optimal bandwidth when it’s greater than $1/6$ (black dotted line), this could explain why we don’t see the empirical best bandwidths follows asymptotic optimal bandwidths exactly. But when the sample size is large enough, we don’t need to worry about the boundary effect anymore. MSE of location a estimates has a better convergence slope than location b.

For location optimal bandwidth $h_1^*$ (colored dotted lines), local linear and local quadratic regressions have exactly the same optimal bandwidth (plotted lines lied on top of each other), this is because of the different Bias term in equation 3.26 and therefore cause the different optimal bandwidth forms for $p$ even and odd cases. Similarly for $p = 2$ and $3$ in optimal height bandwidth $h_2^*$.

The third column of the above plots shows estimated location using two methods with 95% confidence intervals. We can see the estimates all approaches to the true peak location (black dotted line) as the sample size increases, but location a
3.18 is more accurate and is less biased than location method b. Therefore, we use method a, which is estimating the peak location as the point the maximum curve estimate is at in our proposed plug-in approach, which will be introduced in next section.

Figure 3.10: Asymptotic performance of actual best bandwidth of height estimator, for simulated data from Beta(2,6). Top 3 plots are for height a; Bottom 3 plots are for height b.

Similarly, for height estimates, two approaches have similar MSE trend along sam-
ple size. The best bandwidth trend is different from optimal bandwidth trend due to the same reason as in location estimates. Estimated peak height approaches to true peak height very closely, but when the sample size is small, height estimator b is under estimate the value more than estimator a.

It is worth to note that, for peak height estimates, any smoothing method tends to underestimate the true peak, and when we estimate it at an estimated peak location instead of the maximum value, it would underestimate it even more. Therefore, we prefer to use the maximum value of curve estimation instead of curve value at estimated peak location (Muller’s peak estimator) in our plug-in method, details of this part will be discussed in next section.

3.6.2 Data Driven Plug-in Bandwidth Selector

In section 3.5.3, I introduced a plug-in data driven bandwidth selection algorithm for peak location and height estimates using local polynomial regression. In previous section, I have compared two ways of calculating the peak estimates and conclude using method a (peak location as maximum point of curve estimation and peak height as maximum value of curve estimate) is preferred. In this section, we will show how the optimal bandwidth performs using 3 set of simulations.

In optimal bandwidth equation 3.41 to 3.43, for a given kernel function $K()$ and polynomial degree $p$, all terms are known except $\sigma^2$ and $m^{(p+a)}(\theta)$, $a = 1$ or 2. To have a clear view of how the optimal bandwidth performs with $n \to \infty$, we compare results from 3 sets of simulations with different peak sharpness as plotted in Figure 3.11.

- Simulation 1. $m_2(x) = x^5e^{-x}$, peak location $\theta_2 = 5$, peak height $m_2(\theta) \approx 18.75$,
with Gaussian noise \((0, 2^2)\).

- Simulation 2. \(m_1(x) = \frac{x(1-x)^3}{B(2,6)}\), peak location \(\theta_1 = \frac{1}{6}\), peak height \(m_1(\theta) \approx 2.81\), with Gaussian noise \((0, 0.2^2)\).

- Simulation 3. \(m_3(x) = \sin(x) - 0.1 \times x\), peak location \(\theta_3 = \pi/2 - 0.1 \approx 1.47\), peak height \(m_3(\theta) \approx 0.848\), with Gaussian noise \((0, 0.05)^2\).

Figure 3.11: 3 simulated data sets with different peak sharpness, the solid black line is the true curve function. Simulation 1 has a sharper peak (half height width = 26.5%); Simulation 2 has a moderate peak curve (half height width = 34.7%); Simulation 3 has a flatter peak (half height width = 41.4%).

For each simulation, we generate a group of 200 curves from the true function \(m(x)\) with different sample size \(n = 30, 100, 300, 1000, 3000, 1000\), estimating with Epanechnikov kernel, so the optimal bandwidth in Equation 3.41 and 3.43 are known.
Asymptotic MSE vs. Empirical MSE along different sample size

In previous section, we have already seen how MSE decreases when number of points on a curve increases, now let’s see how the MSE changes along bandwidth, when number of points on a curve increased from a small number of infinity. In Figure 3.12, we picked number of points 100 and 10000 to check how MSE of local linear (p=1) peak location estimator performs along the bandwidth. For both theoretically asymptotic (dashed lines) and empirical (solid lines) results, MSEs for all 3 simulations are convex curve with a single lowest MSE point, which is defined as the optimal (best) bandwidth for the given estimator.

Note that, simulation 3 has the flattest true curve function, therefore, it has wide range of bandwidth that gives lowest and close to lowest MSEs. After passing the optimal bandwidth point, the asymptotic MSE increases slowly, while empirical MSE increased drastically. Other than this, the other 2 simulations show very close results of both asymptotic and empirical MSE trend along bandwidth.

The results for local quadratic, cubic and quartic polynomials are similar to the local linear results, saying that from sample size of 100 to 1000 on a curve, both asymptotic and empirical MSEs have same pattern as a convex curve.
Figure 3.12: Asymptotic (dashed lines) and empirical (solid lines) location estimators’ MSEs along bandwidth for 3 simulation sets. Top row shows number of points on a curve being 100, bottom row shows the number being 10000, both have similar pattern as convex curve with an optimal (best) bandwidth at which the MSE is minimized.

Similarly, for peak height estimates, we compared the asymptotic and empirical MSEs. The results are slightly different with location estimator. For height estimators, both empirical and asymptotic MSEs drops very fast when bandwidth is small, and asymptotic MSEs increase slowly afterwards, while the empirical MSEs increase exponentially. This is because the simulated curves have larger variance in height values than location values, therefore, when the bandwidth used in the local polyno-
mial fit increases, MSE increases much faster, especially when the curve peak is as sharp as in that simulation 1. The results for local polynomial with higher degrees are similar to the local linear results showed below.

Figure 3.13: Asymptotic (dashed lines) and empirical (solid lines) height estimators’ MSEs along bandwidth for 3 simulation sets. Top row shows number of points on a curve being 100, bottom row shows the number being 10000, both have similar pattern as convex curve with an optimal (best) bandwidth at which the MSE is minimized.
Optimal vs. Best Bandwidth

For a given local polynomial degree $p$, number of points on a curve $n$, we can calculate the theoretical optimal bandwidth as the bandwidth gives the minimum asymptotic MSE. Similarly, we define the best bandwidth as the best empirical bandwidth that gives the minimum empirical MSE by grid search on bandwidth ranges. In the plot below, 3 columns are for 3 simulations, top row is the bandwidth for location estimator, bottom row is the bandwidth for height estimator. Also, optimal bandwidths are plotted in dashed lines, best bandwidths are plotted in solid lines.
Figure 3.14: Theoretical optimal (dashed lines) and empirical best (solid lines) bandwidths for both peak location and height estimators, color coded for polynomial degree $p = 1, 2, 3$ and 4. 3 columns are for 3 simulation studies.

From the figure above, we can see that when the number of points increases, both optimal and empirical best bandwidths decrease. According to the optimal bandwidths in Equation 3.41 and 3.43, local linear ($p=1$) and quadratic ($p=2$) estimators for peak location have the same optimal bandwidth, local cubic ($p=3$) and quartic ($p=4$) bandwidths for location estimator are the same, while for peak height estimator, local quadratic and cubic bandwidths are the same.
The optimal and best bandwidths show similar downward trends roughly, but there are some exceptions. For example, for simulation 3, location bandwidth, when p=4, empirical best bandwidth (blue solid line) shows an upward trend from n=1000 to 10000, and it is larger than the corresponding theoretical optimal bandwidth. This is possibly due to boundary effect, as the data ranges from 0 to 2.8, with true peak location at around 1.47 (bold black dashed horizontal line), any bandwidth beyond it won’t give accurate estimates since it covers some invalid range in. The issue is similar for simulation 3 height bandwidth (bottom right plot), where local quartic estimate need bandwidth beyond the true peak location, therefore, its empirical best bandwidth is more off from the theoretical optimal bandwidth. While for p <= 3 polynomial fittings, empirical bandwidth are very close to theoretical bandwidth. The small differences could be due to Monte Carlo sampling error, since we only simulated 200 curves.

For simulation 1, location bandwidth for local quadratic fitting (red solid line) shows an abnormal pattern, which is far away from theoretical bandwidth trend. This shows that even we grid search the best bandwidth along the x-axis range, it won’t give always the results close to the truth.
Figure 3.15: Theoretical minimum (dashed lines) and empirical minimum (solid lines) MSE for given polynomial degree $p$ and sample size $n$.

For the theoretical MSE (dashed lines) and empirical MSE (solid lines), Figure 3.15 show how they decreases when sample size increases. Note that the plot is in log scale. By definition, MSEs for location estimator are the same when polynomial degree $p = 1$ and $2$, and also the same for $p = 3$ and $4$, which are showed in the top row. For height estimator, theoretically the MSEs are the same when $p = 2$ and $3$. Empirical best bandwidth are showing downward trend as sample size increases, but not in an accurate way. For example, in simulation 1, local quadratic estimates of peak location, when sample size increases from 30 to 100, MSE increased instead.
Therefore, grid search best bandwidth won’t necessarily give the best results, and it’s computationally expensive when the sample size is large. Therefore, we recommend to use the plug-in bandwidth estimator. The algorithm is presented in previous section, we will show how it performs in simulation studies next.

**Plug-in Bandwidth Selector**

In section 3.5.3, we give a 5-step plug-in bandwidth selection algorithm for peak location and height estimates utilizing optimal bandwidth in equation 3.41 to 3.43. In this section, we use the same simulated curves as in previous sections, and Figure 3.16 shows how the bandwidth estimates using plug-in estimators perform comparing with theoretical optimal bandwidths.

For location bandwidth, both simulation 1 and 2 show the plug-in bandwidths decrease with sample size increases in the same slope as the optimal bandwidth, except a consistent difference. Note location bandwidths, plug-in bandwidth for p=3 and 4 are exactly the same, so they were plotted on top of each other and only show blue (p=4) line. For simulation 3, local cubic and quartic plug-in bandwidth doesn’t decrease as the sample size increase. Similarly, for height bandwidth, simulation 1 and 2 performs much better than simulation 3, which means when the curve peak tend to be flat, using higher order polynomials (p=3 or 4) in the plug-in algorithm won’t give accurate results.
Figure 3.16: Plug-in bandwidth (solid lines) vs. theoretical optimal (dashes lines) bandwidth along sample size from 30 to 10000, color coded in polynomial degrees p.

For the MSE results using plug-in algorithm, we compared them with theoretical minimum MSEs along sample size as well. As shown in Figure 3.17, plug-in MSEs have very similar downward trending as the theoretical minimum MSEs when sample size increases with a small differences. One exception is for simulation 2, height estimator, local linear MSE is higher than theoretical MSE, this is due to its corresponding bandwidth is higher than theoretical values showed in 3.16. Other than this, plug-in MSE is much closer to theoretical MSEs comparing with the empirical best MSE results in Figure 3.15.
Figure 3.17: MSE for Plug-in bandwidth (solid lines) vs. theoretical optimal (dashes lines) bandwidth along sample size from 30 to 10000, color coded in polynomial degrees p.

3.7 Real Data Application

This project is motivated for the repeated experiment measurements from the spectroscopic device, therefore, in this section, we apply the plug-in peak estimation method to the data set of Rhodamine reference solution measurements described at the beginning of this chapter.
The raw measurements are for excitation wavelength 390nm, the log10 scale plot is as below:

![Raw data graph](image)

Figure 3.18: Rhodamine solution repeated measurements, 3 colors for 3 measurement conditions: wet, dry and submerged into the water.

Since we don’t know the true peak location and height for this standard measurement, it’s hard to assess the MSE, but we checked the mean and standard deviation of raw data peak location and height vs. fitted peak location and height.
Table 3.3: LBJ390 peak location estimates

<table>
<thead>
<tr>
<th>LBJ390 - peak location</th>
<th>mean</th>
<th>standard deviation</th>
<th>average bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw data</td>
<td>581.56</td>
<td>1.51</td>
<td>0</td>
</tr>
<tr>
<td>local linear fit</td>
<td>581.61</td>
<td>1.25</td>
<td>28.93</td>
</tr>
<tr>
<td>local quadratic fit</td>
<td>580.63</td>
<td>1.15</td>
<td>28.93</td>
</tr>
<tr>
<td>local cubic fit</td>
<td>581.91</td>
<td>1.12</td>
<td>46.51</td>
</tr>
<tr>
<td>local quartic fit</td>
<td>580.99</td>
<td>1.17</td>
<td>46.51</td>
</tr>
</tbody>
</table>

Table 3.4: LBJ390 peak height estimates

<table>
<thead>
<tr>
<th>LBJ390 - peak height</th>
<th>mean</th>
<th>standard deviation</th>
<th>average bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw data</td>
<td>$1.71 \times 10^3$</td>
<td>201.5</td>
<td>0</td>
</tr>
<tr>
<td>local linear fit</td>
<td>$1.48 \times 10^3$</td>
<td>189.5</td>
<td>17.49</td>
</tr>
<tr>
<td>local quadratic fit</td>
<td>$1.35 \times 10^3$</td>
<td>173.8</td>
<td>50.54</td>
</tr>
<tr>
<td>local cubic fit</td>
<td>$1.35 \times 10^3$</td>
<td>173.8</td>
<td>50.54</td>
</tr>
<tr>
<td>local quartic fit</td>
<td>$1.46 \times 10^3$</td>
<td>177.9</td>
<td>61.36</td>
</tr>
</tbody>
</table>

From the two tables above, we can see that both location and height estimates using local polynomial regression reduce the estimation variations. The higher polynomial fitting gives lower variance estimates, but also need wider bandwidth. For application purpose, if the bandwidth is too wide, we will have boundary effect that doesn’t give accurate results. Therefore, I applied plug-in bandwidth selection method with $p = 2$. 
Figure 3.19: Peak estimates and corresponding bandwidths using local quadratic model. 1, 2 and 3 indicates measurement conditions: dry, wet and submerged.

From Figure 3.19, we see there are systematic differences under 3 measuring conditions, which cannot be eliminated from our local polynomial fitting. To better check the performance of plug-in algorithm applied on the real data, we used measurements from dry condition, and applied local polynomial fitting with degrees from 1 to 4 as plotted below.
Figure 3.20: Peak location and height estimates, together with the corresponding bandwidth used, for polynomial degrees 1 to 4.

For location estimates, all polynomial degrees from 1 to 4 gives similar estimates with very small standard deviation (in Table 3.5). The average bandwidth used are around 12 for local linear and quadratic and 27 for local cubic and quartic estimators. That is to say, higher polynomial degree requires a much wider bandwidth, though they give very similar peak location estimates. Since when bandwidth increase, the chance of getting boundary effect problem is higher, therefore, we recommend to use local linear or quadratic fitting in location estimate.
Table 3.5: Peak location estimates for LBJ dry measurements.

<table>
<thead>
<tr>
<th>LBJ390 - peak location</th>
<th>mean</th>
<th>standard deviation</th>
<th>average bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>local linear fit</td>
<td>580.99</td>
<td>1.31</td>
<td>12.85</td>
</tr>
<tr>
<td>local quadratic fit</td>
<td>580.38</td>
<td>1.26</td>
<td>12.85</td>
</tr>
<tr>
<td>local cubic fit</td>
<td>581.23</td>
<td>1.22</td>
<td>27.78</td>
</tr>
<tr>
<td>local quartic fit</td>
<td>580.56</td>
<td>1.26</td>
<td>27.78</td>
</tr>
</tbody>
</table>

For height estimator, different polynomial degrees give similar peak height estimates, all with very small variance. But local linear fitting requires a much smaller bandwidth (5.6) comparing with higher polynomial degrees. This difference is bigger than location estimates, since the peak height variation is higher than location variation.

Table 3.6: Peak height estimates for LBJ dry measurements.

<table>
<thead>
<tr>
<th>LBJ390 - peak height</th>
<th>mean</th>
<th>standard deviation</th>
<th>average bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>local linear fit</td>
<td>$1.57 \times 10^3$</td>
<td>116.6</td>
<td>5.6</td>
</tr>
<tr>
<td>local quadratic fit</td>
<td>$1.52 \times 10^3$</td>
<td>114.4</td>
<td>26.92</td>
</tr>
<tr>
<td>local cubic fit</td>
<td>$1.52 \times 10^3$</td>
<td>114.4</td>
<td>26.92</td>
</tr>
<tr>
<td>local quartic fit</td>
<td>$1.53 \times 10^3$</td>
<td>115.9</td>
<td>39.63</td>
</tr>
</tbody>
</table>

### 3.8 Discussion and conclusion

In this chapter, we first introduced the problem with the light intensity measurements using the cervical cancer detection device, and proposed to use nonparametric kernel methods to estimate the curves. We briefly reviewed popular kernel methods on curve estimations, and then focused on using local polynomial regression to estimate the
Local polynomial regression has 3 key parameters, the kernel function \( K() \), polynomial degree \( p \) and bandwidth \( h \). Study has shown that the kernel function \( K() \) is not critical, and here we used one of the most popular one, Epanechnikov kernel in both simulation and real data analysis. For polynomial degrees, we considered and compared \( p = 1 \) to 4, because for \( p \leq 5 \), polynomial fitting could lead to unstable and over-fitted results. The critical part is the bandwidth \( h \), which controls the complexity of the model.

In section 3.5, we defined the optimal peak location and height and derived asymptotic bias and variance equation for peak location and height estimates respectively. The asymptotic MSE (AMSE) is the sum of asymptotic bias\(^2\) and variance, therefore, the theoretically optimal bandwidth is the bandwidth that gives the minimum AMSE for peak location and height estimates respectively. In section 3.5.3, we proposed a plug-in bandwidth selection method using the optimal bandwidth formula, and therefore estimate the peak location and height simultaneously. The estimators are asymptotically Normally distributed and independent to each other.

In section 3.6, we first compared performance of two theoretically identical but computationally different estimators for peak location and height, and concluded using maximum point of curves estimation as the peak location estimator and the maximum value of curve estimate as the peak height estimator.

Then we used 3 sets of simulation studies to compare the bandwidths and MSEs of the theoretical optimal, empirical best through grid search and plug-in estimators.
across different polynomial degrees and sample sizes. The empirical results are not always in line with theoretical optimal values, this is showing that realistic problem is not always well modeled by the asymptotic framework. The asymptotic approximations can be quite poor when the sample size is not big enough. So we cannot use asymptotics alone to justify smoothing procedures.

The plug-in bandwidth selection algorithm gives closer to theoretical optimal bandwidth and minimum MSE results, but also suffers from boundary effects, when the optimal bandwidth is larger than true peak location. This could be one of our future work to do, as improve the plug-in bandwidth selection using special boundary kernels.
Chapter 4

Classification with Corrupted Data

In previous chapters, we have been focused on study of the repeated measurements using the spectroscopic device on certain kind of standard (a standard tile or a standard fluorescence solution). All these studies are in preparation for using the device on patient tissue measurement, which is used for making cervical cancer diagnosis. However, when there is only a very small portion of diseased tissue measurements, the classification results are usually not good on test data set. We introduce to use the idea of marginalized corrupted feature [28], adding in ”corrupted data” in the training set to make it more balanced, and therefore improve the predictability of the classifiers. For example, for logistic regression, corrupting the data infinitely large is equivalent to adding in ridge-like penalty term. Simulation study of several classifiers are compared and discussed.
4.1 Literature Review

4.1.1 Marginalized Corrupted Features

There has been studies on intentionally adding noise to the training data set in supervised learning. In [31], [3] and [28], corrupting the training set to obtain regularization and generalization benefits. Since we usually don’t know the true distribution of the features, noises are generated using some analytically convenient distributions, such as Normal distribution. There are researchers adding in noise based on their prior knowledge of the data. Some recent studies also focused on using the model which generates its own inputs as adding in noise from a learned, directly relevant distribution [51]. An application of marginalized corrupted data is presented in [4], which they randomly sample the image tag from the complete set of tags as the corrupted version of the original set and developed an image tagging algorithm.

There are also studies of using marginalized corrupted data in nature language processing (NLP) [56], where they point out that adding in noisy data infinitely is essentially treating the noise as an explicit regularizer, and approximate it using second-order formula instead of generating the noisy data. This coincides with our conclusion in this chapter, which we will explain in details later.

In [2], they showed that training the corrupted data with Gaussian noise is equivalent to a $L_2$ regularization in the low noise limit. But in other cases, [28] pointed out that we need to develop a new objective function for corrupting the data.

There are many studies on different corrupting distributions, for example, for heavy tailed feature distribution, we could drop out features randomly in the learning
procedure ([55] and [47]), to prevent features from co-adaption, and therefore reduce overfitting and improves regression or classification results. Dropout is considered as another way to regularization. But this won’t be our focus here.

4.1.2 Some Classification and Regularization Methods Review

In this section, we will list some popular classification and regularization methods that will be seen in later sections.

- Support Vector Machine (SVM) is one of the most popular machine learning technique for classification. It’s about learning complex structure from data and has very good empirical performance. It maps the data into a higher dimension and constructs an optimal separating hyperplane using kernel tricks. Details of it could be found at [9] and [46].

- Logistic regression is another well known classification method by utilizing logit link function on conditional distribution of the binary response variable, and is estimated using Maximum Likelihood Estimations.

- Regularization is a very powerful tool to solve overfitting problem. In general, it is adding in a penalty term to the loss function such as residual sum of squares. If the penalty term is $L_1$ norm, we call it LASSO ([52]), if it’s $L_2$ norm, we call it ridge regression ([24]), and if it is a linear combination of both, it’s called elastic net ([63]).

We won’t go into details of the techniques above since they are all well know methods with a lot of literature, and we will applying these methods on our marginalized corrupted data sets to illustrate the properties of data corrupting concept.
4.2 Proposed Data Corruption Method

The key idea of marginalized corrupted feature is to adding in data with noise from a fixed distribution, for example, Gaussian noise, and choose predictors that are robust to the noise under the corrupting distribution. The inspiration of this method is, when we have infinite data drawn from the data distribution, any simple classification method could give close to optimal results. In reality, we usually only have a finite sample to train the classifier, that is to say, some of the variations in the data distribution are not captured. Therefore, the classifier would perform worse in the training set than the test set.

Assume we have observations \((y_i, x_i), \ i = 1, ..., n\), the ordinary regression parameter estimator is defined as

\[
\hat{\beta} = \text{argmin} \sum_{i=1}^{n} l(y_i, \beta^T x_i)
\]

with \(l()\) being the loss function. And let’s assume \(\tilde{x}\) is the "corrupted" version of observation \(x\) with corrupting distribution \(p(\tilde{x}|x)\) from a exponential family, and it is unbiased \(E[\tilde{x}|p(\tilde{x}|x)] = x\). Therefore, the corrupted regression parameter estimator is

\[
\hat{\beta}_c = \text{armin} \sum_{i=1}^{n} \int l(y_i, \beta^T \tilde{x})p(\tilde{x}|x_i)d\tilde{x}
\]

If we define the score function as \(\phi(y_i, \beta^T x_i) = \frac{\partial l(y_i, \beta^T x_i)}{\partial \eta} x_i\), with \(\eta = \beta^T x\). Since \(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \phi(y_i, \hat{\beta}_n^T x_i) x_i = 0\) and \(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \int \phi(y_i, \beta_{c,n}^T \tilde{x})\tilde{x}p(\tilde{x}|x_i)d\tilde{x} = 0\), we claim that the empirical ordinary parameter estimator with sample size \(n\), denoted
as $\hat{\beta}_n$, converges to the true value $\beta_0$ in probability with sample size $n \to \infty$, i.e.

$$\hat{\beta}_n \overset{p}{\to} \beta_0$$

Similarly, for corrupted true parameter denoted as $\beta_{c0}$ and corrupted parameter estimator with sample size $n$ denoted as $\hat{\beta}_{c,n}$, we have

$$\hat{\beta}_{c,n} \overset{p}{\to} \beta_{c0}$$

Further,

$$\beta_n \overset{asympt.}{\sim} N(\beta_0, \frac{1}{n}E[\phi^2(Y, \beta_0^TX)XX^T]^{-1})$$

$$\hat{\beta}_{c,n} \overset{asympt.}{\sim} N(\beta_{c0}, W^{-1}(\beta_{c0})V_c(\beta_{c0})W^{-1}(\beta_{c0})/n)$$

where

$$W(\beta_{c0}) = E[\int \nabla\phi(Y, \beta_{c0}^T\tilde{X})\tilde{X}\tilde{X}^Tp(\tilde{X}|X)d\tilde{X}]$$

$$V_c(\beta_{c0}) = E[\int \int \phi(Y, \beta_{c0}^T\tilde{X})\phi(Y, \beta_{c0}^T\tilde{X}')\tilde{X}\tilde{X}'p(\tilde{X}|X)p(\tilde{X}'|X)d\tilde{X}d\tilde{X}']$$

### 4.3 Corruption on Real Data Studies

In previous section, we present the asymptotic properties of the corrupted data parameter estimators. In this section, we will apply corruption on the data with different Gaussian noise, and compare the results with different classification methods.

The data are measurements from the cervical cancer detection machine. Among 798 measurements on individual observations, we label stage CIN2 or worse as dis-
eased case (y=1), CIN1 or better as normal case (y=-1). Predictors include 18 measurement signal to noise ratios for each channel (snr=mean/standard deviation), 45 ratios of the channel intensities (eg. ratio of intensity of White reflectance red channel over UV fluorescence blue channel) and miscellaneous information, such as patient ID, device ID, measurement time etc, among which, the patient age is an important predictor as well.

**SVM Performance on Corrupted Data**

First, we use Support Vector Machine on the corrupted data with different noise distributions to check the effects of corruption distributions. The original data set has 50 diseased and 504 non-diseased patients. We add in Gaussian noise with different variance to generate a corrupted data set with 2500 diseased and 2500 non-diseased cases. The predictors are \( \log_{10}(\text{ratio } 1:9) + \text{I}[\text{age } \geq 45] \).

Table 4.1: SVM performance on corrupted thumbnail data with different corruption distributions.

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Decision value (Score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data</td>
<td>0.4347826</td>
<td>0.7529412</td>
<td>0.1844686</td>
</tr>
<tr>
<td>Corruption (N(0, S_{ij}))</td>
<td>0.826087</td>
<td>0.5235294</td>
<td>-0.1613062</td>
</tr>
<tr>
<td>Corruption (N(0, 0.03))</td>
<td>0.9130435</td>
<td>0.5058824</td>
<td>-0.1984455</td>
</tr>
<tr>
<td><strong>Corruption (N(0, 0.02))</strong></td>
<td><strong>0.9130435</strong></td>
<td><strong>0.5176471</strong></td>
<td><strong>-0.1170514</strong></td>
</tr>
<tr>
<td>Corruption (N(0, 0.01))</td>
<td>0.82608696</td>
<td>0.55294118</td>
<td>-0.08040079</td>
</tr>
</tbody>
</table>

The table above shows sensitivity, specificity and score value of SVM results. Note that these results are based on parameter cost=10, selected from 10-fold CV on original data. The second row is generating noise with variance depends on the actual
variance of the observation. Comparing the original data with the corrupted data, corruption helps to improve the sensitivity significantly, but not necessarily on specificity.

**Classification Methods on Corrupted Data**

Now we compare the performance of different classification and regression shrinkage methods on the original and corrupted data. The predictors are \( \log_{10}(\text{ratio}) + \log_{10}(\text{snr}) + I[\text{age} \geq 45] \). For "Lasso+logistic", we are using Lasso for variable selection first, and apply logistic regression on the selected variables only.

<table>
<thead>
<tr>
<th>Original data (log10ratio+log10snr+age)</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Decision value (Score)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>0.8519</td>
<td>0.5958</td>
<td>0.0224</td>
<td>0.7604</td>
</tr>
<tr>
<td><strong>Lasso</strong></td>
<td><strong>0.8889</strong></td>
<td><strong>0.5838</strong></td>
<td><strong>0.0459</strong></td>
<td><strong>0.7863</strong></td>
</tr>
<tr>
<td>Lasso + logistic</td>
<td>0.8148</td>
<td>0.6227</td>
<td>0.0386</td>
<td>0.7794</td>
</tr>
<tr>
<td>ridge</td>
<td>0.8519</td>
<td>0.5958</td>
<td>0.0504</td>
<td>0.764</td>
</tr>
</tbody>
</table>

The table above shows Lasso regression gives the best AUC of ROC curve, which means it has the best sensitivity and specificity trade-off. While the table below shows the results using same models but on corrupted data instead, the results are different.
Table 4.3: Logistic, lasso and ridge regression on corrupted data.

<table>
<thead>
<tr>
<th>Corrupted data (log10ratio+log10snr+age)</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Decision value (Score)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>0.6667</td>
<td>0.7335</td>
<td>0.0451</td>
<td>0.7351</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.6296</td>
<td>0.7814</td>
<td>0.0618</td>
<td>0.7351</td>
</tr>
<tr>
<td>Lasso + logistic</td>
<td>0.6667</td>
<td>0.7575</td>
<td>0.0508</td>
<td>0.7318</td>
</tr>
<tr>
<td>ridge</td>
<td>0.6296</td>
<td>0.7904</td>
<td>0.0631</td>
<td>0.7482</td>
</tr>
</tbody>
</table>

4.4 Discussion and Conclusion

In this chapter, we present the idea of corrupting the data by adding in Gaussian noise to improve classification results for highly unbalanced data. The corrupted data could capture some of the variations in the data distribution but not displayed in the finite sized samples. We derived the asymptotic properties of the parameter estimator for both the original and corrupted data.

Based on our simulation results of corrupting real data with Gaussian noise of different variances, we see that corrupting with the sample variance itself doesn’t necessarily give the best results. Also, by comparing different classification and regression model results on both original and corrupted data, the latter one doesn’t necessarily give better results. That is to say, data corruption method is improving the classification results when we have infinite amount of corrupted copies, it doesn’t always work when we corrupt a finite sample. Also, corrupting the data with infinitely many is equivalent to adding a ridge-like penalty term in loss function. This concept is discussed in [56]. Further study on this area is one of the popular topics these days.
[1] Matthew Avery. Literature review for local polynomial regression, 2013. 3.3.4, 3.3.4


[3] Chris J.C. Burges and Bernhard Scholkopf. Improving the accuracy and speed of support vector machines. pages 375–381, 1997. 4.1.1


[17] Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press, 1996. 3.3.4, 3.3.4, 3.5.1, 3.5.1


98


[34] Hans-Georg Muller. Adaptive nonparametric peak estimation. The Annals of Statistics, pages 1053–1069, 1989. 3.5.1, 3.5.4, 3.5.4


[38] James O Ramsay. *Functional data analysis*. Wiley Online Library, 2006. 1.2


[51] Erik Talvitie. Model regularization for stable sample rollouts. In *Uncertainty in Artificial Intelligence (UAI)*, 2014. 4.1.1


[54] Lu Wang, Jong Soo Lee, Pierre Lane, E Neely Atkinson, Andres Zuluaga, Michele Follen, Calum MacAulay, and Dennis D Cox. A statistical model for removing inter-device differences in spectroscopy. *Optics express*, 22(7):7617–7624, 2014. 2.3.2, 2.4.1


[58] Yingcun Xia and WK Li. Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. *Journal of multivariate analysis*, 83(2):265–287, 2002. 3.3.4


