NEW PROCEDURES FOR DATA MINING AND MEASUREMENT ERROR MODELS WITH MEDICAL IMAGING APPLICATIONS

by

XIAOFENG WANG

Submitted in partial fulfillment of the requirements

For the degree of Doctor of Philosophy

Dissertation Advisor: Dr. Jiayang Sun

Department of Statistics

CASE WESTERN RESERVE UNIVERSITY

August 2005

CASE WESTERN RESERVE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

We hereby approve the dissertation of

XIAOFENG WANG

candidate for the Doctor of Philosophy degree

Committee Chair: _____
Dr. Jiayang Sun
Dissertation Advisor
Professor, Department of Statistics

Committee: _____
Dr. Sara M. Debanne
Professor, Department of Epidemiology & Biostatistics

Committee: _____
Dr. J. Sunil Rao
Associate Professor, Department of Epidemiology & Biostatistics;
Adjunct Associate Professor, Department of Statistics

Committee: _____
Dr. Kenneth J. Gustafson
Assistant Professor, Department of Biomedical Engineering

Committee: _____
Dr. Mary H. Regier
Adjunct Professor, Department of Statistics

Table of Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to sincerely express my gratitude to my advisor, Dr. Jiayang Sun for being not only an excellent research advisor but also a true mentor. Her professional advice was essential to the completion of this dissertation and has given me a pathway to the future statistical life.

I am very grateful to Dr. Kath Bogie in the Cleveland FES center who brought us interesting clinical research problems and data, and gave me valuable suggestions during my research. My thanks also go to Dr. Sara M. Debanne, Dr. J. Sunil Rao, Dr. Kenneth J. Gustafson and Dr. Mary H. Regier for their serving on the thesis committee and helpful discussions.

A special thanks goes to Dr. Stephen Ganocy for reading the manuscript and providing helpful suggestions, and Yaomin Xu, Meng Xu and Deping Ye for their friendship and help.

Last, but not least, I would like to give my deepest gratitude to my dear wife Wenyan for her love and understanding during the past few years. Her encouragement and support are what made this dissertation possible. My appreciation and love go to my parents Huahui and Xianzhang, and my brother Xiaowei for their dedication and many years of support for now and future.

New Procedures for Data Mining and Measurement Error Models with Medical Imaging Applications

Abstract

by

XIAOFENG WANG

In this dissertation we provide analysis strategies for two research areas: *spatial-temporal data mining* and *measurement error problems*. Motivated by analyzing data from a "Neuromuscular Electrical Stimulation" experiment we develop an efficient procedure for mining spatial-temporal data which combines the following modern and newly developed components: data segmentation and registration, statistical smoothing mapping for identifying "activated" regions and a semiparametric model for detecting spatial-temporal similarities/trends from "large-p-small-n" data sets. For measurement error problems we provide new density and regression estimators for nonparametric errors-in-variables models. The errors can be either homogeneous or nonhomogeneous. In contrast to most existing procedures our new estimators are stable, easy to compute and do not depend on a Fourier transform. The asymptotics of the new estimators is investigated. Our procedures have the potential to become powerful new tools in the image analysis and other fields.

Key words: Spatial-temporal data, medical imaging, registration, smoothing, measurement error models, deconvolution, semiparametrics.

Chapter 1

Introduction

This dissertation provides analysis strategies for two research areas: *spatial-temporal data mining* and *measurement error problems*. Data from many scientific and medical areas such as medical imaging, epidemiology and climatology are often correlated *spatially*. Additionally, if they are collected over time, they may be also correlated in *time*, another dimension; and they are termed spatial-temporal data. Mining spatial-temporal data is challenging in that not only the data exhibit huge dimensionality but also involve both spatial and temporal effects. Measurement error problems constitute another active, rich research area in modern statistics. The effects of measurement error are well-known: the presence of measurement error if ignored can cause unignorable biases in estimated functions. Hence, correcting for such effects is important.

In the spatial-temporal data mining motivated by analyzing data from our "Neuromuscular Electrical Stimulation" (NMES for short) experiment we shall develop an efficient procedure for mining spatial-temporal data. This new procedure is a statistical ensemble built on following modern or newly developed components: (1) data segmentation for separating heterogeneous data and for distinguishing outliers, (2) automatic approaches for

spatial and temporal data registration, (3) statistical smoothing mapping for identifying "activated" regions based on generalized false-discovery-rate controlled p-maps/movies and (4) a semiparametric regression for detecting spatial and temporal similarities/trends from "large-p-small-n" data sets. Our new procedure should be applicable to other types of spatial-temporal data sets beyond those from the NMES experiment. It has the potential to be used in the analysis of time-series images and functional images such as those from fMRI.

In the measurement error problems we provide new density and regression estimators for nonparametric errors-in-variables models. The errors can be either homogeneous or nonhomogeneous. In contrast to most existing procedures our new procedures do not depend on a Fourier transform. The asymptotics of the new estimators is investigated. These estimators are stable, easy to compute and can be applied to many important areas such as imaging deblurring, nonparametric time series and astronomical data analysis.

1.1 Spatial-temporal Data Mining

Remarkable developments of medical and computer technology in the last two decades have enabled scientists and clinicians to collect huge amounts of data in both spatial and temporal dimensions. These types of data have become common in medical imaging, epidemiology, neuroscience, ecology, climatology, environmentology and other areas. Typical spatial-temporal data will be denoted by $y(s,t,n)$, where y is the intensity value at the spatial location $s \in \mathcal{S}$, time $t \in \mathcal{T}$ and for the subject indexed by $n \in \mathcal{N}$. In most applications, $\mathcal{S}$ will be a 1, 2 or 3 dimensional rectangle, indexed by points or pixels $s \in \mathcal{S} = \{1, ..., S\}$; $\mathcal{T} = \{1, 2, ..., T\}$, where T is the number of time points; and $\mathcal{N} = \{1, 2, ..., N\}$, where N is the number of subjects.

In principle, the indexing could be done by continuous variables, but in practice, only a discretized version is observed. In the study of data from the NMES experiment, $N \ll T \ll S$. This is known as *large-p-small-n* data analysis which is a challenging case in current data mining research. We first describe the clinical background and challenges of data analysis from the neuromuscular electrical stimulation experiment.

1.1.1 Clinical Background

Spinal Cord Injury and Pressure Sores

Spinal cord injury (SCI) is damage to the spinal cord that results in a loss of function such as mobility or feeling. Frequent causes of damage are trauma (*e.g.* car accident, gunshot wounds and falls) or disease (*e.g.* polio, spina bifida and Friedreich's Ataxia). Approximately 450,000 people live with SCI in the United States. There are about 10,000 new cases of SCI every year; the majority of them (82%) involve males between the ages of 16-30. These injuries result from motor vehicle accidents (36%), violence (28.9%), or falls (21.2%) (survey data from http://www.spinalinjury.net).

Pressure sores (also called pressure ulcers, bed sores, or decubitus ulcers) are areas of injured skin and tissue. They are usually caused by sitting or lying in one position for too long a period of time. This puts pressure on certain areas of the body which in turn can reduce the blood supply to the skin and the tissues under the skin. When a change in position doesn't occur often enough and the blood supply gets too low, a sore may form. Pressure sores/ulcers are known to be a multi-factorial complication that occurs in many individuals who are wheelchair users due to reduced mobility. Figure 1.1 displays the basic progression of pressure sores.

All individuals with SCI, and particularly those with complete lesions, are considered to be at high risk of pressure sore development throughout

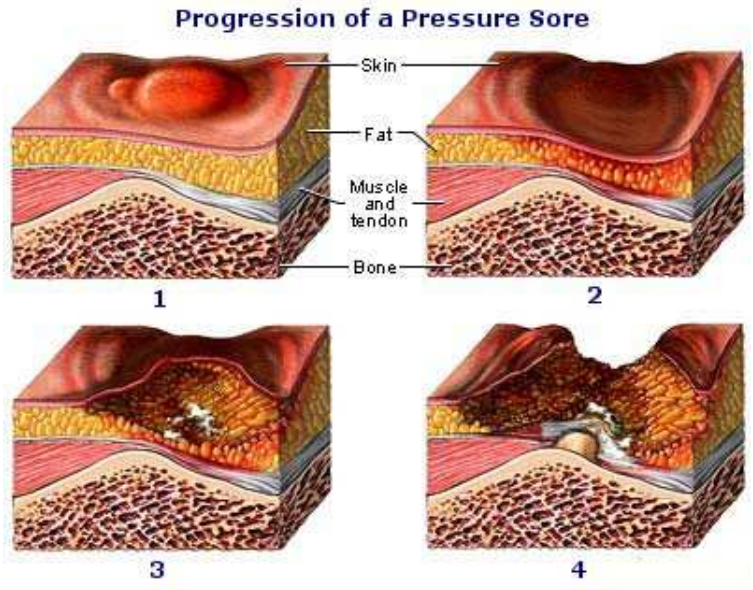


Figure 1.1: *Progression of pressure sores/ulcers (The picture is taken from http://www.spinal-injury.net/pressure-sores-sci.htm). Areas of damaged skin and tissue are developed when sustained pressure - usually from a bed or wheelchair - cuts off circulation to vulnerable parts of the body, especially the skin on the buttocks, hips and heels. Without adequate blood flow, the affected tissue dies.*

their lifetime. This significant secondary complication is the major cause for re–admission to the hospital following primary rehabilitation. Indeed, up to 50% of at-risk elderly individuals may have sitting-induced pressure sores. Individuals with spinal cord injury are also at high risk of pressure ulcer development, with reported community incidence rates in the region of 32% for individuals with chronic SCI (Yarkony and Heinemann, 1995).

Treating pressure sores in the United States has been estimated to cost over $1.33 billion annually, primarily because of the need for prolonged periods of bed rest associated with many methods of treatment. Pressure sores tend to reduce independence and affect many aspects of daily life such as physiological well-being, social interactions, work or college attendance, and need for caregiver time. Further medical complications may also arise, in particular, systemic infections leading to fatality. Approaches to the pre-

vention of pressure sores in high-risk populations can generally be classified as education-focused or device-focused. However, despite the development of many support devices and the introduction of skin care training within many rehabilitation programs the incidence of pressure sores remains unacceptably high (Yarkony and Heinemann, 1995). Furthermore, there remains a significant proportion of the SCI population who exhibit chronic recurrence of tissue breakdown. It is therefore highly beneficial both societally and for the individual to develop effective techniques to reduce the incidence of pressure ulcers and maximize function while sitting.

Neuromuscular Electrical Stimulation

Traditionally, techniques to reduce pressure sore incidence have focused on extrinsic risk factors by providing cushions which improve pressure distribution and educating individuals on the importance of regular pressure relief procedures. There remains a significant number of people with SCI for whom pressure relief cushions are inadequate and/or who are unable to maintain an adequate pressure relief regime. Periodic weight shifting is essential for maintenance of tissue health. Gluteal *neuromuscular electrical stimulation* (NMES) provides a unique technique to produce beneficial changes at the user/support system interface by altering the intrinsic characteristics of the user's paralyzed tissue itself.

For rehabilitation purposes, the effects of NMES on paralyzed muscle can be considered in terms of the activation of paralyzed neuromuscular units. SCI interrupts the normal control of muscles below the lesion which can lead to paralysis. Partially innervated muscles below the level of the lesion will become weak. Thus, muscles controlled by nerves at or below the lesion will be unable to sustain prolonged contractions. An NMES exercise program can be designed to increase both the strength and the fatigue resistance of paralyzed muscles using stimulation patterns that provide repetitive maximal

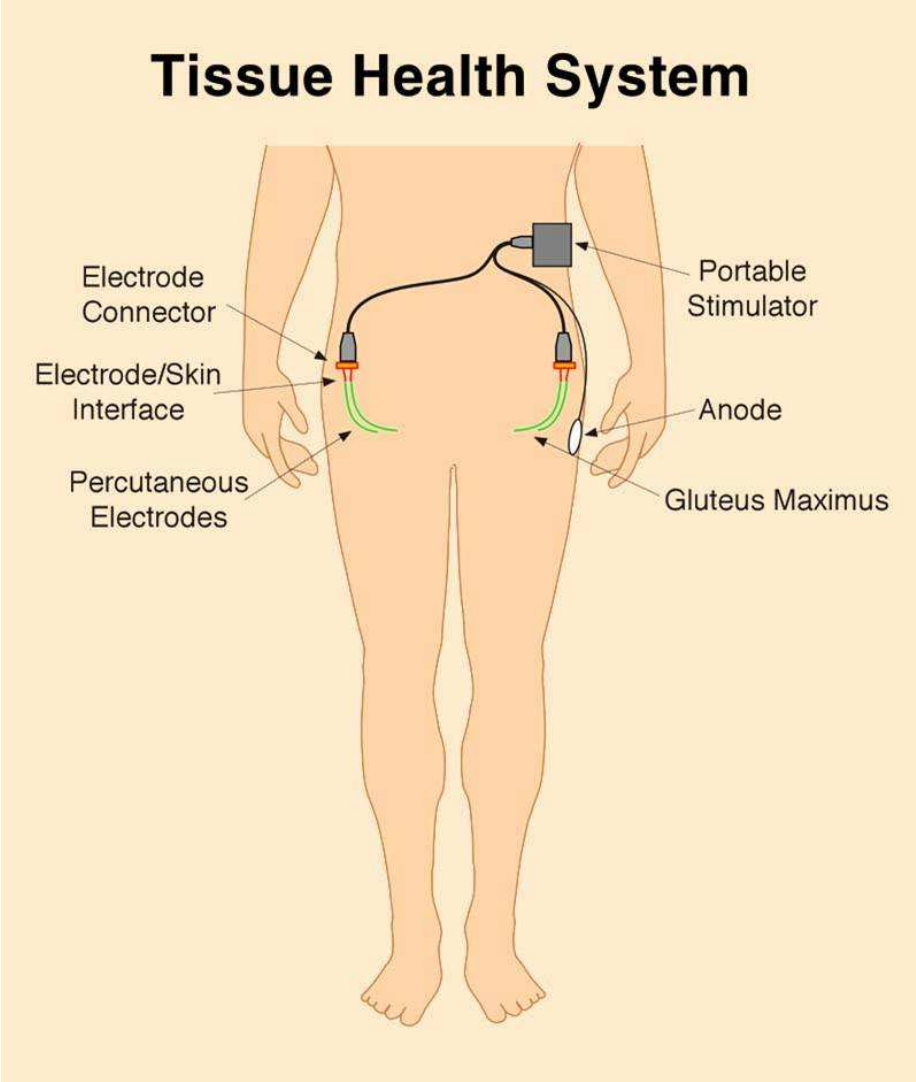


Figure 1.2: *Neuromuscular Electrical Stimulation (NMES) System. The stimulation system consists of four intramuscular electrodes connected to an external battery-powered stimulator. The percutaneous electrodes are implanted bilaterally into the gluteus maximus. The procedure is carried out under local anaesthetic on an out-patient basis.*

contractions to select muscle groups. Concurrently, muscular vascularization will start to increase as early as 4 days after initiating low-frequency electrical stimulation. It has been shown that capillary density can triple in paralyzed muscles after 2 weeks of regular moderately intensive stimulation. These changes in stimulated muscle characteristics may also improve fatigue resistance (Bogie and Triolo, 2003).

The stimulation system comprises four intramuscular electrodes connected to an external battery-powered stimulator which controls the system (Figure 1.2). The percutaneous electrodes are implanted bilaterally into the gluteus maximus. Electrode wires are routed to exit sites on the anterior thigh. The procedure is carried out under local anaesthetic on an out-patient basis. Alternating bilateral stimulation (left/right) is provided at a frequency of 20Hz. 50% active duty cycle for 3 minute period with a 17 minute inter-stimulation period. Total stimulation cycle lasts 20 minutes.

1.1.2 Experimental Method and Data Collection

The primary hypothesis of the study is that chronic use of NMES improves pressure distribution at the seating support area, specifically the reduction of peak pressures over bony prominences due to increased muscle mass area. In addition, chronic NMES will increase vascularity leading to improved tissue blood flow and resulting in improved regional tissue health in individuals with SCI.

Study participants. Repeated assessments of sitting interface pressures were obtained for a group of eight subjects with SCI participating in a study to investigate the use of NMES for standing and transfers. All subjects were full-time wheelchair users at entrance into the study and had sustained traumatic SCI from 13-204 months prior to enrollment. All subjects had completed SCI and were therefore considered to be at increased risk of tissue breakdown, in part due to disuse muscle atrophy of the glutei.

Assessment Protocol. Seating interface pressures were determined using a Tekscan Advanced Clinseat Pressure Mapping System (Tekscan Inc., Boston, Massachusetts). Assessments were carried out prior to commencing regular use of stimulation, to obtain a baseline value, and then at intervals of 3-12 months during their participation in the study, giving an overall time frame of up to five years for repeated assessments of each study participant.

In order to perform an assessment of seating interface pressures the subject transferred out of the wheelchair and a pressure sensor mat was placed over the wheelchair cushion. The sensor mat is comprised of a matrix of pressure sensitive cells (38 rows, 41 columns). The subject then transferred back into the wheelchair and was asked to sit in their customary sitting posture. Care was taken to insure that the sensor mat was not creased or folded under the subject in order to avoid inaccurate high spots. The sensor was then calibrated based on the assumption that 80 percent of the subject's body weight was acting through the seat base. Calibration took less than 20 seconds to complete. Interface pressure data was then collected for 200 seconds at a rate of 2 frames/sec. The subject was then asked to perform a pressure relief procedure and sit back in the same position. The sensor was then recalibrated and a second set of pressure data was collected at the same rate of data collection while left/right alternating gluteal stimulation was applied to provide dynamic side-to-side weight shifting for 200 seconds. Interface pressure data was collected concurrently at a rate of 2 frames/sec. Stimulation was then discontinued and subjects were asked to repeat the pressure relief procedure and sit back in the same position before collecting a third set of interface pressure data with subjects in a quiet sitting posture.

Real-time two-dimensional pressure intensity data at the seating interface are produced with the use of the Tekscan Advanced Clinseat Pressure Mapping System system (Figure 1.3).

Data. In summary, for each subject in each of sessions done over time,

Figure 1.3: *Tekscan Advanced Clinseat Pressure Mapping System. Pressures are measured through a pressure sensor mat on a standard wheel chair support surface.*

	Session																	
Subject	1			2			3			4			5			6		
1	**N**	N	N	N	**S**	N	N	S	N									
2	N	**N**	N	N	**S**	N	N	S	X	X	S	N	N	S	N	N	S	N
3	**N**	N	N	N	**S**	N	N	S	N									
4	**N**	N	N	N	**S**	N	N	S	N									
5	**N**	N	N	N	S	S	S	**Test case**										
6	**N**	N	N	N	**S**	N	N	S	N	N	S	N						
7	**N**	**S**		N	**S**	N	N	S		N	S	N						
8	**N**	N	**S**	N	S	S	N	S	S	N	S							

Legend	Pressure data assessment
N	No stimulation
N	No stimulation - baseline model data
N	No stimulation - after treatment model data
S	Dynamic electrical stimulation
S	Dynamic electrical stimulation - baseline model data
S	Dynamic electrical stimulation - after treatment model data

Figure 1.4: *Data structure in the NMES experiment. There are three sub-data sets in each of assessment sections, under condition: no stimulation, on-off alternation stimulation, and no stimulation, repectively. Each of the sub-data sets consists of sequences of data frames, totaling 400 frames over time.*

our data sets consist of three sub-data sets each of which is under one of three subsequent assessment conditions: no stimulation, on-off alternation stimulation, and no stimulation, as shown in Figure 1.4. Each of the sub-data sets consists of sequences of data frames, totaling 400 frames over time. Each data frame represents spatial pressure intensity over the sitting interface at a certain time point. The numbers of columns and rows correspond to spatial coordinates of a subject's sitting interface.

1.1.3 Spatial-temporal Data Visualization

In order to reach a quantitative understanding the data undoubtedly needs to be analyzed by valid statistical procedures. However, the first step towards qualitative understanding and interpretation of our clinical data is the visualization of the high-dimensional data. It helps us explore data and discover important features.

Data visualization enables us to explore data and information in such a way as to gain understanding and insight into the data. We propose to represent our data as images and movies in the NMES study. For each data frame we can create a grid of colored-scale rectangles with colors corresponding to the values in pressure intensity. Figure 1.5 shows the image representation for one data frame of a subject. Each element of the data frame specifies the color of a rectangular segment in the image. The colorbar in the figure indicates the mapping from data values to colors. The numbers of column and row of the image correspond to the spatial coordinates of the sitting interface of subjects.

In the movie representation, the x-axis and y-axis in three-dimensional Cartesian coordinate system denote the spatial coordinates of the sitting interface of subjects; the z-axis denotes the pressure intensities. Figure 1.6 provides a snapshot of a data movie. The intensities will move over time in the movie. Examples of movies can be found at `http://stat.case.edu/lasr`,

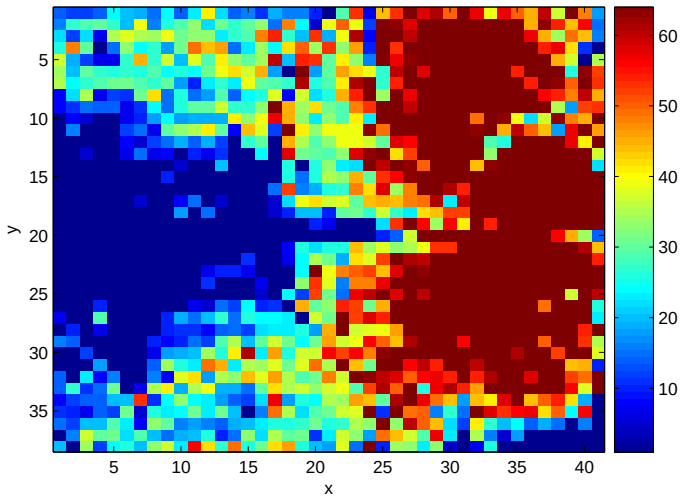


Figure 1.5: *Image representation for one data frame. Each element of the data frame corresponds to a color-scale rectangular segment in the image. The color bar indicates the mapping from data values to colors.*

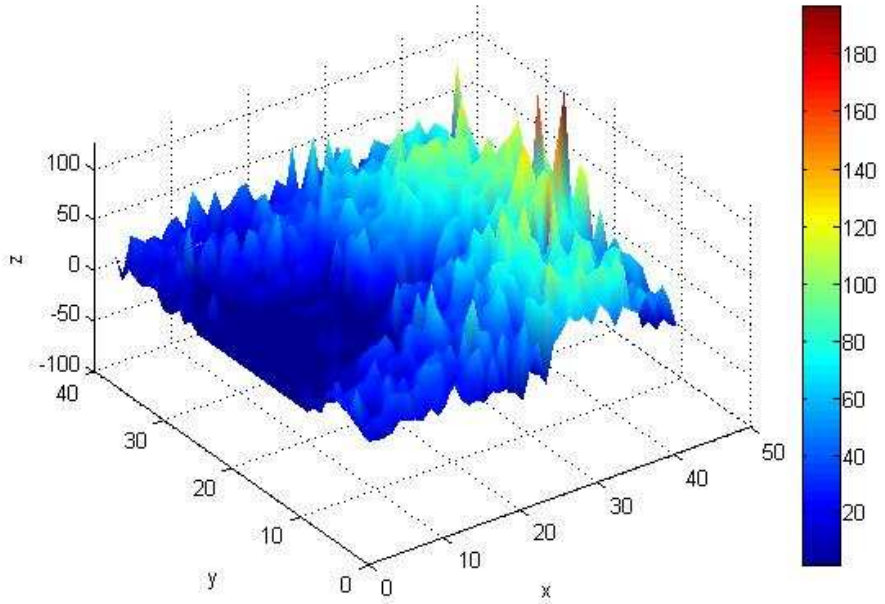


Figure 1.6: *A snapshot of movie representation for one sub-data set. The* x*-axis and* y*-axis denote the spatial coordinates of the sitting interface of subjects, respectively; the* z*-axis denotes the corresponding pressure intensities. The intensities will move over time in the movie.*

in MPEG format. In order to ensure that the analytical method was applied only to data from an inherently stable sitting posture, the initial and final ten frames for each data movie was discarded.

1.1.4 Challenges in the NMES Experiment

Recall that the primary goal of our clinical research protocol is to establish the efficacy of using gluteal NMES for the prevention of pressure ulcers. In order to achieve this objective we must define a valid quantitative method to describe the statistically and clinically significant changes in our outcomes measures, specifically data from seating interface pressure distributions. Thus we need to determine what measure or measures of seating interface pressures will be indicative of an individual's tissue health or risk and what assessment procedures are required to optimize the reliability of repeated measurements. Furthermore, the analytic method derived must include guidelines for identification of improved areas over the sitting interface.

We investigate the effects of long-term gluteal NMES on the intrinsic characteristics of the paralyzed muscles so that the response to loading, including interface pressure distribution when seated in a wheelchair, may be improved. This is generally considered to include reducing peak pressures in the ischial regions and equalizing pressures across the entire interface. Figure 1.7 shows idealized changes in pressure contour across the region of the ischial tuberosities. This is based on comparison *with no electrical stimulation.* Note that the baseline contour shows high mean interface pressures bilaterally in the ischial region which indicates a high risk of local tissue breakdown. Improved pressure distribution with reduced ischial region interface pressures and more evenly distributed seating pressures indicates a lower risk of tissue breakdown.

Clinicians are also interested in exploring the changes of interface pressure distribution during electrical stimulation. Figure 1.8 displays idealized

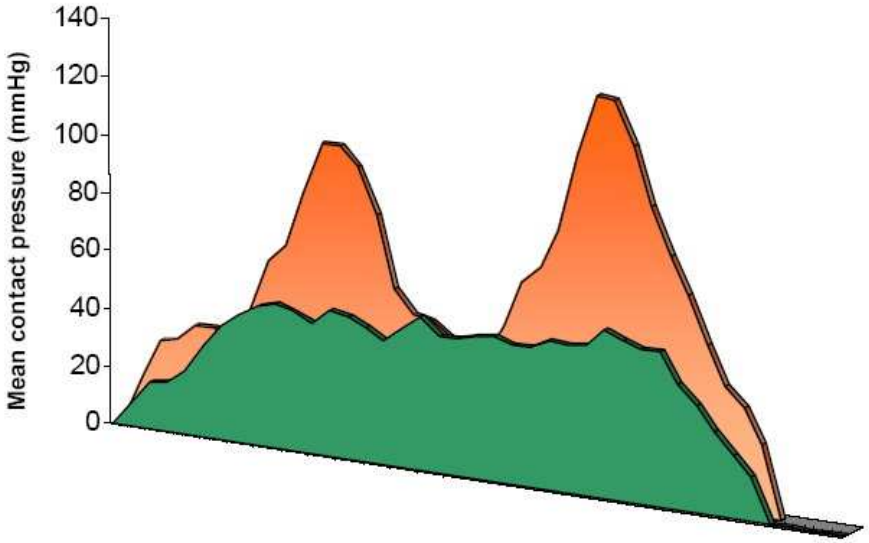


Figure 1.7: *Idealized changes in pressure contour across the region of the ischial tuberosities. 1) Baseline contour shows high mean interface pressures bilaterally in the ischial region, indicating a high risk of local tissue breakdown. 2) Improved pressure distribution with reduced ischial region interface pressures and more evenly distributed seating pressures, indicating a lower risk of tissue breakdown.*

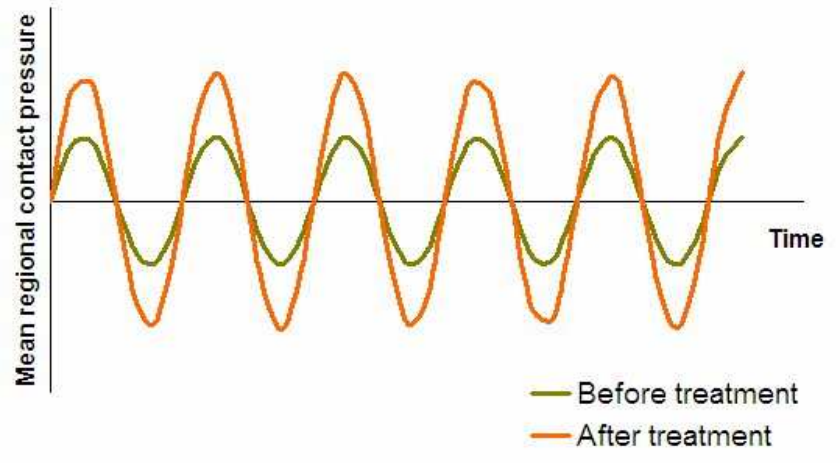


Figure 1.8: *Idealized changes in gluteal pressure variation with electrical stimulation over time. 1) Before treatment: regional interface pressures vary cyclically with applied stimulation 2) After treatment: variations about the mean increase in amplitude due to increasing strength of muscle contraction.*

changes in pressure variation based on comparison *with electrical stimulation over time*. Regional interface pressures vary cyclically with applied stimulation before treatment. Variations about the mean increase in amplitude because of increasing strength of muscle contraction after long-term treatment. In order to show whether this objective has been met over time and/or with different seating setups there must be some basis for comparison between measurements, so that true differences can be determined.

To ascertain the true difference we must overcome the following two challenges: (1) registration for a large sequence of data frames, (2) analysis of large-p-small-n data. These challenges are common in the analysis of high dimensional spatial-temporal data sets.

Registration for a Large Sequence of Data Frames

In the data mining process, the raw data often require some initial processing in order to become useful for further statistical inferences, e.g. filtering, scaling, calibration etc. Unwarping of data frames (or images) is an important stage in the NMES study. Our challenges here are:

1). Data frames recorded at different sessions over time from the same subject may not align spatially because, either the subject did not sit in the same relative position on the sensing mat or with the same posture at each assessment, or the image target regions differ from one session to another.

2). Artificial differences between alternating left/right simulation responses can obscure true differences if the data frames from different phases of the stimulation cycle are not aligned temporally between sessions.

Registration techniques have been remarkably developed in the medical imaging area. However, most existing image registration procedures require a reference image and a similarity measure for each candidate image. They

are not efficient for calibrating a large number of spatial-temporal data sets, such as registering sequences of data frames or movies in pressure mapping. For instance, it is "labor intensive" to identify the landmarks one by one for each data frame when we use corresponding landmark-based registration for thousands of data frames. Developing effective and fast spatial and temporal registration/calibration algorithms for a large volume of spatial-temporal data sets is of interest in this dissertation.

Large-p-small-n Problems

The experimental protocol for the clinical research study produced many time points and three assessment conditions for each subject. Thus, the data obtained from the NMES experiment exhibit a *large-p-small-n* problem; that is, a large number of features (pressure intensities) over space and time relative to a small number of subject samples. As mentioned in the beginning of the chapter, $N \ll T \ll S$ in our data, where N, T, S denote the number of subjects, time points, spatial locations respectively. Moreover, the pressure intensities also exhibit spatial and temporal correlation.

Several characteristics of the data complicate the application of classical statistical methodologies. Traditional statistical approaches usually are based on the assumption that $p < N$. Here $p = S \cdot T \cdot (no.\, of\ frames) \cdot (no.\, of\, sessions)$, so new approaches are needed to handle the complex data.

1.2 Measurement Error Problems

Many practical problems involve density estimation and nonparametric regression from indirect observations such as those in image deblurring, signal processing, image reconstruction in emission tomography and other applications. In the low level microarray data from either the CDNA microarray or Affymetrix GeneChip system, what is observed is an original signal cou-

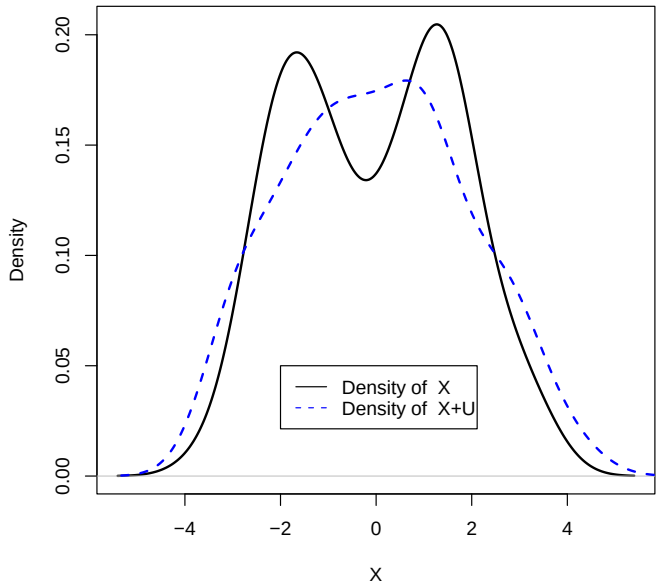


Figure 1.9: *An example to illustrate the effects of measurement errors (take from Sun and Feuerverger (2002)). The solid line denotes the true density curve of X, the dashed line denotes the density curve of $Z = X + U$ where X is measured in error by Z. It is noticed that measurement errors cause bias in the estimated density function.*

pled with a background noise. To obtain an expression measure, the goals here include developing better statistical tools or enhancing algorithms for background correction so that the disease genes can be detected accurately and efficiently. In astronomy, due to great astronomical distances and atmospheric noise, most data are subject to measurement errors. Analyses that ignore measurement errors could be misleading. Figure 1.9 gives an example which illustrates the effects of measurement errors. The example illustrates that even if the true density is bimodal, the density of the data measured with measurement error may be unimodal. Thus, finding efficient deconvolution estimates of the true density is critical.

Measurement error problems are an active, rich research field in statistics.

There is an enormous literature on this topic in linear regression, as summarized by Fuller (1987) and in nonlinear models, as summarized by Carroll et al. (1995).

There are three typical measurement error models as classified by Sun and Feuerverger (2002):

Model I:

$$Y = X + U$$

where we want to recover the density f_X of interest based on observations of Y when direct observation of X is not possible.

Model II:

$$Y = m(X) + \epsilon$$

where the goal is to estimate the regression function $m(X)$ based on observations Y with $Z = X + U$ the covariates measured in error.

Model III:

$$Y(t) = K(x(t)) + U(t)$$

where our intent is to make inferences about the target signal $x(t)$ based on output Y at t and knowledge about K, when K^{-1} can not be easily obtained and when there is non-ignorable random noise $U(t)$.

Each of the three models leads to an ongoing research area. The first model refers to *deconvolution problems* and is also related to imaging deblurring and bump hunting with measurement error. The second is known as *regression with error-in-variables*. The third one is related to *inverse problems* in signal processes or time series. In this dissertation we are concerned with issues arising from nonparametric estimating problems in the first two areas.

More specifically, the fundamental problem here is: for random variables X and Y with domains $\mathcal{X}$ and $\mathcal{Y}$ respectively, we consider the following density estimation from indirect measurements. Let $Y_1, Y_2, \cdots, Y_n$ be n independent observations from a distribution with an unknown density function $g(y)$. Our goal is to estimate another density function $f(x)$ which is related to $g(y)$ via

$$g(y) = \int_x w(y|x) f(x) dx \tag{1.1}$$

where $w(y|x)$ represents the conditional density function of $Y|X$ and is assumed known. Note that if X is a discrete random variable we need to replace the integral in (1.1) by summation. This kind of problem has been studied, for example, in Mendelsohn and Rice (1982), Snyder et al. (1992) and Vardi and Lee (1993), who focused on applications of medical image reconstruction in emission tomography.

In some cases the conditional density in (1.1) depends only on $y - x$, producing a *convolution equation*

$$g(y) = \int_x w(y - x) f(x) dx. \tag{1.2}$$

For example, under model I, w is the density function of U. Estimating f based on a sample from g is a *deconvolution problem* or *statistical inverse problem* in the sense that the sampling distribution is the image of the distribution of interest under a known transformation w:

$$g = \text{image of } f \text{ after transformation specified by } w$$

In that sense estimating f can be interpreted as to first estimate g and then to apply "some inverse transformation" of w to obtain an estimate for f.

Next, under the measurement error model I setting, we generalize to allow that w or the density of U_i does not have to be same for all i. This occurs

when we do not observe X_j but only the random variables $Y_j = X_j + U_j$. The X_j's have a common density f_X (f in (1.2)) and the additive error U_j has density w_j (w in (1.2)) for $j = 1, ..., n$. If $w_j = w_0$ for all j, the errors are *homogeneous*; otherwise, they are *nonhomogeneous*. The question then is: how can we estimate f_X based on the sample of Ys?

In the case of homogeneous errors, the deconvolution literature is vast. The Fourier-type estimates – *Deconvoluting Kernel Density Estimators* have been studied by many researchers. See, for instance, Stefanski and Carroll (1990), Carroll and Hall (1989), Fan (1991), Efromovich (1997), Wand (1998) and Cator (2001). In applications of "deblurring of images", Roy Choudhury (1998) and O'Sullivan and Roy Choudhury (2001) discuss methods for recovering images blurred by Poisson noise where the image plays the role of a density. The challenges of deconvolution problems arise when errors are nonhomogeneous where we literally have only one observation for each error distribution. Sun et al. (2002) proposed new non-Fourier estimators when errors are homogeneous or nonhomogeneous uniform. The new estimators abandon the characteristic functions - there are no Fourier transformations needed in the calculation. Following the successes of the new estimators by Sun et al. (2002), we study the non-Fourier based estimators in the case of homogeneous or nonhomogeneous normal errors and any other arbitrary error distribution.

Another interesting problem that is related to density deconvolution is nonparametric regression with errors-in-variables. Fan and Truong (1993) studied this type of problem and derived Fourier type estimators. Carroll et al. (1995) discussed two applicable functional methods, regression calibration and simulation-extrapolation (SIMEX) in their monograph. As described in the measurement error model II, the predictor X cannot be observed directly. Let (X, Y) denote a pair of random variables for which we are interested in the nonparametric estimating problem of the regression func-

tion $m(x) = E(Y|X = x)$. Due to the measuring mechanism or the nature of the environment, the covariate X_i is observed through $Z_i = X_i + U_i$ where U_i is a measurement error with known density w_i. We shall develop a new non-Fourier regression estimator and study the asymptotics and performance of the new estimator.

1.3 Outline of Rest Chapters

Complex spatial-temporal data usually require the development of an ensemble of (new) statistical tools. We will develop propose a new data-mining technique, the LASR (the abbreviation for longitudinal analysis and self-registration, pronounced "laser") procedure and a semiparametric regression model for a large sequence of spatial-temporal data sets. In the research of measurement error models we will derive new non-Fourier density and regression estimators for both homogeneous error cases and nonhomogeneous error cases.

In chapter 2, we address the data preprocessing issue in our data mining process. Two steps are proposed here, data segmentation and data registration. An optimal threshold method with EM algorithm is used to classify the sitting region in data frames. After reviewing the existing image registration methods, we introduce our new self-registration technique, *Self-Registration by a Line and a Point* (SRLP) for spatial registration incorporated by a fast temporal registration scheme *Intensity-based Correlation Registration* (ICR).

In chapter 3, we focus on the multivariate smoothing techniques and their testing procedures that are used in our data mining process. We propose a *Statistical Smoothing Mapping* (SSM) algorithm for interface pressure analysis. It leads to an efficient procedure for computing false-discovery-rate controlled movies/maps, called FDR movies and FDR maps. The control of the FDR under dependency is studied here to overcome the multiplicity

effect from testing activation pixels simultaneously.

In chapter 4, combining the techniques we developed in the previous chapters, we present a new complete data-mining scheme, the LASR procedure for analyzing a large sequence of spatial-temporal data sets. LASR is shown to be effective in the application to data from the NMES experiments. In order to model the overall treatment effects over subjects, we develop a semiparametric regression model based on Karhunen-Loève expansion and general radial spline technique. The results confirm that NMES improves seating interface pressure distributions thus reducing the risk of developing pressure ulcers.

In chapter 5 we adapt the idea of Sun et al. (2002) and develop fast and non-Fourier based nonparametric estimators of f_X — *3U estimators*, when errors are normal. The new estimators are applicable not only to homogeneous error cases but also to nonhomogeneous error cases. Moreover, because the estimators are inspired from knowledge found in *random number generation* (RNG), the ideas in developing the 3U estimators can be generalized for any arbitrary error distribution and nonhomogeneous case. These estimators are stable and easy to compute – there are no Fourier transformations needed in the calculation. The rates of their optimal estimators are $n^{-1/9}$ for the cumulative distribution function of X and $n^{-1/11}$ for the density distribution function of X. This is in contrast to the slow convergence rates of Fourier deconvolution estimators when errors are either ordinary or super smooth as defined by Fan (1991). We also develop new non-Fourier regression estimators – *SWAP estimators* and study the asymptotics and performance of the new estimators. In chapter 6 we discuss our proposed methods in the research of both spatial-temporal data mining and measurement error models and describe future research issues.

Chapter 2

Image Data Preprocessing: Segmentation and Registration

Data preprocessing is an important step in imaging, astronomy and any data mining applications. It processes raw data to prepare it for another subsequent processing or analyzing procedure.

The goal of data preprocessing is to transform the data into a format that can be more easily and effectively analyzed. The accuracy of statistical inferences will be increased. There are a number of different tools and methods used for preprocessing, such as: sampling, which selects a representative subset from a large population of data (usually done if the sample size is too large to input data all at one time for analysis); transformation, which manipulates raw data to produce a single input that can be analyzed; denoising, which removes noise from data; normalization, which organizes data for more efficient access; and feature extraction, which pulls out specified data that is significant in some particular context.

In the NMES study we propose two steps to pre-process the raw data: (1) *Data Segmentation*: this is a step for data cleaning. We distinguish between the spatial regions of interest and the background in each data frame (pres-

sure mapping) and then remove "spots" from the data sets. Segmentation is important here in that it makes the next step, registration based on random landmarks (estimated from data), more robust. 2) *Data Registration*: this is a step for data calibration. Data acquired by recording the same subject at different times and from different perspectives are in different coordinate systems. Registration is the process of transforming the different sets of data into one coordinate system. Registration is necessary both spatially and temporally in order to be able to compare or model the data obtained from different measurements.

In section 2.1, we develop a data segmentation method for the NMES data sets. In section 2.2, we will introduce the background of registration and review the existing methods for image registration. In section 2.3 we propose new spatial and temporal registration algorithms for the NMES study. The problem of registration error will also be discussed there.

2.1 Data Segmentation

In Figure 1.5, it can be seen that noise and outliers appear outside of the sitting region (*i.e.* the buttock and thigh region). It is critical to detect the edge of the sitting region of subjects and to remove the noise from the sitting region by partitioning the data frame into distinct parts.

Data segmentation here refers to the process of partitioning a data frame or an image into distinct regions by grouping together neighborhood data cells or pixels based on some pre-defined criterion. In other words, our segmentation is a data cell/pixel classification that allows the formation of regions of similarities in the data frame or image.

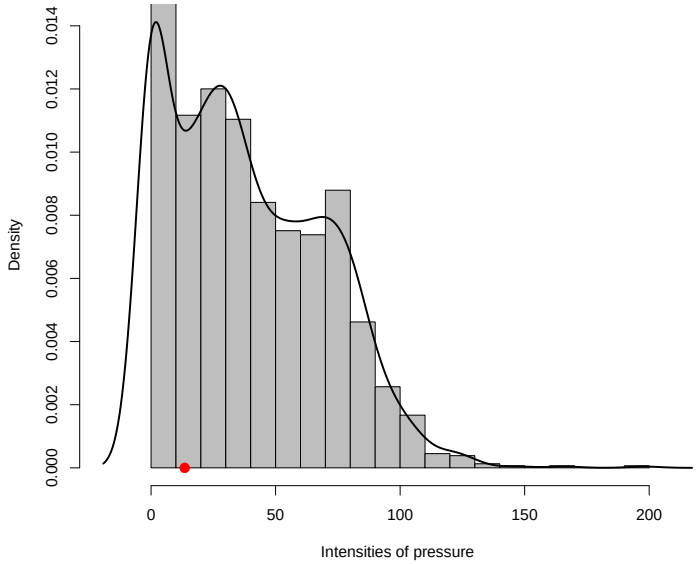


Figure 2.1: *Segmentation by analyzing the histogram and density plot of a data frame. A simple threshold is the red point which corresponds to the first deepest valley point between the first two consecutive major peaks in the density curve.*

2.1.1 Edge Detection by Histograms

We propose a histogram-based classification method to define a threshold to classify a data frame cell-by-cell. The threshold for classifying cells into classes is obtained from the analysis of the histogram or density plot of the data frame. Let $Z(i,j)$ denote the intensity value of the ith row and the jth column of a data frame. In order to remove noise and segment the sitting region of interest, the data frame can be segmented into two classes using an intensity value threshold T such that

$$\tilde{Z}(i,j) = \begin{cases} Z(i,j), & \text{if } Z(i,j) > T; \\ 0, & \text{if } Z(i,j) \leq T. \end{cases}$$

A simple approach is to examine the histogram or the density plot for multi-modal distribution. If the histogram is multi-modal, the threshold can be set to the intensity value corresponding to the first deepest point in the

histogram valley. Figure 2.1 displays the histogram and density plot of the data frame shown in the Figure 1.5. The density plot clearly displays trimodality. The simple approach to determine the threshold T is to find the first deepest valley point between the first two consecutive major peaks in the density curve. The validity of this criterion is equivalent to modeling intensities as from a mixture of two components: background (mostly small values with a unimodal distribution) and signals (which can itself be a unimodal or multimodal distribution.)

2.1.2 Data-driven EM Algorithm and Optimal Thresholding

Rather than determine the threshold visually, it is better to develop a method to find the optimal threshold. Due to the large sample size of the pressure intensities in a data frame, it is reasonable to assume the distribution of the pressure intensities is a finite mixture of m normal components, *i.e.* the density of intensities is

$$f(z) = \sum_{i=1}^{m} \alpha_i \frac{1}{\sigma_i} \phi \left(\frac{z - \mu_i}{\sigma_i} \right) \tag{2.1}$$

where ϕ is the standard normal density, and the parameters are

$$\Theta = (\alpha_1, \cdots, \alpha_m, \mu_1, \cdots, \mu_m, \sigma_1, \cdots \sigma_m),$$

such that $\sigma_i > 0$, $\alpha_i > 0$ and $\sum_{i=1}^{m} \alpha_i = 1$.

The *Expectation-Maximization algorithm* proposed by Dempster et al. (1977), popularly known as the *EM algorithm*, is a broadly applicable approach to the mixture-density parameter estimation problem. Let $\alpha_1, \cdots, \alpha_m$, $\mu_1, \cdots, \mu_m$ and $\sigma_1, \cdots \sigma_m$ denote the unknown parameters. A simple EM algorithm for computing maximum likelihood estimates of $\Theta = (\alpha, \mu, \sigma)$ is as follows.

Algorithm 2.1.1. *EM algorithm for the mixture-density parameter estimation.*

1). Provide "good" initial values of Θ*. (We recommend choosing the values based on a under-smoothed histogram and summary statistics.)*

2). Based on the sample $Z_1, \cdots, Z_n$ *compute*

$$\tau_{ij} = \frac{\alpha_i \frac{1}{\sigma_i}\phi\left(\frac{Z_j-\mu_i}{\sigma_i}\right)}{\sum_{t=1}^{m} \alpha_t \frac{1}{\sigma_t}\phi\left(\frac{Z_j-\mu_t}{\sigma_t}\right)},$$

$$\alpha_i = \frac{1}{n}\sum_{j=1}^{n} \tau_{ij}, \quad \mu_i = \frac{\sum_{i=1}^{n} \tau_{ij} Z_j}{n\alpha}, \quad \sigma_i^2 = \frac{\sum_{i=1}^{n} \tau_{ij}(Z_j-\mu_i)^2}{n\alpha_i}.$$

3). Iteratively repeat step 2 until convergence.

Table 2.1 displays the results of parameter estimation using the EM algorithm for the data frame displayed in Figure 1.5. A mixture with three normal components is used to fit the data. Figure 2.2 shows the estimated density curve by the EM algorithm with the sample histogram. The density plots of three normal components are displayed by the three curves with thin lines. It is clearly seen that the estimated density captures the bumps in the sample histogram successfully. The first normal component shows a high relative frequency due to the large repeat of zeros and nonnegativity of the sample.

After all the parameters are estimated we are able to determine an optimal threshold for data segmentation. The histograms of data in each of frames in the NMES study often exhibit tri-modality or sometimes bimodality. Next we derive the optimal threshold for tri-modal models. The optimal threshold for bimodal or multi-modal models can be similarly derived.

Suppose that the data frame can be classified into two separate regions, the background and the sitting region, where the intensities in the background follow a normal distribution $N(\mu_1, \sigma_1^2)$, the intensities in the sitting

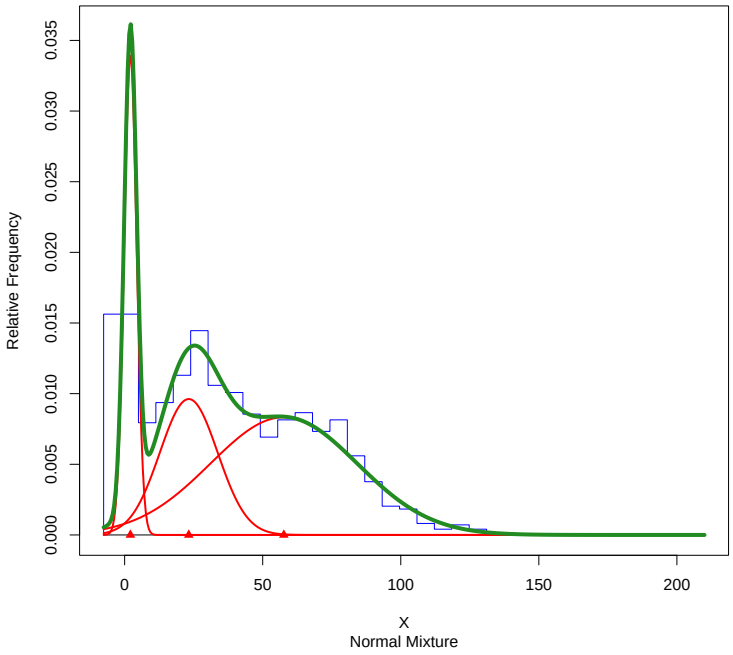


Figure 2.2: *The estimated density plots using the EM algorithm for a finite normal mixture model. The estimated mixture captures the bumps in the sample histogram sucessfully.*

	α	μ	σ
1	0.1923	2.080	2.259
2	0.2524	23.22	10.47
3	0.5553	57.65	26.59

Table 2.1: *Parameters estimation by the EM algorithm for the data frame displayed in Figure 1.5. The data is fitted by a mixture of three normal distributions.*

region follow a mixture of normal distributions with density function $g(z)$, more specifically, we have two components here. The density of intensities in a data frame (2.1) can be written as

$$f(z) = \alpha_1 f_1(z) + \alpha_2 f_2(z) + \alpha_3 f_3(z) = \alpha_1 f_1(z) + \beta g(z),$$

where

$$f_i(z) = \alpha_i \frac{1}{\sigma_i} \phi\left(\frac{z-\mu_i}{\sigma_i}\right),$$
$$\beta = \alpha_1 + \alpha_2, \quad g(z) = \frac{\alpha_1}{\alpha_1+\alpha_2} f_2(z) + \frac{\alpha_2}{\alpha_1+\alpha_2} f_3(z).$$

In this case $\mu_1 < \mu_2, \mu_3$. We define an optimal threshold T so that all cells with an intensity less than or equal to T belong to background region and all cells with an intensity greater than T belong to the sitting region. The probability of misclassifying a cell in the background as a cell in the sitting region is

$$\text{PMC}_1 = \int_T^{+\infty} f_1(z)dz,$$

and the probability of misclassifying a cell in the sitting region as a cell in the background is

$$\text{PMC}_2 = \int_{-\infty}^{T} g(z)dz.$$

Our optimal threshold is to minimize the overall probability of misclassification

$$\min_T \text{PMC} = \min_T \{\alpha_1 \text{PMC}_1 + \beta \text{PMC}_2\},$$

i.e. find the threshold value T that satisfies

$$\min_T \left\{ \alpha_1 \int_T^{+\infty} f_1(z)dz + \alpha_2 \int_{-\infty}^{T} f_2(z)dz + \alpha_3 \int_{-\infty}^{T} f_3(z)dz \right\}.$$

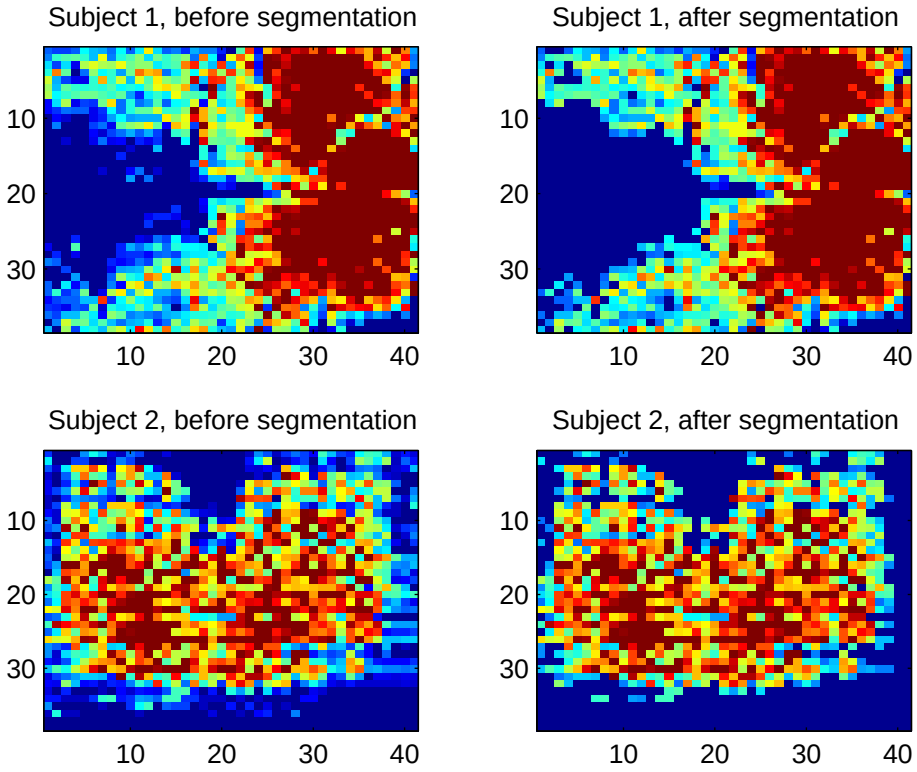


Figure 2.3: *Examples of comparison of images after data segmentation using optimal thresholds. The upper two subplots are for Subject 1 and the lower two subplots are for Subject 2. Note that the sitting regions in the data frames are segmented and the background noises are removed.*

Differentiating PMC with respect to T by *Leibnitz's Rule* and setting $d\text{PMC}/dT = 0$, we have

$$\alpha_1 f_1(T) = \alpha_2 f_2(T) + \alpha_3 f_3(T). \tag{2.2}$$

Since the parameters Θ in (2.2) can be obtained by the EM algorithm, one can easily obtain the optimal threshold by using *Mathematica* or *Maple* software to solve (2.2).

Figure 2.3 shows two examples of our data segmentation methods using the EM algorithm and optimal thresholds. The optimal thresholds in the data frames of subject 1 and subject 2 are 12.7 and 14.3, respectively. The sitting regions in the data frames are clearly segmented and the background noise is removed. This data pre-processing step is valuable in that it will make our next data pre-processing step – data registration much more robust.

2.2 Introduction to Registration

Registration is the process of transforming the different sets of data into the same coordinate system. It has applications in many fields. The past 25 years have seen remarkable developments regarding registration techniques in image analysis. We briefly review the background and existing methods of image registration.

Image registration is the process of systematically placing separate images in a common frame of reference so that the information they contain can be optimally integrated or compared. This plays a central role in analysis, interpretation and visualization of both medical and other images. In many clinical scenarios images from several modalities may be acquired and without a registration the diagnostician's task would be mentally combine or "fuse" this information to draw useful clinical conclusions. This generally requires mental compensation for changes in subject position. Image registration aligns the images and so establishes correspondence between different

features seen on different imaging modalities, allows monitoring of subtle changes in size or intensity over time or across a population, and establishes correspondence between images and physical space in image guided interventions. Registration of an atlas or computer model aids in the delineation of anatomical and pathological structures in medical images as an important precursor to detailed analysis.

More specifically, registration is the determination of a geometrical transformation that aligns points in one view of an object with corresponding points in another view of that object or another object. We use the term "view" generically to include a three-dimensional image, a two-dimensional image, or the physical arrangement of an object in space. Three-dimensional images are acquired by tomographic modalities, such as *computed tomography* (CT), *magnetic resonance imaging* (MRI), *single-photon emission computed tomography* (SPECT) and *positron emission tomography* (PET). In each of these modalities a contiguous set of two-dimensional slices provides a three-dimensional array of image intensity values. Typical two-dimensional images may be x-ray projections captured on film or as a digital radiograph or projections of visible light captured as a photograph or a video frame. In all cases we are concerned primarily with digital images stored as discrete arrays of intensity values. In medical applications, which are our focus, the object in each view will be some anatomical region of the body. Figure 2.4 displays an example of registration procedure using the *iterative closest point* (ICP) algorithm for MR images.

2.2.1 Transformations in Registration

Image registration can be defined as a mapping between two images both spatially and with respect to intensity. Let us define these images as two 2-dimensional (or 3-dimensional) and denote their intensities by I_1 and I_2

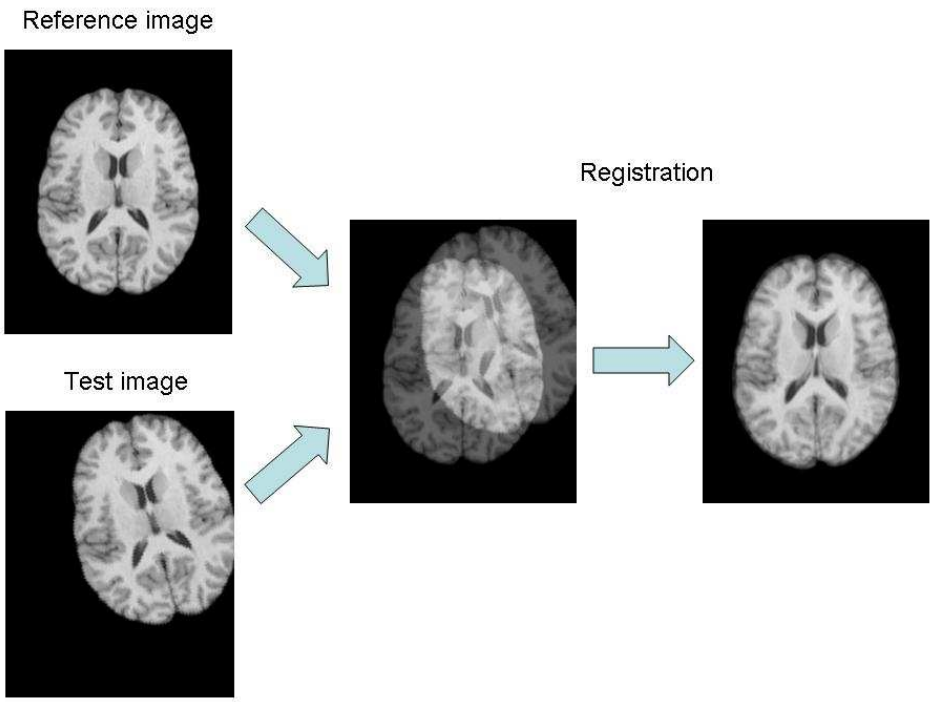


Figure 2.4: *Example of registration procedure using iterative closest point algorithm for MR images.*

respectively. The mapping between images can be expressed as,

$$I_2(\mathbf{x}_2) = g(I_1(T(\mathbf{x}_1)))$$

where $T(\cdot)$ is a 2-dimensional (or 3-dimensional) spatial coordinate transformation and $g(\cdot)$ is an 1-dimensional intensity transformation.

The registration problem involves finding the optimal spatial and intensity transformations so that the images are matched with regard to the misregistration source. However, the intensity transformation is not frequently necessary except in the case where there is a change in sensor type (such as from optical to radar). In this sense we are interested in finding T such that,

$$T : \mathbf{x}_1 \mapsto \mathbf{x}_2 \Leftrightarrow T(\mathbf{x}_1) = \mathbf{x}_2.$$

The primary general transformations for images are rigid, affine, projective, and curved. These are all well-defined mappings of one image onto another.

Rigid Transformations

Rigid transformations are defined as geometrical transformations that preserve all distances. These transformations also preserve the straightness of lines (and the planarity of surfaces) and all nonzero angles between straight lines. Rigid transformations are simple to specify and there are several methods of doing so. In each method there are two components to the specification, a translation and a rotation. In three dimensions there are six parameters which can be defined as translation in the x, y and z directions, and rotations α, β and γ about these three axes. The rigid transformation can be represented as a rotation $\mathbf{R}$ followed by a translation $\mathbf{t}$ that can be applied to any point $\mathbf{x} = (x, y, z)^T$,

$$\mathbf{T}_r(\mathbf{x}) = \mathbf{R}\mathbf{x} + \mathbf{t}$$

where $\mathbf{t} = (t_x, t_y, t_z)^T$ and the rotation matrix $\mathbf{R}$ is constructed from the rotation angles as follows:

$$\mathbf{R} = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix}$$

$$= \begin{bmatrix} \cos\beta\cos\gamma & \cos\alpha\sin\gamma + \sin\alpha\sin\beta\sin\gamma & \sin\alpha\sin\gamma - \cos\alpha\sin\beta\cos\gamma \\ -\cos\beta\sin\gamma & \cos\alpha\cos\gamma - \sin\alpha\sin\beta\sin\gamma & \sin\alpha\cos\gamma + \cos\alpha\sin\beta\sin\gamma \\ \sin\beta & -\sin\alpha\cos\beta & \cos\alpha\cos\beta \end{bmatrix}$$

Affine Transformations

An *Affine transformation* is a non-rigid transformation. It preserves the straightness of lines and the planarity of surfaces. It preserves parallelism but allows angles between lines to change. The affine transformation can be represented as

$$\mathbf{T}_a(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{t}$$

where there is no restriction on the elements a_{ij} of the matrix $\mathbf{A}$.

Affine transformations are sometimes represented by means of *homogeneous coordinates*. In this representation both $\mathbf{A}$ and $\mathbf{t}$ are folded into one 4×4 matrix $\tilde{\mathbf{A}}$,

$$\begin{bmatrix} x_2 \\ y_2 \\ z_2 \\ 1 \end{bmatrix} = \tilde{\mathbf{A}}\mathbf{x} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & t_x \\ a_{21} & a_{22} & a_{23} & t_y \\ a_{31} & a_{32} & a_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ 1 \end{bmatrix}$$

Projective Transformations

Projective transformations are more general non-rigid transformations which preserve the straightness of lines and planarity of surfaces. However, parallelism between straight lines is in general not preserved. Projective transformations can be represented by,

$$\mathbf{T}_{pr}(\mathbf{x}) = (\mathbf{A}\mathbf{x} + \mathbf{t})/(\mathbf{p}\mathbf{x} + \mathbf{u})$$

and can be written in homogeneous coordinates,

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \tilde{\mathbf{P}}\mathbf{x} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & t_x \\ a_{21} & a_{22} & a_{23} & t_y \\ a_{31} & a_{32} & a_{33} & t_z \\ p_x & p_y & p_z & u \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ 1 \end{bmatrix},$$

$$\begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} = \begin{bmatrix} v_1/v_4 \\ v_2/v_4 \\ v_3/v_4 \end{bmatrix}$$

Curved Transformations

Curved transformations are those which do not preserve the straightness of lines. For instance, they may map a straight line onto a curve. The simplest functional form for $\mathbf{T}(\cdot)$ in curved transformations is a polynomial in the components of $\mathbf{x}_1 = (x_1, y_1, z_1)^T$,

$$\mathbf{x}_2 = \mathbf{T}_c(\mathbf{x}_1) = \sum_i^I \sum_j^J \sum_k^K \mathbf{c}_{ijk} x_1^i y_1^j z_1^k$$

where $\mathbf{c}_{ijk}$ is the three-dimensional vector of coefficients for the i, j, k term in the polynomial expression for the three components x_2, y_2, z_2 of $\mathbf{x}_2$.

Other curved transformations such as *cubic spline*, *B-spline*, *thin-plate spline* methods have been widely used for two-dimensional image problems.

2.2.2 Current Image Registration Methods

Different imaging modalities bring complementary information that can be advantageously used to establish a diagnosis or assist the clinician for a therapeutic gesture. To locally compare two or more measurements of different natures a number of registration algorithms have been developed, especially in brain imaging.

Registration is often necessary for 1) integrating information taken from different sensors, 2) finding changes in images taken at different times or

under different conditions, 3) inferring three dimensional information from image in which either the camera or the objects in the scene have moved and 4) for model-based object recognition (Rosenfeld and Kak, 1982). To register two images a transformation must be found so that each point in one image can be mapped to a point in the second. This mapping must "optimally" align the two images where optimality depends on what needs to be matched.

Gerlot and Bizais (1988) have presented a unified description of existing registration methods. They propose the following general registration methodology: 1) extraction of features in each images, 2) pairing of these features, 3) choice of a geometric transformation and estimation of its parameters, and 4) effectuation of this transformation. They classify registration methods into four categories, in which the above four steps are implemented differently: 1) point methods, 2) edge methods, 3) moment methods, and 4) similarity criterion optimization methods. An extensive classification scheme for registration methods has also been presented in van den Elsen et al. (1993). They classify techniques according to a number of criteria: 1) dimensionality (1D vs. 2D vs. 3D), 2) type of features using for registration (intrinsic vs. extrinsic), 3) domain of the transformation (local vs. global), 4) type of transformation (rigid vs. affine vs. projective vs. curved), 5) parameter determination (search vs. closed-form solution), and 6) interaction (interactive vs. semi-automatic vs. automatic). For a detailed survey and review of existing image registration techniques, see Maurer and Fitzpatrick (1993); van den Elsen et al. (1993); Maintz and Viergever (1998); Fitzpatrick et al. (2000); Hill et al. (2001).

The following briefly reviews some of the existing registration methods. We divide our review of registration techniques into two main categories: 1) those based on geometric image features, and 2) those based on voxel similarity measures. The geometric image feature-based methods are divided into registration of a set of points and edges or surfaces. Registration methods

based on voxel similarity measures include intensity difference and correlation methods and methods based on joint entropy or mutual information.

Registration Methods Based on Geometric Features

Point-based Methods

Point-based registration methods, or *corresponding Landmark-based Registration* often use external markers or anatomical landmarks. Corresponding point sets are usually manually defined in the reference and floating images. The advantages of the point-based registration methods are that they can be applied to any imaging modalities where markers or landmarks are visible and that the calculation of the registration parameters between two point sets is usually fast.

A noniterative least squares method can be used to register corresponding point sets. The method uses a *singular value decomposition* (SVD) of a 3×3 covariance matrix to find a unique solution for the registration parameters between two point sets. For example, Algorithm 2.2.1 provides a method for point-based rigid registration. It is desirable to find $\mathbf{R}$ and $\mathbf{t}$ which minimize $\sum_{i=1}^{n} w_i^2 |\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i|^2$ where $\mathbf{x}_i$ and $\mathbf{y}_i$ $(i = 1, \cdots, n)$ are the corresponding landmarks in image $\mathbf{X}$ and $\mathbf{Y}$ respectively, and w_i is some non-negative weighting factor.

Algorithm 2.2.1. *Point-based rigid registration*

1). Calculate the weighted centroid of the landmarks in each image,

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} w_i^2 \mathbf{x}_i}{\sum_{i=1}^{n} w_i^2}, \qquad \bar{\mathbf{y}} = \frac{\sum_{i=1}^{n} w_i^2 \mathbf{y}_i}{\sum_{i=1}^{n} w_i^2}$$

2). Compute the displacement from the centroid to each landmark in each image,

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}, \qquad \tilde{\mathbf{y}}_i = \mathbf{y}_i - \bar{\mathbf{y}}$$

3). compute the weighted covariance matrix,

$$\mathbf{C} = \sum_{i=1}^{n} w_i^2 \tilde{\mathbf{x}}_i \tilde{\mathbf{y}}_i^T$$

4). Perform singular value decomposition of $\mathbf{C}$*,*

$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$*,* $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$*,* $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$*.*
We have $\mathbf{R} = \mathbf{V}\,\mathrm{diag}(1, 1, \det(\mathbf{VU}))\,\mathbf{U}^T$*, and* $\mathbf{t} = \bar{\mathbf{y}} - \mathbf{R}\bar{\mathbf{x}}$*.*

Point-based affine registration, point-based projective registration and point-based curved registration are also available. These landmark-based registrations are widely applied in medical image registration because they allow matching of any imaging modalities in which the positions of markers can be accurately defined.

Surface-based Methods

Corresponding surfaces may be identified and used for registration. In these algorithms corresponding surfaces are delineated in the two imaging modalities and a transformation computed that minimizes some measure of distance between the two surfaces. The first widely used method was the "head-and-hat" algorithm, but the most popular method now is the iterative closest point algorithm.

The *head and hat* algorithm was proposed by Pelizzari et al. (1989) who were the first investigators to apply surface-based registration to a medical problem. They used the algorithm to register CT, MR, and PET images of the head. The "head" is the contours of the surface drawn on a series of slices from one modality; the "hat" is a set of identified points that correspond to the same surface in the other modality. The computer then attempts to fit the hat points on the head contours iteratively. At each iteration the sum of

the squares of the distances between each hat point and the head is calculated and the process continues until the value is minimized. As its name implies, this was first used on images of the head and, especially, the alignment of MR and PET images. Unfortunately, just as there are many ways of placing a real hat on a head, this algorithm can lead to a wrong solution. These types of algorithms tend to fail when the surfaces show symmetries to rotation, which is often the case for many anatomical structures.

The *iterative closest point* (ICP) algorithm, first proposed by Besl and McKay (1992), has been widely applied to surface-based registration of medical images. They presented an algorithm which reduces the general nonlinear minimization problem to an iterative point-based registration problem. The ICP algorithm is a general-purpose, representation-independent, shape-based registration algorithm that can be used with a variety of geometrical objects including point sets, line segment sets, triangle sets and implicit and parametric curves and surfaces.

In the most usual form of this algorithm, one surface is represented by a set of points while the other is represented by a surface made up of many triangular patches or "facets". The algorithm proceeds by finding the closest point on the appropriate triangular patch to each of the points in turn. The closest points form a set and these are registered using the corresponding landmark-based registration and then the residual error is calculated. The closest points are found from this new position and the process is repeated until the residual error decreases to less than some preset value.

The ICP algorithm is described in more detail in the next chapter. It uses more of the available data than landmark identification. It is robust, accurate and has been reported in many applications. Unfortunately, the technique is highly dependent on identification of corresponding surfaces, yet different imaging modalities can provide very different image contrast between corresponding structures.

Registration Methods Based on Similarity Measures

Registration using voxel similarity measures involves calculating the registration transformation **T** by optimizing some measure calculated directly from the voxel values (or pixel values) in the images rather than from geometrical structures such as points or surfaces derived from the images. For registration using voxel similarity measures it is very important to distinguish between registration where images are from the same modality (*intramodality*) and registration where images are from different modalities (*intermodality*).

Intramodality registration using voxel similarity measures

A common reason for carrying out the intramodality registration is to compare images from a subject taken at slightly different times in order to ascertain whether there have been any subtle changes in anatomy or pathology. If there has been no change in the subject we might expect that, after registration and subtraction, there will be no structure in the difference image, just noise. Where there is a small amount of change in the structure we would expect to see noise in most places in the images, with a few regions visible in which there has been some change. There can be considerable clinical benefit to accurately aligning images of the same subject acquired with the same modality at different times in order to detect subtle changes in intensity or shape of a structure. This technique is most widely used for aligning serial MR images of the brain.

One of the simplest voxel similarity measures is the *sum of squared intensity differences* (SSD) between images, which is minimized during registration. It can be shown that this is the optimum measure when two images differ only by Gaussian noise. It is obvious that this will never be the case for intermodality registration. This strict requirement is not often true in intramodality registration either, as noise in medical images such as modulus MRI scans is frequently not Gaussian. The SSD measure makes the implicit

assumption that after registration the images differ only by Gaussian noise. A slightly less strict assumption would be that, at registration, there is a linear relationship between the intensity values in the images. In this case, the optimum similarity measure is the *correlation coefficient* (CC). Another popular intramodality registration, the *ratio image uniformity* (RIU) algorithm was introduced by Woods et al. (1992) for the registration of serial PET studies but has more recently been widely used for serial MR registration (Woods et al., 1998a,b). The algorithm can be thought of as working with a derived ratio image calculated from images A and B. An iterative technique is used to find the transformation T that maximizes the uniformity of this ratio image which is quantified as the normalized standard deviation of the voxels in the ratio image.

Intremodality registration using voxel similarity measures

Because of the similarity of the intensities in the intramodality images being registered the subtraction and correlation techniques described above have an intuitive basis. With intermodality registration the situation is quite different. There is, in general, no simple relationship between the intensities in the images. Any algorithm that is used to register images from two different modalities must be insensitive to modality-specific differences in image intensity associated with the same tissue and also accommodate differences in relative intensity from tissue to tissue.

The first successful application of a voxel similarity-based algorithm to the registration of images from different modalities was the *partitioned intensity uniformity* (PIU) algorithm proposed by Woods et al. (1993) for MR-to-PET registration. The algorithm assumes that at each intensity in the MR image the range of the corresponding PET intensities is small. Implementation involved an almost trivial change to the source code of their previously published RIU technique but proved to be robust for the registration of MR and PET images of the head, provided the scalp was first removed from the MR

images.

Registration can be thought of as reducing the amount of information in the combined image, which suggests the use of a measure of information as a registration metric. The most commonly used measure of information in signal and image processing is *Shannon entropy* H, which is widely used as a measure of information in many branches of engineering. Originally developed as part of information theory in the 1940s (Shannon, 1948a,b), it describes the average information supplied by a set of symbols s whose probabilities are given by $p(s)$,

$$H = -\sum_{s} p(s) \log(p(s)).$$

Initially it seems that image registration has little to do with measuring the amount of information being transmitted down a communication channel. The use of entropy and other information-theoretic measures for image registration came about, however, after inspection of joint histograms and probability density functions. It was proposed by Collignon et al. (1995); Studholme et al. (1995) that the entropy of the joint histogram calculated from images A and $T(B)$ should be iteratively minimized to register these images. Minimizing joint histogram entropy to register images may be considered an extension of PIU minimization.

Joint entropy on its own does not provide a robust voxel similarity measure for all types of image registration. The *mutual information* measure with modifications associated with normalization (Studholme et al., 1996, 1997, 1999) has proved fairly robust and has resulted in fully automated 3D-to-3D rigid-body registration algorithms that are now in widespread use. In maximizing mutual information, we seek solutions that have a low joint entropy together with high marginal entropies. The mutual information is

defined by

$$I(A,B) = H(A) + H(B) - H(A,B) = \sum_a \sum_b p_{AB}(a,b) \log \frac{p_{AB}(a,b)}{p_A(a)p_B(b)}$$

where $H(A), H(B), H(A,B)$ denote the entropy of A, B and the joint entropy A and B, respectively.

2.3 Registrations Procedures for the NMES study

In the NMES study the experimental methodology produced a large volume of data for a relatively small subject population. Interface pressure data stored as discrete arrays of intensity values can be represented as digital images. The data frame or image object was the anatomical seating region of the body, specifically the buttock and thigh region. The current experimental protocol entailed obtaining several data sets from each subject during their participation in the experiment. Since a subject may not sit at the same relative position on the sensor mat or with the same posture as before, or the image target regions may differ from one session to another, it was possible that some spatial change might exist between data sets from different sessions. This is the case in our clinical study as shown in Figure 2.5 and Figure 2.6 where misalignment for some subjects is more obvious than the others. Since the subjects were not restrained in any way during the assessment it was also possible for some change in seating orientation to occur from one assessment condition to another during the same session. Thus in order to determine any changes due to the effect of using NMES we first had to ensure that any changes due solely to seating orientation were fully compensated. This was achieved by spatial registration.

In the middle segment of each session as shown in Figure 1.4, a left/right alternating stimulation is given to a subject. To compare middle segments

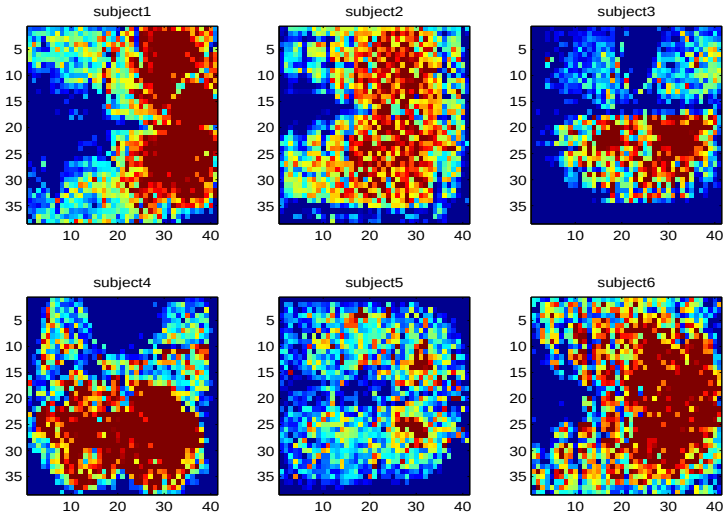


Figure 2.5: *Raw data frames with representation as images for the subjects before treatment*

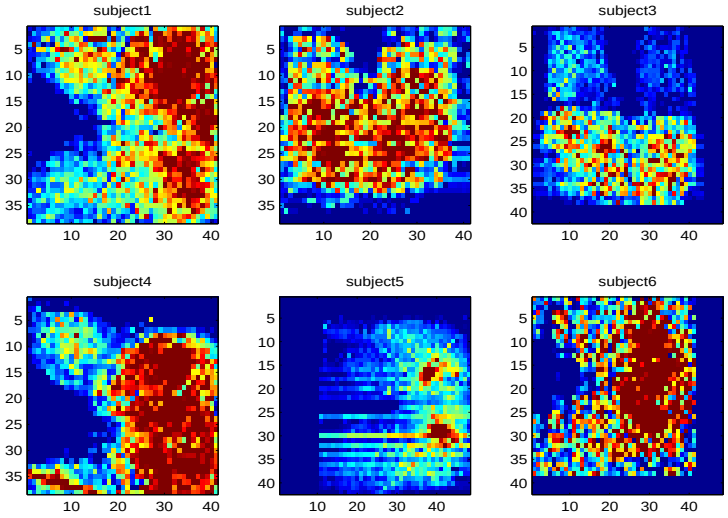


Figure 2.6: *Raw data frames with representation as images for the subjects after treatment*

from two sessions a temporal registration is necessary to avoid artificial differences caused by stimulation cycle phase obscuring true image differences due to treatment.

Next we present methods for developing two types of intra-patient registration, spatial registration and temporal registration. Based on the successful outcome from this study a larger Phase II study with a larger subject population to permit an effective examination of potential differences across the entire seat region between different groups of patients by age, gender and health conditions can be planned. Inter-subject registration might be necessary with a more diverse study population, for example, to quantify specific differences at specific locations of the seating interface region in different groups of patients.

2.3.1 A New Spatial Registration Scheme: SRLP

Figures 2.5 and 2.6 display the first still images from each of six movies (representing six subjects) at the first segment of the first session (before intervention) and at the third segment of the last session (after intervention). Some of these images for identical subjects have different orientations and some of them cover different image target regions. For example, the fourth image in the second row for the fourth subject has been rotated 90 degrees in the last session. The last image for subject 6 has non-overlapping areas between two images. Thus spatial registration is necessary to align images from different sessions. Non-overlapping regions will be chopped out or trimmed during final analysis. Fortunately, the images within one segment in one session, and between different data sets in one session, do not appear to need a spatial registration. Thus, the first (stable) image of the first movie in each session can be used as a reference to register or align movies from different sessions, before we compute difference images or movies for statistical analysis of clinical relevance.

The first step in achieving a spatially registered image pair is to define a coordinate system for each image, thus defining a space for that image. Registration is based on geometrical transformations, which are mappings of points from the space $\mathcal{A}$ of one view to the space $\mathcal{B}$ of a second view. Thus, the transformation $\mathbf{T}$ applied to a point in $\mathcal{A}$ represented by the column vector $\mathbf{a} = (a_i, a_j)^T$ produces a transformed point $\mathbf{a}' = (a_i', a_j')^T$,

$$\mathbf{a}' = \mathbf{T}(\mathbf{a})$$

If the point $\mathbf{b} = (b_i, b_j)^T \in \mathcal{B}$ corresponds to $\mathbf{a}$, then a successful registration will make $\mathbf{a}'$ equal, or approximately equal, to $\mathbf{b}$. Any nonzero displacement $\mathbf{T}(\mathbf{a}) - \mathbf{b}$ is a registration error.

SRLP Algorithm

For spatial registration of seating pressure distribution images, the key is to choose appropriate landmarks and estimate the landmark "midline" of the seating contact area for each patient. The midline and an obvious "end" point in each image will be used as our landmarks for registration which leads to a midline-to-midline and endpoint-to-endpoint alignment. A scale change of images is not expected unless a subject has a significant change in body weight between two sessions. Thus we propose the following algorithm for spatial registration of images based on a midline and an end point.

Algorithm 2.3.1. *Spatial Registration by a line and a point (SRLP)*

1). Determine the midpoints for each image,

$$\text{midpt} = \frac{\text{rowcount}}{2} + \frac{(c_1 - c_2)}{2}$$

where c_1 = the number of non-zero values from the lower half image, c_2 = the number of non-zero values from the upper half image, and the

rowcount is the total number of non-zero values in each column of the image.

2). *Determine the midline. The midline is the regression line estimated by fitting a simple regression to the midpoints.*

3). *Perform a rigid transformation based on the midline, by rotation and translation through matrix* $\mathbf{R}$

$$\begin{bmatrix} a_i' \\ a_j' \\ 1 \end{bmatrix} = \mathbf{R} \begin{bmatrix} a_i \\ a_j \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & u \\ \sin\theta & \cos\theta & v \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_i \\ a_j \\ 1 \end{bmatrix}$$

where $\tan\theta$ *is the slope of the midline and* (u, v) *is the last point of the fitted midline in the image that is to be transformed.*

If the patient is sitting asymmetrically the two halves of the image will have an unequal number of non-zero pixel values. For example, if the patient is leaning toward the lower half of the image there will be more non-zero pixel values in the lower half than in the upper half of the image, i.e $c_1 > c_2$. A positive correction $(c_1 - c_2)/2$ to the rowcount/2 should then be applied so that the location of the midpoint value moves up. After computation of the corrected midpoints, the midline can readily be found through linear regression. In Figure 2.7, the upper graph displays the midline of a patient in one frame; the lower graph displays the images after spatial registration for the same subject.

Remark. *The idea of SRLP is simple, but is highly effective. It allows for self-registering any image, and can correct the bias and save the labor in determining the middle line manually. It is also a consistent algorithm in statistical sense for a random landmark registration problem (as shown in Theorem 2.3.1).*

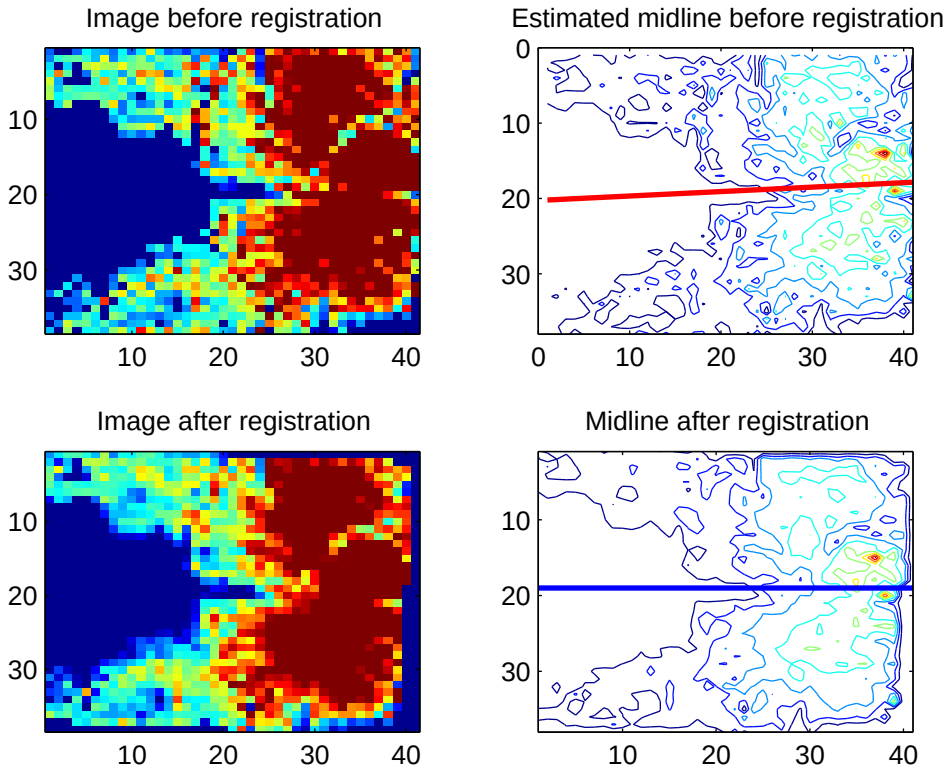


Figure 2.7: *An example of spatial registration by a line and a point (SRLP). The middle line is determined by a simple linear regression and rigid transformation is used in the registration.*

Registration Error of SRLP

As mentioned at the beginning of this section, any nonzero displacement between a transformed point $\mathbf{T}(\mathbf{a})$ and its corresponding point $\mathbf{b}$ is an individual registration error. A naive measure of overall misalignment of a registration is the mean square error (MSE). Overall *registration error* (RE) of SRLP is defined as follows.

$$\mathrm{RE} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} ||\mathbf{T}_{SRLP}(\mathbf{a}_{ij}) - \mathbf{b}_{ij}||^2 \tag{2.3}$$

where $\mathbf{a}_{ij}$ and $\mathbf{b}_{ij}$ $(i, j = 1, \cdots, n)$ are the corresponding points (*i.e.* coordinates of data cells) in spaces $\mathcal{A}$ and $\mathcal{B}$, respectively.

Theorem 2.3.1. *Assume that the intensity values are bounded and we are interested in a bounded domain. The overall registration error of spatial registration by a line and a point tends to zero in probability as the number of pixels increases. In other words, the SRLP is consistent.*

Proof: After SRLP registration, the $\mathbf{a}'_{ij} = (a'_i, a'_j)$ has the representation

$$a'_i = a_i\cos\hat{\theta} - a_j\sin\hat{\theta} + u, \qquad a'_j = a_i\sin\hat{\theta} + a_j\cos\hat{\theta} + \hat{v}$$

where $\tan\hat{\theta} = \hat{\beta}_0$, $\hat{v} = \tan\hat{\theta}u + \hat{\beta}_1$, and $\hat{\beta}_0$, $\hat{\beta}_1$ are the estimates of the slope and intercept of the midline. Notice that u is not an estimated value because the horizonal axis of the last point in the fitted midline keeps immovable.

A perfect registration will make the transformed point equal to $\mathbf{b}_{ij}$, so

$$b'_i = a_i\cos\theta - a_j\sin\theta + u, \qquad b'_j = a_i\sin\theta + a_j\cos\theta + v$$

Then the registration error of SRLP is equal to

$$\mathrm{RE} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big\{[a_i(\cos\hat{\theta} - \cos\theta) - a_j(\sin\hat{\theta} - \sin\theta)]^2 + [a_i(\sin\hat{\theta} - \sin\theta) - a_j(\cos\hat{\theta} - \cos\theta) + (\hat{v} - v)]^2\Big\}.$$

Note that $\hat{\beta}_0$, $\hat{\beta}_1$ are consistent estimators in the midline regression (see details in the Appdenix), and $\theta = g(\beta_1, \beta_2)$, $v = h(\beta_1, \beta_2)$ where g and h are continuous functions. Hence, $\hat{\theta} = g(\hat{\beta}_0, \hat{\beta}_1)$ and $\hat{v} = h(\hat{\beta}_0, \hat{\beta}_1)$ are also consistent by *Slutsky's theorem*. Then by the boundedness of intensities and a_i, a_j, it is easy to see that RE $\to 0$ in probability as $n \to \infty$ or the number of pixels tends to infinity. □

2.3.2 A Temporal Registration Scheme: ICR

As part of the assessment protocol for this study electrical stimulation of the gluteal muscles was applied to produce dynamic weight-shifting from side to side. Temporal registration is required to align stimulation periods (on-off times) for all data sets collected for one subject under the same assessment conditions. Temporally registered data may then be further analyzed to determine the effects of dynamic weight shifting over time, *i.e.* over more than one assessment.

If the intensities in images A and B are linearly related, then the correlation coefficient can be shown to be the ideal similarity measure. Few registration applications will precisely conform to this requirement, but many intra-modality applications, such as aligning on-off signals for two simulation sessions in our case, come sufficiently close for this to be a useful measure.

If there were a registration error, we would expect to see artefactual structure in the difference image resulting from the poor alignment. In this application, various voxel similarity measures suggest themselves. We could, for example, iteratively calculate $\mathbf{T}$ while minimizing the structure in the difference image on the grounds that at correct registration there will be either no structure or a very small amount of structure in the difference image, whereas with increasing misregistration, the amount of structure would increase. The structure could be quantified, for example, by the sum of squares of difference values, or the sum of absolute difference values or the

entropy of the difference image. An alternative, intuitive approach (at least for those familiar with signal processing techniques) would be to find **T** by cross correlation of images **A** and **B**.

Algorithm 2.3.2. *Temporal Registration (ICP)*

1). Discard the first m unstable data frames from each of the sub-data sets with the NMES stimulation (Here we choose $m = 10$*).*

2). For the remaining images A_1*,...,*A_n *and* B_1*,...,*B_n *from the middle segments of two on-off stimulation sessions, compute the correlation coefficient* $cor_{ij}(AB)$ *of* A_i *and* B_{i+j} *for* $i = 1, ..., n-j$ *and* $j = 0, ..., n-1$*. Let*

$$\mathrm{CorAvg}_j = \frac{1}{n-j} \sum_i cor_{ij}(AB).$$

Find j_0 *such that*

$$\mathrm{CorAvg}_{j_0} = \max_j(\mathrm{CorAvg}_j).$$

3). Align images A_i *with* B_{i+j_0}*.*

Chapter 3

Statistical Smoothing Mapping

In the NMES study the primary questions of interest to biomedical researchers are:

- Does the long-term gluteal NMES improve intrinsic characteristics of the paralyzed muscles?
- Can we identify the areas in which interface pressure has significantly improved?

To answer these questions we develop a statistical smoothing mapping algorithm which is inspired from the popular statisical parametric mapping approach in brain imaging. Our algorithm incorporates multivariate nonparametric regression techniques combined with an efficient procedure for computing an "FDR" movie/map to determine whether a change is clinically relevant or merely spurious. In section 3.1, multi-dimensional smoothing techniques are presented and their bandwidth selection and computation aspects are discussed. In section 3.2, hypothesis testing for nonparametric regression is addressed. In section 3.3, the multiple testing problem is discussed. The control of the false discovery rate under dependency is studied there. Finally we propose our statistical smoothing mapping algorithm.

3.1 Multivariate Local Regression

Intuitively, in order to identify improved areas in the sitting region from the pressure data, one can simply take the difference between baseline and treatment data after segmentation and registration. However, it is difficult to draw a conclusion directly from the difference maps because the data are corrupted by random variations in intensity. Nonparametric smoothing techniques are designed to estimate and model the underlying structure. It helps to extract structural elements of variable complexity from patterns of random variation. More precisely, the aim of smoothing here is to remove sampling variability that has no assignable cause, and to make systematic features of the data more apparent which thereby enables us to capture the improved areas from the statistical point of view.

Smoothing becomes more difficult as the dimension of the data set increases. The multivariate local estimation approach that we apply in our study is a powerful tool to be used in high dimensional smoothing. Instead of estimating a constant locally (*i.e.* kernel estimation) one can locally fit a polynomial model. This idea has superior behavior in particular at boundaries (Fan and Gijbels, 1996), which matches our application needs. Moreover, local linear (or quadratic) smoothing not only permits the estimation of the regression function itself but also its derivatives.

In this section we present multivariate smoothing techniques using local polynomial fitting, which includes the Nadaraya-Watson kernel estimator, with a focus on application to the NMES data. The ideas in developing asymptotic results and choosing bandwidth are similar for both local polynomial and Nadaraya-Watson kernel estimators. However, we are only able to introduce those relevant to the dissertation. For a more complete discussion of the subject see the monographs by Scott (1992), Fan and Gijbels (1996) and Wand and Jones (1995).

3.1.1 Multivariate Kernel and Density Estimation

The goal of multivariate nonparametric density estimation is to approximate the probability density function (PDF) $f(\mathbf{x}) = f(x_1, \cdots, x_p)$ of the random variables $\mathbf{X} = (X_1, \cdots, X_p)^T$. The multivariate kernel density estimator in the p-dimensional case is defined as

$$\hat{f}_h(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_1 \cdots h_p} K\left(\frac{X_{i1} - x_1}{h_1}, \cdots, \frac{X_{ip} - x_p}{h_p}\right) \tag{3.1}$$

where $K : \mathbb{R}^p \to \mathbb{R}$ denotes a multivariate kernel function. Note that (3.1) assumes that the bandwidth is a vector of bandwidths $h = (h_1, \cdots, h_p)^T$.

In order to localize in p-dimensions, we need a multivariate kernel. A multivariate kernel function refers to a p-variate function satisfying

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} K(\mathbf{x}) d\mathbf{x} = 1.$$

The second-order kernel requires that

$$\int x_i K(\mathbf{x}) d\mathbf{x} = 0, \qquad i = 1, \cdots, p,$$

and the second-moment be finite. To simplify the notation, we use "$\int$" to indicate multivariate integration over the p-dimensional Euclidean space.

What is the form of the multidimensional kernel function $K(\mathbf{x})$? There are two common approaches for constructing multivariate kernels. One simple way is to use a *product kernel*:

$$K(\mathbf{x}) = \prod_{i=1}^{p} k(x_i)$$

where k denotes a univariate kernel function, for instance, the Epanechnikov kernel $k(x) = 3(1 - x^2)/4 \cdot I(|x| \leq 1)$.

An alternative method is to use a genuine multivariate kernel function $K(x)$ which uses observations from a ball around $\mathbf{x}$ to estimate the PDF at $\mathbf{x}$. This type of kernel is often called the *spherically* or *radially* symmetric kernel since $K(\mathbf{x})$ has the same value for all $\mathbf{x}$ on a sphere around zero. The spherically symmetric kernel is defined as

$$K(\mathbf{x}) = sK(||\mathbf{x}||)$$

where $s = (\int K(||\mathbf{x}||)d\mathbf{x})^{-1}$ is a normalization constant and $||\mathbf{x}|| = (x_1^2 + \cdots + x_p^2)^{1/2}$. Common selections of K include the standard p-dimensional normal density

$$K(\mathbf{x}) = (2\pi)^{-p/2}e^{-||\mathbf{x}||^2/2}$$

and the multivariate Epanechnikov kernel

$$K(\mathbf{x}) = \frac{p(p+2)\Gamma(p/2)}{4\pi^{p/2}}(1 - ||\mathbf{x}||^2)I(||\mathbf{x}||^2 \leq 1).$$

The latter is the optimal kernel according to Fan and Gijbels (1996).

In practice product kernels are recommended. However, for various theoretical studies, general multivariate kernels may be required. The general multivariate kernel estimator includes not only an arbitrary multivariate density as a kernel but also an arbitrary linear transformation of the data.

Let $\mathbf{H}$ be a symmetric positive definite matrix called a bandwidth matrix. The general form for the multivariate density estimator is

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\det(\mathbf{H})}K(\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})) = \frac{1}{n}\sum_{i=1}^{n}K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) \qquad (3.2)$$

where $\det(\cdot)$ denotes the determinant of a square matrix. The localization scheme at a point $\mathbf{x}$ assigns the weight $K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})$ with

$$K_{\mathbf{H}}(\mathbf{x}) = \frac{1}{\det(\mathbf{H})}K(\mathbf{H}^{-1}\mathbf{x})$$

which is analogous to $K_h = K(\cdot/h)/h$ in the one-dimensional case. The bandwidth matrix is introduced to accommodate the dependent structure in the independent variables. In the implementation of our statistical analysis in NMES we take the bandwidth matrix $\mathbf{H}$ to be a diagonal matrix. This accommodates different scales in different independent variables. A further simplification is to take an equal bandwidth h in all dimensions corresponding to $\mathbf{H} = h\mathbf{I}_p$ where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix, assuming that the independent variables have the same scale.

3.1.2 Multivariate Local Regression

Multivariate nonparametric regression aims to estimate the functional relation between a response variable $Y \in \mathbb{R}$ and a multivariate explanatory variable $\mathbf{X} \in \mathbb{R}^p$. In image application, Y is the intensity and $\mathbf{X}$ is the spatial location. Given observations of independent and identically distributed $\mathbb{R}^{p+1}$-valued random vectors $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$, we are interested in the multivariate nonparametric regression problem, estimating the conditional expectation

$$m(\mathbf{x}) = \mathrm{E}(Y|\mathbf{X} = \mathbf{x})$$

without the imposition that $m(\cdot)$ belongs to a parametric family of functions. We assume the model

$$Y_i = m(\mathbf{X}_i) + \varepsilon_i, \qquad i = 1, \cdots, n.$$

where $m(\cdot)$ is an unknown function and ε_i is an error term which represents random errors in the observations or variability from sources not included in the $\mathbf{X}_i$. We further assume that the observations Y_i have constant variance σ^2.

Note that

$$\mathrm{E}(Y|\mathbf{X}) = \int y f(y|\mathbf{x}) dy = \frac{\int y f(y, \mathbf{x}) dy}{f(\mathbf{x})}$$

The denominator can be estimated by the multivariate kernel density estimate (3.1) or (3.2). For the numerator we have

$$\begin{aligned}\int yf(y,\mathbf{x})dy &= \int y\Big(\frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x})K_h(Y_i-y)\Big)dy\\ &=\frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x})\int yK_h(Y_i-y)dy\\ &=\frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x})\int \frac{y}{h}K_h\Big(\frac{Y_i-y}{h}\Big)dy\\ &=\frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x})Y_i\end{aligned}$$

Therefore the multivariate generalization of the Nadaraya-Watson estimator, *multivariate Nadaraya-Watson estimator* is,

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \frac{\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x})Y_i}{\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x})} \tag{3.3}$$

Analogous to the univariate case, the multivariate Nadaraya-Watson estimator is just a weighted sum of the observed responses Y_i. The denominator ensures that the weights sum up to 1. Depending on the choice of the kernel $\hat{m}_{\mathbf{H}}(\mathbf{x})$ is a weighted average of those Y_i such that $\mathbf{X}_i$ lies in a ball or cube around $\mathbf{x}$.

Note that the multivariate kernel regression estimator is based on a local constant approximation; thus it is also called the *multivariate local constant estimator*, that is, it is the solution of

$$\min_{b_0}\sum_{i=1}^{n}\big\{Y_i-b_0\big\}^2 K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x})$$

This kernel estimator has the advantage of being simple to understand intuitively and it is consistent for any smooth function m, provided the density

of the $\mathbf{X}_i$'s satisfies some minimal assumptions. However, it has some disadvantages especially when the design is random. Chu and Marron (1991) discuss this issue in detail. It can be improved by using the local linear approximation,

$$m(\mathbf{X}) \approx m(\mathbf{x}) + m'(\mathbf{x})^T(\mathbf{X} - \mathbf{x})$$

for $\mathbf{X}$ in a local neighborhood of $\mathbf{x}$. This leads to the following least squares problem,

$$\min_{b_0, \mathbf{b}} \sum_{i=1}^{n} \left\{ Y_i - b_0 - \mathbf{b}^T(\mathbf{X}_i - \mathbf{x}) \right\}^2 K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) \tag{3.4}$$

The weight function (kernel function) is defined on the multivariate space, hence observations close to a fitting point $\mathbf{x}$ receive the largest weight. The problem (3.4) is a straightforward weighted least squares problem and, assuming that $\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}}$ is nonsingular, the solution is

$$\begin{bmatrix} \hat{b_0} \\ \hat{\mathbf{b}} \end{bmatrix} = (\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{W}\mathbf{Y}$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & (\mathbf{X}_1 - \mathbf{x})^T \\ \vdots & \vdots \\ 1 & (\mathbf{X}_n - \mathbf{x})^T \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},$$

and $\mathbf{W} = \text{diag}\{K_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{x}), \cdots, K_{\mathbf{H}}(\mathbf{X}_n - \mathbf{x})\}$. The local least squares estimator of $m(\mathbf{x})$ is then

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \mathbf{e}_1^T(\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{W}\mathbf{Y} \tag{3.5}$$

where $\mathbf{e}_1$ is a $(p+1) \times 1$ vector having

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Estimator (3.5), introduced by Stone (1977), has been used in time series analysis for a long time and it is a special case of the robust local regression estimators in Cleveland (1979). The major advantage of (3.5) is that it is very simple to visualize the estimator using the data when estimating m at a point $\mathbf{x}$. The other important advantage is that the asymptotic bias and variance expressions are particularly appealing and appear to be superior to those of the Nadaraya-Watson estimator. This has been demonstrated in the univariate case by Fan (1992, 1993) and in the multivariate case by Ruppert and Wand (1994).

Estimator (3.5) is just one member of a hierarchical class of local least squares kernel estimators since one can locally fit polynomials of arbitrary order. This class includes the multivariate Nadaraya-Watson estimator which corresponds to a local constant fit. Cleveland and Devlin (1988) successfully used a local quadratic fit in several examples, which improved fits obtained by the local quadratic rather than the local linear estimator.

For the multivariate local quadratic estimator we consider the second order Taylor's expansion,

$$m(\mathbf{X}) \approx m(\mathbf{x}) + m'(\mathbf{x})^T(\mathbf{X}-\mathbf{x}) + \frac{1}{2}(\mathbf{X}-\mathbf{x})^T m''(\mathbf{x})(\mathbf{X}-\mathbf{x}).$$

This leads to the problem,

$$\min_{b_0,\mathbf{b},\mathbf{C}} \sum_{i=1}^{n} \left\{Y_i - b_0 - \mathbf{b}^T(\mathbf{X}_i-\mathbf{x}) - \frac{1}{2}(\mathbf{X}_i-\mathbf{x})^T\mathbf{C}(\mathbf{X}_i-\mathbf{x})\right\}^2 K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x}). \quad (3.6)$$

The solution $\hat{m}_{\mathbf{H}}(\mathbf{x})$ is still defined by (3.5) but now $\tilde{\mathbf{X}}$ changes to

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & (\mathbf{X_1}-\mathbf{x})^T & \text{vech}^T\{(\mathbf{X_1}-\mathbf{x})(\mathbf{X_1}-\mathbf{x})^T\} \\ \vdots & \vdots & \vdots \\ 1 & (\mathbf{X_n}-\mathbf{x})^T & \text{vech}^T\{(\mathbf{X_n}-\mathbf{x})(\mathbf{X_n}-\mathbf{x})^T\} \end{bmatrix}$$

where vech$(\cdot)$ returns the vector obtained by eliminating all supradiagonal

elements of the square matrix and stacking the result one column above the other, and $\mathbf{e}_1$ is a $\{1 + p + \frac{1}{2}p(p+1)\} \times 1$ vector.

Remark. *Ruppert and Wand (1994) give a deep discussion about the asymptotic conditional bias and variance of nonparametric regression estimators using locally weighted least squares. The asymptotic conditional variance of the multivariate Nadaraya-Watson estimator, the multivariate local linear estimator and the multivariate local quadratic estimator have the same form,*

$$\mathrm{Var}\{\hat{m}_{\mathbf{H}}(\mathbf{x})|\mathbf{X}\} = \frac{\sigma^2(t)\int K(\mathbf{x})^2 d\mathbf{x}}{n \cdot det(\mathbf{H}) f(\mathbf{x})}\{1 + o_p(1)\}$$

where $f(\cdot)$ denote the true density of $\mathbf{X}$ having support $\mathrm{supp}(f) \subseteq \mathbb{R}^p$ and $\sigma^2(\mathbf{x})$ denotes the variance function $\mathrm{Var}(Y|\mathbf{X})$.

The asymptotic conditional bias of the multivariate local quadratic estimator is $O_p\{(\mathrm{tr}(\mathbf{H}^T\mathcal{H}_m(\mathbf{x})\mathbf{H}))^{3/2}\}$ rather than $O_p\{(\mathrm{tr}(\mathbf{H}^T\mathcal{H}_m(\mathbf{x})\mathbf{H}))\}$ for the multivariate local linear estimator, where $\mathcal{H}_m(\mathbf{x})$ denotes the $p \times p$ Hessian matrix of a sufficiently smooth p-variate function m at $\mathbf{x}$ and $\mathrm{tr}(\cdot)$ is the trace of the matrix.

Fan et al. (1997) point out that the multivariate local linear fit with an Epanechnikov kernel is a best linear estimator and has a minimax efficiency of at least 89.4 % among all estimators.

3.1.3 Bandwidth Selection

Multivariate local polynomial fitting requires a choice for the bandwidth matrix, the degree of the polynomial and the kernel function. The optimization over the bandwidth matrix $\mathbf{H}$ can be cumbersome hence a diagonal bandwidth matrix $\mathbf{H} = \mathrm{diag}\{h_1, \cdots, h_p\}$ (or even $\mathbf{H} = h\mathbf{I}_p$ with appropriate standardization of the data, where $\mathbf{I}_p$ denotes $p \times p$ identity matrix.) is preferred in practice, as is the case in our NMES image analysis.

As with every kernel-type estimator, bandwidth selection in the multivariate local polynomial regression estimation is of great importance. When the bandwidth is too small the resulting curve is too wiggly, reflecting too much of the sampling variability. When the bandwidth is too large the resulting estimate tends to smooth away important features. For this reason data-driven choice of $\mathbf{H}$ has been a key issue in kernel type nonparametric estimation. Theoretically the bandwidth selection problem of multivariate local polynomial regression can be handled the same way as in the one-dimensional case. Two approaches are frequently used: plug-in bandwidths, in particular *rule-of-thumb* bandwidths, and *cross-validation* bandwidths. Manual bandwidth selection or eye-balling method (where one tries several bandwidth values and chooses based on visual examination of the resulting observations and predictions) also may be used; however, it may be time-consuming and rather subjective. Different methods for bandwidth selection can produce rather different values, so bandwidth selection remains a subjective process.

Rule-of-thumb

In data analysis one would like to get a quick idea about how large the amount of smoothing should be. A rule-of-thumb (ROT) bandwidth selection is suitable in such a case. Although it is a rather crude bandwidth selector, it gives a first idea of an appropriate magnitude for the bandwidth parameter. With the local polynomial regression method such a crude bandwidth selector can be obtained by minimizing the mean squared error theoretically, and then use a plug-in estimate to obtain the optimal bandwidth estimate. Consider the asymptotically optimal constant bandwidth which minimizes the asymptotic

Weighted Mean Integrated Square Error (WMISE)

$$\begin{aligned}\text{WMISE} =& E\Big\{\int(\hat{m}_{\mathbf{H}}(\mathbf{x})-m(\mathbf{x}))^2 w(\mathbf{x})f(\mathbf{x})d\mathbf{x}\Big\}\\ =& \int\Big\{\big[Bias\big(\hat{m}_{\mathbf{H}}(\mathbf{x})\big)\big]^2+Var\big(\hat{m}_{\mathbf{H}}(\mathbf{x})\big)\Big\}w(\mathbf{x})f(\mathbf{x})d\mathbf{x}\end{aligned}$$

with $w \geq 0$ some weight function and $f(\mathbf{x})$ the density of $\mathbf{x}$. This leads to a theoretical optimal constant bandwidth. An asymptotically optimal constant bandwidth can be obtained by using the asymptotic expression of conditional bias and variance of the local linear regression estimator. Substituting the estimated value for the optimal bandwidth we obtain the rule of thumb bandwidth selector. Here we refer to Fan and Gijbels (1996) who summarize the bandwidth selection methods and Yang and Tschernig (1999) who investigate the rule-of-thumb bandwidth selector for multivariate local linear regression.

Cross-validation

In the NMES image application we concentrate on the data-driven bandwidth selector least squares *cross-validation* (CV), whose basic idea is to choose H by minimizing the Integrated Squared Error (ISE)

$$\text{ISE} = \int\{\hat{m}_{\mathbf{H}}(\mathbf{x})-m(\mathbf{x})\}^2 w(\mathbf{x})f(\mathbf{x})d\mathbf{x}$$

To motivate the discussion we first consider the *averaged residual sum of squares* (ARSS) as a naive way to assess the goodness of fit

$$\text{ARSS}(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i-\hat{m}_{\mathbf{H}}(\mathbf{X}_i)\}^2 w(\mathbf{X}_i)$$

where the weight function $w(\cdot) \geq 0$, the same as in ROT section, may be used to assign less weight to observations in regions of sparse data (to reduce

the variance in this region) or at the tail of the distribution (to trim away boundary effects). The typical choice of $w(\cdot)$ is the sample density function or just let $w(\cdot) = 1$ for the non-random design as is the case in the NMES image application.

The problem with the ARSS is that Y_i is used in $\hat{m}_{\mathbf{H}}(\mathbf{X}_i)$ to predict itself. As a result, the *averaged squared error* (ASE), a discrete approximation to ISE

$$\text{ASE}(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\{m_{\mathbf{H}}(\mathbf{X}_i) - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)\}^2 w(\mathbf{X}_i)$$

can be made arbitrarily small by letting $\mathbf{H} \to 0$.

For each i, we use the data $\{(\mathbf{X}_j, Y_j), j \neq i\}$ to build a regression function $\hat{m}_{\mathbf{H},-i}(\mathbf{x})$, the *leave-one-out* estimator, and then validate the model by examining the prediction error $Y_i - \hat{m}_{\mathbf{H},-i}(\mathbf{X}_i)$. The least squares cross-validation function

$$\text{CV}(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{\mathbf{H},-i}(\mathbf{X}_i)\}^2 w(\mathbf{X}_i) \tag{3.7}$$

uses the weighted average of squared errors as an overall measure of the effectiveness of the estimation scheme. The least squares cross-validation bandwidth selector is the one that minimizes (3.7). More discussion about cross-validation bandwidth selector is provided in later sections.

Akaike Information Criterion

The third possibility for chosing the bandwidth matrix $\mathbf{H}$ is by using the Akaike (1970) criterion, which balances the goodness of fit with the complexity of the fitted model. This is expressed in the *Akaike Information Criterion* (AIC) function,

$$\text{AIC}(\mathbf{H}) = n\log(\hat{\sigma}^2) + 2\nu_1 \tag{3.8}$$

where

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i))^2$$

$$\nu_1 = \mathrm{tr}(\mathbf{P})$$

$\mathbf{P}$ is called the *hat matrix* or *smoothing matrix* because it maps the vector of observed values into a vector of fitted values and ν_1 is a measure for the degrees of freedom. In a later section we give a detailed description of variance estimation and the degrees of freedom for multivariate local regression.

The small sample behavior for $AIC(\mathbf{H})$ can thereby be improved by replacing the latter component in (3.8). Hurvich et al. (1998) show that the bias corrected AIC avoids the tendency to undersmooth which often occurs when using the classical AIC or *generalized cross-validation* (GCV) (defined by (3.14)). This criterion is given by

$$\mathrm{AIC}_{\mathrm{C}_1}(\mathbf{H}) = n\log(\hat{\sigma}^2) + n\frac{(\delta_1/\delta_2)(n+\nu_1)}{\delta_1^2/\delta_2 - 2}$$

where

$$\delta_1 = \mathrm{tr}\{(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})\}$$

$$\delta_2 = \mathrm{tr}\{[(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})]^2\}$$

Again, a suitable choice for bandwidth matrix $\mathbf{H}$ is obtained by minimizing $\mathrm{AIC}(\mathbf{H})$ or $\mathrm{AIC}_{\mathrm{C}_1}(\mathbf{H})$.

Regardless of the method being used, it can be shown theoretically and in simulations, that the convergence of the bandwidth estimate is slow (Härdle et al., 1988). As a consequence, one should not blindly accept an automatically selected bandwidth but assess the smoothness of the resulting fit $\hat{m}(\cdot)$ visually as well. In principle this means one should try different bandwidths around the optimum to validate the sensitivity of the fit on the bandwidth choice. In practice, we recommend graphical techniques such as cross-validation plots (Loader, 1999) to help us make a decision.

3.1.4 Computational Aspects

A drawback of local polynomial smoothing techniques is the computational difficulties one faces with this approach. The local idea says that one locally fits a model to data. This means that in order to visualize the functional shape of a regression function $m(\cdot)$ one has to estimate $m(\mathbf{x})$ at a number of points $\mathbf{x}$ and then connect the resulting estimates. If the local fit is complex and numerically intensive, then local estimation at a number of points can readily lead to the limits of numerical feasibility.

In principal, because multivariate local regression estimators can be expressed as local polynomial estimators, their computation can be done by any statistical package that is able to run weighted least squares regression. However, when we consider cross-validation bandwidth selection in the multivariate local regression case, this weighted least squares regression has to be performed on all observation points. This can be extremely computationally intensive. Therefore, explicit formulae are extremely useful to improve the algorithm and to save valuable time and resources. In the following, we derive formulae for the multivariate local quadratic estimator and bandwidth selector. Formulae for the multivariate kernel and local linear cases are just special cases of the formulae that we derive.

Consider the sums

$$\mathbf{S}_0 = \mathbf{S}_0(\mathbf{x}) = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}),$$

$$\mathbf{S}_{11} = \mathbf{S}_{11}(\mathbf{x}) = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x}),$$

$$\mathbf{S}_{12} = \mathbf{S}_{12}(\mathbf{x}) = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})\mathrm{vech}\{(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T\},$$

$$\mathbf{S}_{21} = \mathbf{S}_{21}(\mathbf{x}) = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T,$$

$$\mathbf{S}_{22} = \mathbf{S}_{22}(\mathbf{x}) = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})\text{vech}\{(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T\}(\mathbf{X}_i - \mathbf{x})^T,$$

$$\mathbf{S}_{23} = \mathbf{S}_{23}(\mathbf{x}) = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})\cdot$$
$$\text{vech}\{(\mathbf{X_i} - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T\}\text{vech}\{(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T\}^T,$$

and

$$\mathbf{Z}_0 = \mathbf{Z}_0(\mathbf{x}) = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})Y_i,$$

$$\mathbf{Z}_{11} = \mathbf{Z}_{11}(\mathbf{x}) = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})Y_i,$$

$$\mathbf{Z}_{12} = \mathbf{Z}_{12}(\mathbf{x}) = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})\text{vech}\{(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T\}Y_i.$$

Note that $\mathbf{S}_{11}$ and $\mathbf{Z}_{11}$ are $p \times 1$ vectors, $\mathbf{S}_{12}$ and $\mathbf{Z}_{12}$ are $\frac{1}{2}p(p+1) \times 1$ vectors, $\mathbf{S}_{21}$ is a $p \times p$ matrix, $\mathbf{S}_{22}$ is a $\frac{1}{2}p(p+1) \times p$ matrix and $\mathbf{S}_{23}$ is a $\frac{1}{2}p(p+1) \times \frac{1}{2}p(p+1)$ matrix. To simplify the notation let,

$$\mathbf{S}_1 = \begin{bmatrix} \mathbf{S}_{11} \\ \mathbf{S}_{12} \end{bmatrix} = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) \begin{bmatrix} \mathbf{X}_i - \mathbf{x} \\ \text{vech}\{(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T\} \end{bmatrix}$$

$$\mathbf{Z}_1 = \begin{bmatrix} \mathbf{Z}_{11} \\ \mathbf{Z}_{12} \end{bmatrix} = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) \begin{bmatrix} \mathbf{X}_i - \mathbf{x} \\ \text{vech}\{(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T\} \end{bmatrix} Y_i$$

$$\mathbf{S}_2 = \begin{bmatrix} \mathbf{S}_{21} & \mathbf{S}_{22}^T \\ \mathbf{S}_{22} & \mathbf{S}_{23} \end{bmatrix} = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})\cdot$$

$$\begin{bmatrix} (\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T & \text{vech}\{(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T\}(\mathbf{X}_i - \mathbf{x})^T \\ \text{vech}\{(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T\} & \text{vech}\{(\mathbf{X_i} - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T\} \\ \cdot(\mathbf{X}_i - \mathbf{x})^T & \cdot\text{vech}\{(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T\}^T \end{bmatrix}$$

where $\mathbf{S}_1$ and $\mathbf{Z}_1$ are $\{p+\frac{1}{2}p(p+1)\}\times 1$ vectors and $\mathbf{S}_2$ is a $\{p+\frac{1}{2}p(p+1)\}\times\{p+\frac{1}{2}p(p+1)\}$ matrix. Therefore, the multivariate local quadratic estimate can be written as

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \mathbf{e}_1^T \begin{bmatrix} \mathbf{S}_0 & \mathbf{S}_1^T \\ \mathbf{S}_1 & \mathbf{S}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Z}_0 \\ \mathbf{Z}_1 \end{bmatrix}.$$

Applying 2×2 block matrix inversion we derive an explicit expression for the multivariate local quadratic estimator,

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \frac{\mathbf{Z}_0 - \mathbf{S}_1^T\mathbf{S}_2^{-1}\mathbf{Z}_1}{\mathbf{S}_0 - \mathbf{S}_1^T\mathbf{S}_2^{-1}\mathbf{S}_1}. \tag{3.9}$$

Moreover,

$$\begin{aligned}
\mathbf{S}_{0,-i} &= \mathbf{S}_0 - K_{\mathbf{H}}(0),\\
\mathbf{S}_{1,-i} &= \mathbf{S}_1,\\
\mathbf{S}_{2,-i} &= \mathbf{S}_2,\\
\mathbf{Z}_{0,-i} &= \mathbf{Z}_0 - Y_i K_{\mathbf{H}}(0),\\
\mathbf{Z}_{1,-i} &= \mathbf{Z}_1,
\end{aligned}$$

which implies that the leave-one-out estimator is,

$$\hat{m}_{\mathbf{H},-i}(\mathbf{x}) = \frac{\mathbf{Z}_0 - Y_i K_{\mathbf{H}}(0) - \mathbf{S}_1^T\mathbf{S}_2^{-1}\mathbf{Z}_1}{\mathbf{S}_0 - K_{\mathbf{H}}(0) - \mathbf{S}_1^T\mathbf{S}_2^{-1}\mathbf{S}_1}. \tag{3.10}$$

Using (3.9) and (3.10) the cross-validation function changes to,

$$\begin{aligned}
\mathrm{CV}(\mathbf{H}) &= \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{\mathbf{H},-i}(\mathbf{X}_i)\}^2 w(\mathbf{X}_i)\\
&= \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)\}^2 \left\{\frac{Y_i - \hat{m}_{\mathbf{H},-i}(\mathbf{X}_i)}{Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)}\right\}^2 w(\mathbf{X}_i)\\
&= \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)\}^2 \left\{1 - \frac{K_{\mathbf{H}}(0)}{\mathbf{S}_0(\mathbf{X}_i) - \mathbf{S}_1(\mathbf{X}_i)^T\mathbf{S}_2(\mathbf{X}_i)^{-1}\mathbf{S}_1(\mathbf{X}_i)}\right\}^{-2} w(\mathbf{X}_i).
\end{aligned} \tag{3.11}$$

Notice that running weighted least-squares regression point by point is avoided if (3.11) is used for finding the CV bandwidth. The computational loading is then remarkably reduced.

Let us denote $\mathbf{P}$ as the *hat matrix* which maps the vector of observed values into a vector of fitted values. $\mathbf{P}$ is the $n \times n$ matrix with rows $\mathbf{p}(\mathbf{X}_i)^T$, where

$$\mathbf{p}(\mathbf{x})^T = (p_1(\mathbf{x}), \cdots, p_n(\mathbf{x})) = \mathbf{e}_1^T(\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{W}. \tag{3.12}$$

Therefore,

$$\begin{bmatrix} \hat{m}(\mathbf{X}_1) \\ \vdots \\ \hat{m}(\mathbf{X}_n) \end{bmatrix} = \begin{bmatrix} \mathbf{p}(\mathbf{X}_1)^T \\ \vdots \\ \mathbf{p}(\mathbf{X}_n)^T \end{bmatrix} \mathbf{Y} = \mathbf{P}\mathbf{Y}.$$

The diagonal element p_{ii} of the hat matrix $\mathbf{P}$ is defined as the *leverage* or *influence* of the ith data point, which measures the sensitivity of the fitted values to the individual data points. We denote $l(\mathbf{X}_i) = p_i(\mathbf{X}_i) = p_{ii}$ as the leverage of the ith data point. It is easy to verify that in multivariate local quadric regression,

$$l(\mathbf{X}_i) = \frac{K_{\mathbf{H}}(0)}{\mathbf{S}_0(\mathbf{X}_i) - \mathbf{S}_1(\mathbf{X}_i)^T\mathbf{S}_2(\mathbf{X}_i)^{-1}\mathbf{S}_1(\mathbf{X}_i)}.$$

This leads to, by (3.11)

$$\mathrm{CV}(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)}{1 - l(\mathbf{X}_i)}\right\}^2 w(\mathbf{X}_i). \tag{3.13}$$

Equation (3.13) has the same expression as in the univariate nonparametric regression case (Simonoff, 1996, chap. 5). The motivation for *generalized cross-validation* (GCV) criterion, first proposed by Craven and Wahba (1979), follows from the approximation of (3.13) by simply replacing $l(\mathbf{X}_i)$ by the average value $\mathrm{tr}(\mathbf{P})/n$,

$$\mathrm{GCV}(\mathbf{H}) = \frac{n\sum_{i=1}^{n}(Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i))^2}{(n - tr(\mathbf{P}))^2} w(\mathbf{X}_i). \tag{3.14}$$

3.2 Statistical Tests and Confidence Regions

Almost all nonparametric regression techniques are weighted averages of the response observations Y_i, where the weights depend on the technique and on the distance between $\mathbf{x}$ and $\mathbf{X}_i$ scaled by a smoothing parameter h or $\mathbf{H}$. In multivariate local estimation, the estimated function also can be written as a linear combination of the response observations,

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \sum_{i=1}^{n} p_i(\mathbf{x}) Y_i \tag{3.15}$$

where $\mathbf{p}(\mathbf{x})^T = (p_1(x), \cdots, p_n(x))$ is defined by (3.12).

The weighted average expression leads to compact forms for the mean and variance of the local estimator,

$$\mathrm{E}(\hat{m}(\mathbf{x})) = \sum_{i=1}^{n} p_i(\mathbf{x}) m(\mathbf{X}_i), \tag{3.16}$$

$$\mathrm{Var}(\hat{m}(\mathbf{x})) = \sigma^2 \sum_{i=1}^{n} p_i(\mathbf{x})^2 = \sigma^2 ||p(\mathbf{x})||^2. \tag{3.17}$$

So far, our discussion has focused on function estimation and choosing the amount of smoothing. In this section we discuss some other inference topics including inference about the true mean function $m(\mathbf{x})$, asymptotic normality and goodness-of-fit test.

3.2.1 Degrees of Freedom and Variance Estimation

The degrees of freedom of the local regression provide a generalization of the number of parameters in a parametric model. The usefulness of the degrees of freedom is to provide a measure of the complexity of the fitted function and the amount of smoothing that is comparable between different estimates

applied to the same data. The two kinds of degrees of freedom are defined as follows,

$$\nu_1 = \sum_{i=1}^{n} l(\mathbf{X}_i) = \mathrm{tr}(\mathbf{P}),$$
$$\nu_2 = \sum_{i=1}^{n} ||p(\mathbf{X}_i)||^2 = \mathrm{tr}(\mathbf{P}^T\mathbf{P}). \tag{3.18}$$

For a parametric regression model the two degrees of freedom are identical and usually equal to the number of parameters. For local regression models they are often not equal and have $1 \leq \nu_1 \leq \nu_2 \leq n$.

The degrees of freedom and variance estimation have already been discussed in Section 3.1.3. There $\hat{\sigma}^2$ is defined as

$$\hat{\sigma}^2_N = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i))^2 = \frac{1}{n}\mathbf{Y}^T(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})\mathbf{Y}$$

which can be viewed as a *naive estimate* of the error variance σ^2. However, in analogy with parametric regression, the *residual variance estimate* is often used in practice. Consider the expected residual sum-of-squares,

$$\begin{aligned}
\mathrm{E}\sum_{i=1}^{n}\Big[Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)\Big]^2 &= \sum_{i=1}^{n}\Big[\mathrm{E}\big(Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)\big)\Big]^2 + \sum_{i=1}^{n}\mathrm{Var}\big(Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)\big) \\
&= \sum_{i=1}^{n}\Big[\mathrm{Bias}\big(\hat{m}_{\mathbf{H}}(\mathbf{X}_i)\big)\Big]^2 + \sum_{i=1}^{n}\mathrm{Var}\big(Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)\big).
\end{aligned}$$

Note that by (3.15), (3.17) and independence of Y_i, we have

$$\begin{aligned}
\mathrm{Var}\big(Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)\big) &= \mathrm{Var}(Y_i) - 2\mathrm{Cov}\big(Y_i, \hat{m}_{\mathbf{H}}(\mathbf{X}_i)\big) + \mathrm{Var}\big(\hat{m}_{\mathbf{H}}(\mathbf{X}_i)\big) \\
&= \sigma^2 - 2\sum_{j=1}^{n} p_j(\mathbf{X}_i)\mathrm{Cov}(Y_i, Y_j) + \sigma^2||p(\mathbf{X}_i)||^2 \\
&= \sigma^2(1 - 2p_i(\mathbf{X}_i) + ||p(\mathbf{X}_i)||^2).
\end{aligned}$$

Hence,

$$\mathrm{E}\sum_{i=1}^{n}\Big[Y_i-\hat{m}_{\mathbf{H}}(\mathbf{X}_i)\Big]^2=\sum_{i=1}^{n}\Big[\mathrm{Bias}\big(\hat{m}_{\mathbf{H}}(\mathbf{X}_i)\big)\Big]^2+\sigma^2(n-2\nu_1+\nu_2).$$

This motivates the residual variance estimate of the error variance using the normalized residual sum of squares,

$$\hat{\sigma}_R^2=\frac{1}{n-2\nu_1+\nu_2}\sum_{i=1}^{n}\Big[Y_i-\hat{m}_{\mathbf{H}}(\mathbf{X}_i)\Big]^2. \tag{3.19}$$

Similar to parametric regression, $(n-2\nu_1+\nu_2)$ is called *the residual degrees of freedom*. Notice that the expectation of residual variance estimate $\hat{\sigma}_R^2$ is

$$\mathrm{E}(\hat{\sigma}_R^2)=\sigma^2+\frac{1}{n-2\nu_1+\nu_2}\sum_{i=1}^{n}\Big[\mathrm{Bias}\big(\hat{m}_{\mathbf{H}}(\mathbf{X}_i)\big)\Big]^2.$$

Obviously, $\hat{\sigma}_R^2$ is unbiased only if the estimate $\hat{m}(\mathbf{x})$ is unbiased.

3.2.2 Hypothesis Testing

Typical questions of nonparametric estimates that arise in NMES study are:

1). Is there indeed significant pressure improvement at location $\mathbf{x}$?

2). Is there any impact of $\mathbf{X}$ on Y, *i.e.* is $m(\mathbf{x})=0$, $\forall\mathbf{x}$? If yes, is the estimated function significantly different from the traditional parameterization (e.g. the linear model)?

To answer these questions, we shall next discuss the *t-type test* and the *goodness-of-fit test* in the nonparametric regression context.

T-type Test

We are interested in testing the following hypothesis in order to answer question 1):

$$H_0: m(\mathbf{x})=0 \qquad \text{vs.} \qquad H_1: m(\mathbf{x})>0. \tag{3.20}$$

In order to derive a proper test statistic, let us assume that $\hat{m}(\mathbf{x})$ is the unbiased estimate of $m(\mathbf{x})$. Also, assume the ε_i are normally distributed with mean zero and variance σ^2. The multivariate local estimate $\hat{m}(\mathbf{x})$ has the distribution

$$\frac{\hat{m}(\mathbf{x}) - m(\mathbf{x})}{\sigma||p(\mathbf{x})||} \sim N(0,1),$$

and an approximate confidence band for $m(\mathbf{x})$ is

$$\left(\hat{\mu}(\mathbf{x}) - z_{1-\frac{\alpha}{2}}\sigma||p(\mathbf{x})||, \hat{\mu}(\mathbf{x}) + z_{1-\frac{\alpha}{2}}\sigma||p(\mathbf{x})||\right).$$

If σ^2 is unknown we can replace it by the residual variance estimate $\hat{\sigma}_R^2$. So we have

$$\frac{\hat{m}(\mathbf{x}) - m(\mathbf{x})}{\hat{\sigma}_R||p(\mathbf{x})||} \sim T(n - 2\nu_1 + \nu_2),$$

if $\hat{m}$ and $\hat{\sigma}_R^2$ are independent and $(n - 2\nu_1 + \nu_2)\hat{\sigma}_R^2/\sigma_R^2 \sim \chi^2(n - 2\nu_1 + \nu_2)$.

Unfortunately the assumption that $\hat{m}(\mathbf{x})$ is unbiased is seldom true (for instance $\hat{m}(\mathbf{x})$ is biased in the multivariate local regression) even though it might be reasonable to assume the bias is negligible with small bandwidth. Moreover, $\mathbf{P}$ is no longer a projection operator in the multivariate local regression. Recall that the estimate vector $\mathbf{PY}$ and the residual vector $(\mathbf{I} - \mathbf{P})\mathbf{Y}$ are independent in the parametric least squares theory, *i.e.*

$$\text{cov}((\mathbf{I} - \mathbf{P})\mathbf{Y}, \mathbf{PY}) = \sigma^2(\mathbf{I} - \mathbf{P})^T\mathbf{P} = 0,$$

where $\mathbf{P}^T\mathbf{P} = \mathbf{P}^2 = \mathbf{P}$. The property does not hold in the multivariate local regression. Therefore the T distribution above is incorrect. To solve the dilemma consider alternative estimates of σ^2. Note that, because of unbiasedness,

$$\begin{aligned}\text{E(RSS)} &= \text{E}\left\{\sum_{i=1}^{n}\left(Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)\right)^2\right\} \\ &= \text{E}\{\mathbf{Y}^T(\mathbf{I} - \mathbf{P})^T(\mathbf{I} - \mathbf{P})\mathbf{Y}\} = \sigma^2\text{tr}[(\mathbf{I} - \mathbf{P})^T(\mathbf{I} - \mathbf{P})],\end{aligned}$$

and we can estimate $\hat{\sigma}^2$ by

$$\hat{\sigma}^2 = \frac{\mathbf{Y}^T(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})\mathbf{Y}}{\text{tr}[(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})]}. \tag{3.21}$$

It is difficult to find the exact distribution $\hat{\sigma}^2$ but we can approximate it by the distribution of a quadratic form in normal variables. In fact, using the eigenvalue decomposition technique,

$$\frac{1}{\sigma^2}\mathbf{Y}^T(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})\mathbf{Y} = \sum_{i=1}^{n}\lambda_i U_i^2,$$

where λ_i are the eigenvalues of $(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})$ and U_i are independent standard normal variables leads to a χ^2 distribution. Finding the exact eigenvalues is not easy due to the large dimension of hat matrix. The degrees of freedom can be chosen so that the first two moments of the approximating distribution match those of the distribution of the quadratic form (Kendall et al., 1998). Recall that a χ^2 distribution with ν degrees of freedom has mean ν and variance 2ν. Let us denote $\delta_1 = \text{tr}\{(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})\}$ and $\delta_2 = \text{tr}\{[(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})]^2\}$. It is easy to verify that

$$\text{E}(\hat{\sigma}^2) = \sigma^2\delta_1, \qquad \text{Var}(\hat{\sigma}^2) = 2\sigma^4\frac{\delta_2}{\delta_1^2}.$$

Therefore, the first two moments of $(\delta_1^2\hat{\sigma}^2)/(\delta_2\sigma^2)$ match those of a χ^2 distribution with δ_1^2/δ_2 degree of freedom. Using this approximate χ^2 distribution we have the following approximation

$$\frac{\hat{m}(\mathbf{x}) - m(\mathbf{x})}{\hat{\sigma}||p(\mathbf{x})||} \sim T(\delta_1^2/\delta_2). \tag{3.22}$$

We can use this result to construct an approximate t test for our hypothesis and also get approximate confidence bands,

$$\left(\hat{\mu}(\mathbf{x}) - t_{\frac{\alpha}{2}}\hat{\sigma}||p(\mathbf{x})||, \hat{\mu}(\mathbf{x}) + t_{1-\frac{\alpha}{2}}\hat{\sigma}||p(\mathbf{x})||\right).$$

From (3.22) it is easy to construct the test statistic for hypothesis (3.20). The approximate t statistic is

$$T(\mathbf{x}_i) = \frac{\hat{m}(\mathbf{x}_i)}{\hat{\sigma}||p(\mathbf{x}_i)||} \tag{3.23}$$

where $\hat{m}(\mathbf{x}_i)$ and $\hat{\sigma}$ are computed by (3.15) and (3.21) respectively. The null hypothesis H_0 will be rejected if $T(\mathbf{x}_i) > t_{1-\alpha}(\delta_1^2/\delta_2)$ with a given significance level α.

Remark 1. *Although the above test statistic T is a "t-type test" statistic, it differs from the conventional t test in the following way. It is a weighted average of y values in a neighborhood of x, while the standard t test statistic is a simple average of an independent and identically distributed sample (divided by an appropriately estimated standard deviation). Therefore, $T(x)$ and $T(x')$ are often correlated if x and x' are not far away.*

Remark 2. *The discussions above are based on normal error assumption. What happens for tests with non-normal data? By the central limit theorem, under some regularity conditions below,*

$$\frac{\hat{m}(\mathbf{x}) - m(\mathbf{x})}{\sigma||p(\mathbf{x})||} \to N(0,1).$$

The Lindeberg condition maintains that the central limit theorem holds if the maximum contribution of any single observation converges to 0. In smoothing context, we need

$$\max_{1\le i\le n} \frac{|p_i(\mathbf{x})|}{||p(\mathbf{x})||} \to 0$$

In practice, a sufficient and necessary condition for most distributions is that $nh^p \to \infty$ where p is the dimension of $\mathbf{x}$. This is the same condition as that required for $\mathrm{Var}(\hat{m}(\mathbf{x})) \to 0$. Recalling usual asymptotic bandwidth conditions, we want both the bias and variance to converge to zero as $n \to \infty$.

Based on bias considerations we require $h^p \to 0$ *so that local polynomial approximation becomes better; based on variance considerations, we want more data within the smoothing window,* i.e. $nh^p \to \infty$.

The central limit theorem justifies that confidence intervals are asymptotically valid when data is not normal. However, linear smoothing will not be robust since we know that the sample average is not a robust estimate of location when the distribution has heavy tails.

Goodness of Fit

Can the mean function be adequately described by a constant, or is there really a regression effect?

Let $\mathcal{X}$ be the domain of interest, consider testing the following hypotheses:

$$H_0 : m(\mathbf{x}) = C, \ \forall \mathbf{x} \in \mathcal{X} \qquad H_1 : \textit{otherwise}; \tag{3.24}$$

or more generally, consider whether the target regression function significantly differs from a linear regression function. The hypothesis testing problem can be stated as

$$\begin{aligned} &H_0 : m(\mathbf{x}) = b_0 + \mathbf{b}_1\mathbf{x}, \ \textit{for some } b_0, \mathbf{b}_1, \forall \mathbf{x} \in \mathcal{X} \\ &H_1 : \textit{otherwise}. \end{aligned} \tag{3.25}$$

Analagous to the theory of linear models, an F-ratio can be formed by residual sums of squares from both the null and alternative models (Cleveland and Devlin, 1988; Loader, 1999). Under the null model the parametric least squares estimate is used. Consider the residual sums of squares:

$$RSS_0 = \sum_{i=1}^{n} (Y_i - (\hat{b}_0 + \hat{\mathbf{b}}_1\mathbf{x}))^2 = \mathbf{Y}^T\mathbf{R}_0\mathbf{Y},$$

$$RSS_1 = \sum_{i=1}^{n} (Y_i - \hat{m}(\mathbf{x}))^2 = \mathbf{Y}^T\mathbf{R}_1\mathbf{Y}.$$

where

$$\mathbf{R}_0 = (\mathbf{I} - \mathbf{P}_0)^T(\mathbf{I} - \mathbf{P}_0),$$

$$\mathbf{R}_1 = (\mathbf{I} - \mathbf{P}_1)^T(\mathbf{I} - \mathbf{P}_1).$$

$\mathbf{P}_0$ and $\mathbf{P}_1$ are hat matrices for the parametric fit and local fit respectively. Let $\nu_1 = \text{tr}(\mathbf{R}_0 - \mathbf{R}_1)$, $\nu_2 = \text{tr}[(\mathbf{R}_0 - \mathbf{R}_1)^2]$, $\delta_1 = \text{tr}(\mathbf{R}_1)$ and $\delta_2 = \text{tr}[(\mathbf{R}_1)^2]$. Then the F-ratio statistic is

$$F = \frac{(RSS_0 - RSS_1)/\nu_1}{RSS_1/\delta_1}$$

Its distribution is approximated by an F distribution with ν_1^2/ν_2 and δ_1^2/δ_2 degrees of freedom. An α-level test of (3.25) rejects H_0 if $F \geq F_{1-\alpha}(\nu_1^2/\nu_2, \delta_1^2/\delta_2)$.

Bootstrap Method

The F-tests discussed above are approximate, based on the two-moment χ^2 approximation for the numerator and the denominator. Additionally, the degrees-of-freedom computations require expensive computations. For this reason approximations of the critical values corresponding to the finite sample distribution can be used. The most popular way to approximate this finite sample distribution is via a resampling scheme: simulate the distribution of your test statistic under the hypothesis (*i.e.* "resample") and determine the critical values based on that simulated distribution. This method is called a Monte Carlo method or *bootstrap*, depending on how the distribution of the test statistic can be simulated.

For our current testing problem we propose the following nonparametric bootstrap approach to evaluate the p-value of the test.

Algorithm 3.2.1. *Nonparametric Bootstrap*

1). Compute the estimate of the regression function $\hat{m}(\cdot)$ under the null hypothesis and construct the residuals $\varepsilon_i = Y_i - \hat{m}(\mathbf{X}_i)$; calculate our

test statistic:

$$T = \frac{RSS_0 - RSS_1}{RSS_1}.$$

2). Generate the bootstrap residuals $\{\varepsilon_i^*\}_{i=1}^n$ *from the empirical distribution of the centered residuals* $\{\varepsilon_i - \bar{\varepsilon}\}_{i=1}^n$ *where* $\bar{\varepsilon} = \sum_{i=1}^n \varepsilon_i/n$.

3). Define $Y_i^* = \hat{m}(\mathbf{X}_i) + \varepsilon_i^*$ *and compute the test statistic* T^* *based on the re-sampled data.*

4). Step 2 and 3 are repeated a large number of times. Reject the null hypothesis H_0 *if* T *is greater than the upper* α*-point of the distribution of* T^*.

3.3 Multiple Testing Problem

An important and common question in the NMES experiment is the identification of spatial locations where the intensity of spatial locations are significantly different in treatment versus baseline. Recall in Figure 1.7 and 1.8 we already showed the idealized changes in pressure contour across the region of the ischial tuberosities. How can we identify those pressure-changed regions? This biomedical question can be restated as a multiple hypothesis test, or the simultaneous test for each spatial compartment of the null hypothesis.

A common approach to identifying active spatial compartments in the NMES data is to perform compartment-wise hypothesis "t-type" tests (which we have stated in section 3.2.2) after performing bivariate local smoothing over the data. At each spatial compartment the null hypothesis is that there is no pressure difference between baseline and treatment. The compartments for which the test statistics exceed the threshold are then classified as active. Images of statistics can be created which assess evidence for an experimental effect. This approach has proved reasonably effective for a wide variety of

testing methods. However, a basic problem remains: how to choose the threshold?

In any testing situation two types of errors can be committed. A *false positive*, or *Type I error*, is committed by declaring that the pressure between baseline and treatment in one compartment is significantly different when it isn't. A *false negative*, or *Type II error*, is committed when the test fails to identify a truly differential spatial compartment. When one uses naive thresholds for the individual tests, ignoring the fact that many tests are being performed, the probability that there will be false positives among all the tests becomes very high. This is well-known as the *multiple testing problem*.

For example, there are 1600 hypotheses; if a significance level of 0.05 pointwise procedure is used for each hypothesis, there will be $0.05 \times 1600 = 80$ false positives even if the null hypotheses are true. Actually, a p-value of 0.01 for one compartment among a list of several thousands will no longer correspond to a significant finding, as it is inevitable that such small p-values will occur by chance when considering a large enough set of compartments. This is the well-known *multiplicity* problem which must be solved.

Special problems arising from the multiplicity aspect include defining an appropriate *overall* Type I error rate and devising powerful multiple testing procedures which control this error rate and account for the joint distribution of the test statistics. In the following we illustrate the basic background of multiple testing, describe the Benjamini and Hochberg (1995) step-up procedure for (strong) control of the *false discovery rate* (*i.e.* BH-FDR procedure) that we will use in NMES data analysis, and discuss the validity of the BH-FDR procedure under dependency in our multiple t-type tests.

3.3.1 Background

Consider the problem of testing simultaneously m null hypotheses H_{0i}, $i = 1, \cdots, m$, and denote by R the number of rejected hypotheses. Each test can be classified into one of four types, depending on whether or not the pixel (or data cell) is truly active and whether or not it is declared active, as shown in table 3.1. U is the number of null hypotheses correctly classified as true; V is the number of null hypotheses incorrectly classified as false; T is the number of null hypotheses incorrectly classified as true; S is the number of null hypotheses correctly classified as false. The specific number m is assumed to be known in advance, the numbers m_0 and $m_1 = m - m_0$ (of true and false null hypotheses) are unknown parameters, R is an observable random variable, and U, V, T, and S are unobservable random variables.

In the NMES data analysis there is a null hypothesis H_{0i} for each special compartment i and rejection of H_{0i} corresponds to declaring that the pressure in this compartment i is significantly improved. For each compartment i the null hypothesis H_{0i} is tested based on a statistic T_i, where t_i denotes a realization of the random variable T_i. To simplify matters, and unless specified otherwise, we further assume that the null H_{0i} is rejected for large values of T_i (we have one-sided hypotheses in the NMES case). In general, one would like to minimize the number V of false positives and the number T of false negatives. The standard approach is to pre-specify an acceptable Type I error rate α and seek tests which minimize the Type II error rate, *i.e.* maximize power, within the class of tests with Type I error rate α.

Type I Error Rates

When testing a single hypothesis H_{01}, the probability of a Type I error, *i.e.*, of rejecting the null hypothesis when it is true, is usually controlled at a designated level α. This can be achieved by choosing a critical value c_α such

	H_0 is not rejected	H_0 is rejected	
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	m_1
	$m - R$	R	m

Table 3.1: *Cross-classification in m Simultaneous Tests*

that $Pr(T_1 \geq c_\alpha|\mathrm{H}_{01}) \leq \alpha$ and rejecting H_{01} when $T_1 \geq c_\alpha$. In order to control for the multiplicity effect, alternative Type I error rates given below are the most standard (Hochberg and Tamhane, 1987; Ge et al., 2003).

- *Per-family error rate (PFER).* The PFER is defined as the expected number of Type I errors,

$$\mathrm{PFER} = \mathrm{E}(V)$$

.

- *Per-comparison error rate (PCER).* The PCER is defined as the expected proportion of Type I errors,

$$\mathrm{PCER} = \mathrm{E}(V)/m$$

.

- *Family-wise error rate (FWER).* The FWER is defined as the probability of at least one Type I error,

$$\mathrm{FWER} = p(V \geq 1)$$

.

- *False discovery rate (FDR).* The FDR is the expected proportion of Type I errors among the rejected hypotheses,

$$\text{FDR} = \text{E}(Q),$$

where by definition

$$Q = \begin{cases} V/R, & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$

In general, for a given multiple testing procedure, PCER $\leq$ FDR $\leq$ FWER $\leq$ PFER. Thus, for a fixed criterion α for controlling the Type I error rates, the order reverses for the number of rejections R. Procedures controlling the PFER are generally more conservative than those controlling either the FDR or the FWER or the PCER, and procedures controlling the FWER are more conservative than those controlling the FDR or the PCER. FWER had reigned in the field of multiple comparison procedures (MCP) until Benjamini and Hochberg (1995) proposed FDR. Much research has been generated since then.

Strong Control and Weak Control

A fundamental, yet often ignored distinction, is that between strong and weak control of the Type I error rate. *Strong control* refers to control of the Type I error rate under any arbitrarily combination of true and false hypotheses, *i.e.*, any value of m_0. In contrast, *weak control* refers to control of the Type I error rate only when all the null hypotheses are true, *i.e.*, under the *complete null hypothesis* with $m_0 = m$. In this case, we have $FDR = FWER$ (since $S = 0$, $V = R$). This is called *FDR control FWER weakly*.

In the NMES study, where it is very unlikely that no spatial compartments have differential pressure, it is particularly important to have strong control of the Type I error rate.

Adjusted p-values

Given any testing procedure, we can define the *adjusted p-value* corresponding to the test of a single hypothesis H_{0i} as the level of the entire testing procedure at which H_{0i} would just be rejected. The adjusted p-value with FDR controlling for hypothesis H_{0i} (Yekutieli and Benjamini, 1999) is

$$\tilde{p}_i = \inf \{\alpha \in [0,1] : H_{0i} \text{ is rejected at FDR} = \alpha\}. \tag{3.26}$$

Similarly, the adjusted p-value with FWER controlling for hypothesis H_{0i} is

$$\tilde{p}_i = \inf \{\alpha \in [0,1] : H_{0i} \text{ is rejected at FWER} = \alpha\}. \tag{3.27}$$

The corresponding random variables for unadjusted (or raw) and adjusted p-values are denoted by p_i and $\tilde{p}_i$, respectively. Hypothesis H_i is then rejected, *i.e.* at FDR or FWER α, if $\tilde{p}_i \leq \alpha$ depending on which $\tilde{p}_i$ is used in (3.26) or (3.27).

As in the single hypothesis case an advantage of reporting adjusted p-values, as opposed to only rejection or not of the null hypotheses, is that the level of the test does not need to be determined in advance. Some multiple testing procedures are also most conveniently described in terms of their adjusted p-values.

Stepwise Procedures

One usually distinguishes among three types of multiple testing procedures: single-step, step-down, and step-up procedures. In *single-step* procedures, equivalent multiplicity adjustments are performed for all hypotheses, regardless of the ordering of the test statistics or unadjusted p-values. That is, each hypothesis is evaluated using a critical value that is independent of the results of tests of other hypotheses. Improvement in power, while preserving Type I error rate control, may be achieved by *stepwise procedures*, in which rejection

of a particular hypothesis is based not only on the total number of hypotheses, but also on the outcome of the tests of other hypotheses. In *step-down* procedures, the hypotheses corresponding to the *most* significant test statistics (*i.e.*, smallest unadjusted p-values or largest absolute test statistics) are considered successively, with further tests depending on the outcomes of earlier ones. As soon as one hypothesis is accepted, all remaining hypotheses are accepted. In contrast, for *step-up* procedures, the hypotheses corresponding to the *least* significant test statistics are considered successively, again with further tests depending on the outcomes of earlier ones. As soon as one hypothesis is rejected, all remaining hypotheses are rejected.

3.3.2 False Discovery Rate under Dependency

The classical approach to multiple testing calls for strong control of the FWER (e.g. Bonferroni procedure). However, the Bonferroni procedure is too conservative when the number of hypotheses is very large. The conservativeness of the Bonferroni procedure comes from two sources: (1) the Bonferroni procedure was based on a very conservative *upper bound* for the FWER:

$$FWER = P(\cup A_i) \leq \sum P(A_i) = \alpha;$$

(2) FWER is a more stringent error test than FDR as illustrated on page 81.

Remark. *To overcome (1), there are sharper upper bounds for FWER developed for finite m cases (see Hsu (1996) and reference therein); there are also exact and accurate approximations by tube formulae such as those shown in Sun (1991, 1993, 2001). As to when to use FWER or FDR as allowed in (2), see Zhang (2005, section 4.2). In this thesis, we opt for FDR.*

The step-up procedure for strong control of the false discovery rate introduced by Benjamini and Hochberg (1995) is easily implemented, even for very large data sets, which can be less conservative than FWER.

Algorithm 3.3.1. *Benjamini & Hochberg step-up procedure for strong control of the false discovery rate.*

1). Select a desired FDR level α between 0 and 1. This is the maximum false discovery rate that the researcher is willing to tolerate.

2). For the m hypothesis tests, compute the raw p-values $p_1, \cdots, p_m$.

3). Order the p-values from smallest to largest:

$$p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}.$$

4). Set $p_{(0)} = 0$,

$$k_{BH} = max\{0 \leq k \leq m : p_{(k)} \leq \frac{\alpha k}{m c_m}\}$$

where c_m is a predetermined constant described below.

5). Declare that the null hypotheses H_{0k} are rejected if $p_k \leq p_{(k_{BH})}$.

Alternatively, we can compute the adjusted p-values for the BH-FDR step-up procedure above,

$$\tilde{p}_{(i)} = \min_{k=i,...,m} \Big\{ \min\Big(\frac{m c_m}{k} \; p_{(k)}, 1 \Big) \Big\}. \tag{3.28}$$

Hypothesis H_{0k} will be rejected, at FDR level α, if $\tilde{p}_k \leq \alpha$.

The choice of the constant c_m depends on assumptions about the joint distribution of the test statistics of the hypotheses family. When the test statistics are independent we have $c_m = 1$.

Returning to the NMES study, it is observed that the approximate T statistics of the multiple tests are dependent as they are from the estimate regression function. Is the BH-FDR procedure appropriate to our case? Benjamini and Yekutieli (2001) showed that the BH-FDR procedure is valid under "positive regression dependency on subsets" (PRDS). They also proposed

a simple conservative modification of the procedure which controls the false discovery rate for arbitrary dependence structures by letting $c_m = \sum_{i=1}^{m} 1/i$. Note that $\sum_{i=1}^{m} 1/i \approx \ln m + \gamma$ where γ is the Euler's constant. For a large number m of hypotheses, the penalty in this conservative procedure is about $\log m$, as compared to the Benjamini and Hochberg (1995) procedure, which can be still too large and can be more conservative than the tube methods or random field methods by Sun and Loader (1994), and Sun (2001).

Rather than using this conservative procedure with a factor $\ln m + \gamma$, we prove that the joint distribution of the approximate T test statistics is PRDS on the subset of test statistics corresponding to true null hypotheses, and thereby the BH-FDR procedure (1995) is still valid.

Recall that a set D is called increasing if $x \in D$ and $y \geq x$, implies that $y \in D$ as well. The following property is called *positive regression dependency on each one from a subset* I_0, or PRDS on I_0 (Benjamini and Yekutieli, 2001).

Property 3.3.1 (PRDS)**.** *For any increasing set D, and for each $i \in I_0$, $P(X \in D | X_i = x)$ is nondecreasing in x.*

Proposition 3.3.1 (PRDS of test statistics in multivariate local regression)**.** *Consider a vector of test statistics $\mathbf{T} = (T_1, T_2, \cdots, T_m)^T$. Each T_i tests the hypothesis $m(x_i) = 0$ against the alternative $m(x_i) > 0$ for $i = 1, \cdots, m$, where T_i is defined by (3.23). The distribution of $\mathbf{T}$ is PRDS over I_0, the set of true null hypotheses.*

Proof: Let $\mathbf{U} = (U_1, \cdots, U_m)^T$ where $U_i = \hat{m}(\mathbf{x}_i)/||p(\mathbf{x}_i)||$. We first verify that $\mathbf{U}$ is PRDS on a subset I_0. By (3.15), for any $i \neq j$,

$$\begin{aligned}\operatorname{cov}(U_i, U_j) &= \frac{\operatorname{cov}\big(\sum_{t=1}^{n} p_t(\mathbf{x}_i)Y_t, \sum_{k=1}^{n} p_k(\mathbf{x}_j)Y_k\big)}{||p(\mathbf{x}_i)|| \cdot ||p(\mathbf{x}_j)||} \\ &= \frac{\sum_{t=1}^{n}\sum_{k=1}^{n} \operatorname{cov}\big(p_t(\mathbf{x}_i)Y_t, p_k(\mathbf{x}_j)Y_k\big)}{||p(\mathbf{x}_i)|| \cdot ||p(\mathbf{x}_j)||} \\ &= \frac{\sigma^2 \sum_{t=1}^{n} p_t(\mathbf{x}_i)p_t(\mathbf{x}_j)}{||p(\mathbf{x}_i)|| \cdot ||p(\mathbf{x}_j)||} > 0\end{aligned}$$

Under the normality assumption of errors, $\mathbf{U}$ follows a multivariate normal distribution with the covariance matrix having positive elements. Then $\mathbf{U}$ is PRDS on a subset I_0 because the conditional distribution $\mathbf{U}_{(i)}$ given $U_i = u_i$ increases stochastically as u_i increases (where $\mathbf{U}_{(i)}$ denotes the remaining $m-1$ test statistics except U_i).

Since $\hat{\sigma}^2$ approximately follows a χ^2 distribution, we let $V = 1/\hat{\sigma}$. Then for $j = 1, \cdots, m$ the components of $\mathbf{T}$, $T_j = U_j V$ are strictly increasing continuous functions of the coordinates U_j and of V. Therefore, $\mathbf{U}$ is PRDS on I_0 by applying Lemma 3.1 of Benjamini and Yekutieli (2001). □

3.4 Statistical Smoothing Mapping

In statistical methods of brain imaging (e.g. MRI), one of the most common analysis approaches currently in use, called *statistical parametric mapping* (SPM) (Friston et al., 1995; Friston, 2004), analyzes each voxel's change independently of the others and builds a map of statistic values for each voxel. The significance of each voxel can be ascertained statistically with a Students t-test, an F-test, a correlation coefficient, etc. SPM is widely used to identify functionally specialized brain regions and is the most prevalent approach to characterizing functional anatomy and disease-related changes. The success of SPM is due largely to the simplicity of the idea. Namely, one analyzes each and every voxel using any standard (univariate) statistical

parametric test. The resulting statistical parameters are assembled into an image – the SPM.

Inspired by the SPM, we propose a statistical smoothing mapping (SSM) procedure in the NMES data analysis. The approach is called SSM because we use multivariate smoothing techniques on the data and our test statistics are constructed based on multivariate nonparametric regression. Since we are comparing many voxel values *simultaneously* across the entire image, the multiplicity of these tests must be adjusted to overcome an overall false-positive error rate. Our significance threshold for deciding which voxel is significantly different (between two sessions) will be chosen with a BH-FDR controlling procedure that accounts for the multiplicity of tests. Then an FDR map can be built to provide the significance of voxels. Those with p-values less than the BH critical value are the points or areas for which stimulation has had a significant effect (difference) in terms of measurements.

Let $\tilde{x} = (x_1, x_2)$ denote a cell (or pixel) of a data frame. Then $r_{\tilde{x},C}$, $r_{\tilde{x},T}$ denote the intensities of the images before treatment and after treatment. We propose the following statistical smoothing mapping algorithm.

Algorithm 3.4.1. *Statistical smoothing mapping*

1). Compute the difference map,

$$y_{\tilde{x}} = r_{\tilde{x},T} - r_{\tilde{x},C}$$

which is the cell-by-cell subtraction data frame of the differences in r before treatment and after treatment.

2). Smooth $y_{\tilde{x}}$ by multivariate local polynomial regression.

3). Compute adjusted p-values using the BH-FDR controlling procedure. Generate an FDR map based on the adjusted p-values.

Examples to implement the SSM algorithm will be discussed in the next chapter.

Chapter 4

Mining Spatial-temporal Data

This chapter consists of two parts. The first describes a new data-mining technique, *longitudinal analysis with self-registration* (LASR) procedure for interface pressure "intra-subject" data that is based on the techniques described in chapters 2 and 3. The second describes semiparametric regression for modeling spatial-temporal trend for interface pressure "inter-subject" data.

4.1 LASR – A New Data Mining Procedure

4.1.1 LASR

The assessment protocol produced multiple large volume data files for a relatively small number of subjects. In statistical terms this represents a "huge-p, small-n" problem. Further complexity was added to the analytical process because, although the subjects are seated carefully at each assessment, it is often not feasible to ensure a true reproduction of seating posture on each visit. In assessing the effects of dynamic stimulation over time it is also necessary to ensure that comparison is made between pressure maps obtained at the same phase of stimulation, e.g. when left gluteal stimulation is on.

Thus the challenges to be met include both spatial registration to align static pressure maps obtained at different times and temporal registration to ensure dynamic pressure maps are synchronous. The LASR algorithm uses a multi-stage procedure to sequentially address these challenges (see Figure 4.1).

Longitudinal Analysis with Self-Registration (LASR) Procedure

Step 1: *Segment* all images by the EM algorithm. We distinguish the spatial regions of interest from the background in each data frame and then remove background noise and outliers from the data sets.

Step 2: *Spatially register* all images via our newly developed self-registration scheme. The self-registration algorithm is built on an end point and a middle line estimated by a regression analysis applied on "apparent middle points" computed from each column of an image. This step is done automatically for all images so that all registered images have the middle line placed horizontally in the middle of each image and the end point at the same location.

Step 3: If both movies are static movies, go to Step 4; if both are dynamic movies, *temporally register* the spatially-registered movies. The temporal registration is based on a fast algorithm to maximize the correlations between images from two candidate movies, frame-by-frame so that the left side that is stimulated in one movie is compared with the left-side stimulated image in another movie (See movies at `http://stat.case.edu/lasr/`).

Step 4: *Create difference images and movies* by taking differences pixel-by-pixel (and frame-by-frame) between two sessions that are potentially of clinical interest.

Step 5: *Filter* the difference images. The nonparametric filtering procedure used is a local-polynomial smoothing technique which is suitable for a great variety of images.

Step 6: *Create T image/maps and movies.* T images are obtained by computing a t-type test statistic at each pixel in the spirit of a two-sample t

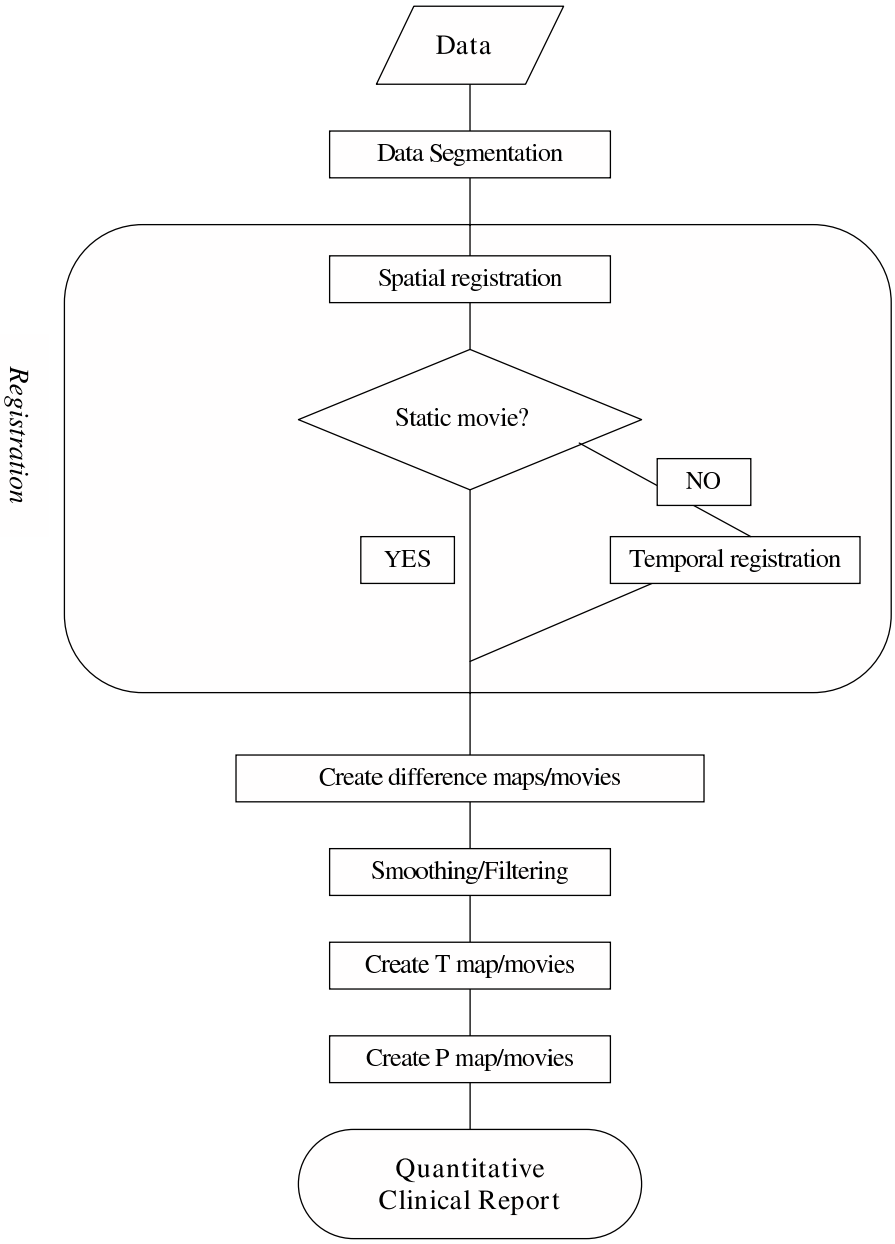


Figure 4.1: *LASR procedure flow chart*

test. However, it differs from the standard t test in the following way. Our test statistic at each pixel x is $T_x = D_x/S_x$, where D_x is the pixel value of a filtered difference image from Step 5, i.e. a *weighted* average of the difference values in a *neighborhood of* x from a difference image in Step 4 (versus a *simple* average of an *independent and identically distributed* sample drawn at the *same* location x, in a two-sample t test statistic), and S_x is an appropriately estimated standard deviation of D_x.

Step 7: *Compute FDR-controlled P maps and movies.* Based on the T images and movies, we can compute p-values at all pixels. Each of the p-values allows us to decide if two images are significantly different at that pixel. The BH-FDR method is applied to adjust the p-values. If a p-value p at x is less than the critical value derived from a 0.05 FDR-controlled procedure, we change the pixel value to $1 - p$; if p is greater than the FDR cut-off value, the pixel value is set to zero. These resulting FDR-controlled P maps or movies show which areas (the elevated areas) show improvement of interface pressures (implying improved tissue health).

In summary, the LASR output map gives a graphical representation of statistically significant pressure changes across the entire mapped region, *i.e.* it helps us to decide if the NMES is effective at a particular region, with an FDR no more than 0.05, an analogy of P-value for simultaneously comparing differences at many locations (e.g. pixels). The algorithm is applied frame-by-frame to aligned pressure data sets. LASR maps can thus be viewed as single frame "snapshots", suitable for comparison of static seating postures, or as videos for comparison of dynamically changing pressures.

4.1.2 Statistical Analyses and Results

In this section we present two typical analyses and results for both static and dynamic mappings.

Static pressure mapping

Subject A. Pressure mapping assessment for subject A "appeared" to show reasonable spatial alignment (Figure 4.2a), but a spatial registration was still conducted to align images and correct any differences in alignment that are not visually obvious. Qualitative evaluation of baseline/post-treatment pressure maps appeared to indicate some positive changes in pressure distribution over time, i.e. ischial region pressures appeared to decrease. However this could not be shown to be statistically significant without further detailed analysis.

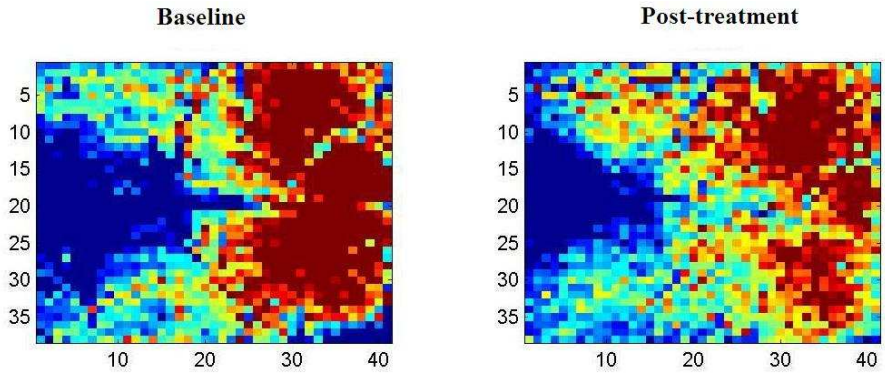
After applying the LASR algorithm to assess changes between baseline and post- treatment interface pressure data sets it could be seen that pressures were reduced bilaterally over time (Figure 4.2b). The left sacro-ischial region was more extensively affected than the right side.

Subject B. Pressure mapping assessment for subject B showed poor spatial alignment, with both translation and rotation occurring between the baseline and post-treatment images (Figure 4.3a). Qualitative evaluation of longitudinal changes could not readily be performed.

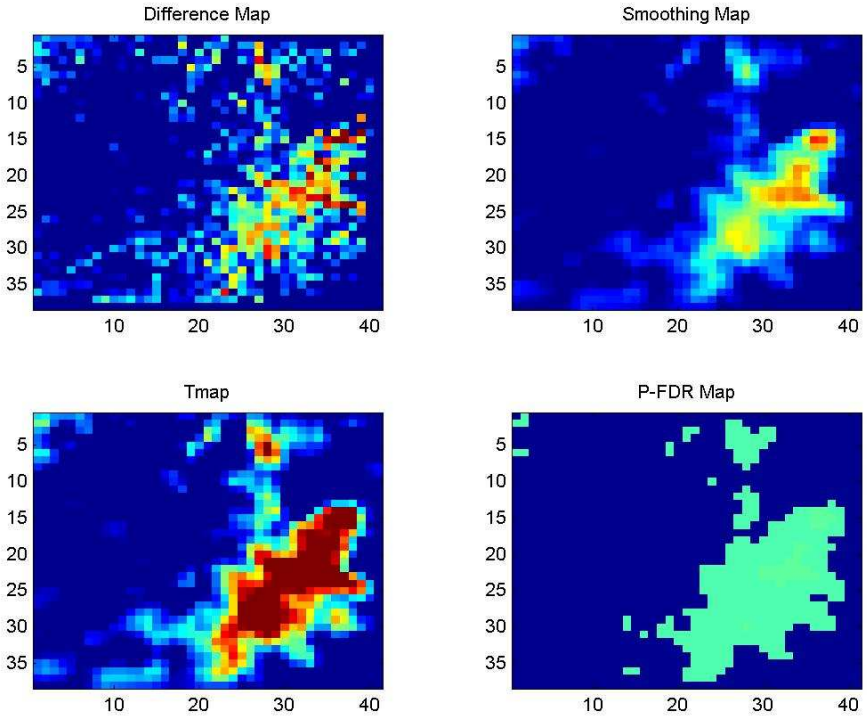
After applying the LASR algorithm to assess changes between baseline and post- treatment interface pressure data sets it could be seen that pressures were reduced bilaterally over time (Figure 4.3b). The left and right sacro-ischial regions were equally affected.

Dynamic pressure mapping

In developing the temporal registration stage of the LASR algorithm it was assumed that the pressure variations exhibited a regular periodicity. This allowed them to be brought into phase (temporally registered) for direct inter-assessment comparison. Dynamic changes in interface pressure distributions can then be presented in a video format, comparable to a motion analysis

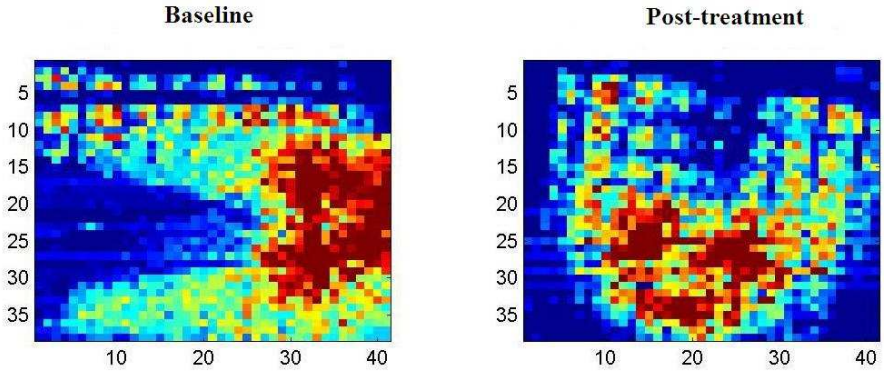


a: *Unprocessed pressure data maps, assessments repeated at a 6-month interval*

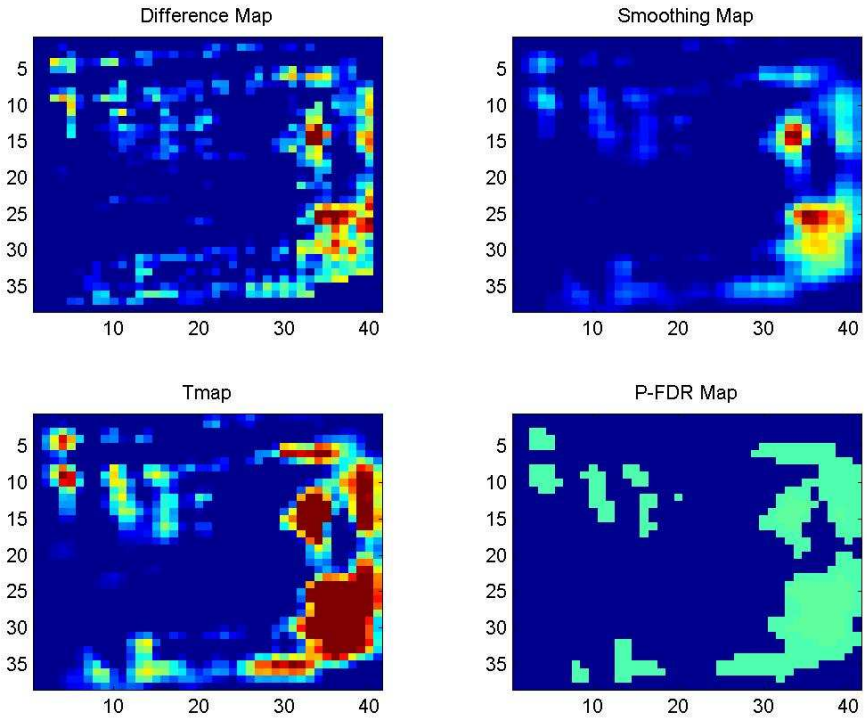


b: *LASR analysis of long-term changes in static mode seated pressure distribution*

Figure 4.2: *Pressure mapping analysis: subject A. LASR analysis results identify the regions of the pressure reduction. The left sacro-ischial region was more extensively affected than the right side*



a: *Unprocessed pressure data maps, assessments repeated at a 6- month interval*



b: *LASR analysis of long-term changes in static mode seated pressure distribution*

Figure 4.3: *Pressure mapping analysis: subject B. LASR analysis results identify the regions of the pressure reduction. The left and right sacro-ischial regions were equally affected.*

output.

In the current study the effects of dynamic gluteal NMES were assessed using real-time interface pressure mapping. Over several months of regular use it was proposed that the response to gluteal stimulation would increase as the stimulated muscles became stronger. Application of the LASR algorithm to initial stimulation data sets and response after 6 months of regular use showed significant changes in interface pressures for both subjects. Subject A showed changes predominantly on the left side, under the thigh region as well as the ischial region, with some areas of change also occurring in the right ischial region. Subject B showed changes bilaterally in the ischial region. Relevant LASR movies can be viewed at stat.case.edu/lasr/ or sun.case.edu/lasr/.

One clear advantage of examining FDR movies over FDR maps is to help decide which 5% of reported activations are most likely to be the false ones. This is because the false ones will not persistently appear to be significant over time (see our difference- and FDR- movies on stat.case.edu/lasr/ or sun.case.edu/lasr/). As shown on our LASR webpage, those in the upper left corner (reflecting the lower right thigh) of subject A and in the middle right for subject B (reflecting the sacral area), are most likely to be false positives. Note that for subject A we compared the baseline session with the third session in producing both the static and dynamic data movies. For subject B we did not have the baseline dynamic data, so the difference movie for the dynamic data is taken between the second and the third session, while the difference movie for the static data is taken between the first and third session. Nevertheless, looking at both dynamic and static P-movies for subject B, we still have more information (than if we had no movies) to decide which 5% of reported activations might be false ones.

4.2 Semiparametric Regression for the Spatial-Temporal Data

The LASR procedure above gives us a complete solution for analyzing the "intra-subject" data. Note that in the NMES study we have two "time" variables; one is the time over data frames in each sub-data set; the other is the time at intervals of roughly 12 months during the patient's participation in the study. Clinicians also would like to know the overall treatment effect or temporal trend of roughly 12 months (*i.e.* the second time variable) for all the subjects. Modeling the temporal trend from this large-p-small-n data is not easy. In order to solve the problem we first implement a further mining step – *data reduction* after data segmentation and registration. Recall that there are 400 frames in each sub-data set. To increase the efficiency of modeling overall treatment effect we reduce the huge data set to a smaller representative. That is, we take an average over 400 frames for each sub-data set. A single data frame is obtained at each assessment for each subject. We refer to the summarized data as "inter-subject" data. Our semiparametric regression model described below is proposed for analyzing this data.

Linear regression can be applied to the "inter-subject" data by modeling the intensities y_i as a function of spatial location $\mathbf{s}_i = (s_{1i}, s_{2i})$ and treatment t_i:

$$y_i = \beta_0 + \beta_1 s_{1i} + \beta_2 s_{2i} + \beta_3 t_i + \beta_4 t_i s_{1i} + \beta_5 t_i s_{2i} + \varepsilon_i$$

where y_i is the intensity value of subject i, s_{1i} and s_{2i} are the coordinates of y_i in the data frames, t_i is the dummy variable denoted before treatment or after treatment (t can be a continuous variable if one has enough assessments over the time) and ε_i is the normally distributed random error.

The linear regression model assumes independence of ε_i's which makes much of the statistical theory tractable. However, the data in the NMES study involves both spatial-correlation structures and temporal-correlation

structures. The spatial dependence is present in all directions and becomes weaker as data locations become more dispersed. Models that involve these dependencies are often more realistic than the ones ignoring the dependencies. Moreover, it is not reasonable that we assume there be only simple linear relations between spatial coordinates and the corresponding intensity values. To control the spatial heterogeneity (correlation) we introduce a Gaussian random field $Z(\mathbf{s})$ where $\mathbf{s}_i = (s_{1i}, s_{2i})$ to replace location covarites s_1, s_2 in the linear predictor. The new model is

$$y_i = \beta_0 + \beta_3 t_i + \beta_4 t_i s_{1i} + \beta_5 t_i s_{2i} + Z(\mathbf{s}_i) + \varepsilon_i \tag{4.1}$$

where Z is a stationary Gaussian random field on a bounded region $\mathcal{S} \in \mathbb{R}^2$ with mean $\mathrm{E}\{Z(\mathbf{s})\} = 0$ and isotropic covariance

$$\mathrm{cov}\{Z(\mathbf{s}), Z(\mathbf{s}')\} = \sigma^2\gamma(\|\mathbf{s} - \mathbf{s}'\|) = \sigma^2\gamma(\mathbf{r}) \tag{4.2}$$

with $\|\cdot\|$ denoting Euclidean norm (see Appendix 3 for the definition of random fields). Conditioning on Z the intensities y_i are independent normal observations.

Our approach to solve the spatial-temporal model is a semiparametric method based on mixed models, Karhunen-Loève expansion and regression splines. In the next subsection we introduce *reproducing kernel Hilbert space* (RKHS) and *Karhunen-Loève expansion* on a fairly elementary level with special emphasis on characteristics relevant to our solution of the semiparametric model. A more comprehensive discussion about RKHS and Karhunen-Loève expansion can be found in the monographs by Wahba (1990); Gu (2002); Adler (1981) and Berlinet and Thomas-Agnan (2004).

4.2.1 RKHS and Karhunen-Loève Expansion

We first give the definition of *Hilbert space*.

Definition 4.2.1 (Hilbert space)**.** *Every inner product* $\langle \cdot, \cdot \rangle$ *on a linear space* H *gives rise to a norm* $|| \cdot ||$ *as*

$$\|x\| = \sqrt{\langle x, x \rangle}.$$

We call H *a Hilbert space if it is complete with respect to this norm. Completeness in this context means that any Cauchy sequence of elements of the space converges to an element in the space, in the sense that the norm of differences approaches zero.*

RKHS

A reproducing kernel Hilbert space (RKHS) is, first of all, a Hilbert space. Intuitively speaking, an RKHS is a space of functions with the nice property that if a function f is close to a function g in the sense of the distance derived from the inner product, then the values $f(\mathbf{x})$ are close to the values $g(\mathbf{x})$. Among Hilbert spaces of functions an RKHS is characterized by the property that the evaluation of functions at a fixed point $\mathbf{x}$, $f \mapsto f(\mathbf{x})$ is a continuous mapping.

Definition 4.2.2 (reproducing kernel Hilbert space)**.** *Consider a Hilbert space* $\mathcal{H}$ *of functions on domain* $\mathcal{X}$*. If the evaluation functional* $\delta_{\mathbf{x}} f = f(\mathbf{x})$ *is continuous in* $\mathcal{H}$*,* $\forall x \in \mathcal{X}$*, then* $\mathcal{H}$ *is called a reproducing kernel Hilbert space.*

This implies that there exists a kernel $C(\mathbf{x}, \mathbf{x}')$ s.t. $C(\cdot, \mathbf{x}) \in \mathcal{H}$ for all $\mathbf{x} \in \mathcal{X}$ and

$$f(\mathbf{x}) = \delta_{\mathbf{x}} f = \langle C(\cdot, \mathbf{x}), f \rangle \tag{4.3}$$

for all $f \in \mathcal{H}$ where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{H}$. By the Riesz representation theorem (Gu, 2002, page 27) there exists a unique representer $C_{\mathbf{x}} \in \mathcal{H}$ such that (4.3) holds with $C(\cdot, \mathbf{x}) = C_{\mathbf{x}}$. It can be seen that the kernel C is positive semidefinite (Berlinet and Thomas-Agnan, 2004, chapter 1). C

is called a *reproducing kernel* (RK) of $\mathcal{H}$. The reproducing kernel of such a space $\mathcal{H}$ is a function of two variables $C(\mathbf{y}, \mathbf{x})$ with the property that for fixed x, the function of $\mathbf{y}$, $C(\mathbf{y}, \mathbf{x})$, denoted by $C(\cdot, \mathbf{x})$ belongs to $\mathcal{H}$ and represents the evaluation function at the point $\mathbf{x}$. Note that

$$\langle C_{\mathbf{x}}, C_{\mathbf{x}'} \rangle = \langle C(\cdot, \mathbf{x}), C(\cdot, \mathbf{x}') \rangle = C(\mathbf{x}, \mathbf{x}'). \tag{4.4}$$

In essence the RKHS is made up of functions that have about the same smoothness properties that $C(\mathbf{s}, \mathbf{t})$ has, as a function in $\mathbf{t}$ for fixed $\mathbf{s}$, or vice versa. Let us consider

$$S = \left\{ u : \mathcal{X} \to \mathbb{R} : u(\cdot) = \sum_{i=1}^{n} a_i C(\mathbf{s}_i, \cdot),\ a_i \text{ real},\ \mathbf{s}_i \in \mathcal{X}, n \geq 1 \right\}.$$

The inner product on S is

$$\begin{aligned} \langle u, v \rangle &= \langle \sum_{i=1}^{n} a_i C(\mathbf{s}_i, \cdot), \sum_{j=1}^{m} b_i C(\mathbf{t}_j, \cdot) \rangle \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j C(\mathbf{s}_i, \mathbf{t}_j) \geq 0. \end{aligned}$$

Note that by the reproducing kernel property

$$\begin{aligned} \langle u, C(\mathbf{t}, \cdot) \rangle &= \langle \sum_{i=1}^{n} a_i C(\mathbf{s}_i, \cdot), C(\mathbf{t}, \cdot) \rangle \\ &= \sum_{i=1}^{n} a_i C(\mathbf{s}_i, \mathbf{t}) \\ &= u(\mathbf{t}). \end{aligned} \tag{4.5}$$

For the sake of exposition assume that the covariance function, C, is positive definite (rather than merely positive semidefinite) so that $\langle u, u \rangle = 0$ if and only if $u(\mathbf{t}) \equiv 0$. In this case (4.5) defines a norm $\|u\| = \langle u, u \rangle^{\frac{1}{2}}$.

For $\{u_n\}_{n\geq 1}$ a sequence in S we have

$$\begin{aligned} |u_n(\mathbf{t}) - u_m(\mathbf{t})| &= |\langle u_n - u_m, C(\mathbf{t}, \cdot)\rangle| \\ &\leq \|u_n - u_m\| \cdot \|C(\mathbf{t}, \cdot)\| \\ &\leq \|u_n - u_m\| C(\mathbf{t}, \mathbf{t}), \end{aligned}$$

where the last line follows directly from (4.4) and (4.5). Hence it follows that if $\{u_n\}$ is Cauchy in $\|\cdot\|_{\mathcal{H}}$ then it is pointwise Cauchy. The closure of S under this norm is a space of real-valued functions, denoted by $\mathcal{H}(C)$, the RKHS of C, since every $u \in \mathcal{H}(C)$ satisfies (4.5) by the *separability* of $\mathcal{H}(C)$ (Adler, 1981). (The separability of $\mathcal{H}(C)$ follows from the separability of $\mathcal{X}$ and the assumption that C is continuous.)

As a concrete example of RKHS take $\mathcal{X} = \{1, 2, \cdots, n\}$ and f to be centered Gaussian process with covariance matrix $C = (c_{ij}), c_{ij} = E\{f_i f_j\}$. Let $C^{-1} = (c^{ij})$ denote the inverse of C, which exists by positive definiteness. Then the RKHS of f is made up of all n-dimensional vectors $u = (u_1, u_2, \cdots, u_n)$ with inner product

$$\langle u, v\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} u_i c^{ij} v_j.$$

Karhunen-Loève Expansion

It is noted that $\mathcal{L}_2(\mu)$ is not an RKHS in general, but for many kernels C it contains a (unique) RKHS as a subspace (Gu, 2002; Seeger, 2004). Recall that $\mathcal{L}_2(\mu)$ contains all functions $f : \mathcal{X} \to \mathbb{R}$ for which

$$\int f(\mathbf{x})^2 \, d\mu(\mathbf{x}) < \infty$$

holds. The standard inner product is

$$(f, g) = \int f(\mathbf{x}) g(\mathbf{x}) \, d\mu(\mathbf{x}).$$

Often μ is taken to be an indicator function of a compact set such as the unit hypercube. A positive semidefinite $C(\mathbf{x}, \mathbf{x}')$ can be regarded as a kernel (or represent) of a positive semidefinite linear operator $\mathcal{C}$ in the sense

$$(\mathcal{C}f)(\mathbf{x}) = \langle C(\cdot, \mathbf{x}), f \rangle.$$

Now, ϕ is an eigenfunction of C with eigenvalue $\lambda \neq 0$ if

$$(\mathcal{C}\phi)(\mathbf{x}) = \langle C(\cdot, \mathbf{x}), \phi \rangle = \lambda\phi.$$

For $\mathcal{C}$ all eigenvalues are real and non-negative. Furthermore, suppose $\mathcal{C}$ is continuous and

$$\int C(\mathbf{x}, \mathbf{x}')^2 \, d\mu(\mathbf{x}) \, d\mu(\mathbf{x}') < \infty.$$

For simplicity take $\mathcal{X} = [0, 1]^N$. Let $\lambda_1 \geq \lambda_2 \geq \cdots$, and $\phi_1, \phi_2, \ldots$, be, respectively, the eigenvalues and normalized eigenfunctions of operator $\mathcal{C} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ defined by $(\mathcal{C}\psi)(\mathbf{t}) = \int_T C(\mathbf{s}, \mathbf{t})\psi(\mathbf{s}) \, d\mathbf{s}$. That is, the λ_n and ψ_n solve the integral equation

$$\int_{\mathcal{X}} C(\mathbf{s}, \mathbf{t})\psi(\mathbf{s}) \, ds = \lambda\psi(\mathbf{t}), \tag{4.6}$$

with the normalization

$$\int_{\mathcal{X}} \psi_n(\mathbf{t})\psi_m(\mathbf{t}) \, d\mathbf{t} = \begin{cases} 1 & m = n \\ 0 & m \neq n \end{cases}$$

These eigenfunctions lead to a natural expansion of C, known as Mercer's Theorem.

Theorem 4.2.1 (Mercer)**.** *Let C, $\{\lambda_n\}_{n \geq 1}$ and $\{\psi_n\}_{n \geq 1}$ be as above. Then*

$$C(\mathbf{s}, \mathbf{t}) = \sum_{n=1}^{\infty} \lambda_n \psi_n(\mathbf{s})\psi_n(\mathbf{t}), \tag{4.7}$$

where the series converges absolutely and uniformly on $[0, 1]^N \times [0, 1]^N$.

The Mercer theorem leads to an important representation of a zero-mean Gaussian random field $Z(\mathbf{s})$ with covariance function C – the *Karhunen-Loève expansion*,

$$Z(\mathbf{s}) = \sum_{n=1}^{\infty} \lambda_n^{\frac{1}{2}} \psi_n(\mathbf{s}) \xi_n \tag{4.8}$$

where $\varphi_n = \lambda_n^{\frac{1}{2}} \psi_n$ is an orthonormal expansion for the RKHS $\mathcal{H}(C)$ and ξ_n are i.i.d $N(0,1)$. Sun (1993) gave general conditions for (4.8) to exist for a fairly arbitrary smooth Gaussian random field.

4.2.2 Mixed Modeling

From the discussion of the Karhunen-Loève expansion above, every Gaussian random field whose covariance function satisfies weak constrains (specified by Sun (1993) which includes all smooth second order stationary process) can be written as

$$Z(\mathbf{s}) = \sum_{l=1}^{K} \widetilde{Z}_l(\mathbf{s}) u_l = \widetilde{\mathbf{Z}}(\mathbf{s})^T \mathbf{u} \tag{4.9}$$

where $\mathbf{u} = (u_1, \cdots, u_K)^T$ are random variables following $N(0, \sigma_u^2 \mathbf{I})$, where K can be finite or infinite.

To develop this view we consider the Karhunen-Loève expansion for $Z(\mathbf{s})$. Under mild conditions on the covariance function $C(\mathbf{s}, \mathbf{s}')$ of $Z(\mathbf{s})$ we can construct a sequence

$$\sum_{l=1}^{K} \lambda_l^{\frac{1}{2}} \psi_l(\mathbf{s}) \xi_l,$$

which converges to $Z(\mathbf{s})$ in quadratic mean ($K \to \infty$). Here the ξ_l are i.i.d. $N(0,1)$ variables. ψ_l are orthonormal eigenfunctions of the operator induced by C with corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, $\sum_{l \geq 1} \lambda_l^2 < \infty$,. Thus, if $\mathbf{u} = \left[\sigma_u \xi_l\right]_{1 \leq l \leq K}$ and $\widetilde{\mathbf{Z}}(\mathbf{s})^T = \left[\lambda_l^{\frac{1}{2}} \psi_l(\mathbf{s}) / \sigma_u\right]_{1 \leq l \leq K}$, then $\widetilde{\mathbf{Z}}(\mathbf{s})^T \mathbf{u} \to$

$Z(\mathbf{s})$ in quadratic mean.

The random field models of interest to us have their origin in spline smoothing techniques and penalized likelihood estimation. Also, for low-dimensional input spaces spline kernels are widely used due to the favorable approximation properties of splines and their computational advantages. Spline smoothing is a special case of penalized likelihood methods which provides another view on the reproducing kernel via the Green's function of a penalization operator. We refer the monographs by Gu (2002) and Wahba (1990) who give excellent discussions of spline techniques from the RKHS perspective.

Returning to our model (4.1), by the Karhunen-Loève expansion of the random field $Z(\mathbf{s})$ our model can be rewritten as a standard linear mixed model, *i.e.*

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \widetilde{\mathbf{Z}}\mathbf{u} + \boldsymbol{\varepsilon} \tag{4.10}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_5)^T$, $\mathbf{u} = (u_1, \cdots, u_K)^T$,

$$\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2\mathbf{I} & 0 \\ 0 & \sigma_\varepsilon^2\mathbf{I} \end{bmatrix} \right).$$

$\widetilde{\mathbf{Z}}$ can be derived by using a spatial extension of the penalized spline. Our choice of the penalized spline is the *general radial spline* (Ruppert et al., 2003) which corresponds to the *thin plate spline* family. The low-rank radial spline is computed based on a matrix of correlation functions $\mathbf{C} = \left[C(\|\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'}\|) \right]_{1 \le k,k' \le K}$, where $C(\mathbf{r})$ is a radially symmetric function that approximates $\gamma(\mathbf{r})$ (defined by the correlation function (4.2)); $\boldsymbol{\kappa}$ are knots and $K < n$ is the number of the knots. The radial centers for the correlation are a set of K knots. Following Ruppert et al. (2003, page 254) we model the correlation by using the function

$$C(\mathbf{r}) = \|\mathbf{r}\|^2 \log \|\mathbf{r}\|.$$

Singular value decomposition of the $K \times K$ matrix

$$\mathbf{C} = \left[\|\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'}\|^2 \log \|\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'}\| \right]_{1 \le k,k' \le K}$$

yields

$$\mathbf{C} = \mathbf{U}\text{diag}(\mathbf{d})\mathbf{V}^T,$$

where $\mathbf{d}$ consists of singular values $\mathbf{C}$: $\{\lambda_l : \lambda_1 > \lambda_2 > \cdots > \lambda_K > 0\}$. The matrix square root of C is then

$$\mathbf{C}^{1/2} = \mathbf{U}\text{diag}(\sqrt{\mathbf{d}})\mathbf{V}^T.$$

Notice that for each data frame in the NMES study we have inputs $y_1, \cdots, y_S \in \mathbb{R}$ and two-dimensional spatial locations $\mathbf{s}_1, \cdots, \mathbf{s}_S \in \mathbb{R}^2$. Now we can define the $S \times K$ matrix $\widetilde{\mathbf{Z}}$ as

$$\widetilde{\mathbf{Z}} = \mathbf{Z}_K \mathbf{C}^{-1/2},$$

where

$$\mathbf{Z}_K = \left[\|\mathbf{s}_i - \boldsymbol{\kappa}_{k'}\|^2 \log \|\mathbf{s}_i - \boldsymbol{\kappa}_k\| \right]_{1 \le i \le S,\, 1 \le k \le K}$$

The choice of K will be discussed later.

Therefore, the spline technique helps us to find the Karhunen-Loève expansion of the random field. Then the form of model (4.9) allows fitting through standard mixed model software. Our spatial-temporal model can then be obtained by applying (restricted) maximum likelihood to $\boldsymbol{\beta}$, σ_u^2 and σ_ε^2 and best prediction to $\mathbf{u}$.

Test for the Random field

A hypothesis of interest in the spatial-temporal model is that there is no random field effect, *i.e.*

$$H_0 : Z(\mathbf{s}) = 0 \qquad v.s. \qquad H_1 : otherwise$$

This hypothesis is equivalent to

$$H_0 : \sigma_u^2 = 0 \qquad v.s. \qquad H_1 : \sigma_u^2 > 0 \tag{4.11}$$

Thus the problem is converted to variance component testing in linear mixed models. In general linear model, if $L(\boldsymbol{\theta})$ be the likelihood of the parameter vector $\boldsymbol{\theta}$ based on the data, the classical likelihood ratio test is under H_0,

$$-2\log\frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})} \sim \chi^2_\nu, \tag{4.12}$$

where $\hat{\boldsymbol{\theta}}_0 = (\hat{\beta}_0, \hat{\sigma}^2_{\varepsilon_0})$ and $\hat{\boldsymbol{\theta}} = (\hat{\beta}, \hat{\sigma}^2_\varepsilon, \hat{\sigma}^2_u)$ are the maximum likelihood estimates of $\boldsymbol{\theta}$ under the null model and unrestricted model, respectively. The degrees of freedom ν is the difference between the number of parameters in the unrestricted model and the null model. However, (4.12) assumes that the parameter of interest is not on the boundary of its parameter space. This assumption is violated for hypothesis test (4.11) since the parameter space for σ_u^2 is $[0, \infty)$. A correction for the asymptotic distribution under H_0 of (4.11) (Self and Liang, 1987) is:

$$-2\log\frac{L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2_\varepsilon)}{L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2_\varepsilon, \hat{\boldsymbol{\sigma}}^2_u)} \sim \frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1. \tag{4.13}$$

Choice of smoothing parameter

The low-rank radial spline technique uses a truncation of the decomposed smoothing basis to compute the covariates of the random effect in the mixed model. This truncation corresponds to the K-rank approximation to the covariance matrix of the random field. Choosing K is critical in that K determines the accuracy of how well the spline approximates the true covariance. We suggest a model selection approach for choosing the rank K, i.e. use the AIC of the fitted model. In practice the model is fitted by using increasing numbers of K; the optimal K is chosen by minimizing the AIC.

Effect	Treatment	Estimate	Standard Error	$Pr > \|t\|$
Intercept		17.89	0.96	<0.0001
Treatment	Before	5.83	1.58	0.009
Treatment	After	0	.	.
Trt*s_1	Before	0.88	0.09	<0.0001
Trt*s_1	After	0	.	.
Trt*s_2	Before	-0.35	0.19	0.01
Trt*s_2	After	0	.	.

Table 4.1: *Statistical results for semiparametric model fitting. The solution for fixed effects is given here. The treatment effect is significant indicating the efficiency of NMES.*

Statistical Results

Table 4.1 displays the solution for the fixed effects of our semiparametric model. *MIXED Procedure* in SAS software is used to solved the mixed models. Each exploratory variable is significant. Note that treatment effect is highly significant indicating the efficiency of NMES on paralyzed muscles. The optimal number of knots of the radial splines is $K = 20$ by the AIC criterion.

From this study it is seen that the semiparametric model we proposed is an effective tool for the analysis of spatial-temporal data and allows a fast processing of large databases. Spatial models are a more recent addition to the statistics literature (Cressie, 1993). Image processing, epidemiology, ecology, geology, forestry, astronomy, climatology or simply any discipline that works with data collected from different spatial locations, need to develop models that indicate when there is dependence between measurements at different locations. Our proposed semiparametric model is applicable to these research areas and can easily be implemented by standard software.

Chapter 5

Measurement Error Problems

In this chapter we study the estimation problems of nonparametric densities and regression functions in which variables are measured with error. Non-Fourier based estimators are developed in the case of both homogeneous and nonhomogeneous normal errors. The asymptotics of the new estimators are investigated.

5.1 Density Estimation for Data with Measurement Errors

The problem of nonparametric estimation of curves such as probability densities and regression functions in the presence of measurement error has been studied considerably in the literature. In the density estimation setting the problem has been stated in Chapter One. Let X be the measurement of interest and U the measurement error. Assume that X and U are continuous independent random variables, with X having densities f and U having densities w. Then the random variable $Y = X + U$ has density $g = f * w$ where $*$ is the convolution operator. The problem of estimating f from a sample $Y_1, \cdots, Y_n$ is referred to as the deconvolution density estimation problem. The usual

procedure is by a Fourier inversion. Let ϕ_X, ϕ_U, ϕ_Y denote the characteristic function of X, U and Y, respectively. Then $\phi_Y(t) = \phi_X(t)\phi_U(t)$. So an inverse Fourier transformation leads to,

$$f(x) = \frac{1}{2\pi}\int e^{-itx}\phi_X(t)dt = \frac{1}{2\pi}\int e^{-itx}\frac{\phi_Y(t)}{\phi_U(t)}dt. \tag{5.1}$$

An estimator of $f(x)$ can be obtained by substituting $\phi_Y(t)$ in (5.1) by its estimate

$$\hat{\phi}_Y = \frac{1}{n}\sum_{i=1}^{n} e^{itY_i},$$

and $\phi_U(t)$ by its explicit expression (assumed known); let's call the resulting estimate the plug-in estimate. However, in practice this plug-in estimate is unstable because its characteristic function has large fluctuations at tails.

To avoid this defect, a "tamper" function $W(h_n t)$ is inserted into the integral (5.1) where $h_n \to 0$ is an appropriately selected tuning parameter. When $W = \phi_K$, the characteristic function of a kernel function K such that $\phi_K(0) = 1$, the plug-in estimator with $W = \phi_K$ is the following deconvolving kernel density estimator introduced by Stefanski and Carroll (1990),

$$\hat{f}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K^*\left(\frac{x - Y_i}{h_n}\right), \tag{5.2}$$

where

$$K^*(z) = \frac{1}{2\pi}\int e^{-itz}\frac{\phi_K(t)}{\phi_U(t/h_n)}dt \tag{5.3}$$

is called the deconvolving kernel.

Observe that by (5.2) the deconvolving kernel estimate (DKE) is just an ordinary kernel estimate but with specific kernel function equal to (5.3) and the bandwidth h_n. The convergence rate for DKE can be very slow for some error distributions. Especially, when errors belong to the normal family, the

convergence rate is only $O\big((\log n)^{-1/2}\big)$ (Zhang, 1990; Fan, 1991). In addition, the Fourier estimate above is based on the assumption that errors are homogeneous. Therefore, in this chapter we look for "non-Fourier" estimators that are effective and that work for both homogeneous and nonhomogeneous errors.

Sun et al. (2002) considered non-Fourier estimators of density estimation in the uniform error case, which opened a completely new line of attack. In the homogeneous error case, $U_i \sim \mathcal{U}(-\theta, \theta)$ independently. It is easy to see that the common density g of an independent simple $Y_1, \cdots, Y_n$ is

$$g(y) = \frac{1}{2\theta}[F(y+\theta) - F(y-\theta)].$$

where F is the distribution of X. Therefore, F can be recovered from the density g by either a "left" series representation F_- or a "right" series representation F_+ defined below,

$$F_-(x) \triangleq 2\theta \sum_{t=0}^{\infty} g\left(x - (2t+1)\theta\right), \tag{5.4}$$

$$F_+(x) \triangleq 1 - 2\theta \sum_{t=0}^{\infty} g\left(x + (2t+1)\theta\right). \tag{5.5}$$

In another word,

$$F(x) = F_-(x) = F_+(x)$$

provided the both series converge.

Moreover, the equations (5.4) and (5.5) lead to expressions for the density function f,

$$f_-(x) = 2\theta \sum_{t=0}^{\infty} g'\left(x - (2t+1)\theta\right), \tag{5.6}$$

$$f_+(x) = -2\theta \sum_{t=0}^{\infty} g'\left(x + (2t+1)\theta\right), \tag{5.7}$$

so that a usual density estimation of g can be used to estimate F and f:

$$\hat{F}_-(x) = 2\theta \sum_{t=0}^{m_n} \hat{g}\left(x - (2t+1)\theta\right), \qquad \hat{F}_+(x) = 1 - 2\theta \sum_{t=0}^{m_n} \hat{g}\left(x + (2t+1)\theta\right),$$

$$\hat{f}_-(x) = 2\theta \sum_{t=0}^{m_n} \hat{g}'\left(x - (2t+1)\theta\right), \qquad \hat{f}_+(x) = -2\theta \sum_{t=0}^{m_n} \hat{g}'\left(x + (2t+1)\theta\right),$$

where $m_n \to \infty$ and $\hat{g}$ can be a typical kernel estimator

$$\hat{g}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - Y_i}{h}),$$

or any reasonable nonparametric estimator. Here the kernel K is assumed to have the properties of $\int K = 1, \int K^2 < \infty$ and $\int xK(x)dx = 0$.

In the nonhomogeneous case a similar procedure leads to estimators:

$$\hat{F}_-(x) = 2\sum_{t=0}^{m_n} \hat{g}_-(x,t), \qquad \hat{F}_+(x) = 1 - 2\sum_{t=0}^{m_n} \hat{g}_+(x,t),$$

$$\hat{f}_-(x) = 2\sum_{t=0}^{m_n} \hat{g}'_-(x,t), \qquad \hat{f}_+(x) = -2\sum_{t=0}^{m_n} \hat{g}'_+(x,t),$$

where $m_n \to \infty$ and a kernel-type estimator $\hat{g}_-$ and $\hat{g}_+$ are

$$\hat{g}_+ = \frac{1}{nh} \sum_{i=1}^{n} \theta_i K\left(\frac{x - 2(t+1)\theta_i - Y_i}{h}\right),$$

$$\hat{g}_- = \frac{1}{nh} \sum_{i=1}^{n} \theta_i K\left(\frac{x + 2(t+1)\theta_i - Y_i}{h}\right).$$

The estimators above are stable and easy to compute - there are no Fourier transformations needed in the calculation. Sun et al. (2002) also show that the rates of their optimal estimators are $n^{-2/5}$ for the cumulative distribution and $n^{-1/5}$ for the density of X. This is in contrast to the slow convergence rates of Fourier deconvolution estimators when errors are either ordinary or super smooth as defined by Fan (1991).

5.1.1 3U Deconvolving Density Estimators

Following the sucess of Sun et al. (2002), in this section we study the density estimation problem when measurement errors are approximately normal. Our idea is inspired from random number generation (RNG). Most RNG are built on the uniform RNG. For example, normal pseudo random deviates can be generated well by the sum of 12 uniform random variables (Gentle, 2003) or by the popular Marsaglia-Bray algorithm. The Marsaglia-Bray composition method (Ripley, 1987, page 84) for normal RNG generates 97.45% normal random variables using linear combinations of 2 or 3 uniforms and another (99.73-97.45)% normal random variables using a rejection method based on 2 uniforms. In summary, more than 97% times, a sum of 3 uniforms provides adequate approximation to a normal RNG.

Indeed, there is little visual difference between the densities of the standard normal and a rescaled sum of three uniforms. Figure 5.1 shows histograms and density plots for random numbers from normal and 3 uniforms. The upper two subplots are from standard normal random numbers and the lower two subplots are from a sum of three uniforms (*i.e.* generate X from $X = 2(U_1 + U_2 + U_3) - 3$ where U_1, U_2, U_3 are from $\mathcal{U}(0, 1)$). We also implement a simple Monte Carlo simulation to compare the two distributions. 1000 random numbers are generated from standard normal and the sum of 3 uniforms separately. Then Kolmogorov-Smirnov test is used to test the normality of each sample. We repeat the procedure 100 times. Table 5.1 shows that the average p-values are very large in both distributions which indicates the samples are not significantly different from the standard normal. Among the 100 tests, only 7 and 5 samples are rejected at level 5% in the random number generations of the normal and the 3 uniforms, respectively.

In practice, any model including normal assumption of measurement error is only approximately true. So, in the case of measurement error models

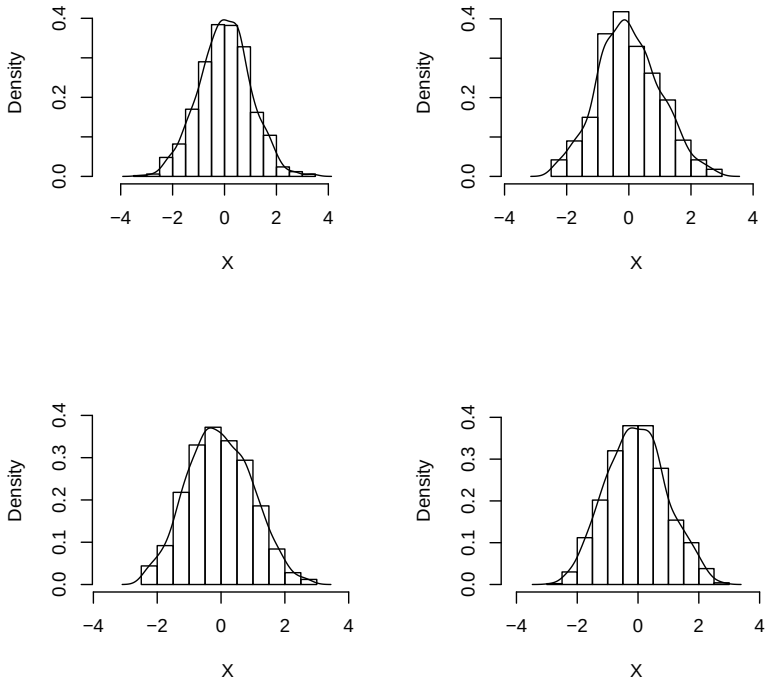


Figure 5.1: *Histograms and density plots for random numbers from normal and 3 uniforms. It is hard to distinguish the visual difference between the densities of the standard normal and a rescaled sum of three uniforms. The upper plots are from normal; the lower plots are from a sum of the three uniforms.*

Distribution	Average p-value	Percentage of rejection
3 uniforms	0.411	7/100
Normal	0.549	5/100

Table 5.1: *Monte Carlo simulation to compare random numbers from normal and 3 uniforms: Kolmogorov-Smirnov test is used to test the normality for each sample. 1000 random numbers are generated for each distribution in each time and the procedure is repeated 100 times.*

with normal error we propose to use the mixtures of sums of 3 uniforms to approximate the normal errors. Let us first focus on the homogeneous normal errors. Consider $Y = X + E$, where X and E are independent and $E \overset{appr}{\sim} \mathcal{N}(0, \sigma^2)$. We would like to estimate the cumulative distribution function (CDF) and probability density function (PDF) of X. Note that $E = \sigma V$ where $V \overset{appr}{\sim} \mathcal{N}(0, 1)$. By the Marsaglia-Bray algorithm, we consider the following approximation,

$$E \approx \sigma \left(2(U_1 + U_2 + U_3) - 3\right)$$

where $U_1, U_2, U_3 \sim \mathcal{U}(0, 1)$ independently.

Let $\widetilde{U}_1 = \sigma(2U_1 - 1)$, $\widetilde{U}_2 = \sigma(2U_2 - 1)$, $\widetilde{U}_3 = \sigma(2U_3 - 1)$, we have

$$Y = X + E \approx X + \widetilde{U}_1 + \widetilde{U}_2 + \widetilde{U}_3$$

where $\widetilde{U}_1, \widetilde{U}_2, \widetilde{U}_3 \sim U(-\sigma, \sigma)$ independently.

Consider the 3-fold estimating procedure of Sun et al. (2002):

$$\begin{cases} Y = Y_1 + \widetilde{U_3} \\ Y_1 = Y_2 + \widetilde{U_2} \\ Y_2 = X + \widetilde{U_1} \end{cases}$$

Denote $g_1(y), g_2(y)$ as the PDF of Y_1, Y_2 and apply (5.6) and (5.4),

$$\begin{cases} f(y) &= 2\sigma \sum_{t_3=0}^{\infty} g_2' \left(y - (2t_3 + 1)\sigma\right) \\ g_2(y) &= 2\sigma \sum_{t_2=0}^{\infty} g_1' \left(y - (2t_2 + 1)\sigma\right) \\ g_1(y) &= 2\sigma \sum_{t_1=0}^{\infty} g' \left(y - (2t_1 + 1)\sigma\right) \end{cases}$$

Hence the CDF and PDF of x are,

$$F_-(x) = 8\sigma^3 \sum_{t_3=0}^{\infty} \sum_{t_2=0}^{\infty} \sum_{t_1=0}^{\infty} g'' \left(x - (2t_1 + 2t_2 + 2t_3 + 3)\sigma\right) \tag{5.8}$$

$$f_-(x) = 8\sigma^3 \sum_{t_3=0}^{\infty} \sum_{t_2=0}^{\infty} \sum_{t_1=0}^{\infty} g''' \left(x - (2t_1 + 2t_2 + 2t_3 + 3)\sigma\right). \tag{5.9}$$

Similarly, if we apply (5.7) and (5.5), we obtain:

$$F_+(x) = 1 - 8\sigma^3 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} g''\left(x + (2t_1 + 2t_2 + 2t_3 + 3)\sigma\right) \quad (5.10)$$

$$f_+(x) = -8\sigma^3 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} g'''\left(x + (2t_1 + 2t_2 + 2t_3 + 3)\sigma\right). \quad (5.11)$$

Then a usual density estimate of g can be used to estimate (5.8) – (5.11). Our three-fold estimators are:

$$\widetilde{F}_-(x) = 8\sigma^3 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} \hat{g}''\left(x - (2t_1 + 2t_2 + 2t_3 + 3)\sigma\right) \quad (5.12)$$

$$\widetilde{f}_-(x) = 8\sigma^3 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} \hat{g}'''\left(x - (2t_1 + 2t_2 + 2t_3 + 3)\sigma\right) \quad (5.13)$$

$$\widetilde{F}_+(x) = 1 - 8\sigma^3 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} \hat{g}''\left(x + (2t_1 + 2t_2 + 2t_3 + 3)\sigma\right) \quad (5.14)$$

$$\widetilde{f}_+(x) = -8\sigma^3 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} \hat{g}'''\left(x + (2t_1 + 2t_2 + 2t_3 + 3)\sigma\right) \quad (5.15)$$

where

$$\hat{g}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - Y_i}{h}) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - Y_i).$$

Note that we use $K_h(x) = K(x/h)/h$ thereafter, h is a bandwidth, and K is a symmetric kernel with a finite variance such that $\int K = 1, \int K^2 < \infty$ and $\int xK(x)dx = 0$. We call these three-fold estimators *3U deconvolving estimators*.

5.1.2 Asymptotic Performance

We explore the asymptotic behavior of the 3U deconvolving estimators in this section under some regularity conditions. First we state the regularity

conditions that are inherited from Sun et al. (2002).

Condition 1 (Kernel). *The kernel K is nonnegative such that $\int K = 1$, $\int K^2 < \infty$, $\int xK(x)dx = 0$, $\int K''(x)dx = 0$ and $\int xK''(x)dx = 0$.*

Condition 2 (Density). *The density g has continuous fourth derivatives and satisfies that $\sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty}|g''(x-(2t_1+2t_2+2t_3+3)\sigma)|^{1/2}$ converges uniformly in x.*

Condition 3 (CDF). *The cumulative distribution function F of X is twice integrable.*

Condition 4 (Errors). *The error $E \stackrel{D}{=} \widetilde{U}_1 + \widetilde{U}_2 + \widetilde{U}_3$ exactly, where "$\stackrel{D}{=}$" denotes the equal in distribution; and $\widetilde{U}_1$, $\widetilde{U}_2$, $\widetilde{U}_3$ are identically independently distributed from $\mathcal{U}(-\sigma,\sigma)$.*

Remark. *A typical kernel that satisfies condition 1 is the Gaussian kernel $K(x) = (2\pi)^{1/2}e^{-x^2/2}$. For a discussion of condition 2 see Sun et al. (2002).*

Theorem 5.1.1. *Under Conditions 1 ,2, 3 and 4,*

$$\begin{aligned} \mathrm{E}\widetilde{F}_-(x) &= \mathrm{E}\widetilde{F}_+(x) = K_h * F(x) \\ \mathrm{E}\widetilde{f}_-(x) &= \mathrm{E}\widetilde{f}_+(x) = K_h * f(x) \end{aligned}$$

where $*$ is the convolution operator.

Proof: We shall prove $\mathrm{E}\widetilde{F}_-(x) = K_h * K(x)$. One can easily obtain the other results similarly. To simplify the notation we denote $\tilde{t} = 2(t_1+t_2+t_3)+3$. Note that

$$\begin{aligned} \mathrm{E}\widetilde{F}_-(x) &= 8\sigma^3 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[K_h''(x-\sigma\tilde{t}-Y_i)\right] \\ &= 8\sigma^3 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty}\mathrm{E}\left[K_h''(x-\sigma\tilde{t}-Y_1)\right] \\ &= 8\sigma^3 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty}\int K_h''(x-\sigma\tilde{t}-y)g(y)\,dy. \end{aligned} \tag{5.16}$$

Let $s = x - \sigma\tilde{t} - y$ and

$$H(x) = 8\sigma^3 \sum_{t_3=0}^{\infty} \sum_{t_2=0}^{\infty} \sum_{t_1=0}^{\infty} g(x - \sigma\tilde{t}), \tag{5.17}$$

then

$$\mathrm{E}\widetilde{F}_-(x) = \int K_h''(s) H(x-s)\, ds.$$

Note that $F(x) = H''(x)$ and by condition 1 and using partial integration, we show the assertion. □

Remark. *It is clear that* $K_h * F(x) \to F(x)$ *and* $K_h * f(x) \to f(x)$ *as* $h \to 0$.

Theorem 5.1.2. *Under Conditions 1 ,2, 3 and 4, we have*

$$\sqrt{nh^5}\left[\widetilde{F}_-(x) - \mathrm{E}\widetilde{F}_-(x)\right] \xrightarrow{\mathcal{D}} N\left(0,\ 8\sigma^3 H(x)\|K''\|^2\right)$$

as $h \to 0, nh^5 \to \infty$, *where* $H(x) = \int_{-\infty}^{x}\int_{-\infty}^{s} F(v)\, dv\, ds$ *and* $\|K''\|^2 = \int K''(x)^2\, dx$ *is the* L^2 *norm of* K''.

Proof: By equation (5.16) we then can write

$$\sqrt{nh^5}\left[\widetilde{F}_-(x) - \mathrm{E}\widetilde{F}_-(x)\right] = \sum_{k=1}^{n} \xi_{nk}$$

with

$$\xi_{nk} = 8\sigma^3\sqrt{\frac{h^5}{n}} \sum_{t_3=0}^{\infty} \sum_{t_2=0}^{\infty} \sum_{t_1=0}^{\infty} \left(K_h''(x - \sigma\tilde{t} - Y_k) - \int K_h''(x - \sigma\tilde{t} - y) g(y)\, dy \right).$$

By the *Lindeberg-Feller central limit theorem* for triangular arrays (Ferguson, 1996, Section 5), we have

$$\frac{\sum_{k=1}^{n} \xi_{nk}}{s_n} \xrightarrow{\mathcal{D}} N(0,1),$$

where

$$
\begin{aligned}
s_n^2 &= \sum_{k=1}^{n} \mathrm{Var}\xi_{nk} = nE\xi_{n1}^2 \\
&= (8\sigma^3)^2 h^5 \mathrm{Var}\left(\sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} K_h''(x-\sigma\tilde{t}-Y_1) \right) \\
&= (8\sigma^3)^2 h^5 \int \left(\sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} K_h''(x-\sigma\tilde{t}-y) \right)^2 g(y)\,dy \\
&\qquad\qquad - (8\sigma^3)^2 h^5 \left(\int \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} K_h''(x-\sigma\tilde{t}-y) g(y)\,dy \right)^2 \\
&= 8\sigma^3(A_1 - A_2^2).
\end{aligned}
$$

Because K follows condition 1 and for sufficiently small h, we have,

$$
\begin{aligned}
A_1 &= 8\sigma^3 h^5 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} \int K_h''(x-\sigma\tilde{t}-y)^2 g(y)\,dy \\
&= 8\sigma^3 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} \int K''(z)^2 g(x-\sigma\tilde{t}-hz)\,dz \qquad (Let\ z = (x-\sigma\tilde{t}-y)/h)
\end{aligned}
$$

Notice that by (5.8) and (5.17),

$$
\begin{aligned}
8\sigma^3 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} g(x-\sigma\tilde{t}-hz) &= \int_{-\infty}^{x} \left(\int_{-\infty}^{s} F(v-hz)\,dv \right) ds \\
&= H(x-hz).
\end{aligned}
$$

By interchanging the order between integral and summation,

$$
A_1 = \int K''(z)^2 H(x-hz)\,dz \longrightarrow H(x) \int K''(z)^2\,dz = H(x)\|K''\|^2
$$

as $h \to 0$, by dominated convergence. On the other hand,

$$\begin{aligned} A_2 &= \sqrt{8\sigma^3 h^5} \sum_{t_3=0}^{\infty} \sum_{t_2=0}^{\infty} \sum_{t_1=0}^{\infty} \int \left[K_h''(x - \sigma\tilde{t} - y)\right]^2 g(y)\, dy \\ &= \sqrt{8\sigma^3 h} \sum_{t_3=0}^{\infty} \sum_{t_2=0}^{\infty} \sum_{t_1=0}^{\infty} \int K''(z)^2 g(x - \sigma\tilde{t} - hz)\, dz \\ &= \sqrt{h/(8\sigma^3)} \int K''(z)^2 H(x - hz)\, dz \end{aligned}$$

which is of smaller order than the first term. Therefore, $s_n^2 \to 8\sigma^3 H(x)\|K''\|^2$.

Now let us verify the *Lindeberg condition*. For every $\varepsilon > 0$,

$$\frac{1}{s_n^2} \sum_{k=1}^{n} \mathrm{E}\left[\xi_{nk}^2 I\{|\xi_{nk}| > \varepsilon s_n\}\right] = \frac{1}{\mathrm{E}\xi_{n1}^2} \mathrm{E}\left[\xi_{n1}^2 I\{|\xi_{n1}| > \varepsilon\sqrt{n}\mathrm{E}\xi_{n1}^2\}\right] \tag{5.18}$$

Since $\mathrm{E}\xi_{n1}^2 < \infty$ and $I\{|\xi_{n1}| > \varepsilon\sqrt{n}\mathrm{E}\xi_{n1}^2\} \to 0$ as $n \to \infty$, by the dominated convergence theorem, (5.18) goes to zero. Thus, the Lindeberg condition is satisfied. □

To see how good the $\widetilde{F}_-$ is to the target function F, next we see how close $\mathrm{E}\widetilde{F}_-$ is to F.

Theorem 5.1.3. *Suppose that $f(x)$ is twice differentiable at x. Then under conditions 1,2 and 3*

$$\lim_{h \to 0} \left\{ h^{-2} \left(E\widetilde{F}_-(x) - F_-(x) \right) \right\} = \frac{f''(x) K_2}{2},$$

where $K_2 = \int z^2 K(z)\, dz$.

Proof: From (5.16),

$$\begin{aligned}
E\widetilde{F_-}(x) &= 8\sigma^3 \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} \int K_h''(x-\sigma\tilde{t}-y)g(y)\,dy \\
&= 8\sigma^3 h^{-3} \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} \int K''\left(\frac{x-\sigma\tilde{t}-y}{h}\right) g(y)\,dy \\
&= 8\sigma^3 h^{-2} \sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty} \int K''(z)g(x-\sigma\tilde{t}-hz)\,dz \\
&= h^{-2}\int K''(z)H(x-hz)\,dz \\
&= h^{-2}\int K''(z)\Big\{H(x)-hzH'(x)+h^2z^2\frac{H''(x)}{2!} \\
&\qquad\qquad -h^3z^3\frac{H'''(x)}{3!}+h^4z^4\frac{H^{(4)}(x-\eta hu)}{4!}\Big\}\,dz.
\end{aligned}$$

We use a Taylor series expansion in the last step. Note that K is a symmetric PDF with zero mean and satisfies the condition 1 (a typical choice is a standard normal kernel). Using partial integration,

$$E\widetilde{F_-}(x) \approx 0+0+F_-(x)+0+\frac{h^2H^{(4)}(x-\eta hu)}{2}\int z^2K(z)\,dz.$$

Then

$$h^{-2}\left(E\widetilde{F_-}(x)-F_-(x)\right) \longrightarrow \frac{f''(x)}{2}\int z^2K(z)\,dz$$

as $h \to 0$. □

Now, combining Theorems 5.1.2 and 5.1.3, we have the convergence of $\widetilde{F}_-$ to the target function F below.

Theorem 5.1.4. *If $h_n \sim c \cdot n^{1/9}$, for some $c > 0$, then,*

$$n^{2/9}\left\{\widetilde{F}_-(x)-F(x)\right\} \xrightarrow{\mathcal{D}} N(\lambda,\rho^2)$$

as $n \to \infty$. *Where*

$$\lambda = \frac{c^2 f''(x) K_2}{2}$$
$$\rho^2 = 8\sigma^3 c^{-5} H(x) \|K''\|^2.$$

Proof: From theorem 5.1.2 and 5.1.3 it is easy to verify that when $h_n \sim cn^{1/9}$.

$$\begin{aligned} n^{2/9}\left\{\widetilde{F}_-(x) - F(x)\right\} &= c^{-5/2}(nh^5)^{1/2}\left\{\widetilde{F}_-(x) - E\widetilde{F}_-(x)\right\} \\ &\quad + n^{2/9}\left(E\widetilde{F}_-(x) - F(x)\right) \\ &\xrightarrow{\mathcal{D}} N(\lambda, \rho^2) \end{aligned}$$

as $n \to \infty, h \to 0$, and $nh \to \infty$ by Slutsky's Theorem. □

Corollary 5.1.1. *The bandwidth* h *that minimizes the asymptotic mean square error of* $\widetilde{F}_-(x)$ *in estimating* $F(x)$ *has a rate* $n^{-1/9}$. *The optimal bandwidth is given by* $h_{opt} = c_{opt} \cdot n^{-1/9}$ *with*

$$c_{opt} = \left(\frac{40\sigma^3 H(x)\|K''\|^2}{\left(f''(x)K_2\right)^2}\right)^{1/9}.$$

Proof: To obtain the optimal bandwidth minimize the asymptotic mean square error, *i.e.*, the sum of the variance and the squared bias in theorem 5.1.4, as a function of c,

$$\min_c\{\lambda^2 + \rho^2\} = \min_c\left\{\frac{1}{4}c^4\left[f''(x)K_2\right]^2 + 8\sigma^3 c^{-5} H(x)\|K''\|^2\right\}$$

which leads to c_{opt}. □

In practice one can use "truncated sums" in the estimators. That is, for

a large number m_n,

$$\hat{F}_-(x) = 8\sigma^3 \sum_{t_3=0}^{m_n}\sum_{t_2=0}^{m_n}\sum_{t_1=0}^{m_n} \hat{g}''\left(x-(2t_1+2t_2+2t_3+3)\sigma\right) \tag{5.19}$$

$$\hat{f}_-(x) = 8\sigma^3 \sum_{t_3=0}^{m_n}\sum_{t_2=0}^{m_n}\sum_{t_1=0}^{m_n} \hat{g}'''\left(x-(2t_1+2t_2+2t_3+3)\sigma\right) \tag{5.20}$$

$$\hat{F}_+(x) = 1-8\sigma^3 \sum_{t_3=0}^{m_n}\sum_{t_2=0}^{m_n}\sum_{t_1=0}^{m_n} \hat{g}''\left(x+(2t_1+2t_2+2t_3+3)\sigma\right) \tag{5.21}$$

$$\hat{F}_+(x) = -8\sigma^3 \sum_{t_3=0}^{m_n}\sum_{t_2=0}^{m_n}\sum_{t_1=0}^{m_n} \hat{g}'''\left(x+(2t_1+2t_2+2t_3+3)\sigma\right) \tag{5.22}$$

Corollary 5.1.2. *Under conditions 1 – 4,*

i) if $h \to 0$ and $m_n = [\frac{x}{6\sigma}]^+ + \frac{k_n}{6\sigma}$, then $\hat{F}_-(x)$ is asymptotically unbiased, where $[\cdot]^+$ indicate the positive integer part and $k_n \to \infty$ is independent of x.

ii) if further, $h \to 0$ and $nh^5 \to \infty$, the mean square error of $\hat{F}_-(x)$ is asymptotically equal to zero.

Proof: It is easy to see that

$$\begin{aligned}\hat{F}_-(x)-F(x) &= 8\sigma^3 \sum_{t_3=0}^{m_n}\sum_{t_2=0}^{m_n}\sum_{t_1=0}^{m_n}\left[\hat{g}''(x-\sigma\tilde{t})-g''(x-\sigma\tilde{t})\right] \\ &\qquad -8\sigma^3 \sum_{t_3=m_n+1}^{\infty}\sum_{t_2=m_n+1}^{\infty}\sum_{t_1=m_n+1}^{\infty} g''(x-\sigma\tilde{t}) \\ &\triangleq A_1 - A_2.\end{aligned}$$

The last term of the above equation has,

$$\begin{aligned} A_2 &= 8\sigma^3 \sum_{t_3=m_n+1}^{\infty} \sum_{t_2=m_n+1}^{\infty} \sum_{t_1=m_n+1}^{\infty} g''(x-\sigma\tilde{t}) \\ &= 8\sigma^3 \sum_{t_3=m_n+1}^{\infty} \sum_{t_2=m_n+1}^{\infty} \sum_{t_1=m_n+1}^{\infty} g''\left(x-6\sigma(m_n+1)-\sigma\tilde{t}\right) \\ &= F\left(x-6\sigma(m_n+1)\right) \leq F(-k_n) \to 0 \end{aligned}$$

uniformly as $n \to \infty$.

$$\begin{aligned} \mathrm{E}A_1 &= 8\sigma^3 \sum_{t_3=0}^{m_n} \sum_{t_2=0}^{m_n} \sum_{t_1=0}^{m_n} \left[\int K_h''(x-\sigma\tilde{t}-y)g(y)\,dy - g''(x-\sigma\tilde{t}) \right] \\ &= 8\sigma^3 \sum_{t_3=0}^{m_n} \sum_{t_2=0}^{m_n} \sum_{t_1=0}^{m_n} \left[\frac{1}{h^2} \int K''(s)g(x-\sigma\tilde{t}-sh)\,ds - g''(x-\sigma\tilde{t}) \right]. \end{aligned}$$

Consider a Taylor series expansion of $g(x-\sigma\tilde{t}-sh)$ in $x-\sigma\tilde{t}$ and notice that K is a "Normal" kernel with mean zero by partial integration

$$\begin{aligned} & \frac{1}{h^2} \int K''(s)g(x-\sigma\tilde{t}-sh)\,ds \\ = \; & \frac{1}{h^2} \int K''(s) \Big\{ g(x-\sigma\tilde{t}) - g'(x-\sigma\tilde{t})sh + \frac{1}{2}g''(x-\sigma\tilde{t})s^2h^2 \\ & \qquad -\frac{1}{6}g'''(x-\sigma\tilde{t})s^3h^3 + \frac{1}{24}g^{(4)}(x-\sigma\tilde{t})s^4h^4 + o(s^4h^4) \Big\}\, ds \\ = \; & 0 - 0 + g''(x-\sigma\tilde{t}) - 0 + \frac{h^2}{2}g^{(4)}(x-\sigma\tilde{t}) \int s^2K(s)\,ds + o(h^2). \end{aligned}$$

Hence,

$$\mathrm{E}A_1 = \frac{K_2h^2}{2}\left(f'(x)-o(1)\right)\left(1+o(1)\right) \to 0$$

as $h \to 0$.

To show *ii)*, notice that, by the asymptotic unbiasedness of the estimate,

$$\begin{aligned}\mathrm{Var}\hat{F}_-(x) &= \mathrm{Var}\left(8\sigma^3\sum_{t_3=0}^{m_n}\sum_{t_2=0}^{m_n}\sum_{t_1=0}^{m_n}\frac{1}{n}\sum_{i=1}^{n}K_h''(x-\sigma\tilde{t}-Y_i)\right)\\ &= \frac{8\sigma^3}{n}(B_1+B_2)-o(1),\end{aligned}$$

where

$$\begin{aligned}B_1 &= 8\sigma^3\int\sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty}\left[K_h''(x-\sigma\tilde{t}-y)\right]^2 g(y)\,dy\\ &= \frac{8\sigma^3K''(s)^2}{h^5}\sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty}g(x-\sigma\tilde{t})\left(1+o(h)\right).\\ B_2 &= 8\sigma^3\sum_{\tilde{t}_1\neq\tilde{t}_2}\int K_h''(x-\sigma\tilde{t}_1-y)K_h''(x-\sigma\tilde{t}_2-y)g(y)\,dy\\ &\leq 8\sigma^3\sum_{\tilde{t}_1\neq\tilde{t}_2}\left[\int\left[K_h''(x-\sigma\tilde{t}_1-y)\right]^2 g(y)\,dy\right]^{\frac{1}{2}}\left[\int\left[K_h''(x-\sigma\tilde{t}_2-y)\right]^2 g(y)\,dy\right]^{\frac{1}{2}}\\ &= \frac{8\sigma^3K''(s)^2}{h^5}\sum_{\tilde{t}_1\neq\tilde{t}_2}\left[g(x-\sigma\tilde{t}_1)^{\frac{1}{2}}g(x-\sigma\tilde{t}_2)^{\frac{1}{2}}\left(1+o(h)\right)\right].\end{aligned}$$

Therefore, as $nh^5\to 0$, $Var\hat{F}_-(x)\to 0$. □

Now consider cases with *nonhomogeneous normal errors*. We have $Y_i = X_i + E_i$ where X_i is from F, E_i is normally distributed on $\mathcal{N}(0,\sigma_i^2)$ and independent of X_i, for $i = 1, 2, \cdots, n$. Analogous to the homogeneous case we use a sum of three uniforms to approximate the normal errors

$$Y_i = X_i + E_i \approx X_i + \tilde{U}_{i1} + \tilde{U}_{i2} + \tilde{U}_{i3},$$

where $\tilde{U}_{i1}, \tilde{U}_{i2}, \tilde{U}_{i3} \sim \mathcal{U}(-\sigma_i, \sigma_i)$ independently.

Following the derivation for the nonhomogeneous uniform error case in Sun et al. (2002) we also derive the three-fold estimators of the CDF of X,

$$F(x) = 8\sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty}\widetilde{g}''(x)$$

where $\widetilde{g}(x) = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^3 g_i\left[x-\sigma_i(2t_1+2t_2+2t_3+3)\right]$. The $3U$ deconvolving estimators in nonhomogeneous case are:

$$\begin{aligned}\hat{F}_-(x) &= 8\sum_{t_3=0}^{m_n}\sum_{t_2=0}^{m_n}\sum_{t_1=0}^{m_n}\hat{g}''(x)\\ \hat{f}_-(x) &= 8\sum_{t_3=0}^{m_n}\sum_{t_2=0}^{m_n}\sum_{t_1=0}^{m_n}\hat{g}'''(x).\end{aligned}$$

where $m_n \to \infty$ and a kernel type estimator of $\hat{g}(t)$ is

$$\begin{aligned}\hat{g}(x) &= \frac{1}{nh}\sum_{i=1}^{n}\sigma_i^3 K\left(\frac{x-\sigma_i(2t_1+2t_2+2t_3+3)-Y_i}{h}\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\sigma_i^3 K_h\left(x-\sigma_i(2t_1+2t_2+2t_3+3)-Y_i\right).\end{aligned}$$

A similar consideration leads to

$$\begin{aligned}\hat{F}_+(x) &= 1-\frac{8}{n}\sum_{t_3=0}^{m_n}\sum_{t_2=0}^{m_n}\sum_{t_1=0}^{m_n}\sigma_i K_h''\left(x+\sigma_i^3(2t_1+2t_2+2t_3+3)-y_i\right)\\ \hat{f}_+(x) &= -\frac{8}{n}\sum_{t_3=0}^{m_n}\sum_{t_2=0}^{m_n}\sum_{t_1=0}^{m_n}\sigma_i^3 K_h'''\left(x+\sigma_i(2t_1+2t_2+2t_3+3)-y_i\right)\end{aligned}$$

Condition 2 $'$ (Density)**.** *The density g has continuous fourth derivatives and satisfies that $\sum_{t_3=0}^{\infty}\sum_{t_2=0}^{\infty}\sum_{t_1=0}^{\infty}|g_i''(x-(2t_1+2t_2+2t_3+3)\sigma)|^{1/2}$ converges uniformly in x.*

Condition 4 ′ (Errors)**.** *The error* $E_i \overset{D}{=} \widetilde{U}_{i1} + \widetilde{U}_{i2} + \widetilde{U}_{i3}$ *exactly, where* "$\overset{D}{=}$" *denotes the equal in distribution; and* $\widetilde{U}_{i1}$, $\widetilde{U}_{i2}$, $\widetilde{U}_{i3}$ *are identically independently distributed from* $\mathcal{U}(-\sigma_i, \sigma_i)$. *There is a finite number* $M > 0$ *such that* $\sum_{i=1}^{n} \sigma_i^2/n < M$.

Given the additional conditions above, we can show the following asymptotic properties analogous to the homogeneous case. Most parts of the proofs are similar to the derivations above.

Corollary 5.1.3. *Under conditions 1, 2 ′, 3 and 4 ′,*

i) If $h \to 0$ *and* $m_n = [\frac{x}{6\sigma_0}]^+ + \frac{k_n}{6\sigma_0}$, *then* $\hat{F}_-(x)$ *and* $\hat{f}_-(x)$ *are asymptotically unbiased, where* $0 < \sigma_0 = min_i \sigma_i$; $[\cdot]^+$ *indicates the positive integer part and* $k_n \to \infty$ *is independent of* x.

ii) If further, $nh^5 \to \infty$ *as* $h \to 0$, *the mean square error of* $\hat{F}_-(x)$ *is asymptotically equal to zero; if further,* $nh^7 \to \infty$ *as* $h \to 0$, *the mean square error of* $\hat{f}_-(x)$ *is asymptotically equal to zero.*

iii) Under the condition in i) and ii) the bandwidth h *for* $\hat{F}_-(x)$ *and* $\hat{f}_-(x)$ *that minimizes the asymptotic mean square error has rate of* $n^{-1/9}$ *and* $n^{-1/11}$ *respectively.*

5.1.3 Simulation

In the simulation study we consider the true distribution to be $Gamma(2, 2)$ and the error distribution to be $N(0, 1)$. Our kernel is the standard Gaussian kernel. The sample size is 100 and $m_n = 50$.

Bandwidth selection in nonparametric estimation is always an important issue as discussed in Chapter 3. Sun et al. (2002) suggest two ways of computing the bandwidth of their estimators, bootstrap method and the one based on Silverman's rule of thumb. In our simulation we simply compute the bandwidth h_w from the observations z_i only, using $h_w =$

$0.9n^{-1/5}\min(SD, IQR/1.34)$ (Silverman, 1986, page 48) where SD is the standard deviation of the data and IQR is the interquartile range of the data. Then we use a graphical method to select the bandwidth. For example, we compute the estimates using a sequence of $h_i = c_i h_w$, where c_i is some specified constant. Thus, we choose the optimal one based on graphics.

Figure 5.2 displays some plots based on our simulation results. $\hat{F}_-$ and $\hat{f}_-$ are used to compute the corrected CDF and PDF. The left upper subplot is the true density from a Gamma random sample data. The right upper subplot is the density plot for the data with normal measurement error. The left lower subplot displays the density of our corrected estimate using 3U deconvolving estimators. It is clear that the 3U deconvolving estimate captures the location and bumps of the true density successfully by comparing it with the plot of the contaminated sample, although there is some over-estimation in the right tail of the density. The right lower subplot shows the comparison of CDF curves. The red dash line is the CDF curve of the contaminated sample. The blue solid line is the CDF curve based on our 3U deconvolving estimate which almost matches the true CDF except in the tails.

Regarding the problem of tail-effects, we expect the variance reduction estimators, which are similar as Sun et al. (2002), to have better performance than $\hat{F}_-$ and $\hat{f}_-$. Chapter 6 discusses this further.

5.2 Nonparametric Regression with Errors in Variables

Deconvolution problems arise in a variety of situations in statistics. A very interesting and challenging problem is related to nonparametric regression when the predictor X cannot be observed directly. More specifically, let (X, Z) denote a pair of random variables and consider the problem of estimating the regression function $m(x) = E(Z|X = x)$. Due to the measurement

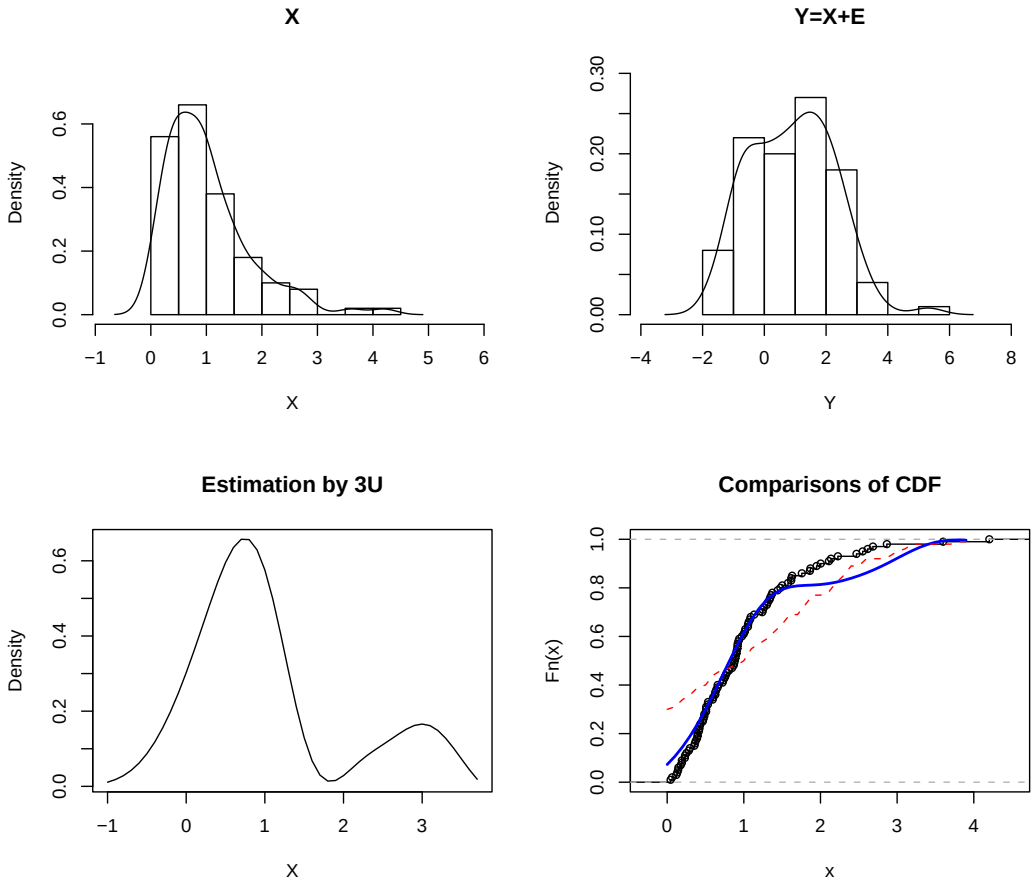


Figure 5.2: *Simulation for 3U deconvolving estimators. The true data are from Gamma(2,2) and the measurement errors are from N(0,1). The 3U deconvolving estimates capture the location and bumps of the true density and CDF successfully.*

mechanism or the nature of the environment, the variable X is not directly observable, instead what is observable is $Y = X+U$, called X measured with error. Let $Y_i = X_i + U_i$ where U_i is random error with either known density h in the homogeneous case or h_i in the nonhomogeneous case. In the section we develop new *SWAP regression estimators* for data with measurement errors and study the asymptotics of the new estimators. Here we focus on the case with homogeneous uniform errors. Estimators in the case of other error distributions and nonhomogeneous error cases will be studied in the future.

Remark. *The new estimators for the nonparametric regression with errors in variables are named "SWAP" estimators here because this type of estimators were initially proposed by Sun and Woodroofe for uniform errors, then were generalized by Sun and Wang here for other errors and the form of the new estimators has some similarity to the form of Shannon Weighted Average Procedure.*

5.2.1 SWAP Estimators

Since the variables $X_1, ..., X_n$ are not observable, the estimator $\hat{f}_n(x)$ can be constructed from our non-Fourier Deconvolution estimators. When the errors have a uniform distribution on $[-\theta, \theta]$, independently of X_i, for $i = 1, ..., n$, the non-Fourier Deconvolution estimator (Sun et al., 2002) is,

$$\hat{f}(x) = \frac{2\theta}{nh^2} \sum_{k=0}^{m_n} \sum_{i=1}^{n} K' \left(\frac{x - (2k+1)\theta - Y_i}{h} \right) \tag{5.23}$$

where $m_n \to \infty$.

Note that equation(5.23) can be rewritten in the kernel form:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \widetilde{K} \left(\frac{x - Y_i}{h} \right), \tag{5.24}$$

with

$$\widetilde{K}(x) = \frac{2\theta}{h} \sum_{k=0}^{m_n} K' \left(x - \frac{(2k+1)\theta}{h} \right). \tag{5.25}$$

$\widetilde{K}$ has many properties of an ordinary kernel. For example, it is easy to show that $\int \widetilde{K}(x)dx = 1$ implies that $\int \hat{f}(x)dx = 1$. Appealing to these facts, we propose the following non-Fourier *SWAP estimator* involving error-in-variables,

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} \widetilde{K}\left(\frac{x-Y_i}{h}\right) Z_j}{\sum_{i=1}^{n} \widetilde{K}\left(\frac{x-Y_i}{h}\right)}. \tag{5.26}$$

5.2.2 Asymptotic Performance

Consider our non-Fourier kernel estimator:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} \widetilde{K}(\frac{x-Y_i}{h}) Z_i}{\sum_{j=1}^{n} \widetilde{K}(\frac{x-Y_i}{h})} = \frac{\frac{1}{nh} \sum_{i=1}^{n} \widetilde{K}(\frac{x-Y_i}{h}) Z_i}{\hat{f}(x)}. \tag{5.27}$$

Note that we are interested in estimating the true regression function

$$m(x) = E(Z|X = x) = \frac{\int z f(x,z)\, dz}{f(x)}. \tag{5.28}$$

Here $f(x,z)$ denotes the joint density of (X,Z) and $f(x)$ the marginal density of X. In the model with errors-in-variables we denote $g(y,z)$ as the joint density of (Y,Z). By the independence of U and (X,Z), and $Y = X + U$ where $U \sim U(-\theta,\theta)$ we have,

$$g(y,z) = \int_{-\theta}^{\theta} f(y-u,z) \frac{1}{2\theta}\, du. \tag{5.29}$$

The fact that the numerator and denominator of the statistic $\hat{m}(x)$ are both random variables presents added difficulty to the problem. In order to study the expectation and variance of $\hat{m}(x)$ let us denote

$$r(x) = \int z f(x,z)\, dz = m(x) f(x), \tag{5.30}$$

then

$$\hat{r}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} \tilde{K}(\frac{x - Y_i}{h})Z_i = \frac{1}{n}\sum_{i=1}^{n} \tilde{K}_h(x - Y_i)Z_i$$

The regression function estimate is thus given by

$$\hat{m}_h(x) = \frac{\hat{r}_h(x)}{\hat{f}_h(x)}.$$

Lemma 5.2.1. *If $r(x)$ has continuously second derivatives, then the expectation of $\hat{r}_h(x)$ asymptotically converges to $u * r(x)$ as $h \to \infty$, where $*$ is the convolution operator and u is the density function of the error distribution.*

Proof: Notice that

$$\begin{aligned}
\mathrm{E}\hat{r}_h(x) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} \widetilde{K}_h(x - Y_i)Z_i\right) \\
&= \mathrm{E}\left(\widetilde{K}_h(x - Y_1)Z_1\right) \\
&= \int\int z\widetilde{K}_h(x - y)g(y, z)\,dy\,dz \\
&= \int\int z\widetilde{K}_h(x - y)\left(\int_{-\theta}^{\theta} f(y - u, z)\frac{1}{2\theta}\,du\right)\,dy\,dz \\
&= \frac{1}{2\theta}\int_{-\theta}^{\theta}\left[\int \widetilde{K}_h(x - y)\left(\int z f(y - u, z)\,dz\right)\,dy\right]\,du \\
&= \frac{1}{2\theta}\int_{-\theta}^{\theta}\left[\int \widetilde{K}_h(x - y)f(y - u)m(y - u)\,dy\right]\,du \\
&= \frac{1}{2\theta}\int_{-\theta}^{\theta}\left[\int \widetilde{K}_h(x - y)r(y - u)\,dy\right]\,du \\
&= \frac{1}{2\theta}\int_{-\theta}^{\theta}\left[\int \widetilde{K}(s)r(x - sh - u)\,ds\right]\,du.
\end{aligned}$$

The above derivation is by using (5.28) and (5.30), and by changing variable $s = (x - y)/h$.

Consider the Taylor series expansion of $r(x-sh-u)$ in $x-u$. Then

$$\begin{aligned}
\mathrm{E}\hat{r}_h(x) &= \frac{1}{2\theta}\int_{-\theta}^{\theta}\left[\int \tilde{K}(s)\left(r(x-u)-shr'(x-u)+o(s^2h^2)\right)\,ds\right]du \\
&\to \frac{1}{2\theta}\int_{-\theta}^{\theta}\left[\int \tilde{K}(s)r(x-u)\,ds\right]du \\
&= \frac{1}{2\theta}\int_{-\theta}^{\theta} r(x-u)\,du = u*r(x).
\end{aligned}$$

□

Theorem 5.2.1. *Under modest conditions,*

$$\hat{m}(x) \xrightarrow{p} \frac{u*r(x)}{r(x)}m(x).$$

as $n\to\infty$, $h\to 0$ *and* $nh\to\infty$.

Proof: First let us show $\mathrm{Var}\,(\hat{r}_h(x))\to 0$. We denote $q(x)=\int z^2 f(x,z)\,dz$. Note that

$$\begin{aligned}
\mathrm{E}\left[\widetilde{K}_h(x-Y)Z\right]^2 &= \int \widetilde{K}_h^2(x-y)z^2 g(y,z)\,dy\,dz \\
&= \int \widetilde{K}_h^2(x-y)z^2\left(\int_{-\theta}^{\theta} f(y-u,z)\frac{1}{2\theta}\,du\right)dy\,dz \\
&= \frac{1}{2\theta}\int_{-\theta}^{\theta}\left[\int \widetilde{K}_h^2(x-y)\left(\int z^2 f(y-u,z)\,dz\right)dy\right]du \\
&= \frac{1}{2\theta}\int_{-\theta}^{\theta}\left[\int \widetilde{K}_h^2(x-y)q(y-u)\,dy\right]du \\
&= \frac{1}{2\theta}\int_{-\theta}^{\theta}\frac{1}{h}\left[\int \widetilde{K}^2(s)q(x-sh-u)\,ds\right]du \\
&= \frac{1}{h}\left[\frac{1}{2\theta}\int_{-\theta}^{\theta}\left(q(x-u)\int \widetilde{K}^2(s)\,ds+O(h)\right)du\right]
\end{aligned}$$

Then combining with the lemma 5.2.1 we have

$$\begin{aligned}\text{Var}\left(\hat{r}_h(x)\right) &= \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{K}_h(x-Y_i)Z_i\right] \\ &= \frac{1}{n}\text{Var}\left[\tilde{K}_h(x-Y_1)Z_1\right] \\ &= \frac{1}{n}\left[\frac{1}{h}O(1)+O(h)+O(1)\right] \longrightarrow 0\end{aligned}$$

as $h \to 0$, $n \to \infty$, $nh \to \infty$. Hence, $\hat{r}_h(x)$ converges to $u * r(x)$ as $nh \to \infty$ by lemma 5.2.1. Note that the denominator of $\hat{m}_n(x)$ is the non-Fourier deconvolution density estimate $\hat{f}_h(x)$, which is consistent for same asymptotic of h . Using *Slutzky's theorem*, we obtain

$$\hat{m}(x) = \frac{\hat{r}_h(x)}{\hat{f}_h(x)} \xrightarrow{p} \frac{u * r(x)}{f(x)} = \frac{u * r(x)}{r(x)} m(x).$$

□

5.2.3 Simulations

In the simulation study we consider the true regression function to be $z = sin(x)$. We generate x_i from $U(0, 10)$ and random noise ε_i from $N(0, 1)$. Then our observed z_i is equal to $sin(x_i)+\varepsilon_i$. The measurement error u_i is generated from a uniform distribution $U(-1, 1)$, so the contaminated covariate is $y_i = x_i + u_i$.

Figure 5.3 displays our simulation results. The solid black line is the true function $z = sin(x)$. The short, red dashed line is the kernel regression estimate with the contaminated data, *i.e.* $\hat{z} = \hat{m}(y)$. The long, blue dashed line is based on our error-corrected SWAP estimate. The kernel of our estimate is still the standard Gaussian kernel. The sample size is 100 and $m_n = 50$. Similarly, for our simulation in density estimation, we compute the bandwidth h_w from the observations z_i only, using $h_w = 0.9n^{-1/5}\min(SD, IQR/1.34)$

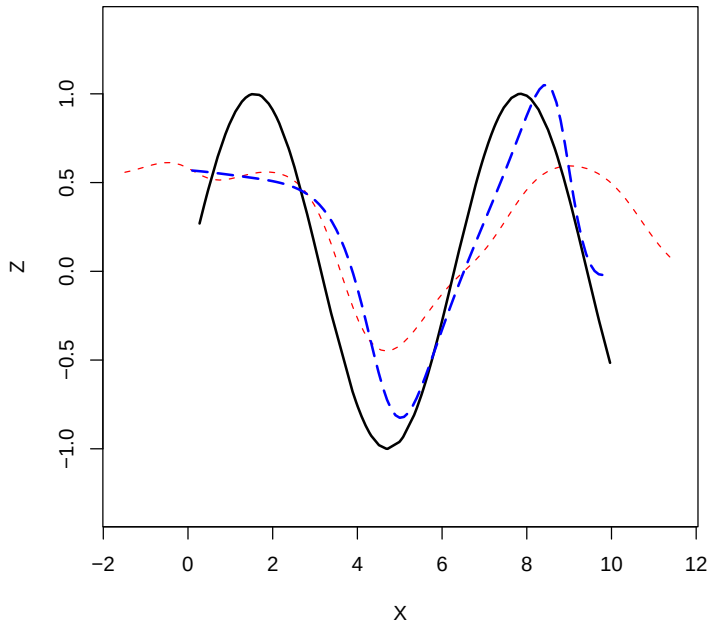


Figure 5.3: *Simulation for non-Fourier SWAP regression estimators. The solid, black line is the true function* $z = \sin(x)$*; the short, red dashed line is the nonparametric regression estimate for the contaminated data; the long, blue dashed line is our error-corrected SWAP estimate which is much closer than the true function.*

where SD is the standard deviation of the data and IQR is the interquartile range of the data. Then we use a graphical method to determine the bandwidth by specifying a sequence of $h_i = c_i h_w$.

It is clearly seen that the non-Fourier SWAP estimate corrected the regression line based on the contaminated sample. It is much closer to the true regression function. The left tail of the regression line is slightly misestimated, but overall the corrected estimate captures the trend of the true function.

Our second stimulation example uses the famous "ethanol" data (Simonoff, 1996, page 134). The ethanol data frame contains 88 sets of measurements for variables from an experiment in which ethanol was burned in a

single cylinder automobile test engine. The covariate is the *equivalence ratio* at which the engine was run - a measure of the richness of the air/ethanol mix. The response is the *concentration of nitric oxides* in the engine exhaust, normalized by engine work.

We artificially add measurement errors to the covariate, where the measurement errors are generated from the uniform distribution $U(-0.3, 0.3)$. Figure 5.4 displays the comparisons of the regression lines. The solid, black line is the nonparametric kernel regression estimate based on the original data. The short, red dashed line is the kernel regression estimate based on the contaminated data. The long, blue dashed line is our error-corrected SWAP estimate. The kernel of our estimate is still the standard Gaussian kernel and m_n is set to 50. The regression line based on the contaminated data is fairly different from the true line, while our corrected line is very close to the true line. Although the corrected line slightly shifts toward the right of the true line, it captures the shape of the true line successfully.

Remark (Bias correction estimator)**.** *The SWAP estimator we developed in previous sections has already had good performance on the error correction of regression function. However, it still has some tail-effects as shown in the simulations. This effect is due to the bias of our estimator $\hat{m}(x)$ in (5.26) as proved in theorem 5.2.1.*

In the proof of lemma 5.2.1, we obtained the following equation

$$\mathrm{E}\hat{r}_h(x) = \frac{1}{2\theta}\int_{-\theta}^{\theta} r(x-u)\,du.$$

This inspires us to derive an unbiased estimator of $\hat{r}(x)$. Let $R(x) = \int_{-\infty}^{x} r(x)\,dx$, then

$$\mathrm{E}\hat{r}_h(x) = \frac{R(x+\theta) - R(x-\theta)}{2\theta}$$

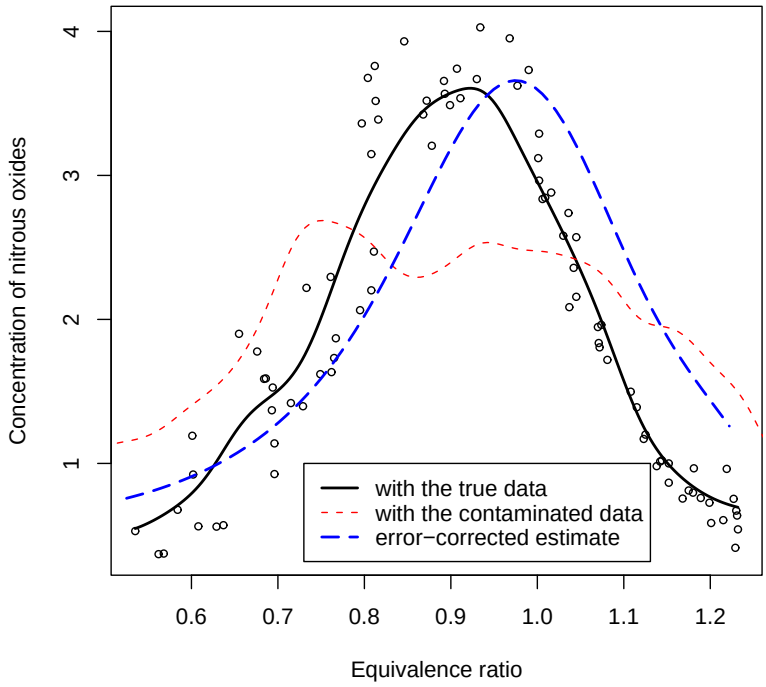


Figure 5.4: *Simulation study for ethanol data. The solid, black line is the nonparametric regression estimate for the original data; the short, red dashed line is the nonparametric regression estimate for the contaminated data; the long, blue dashed line is our error-corrected SWAP estimate. Our estimate successfully captures the shape of the true function.*

Using the same skill described in the first section of this chapter, we have

$$R(x) \approx 2\theta \sum_{t=0}^{\infty} \mathrm{E}\hat{r}_h\big(x - (2t+1)\theta\big)$$

then

$$r(x) \approx 2\theta \sum_{t=0}^{\infty} \mathrm{E}\hat{r}'_h\big(x - (2t+1)\theta\big).$$

So, a revised estimator of $m(x)$ *is*

$$\widetilde{m}(x) = \frac{\tilde{r}(x)}{\hat{f}(x)} \tag{5.31}$$

where

$$\tilde{r}(x) = 2\theta \sum_{t=0}^{\infty} \hat{r}'_h\big(x - (2t+1)\theta\big).$$

Note that $\widetilde{m}(x)$ *is an asymptotical unbiased estimator of* $m(x)$ *because it is easy to see that* $\mathrm{E}\tilde{r}(x) \to r(x)$ *and then*

$$E(\widetilde{m}(x)) \to \frac{r(x)}{r(x)/m(x)} = m(x),$$

as $n \to \infty$, $h \to 0$ *and* $nh \to \infty$.

The stimulation and other asymptotical properties will be studied in the future.

Chapter 6

Discussion and Further Issues

Both spatial-temporal data mining and measurement error problems are rich research areas in modern statistics. In this chapter, we discuss the applications of our methods and address future research issues in these two areas.

6.1 Applications of LASR

The development of the multi-stage statistical LASR algorithm allows both clinicians and researchers to derive more useful, objective information from pressure maps, such as the location of significant pressure changes or the relative efficacy of pressure relief procedures. Furthermore, spatial registration allows global analysis of pre- and post-intervention differences without any subjective bias in selecting areas of interest.

In the specific study of the effects of gluteal NMES it was found that subjects who received a gluteal stimulation system showed statistically significant changes in ischial region pressure over time, when baseline/post-treatment comparisons were made. The region of significant change was not symmetrical in all cases which reflects the asymmetric nature both of gluteal muscle recruitment area and contractile responses. However, in the two cases where

we did not have baseline data, pressure distributions obtained in the first session for which we had data (already after at least initial conditioning or treatment) and the last session of the assessment we had, did not show significant changes. This implies that for these two subjects the majority of the intrinsic changes in tissue characteristics occurred acutely, during early treatment, and that continued regular use of gluteal stimulation maintains these improved responses. This motivates us, given that NMES is effective as shown in this dissertation, to study the length of treatment suitable for each patient, and to examine if the experience of these two patients applies to other subjects after a "critical" time point.

The last two decades have seen remarkable developments in medical imaging technology. Universities and industries have made huge investments in inventing and developing the technology needed to acquire images from multiple imaging modalities. Medical images are increasingly widely used in health care and biomedical research; a wide range of imaging modalities is now available. The clinical significance of medical imaging in the diagnosis and treatment of diseases is overwhelming. While planar X-ray imaging was the only radiological imaging method in the early part of the last century, several modern imaging techniques are available today for the acquisition of anatomical, physiological, metabolic and functional information from the human body. The commonly used medical imaging modalities capable of producing multidimensional images for clinical applications are: X-ray Computed Tomography (X-ray CT), Magnetic Resonance Imaging (MRI), Single Photon Emission Computed Tomography (SPECT), Positron Emission Tomography (PET) and Ultrasound (US).

It should be noted that these modern imaging methods involve sophisticated instrumentation and equipment which employ high-speed electronics and computers for data collection. Spatial-temporal image data occur in a broad range of medical imaging applications. It is now common for patients

to be imaged multiple times, either by repeated imaging with a single modality, or by imaging with different modalities. It is also common for patients to be imaged *dynamically*, that is, to have sequences of images acquired, often at many frames per second. The ever increasing amount of image data acquired makes it more and more desirable to relate more than one statistical tool to assist in extracting relevant clinical information.

Application of the LASR algorithm enhances data extraction and acquires statistical inferences from complex spatial-temporal data sets, as shown in the NMES study. Thus the LASR analytical methodology has the potential to become a powerful new tool in the field of image analysis. It should also be noted that our spatial registration technique (with random landmarks) has wide potential applications, even beyond the field of clinical care. Other potential clinical applications include images of soft tissues, which may not include bony landmarks. Applications could include situations where an imaged object may change dimensions and/or orientation over time.

6.2 I-Map – FDR Ratio Mapping

The multiple testing problems of multivariate local regression were studied in Section 3 of Chapter 3. Benjamini & Hochberg's step-up procedure for strong control of the false discovery rate was used to control the multiplicity in the tests of the NMES study. We declared the FDR level to be 0.05, which means we were subject to 5% false discovery that the pressure between baseline and treatment (in one compartment) was significantly different when it actually wasn't. It is of interest to identify these 5% falsely discovered locations for clinicians and researchers. Replication of data frames over time for each assessment enables us to identify them. We propose an FDR ratio mapping algorithm as follows.

Algorithm 6.2.1. *FDR ratio mapping algorithm.*

1). Based on the LASR procedure we obtaine m adjusted p-values, $p_{i1}, \cdots, p_{im}$ in the data frame i. Compute

$$\tilde{I}_{ij} = \begin{cases} 1, & \text{if } p_{ij} < \alpha, \\ 0, & \text{if } p_{ij} \geq \alpha. \end{cases}$$

where $i = 1, \cdots, n$ (n is the total number of the data frames), $j = 1, \cdots, m$, α is the pre-determined FDR level, and m is the number of compartments or pixels.

2). Compute I-values which are defined as

$$I_j = \frac{\sum_{i=1}^{n} \tilde{I}_{ij}}{n},$$

where $j = 1, \cdots, m$.

3). Compute the α-quantile q_α for the sequence $A = \{I_j > 0\}$.

4). Declare that the null hypothesis H_{0k} (i.e. the hypothesis test at location k) is falsely rejected if $0 < I_k < q_\alpha$. Then generate a map based on I-values and declared results.

As an example, we generate an I-map for one subject in the NMES study by applying the FDR ratio mapping algorithm above. Figure 6.1 shows an I-map in the case of static pressure mapping. The light blue blocks indicate the true discovered regions that the subject's pressures in these regions are significantly improved. The red blocks are the locations that were falsely discovered as significantly improved locations, which happened in a single frame. The simulation experiment will be conducted in a future paper to assess the validity of this I-map; the asymptotic properties of it will also be studied.

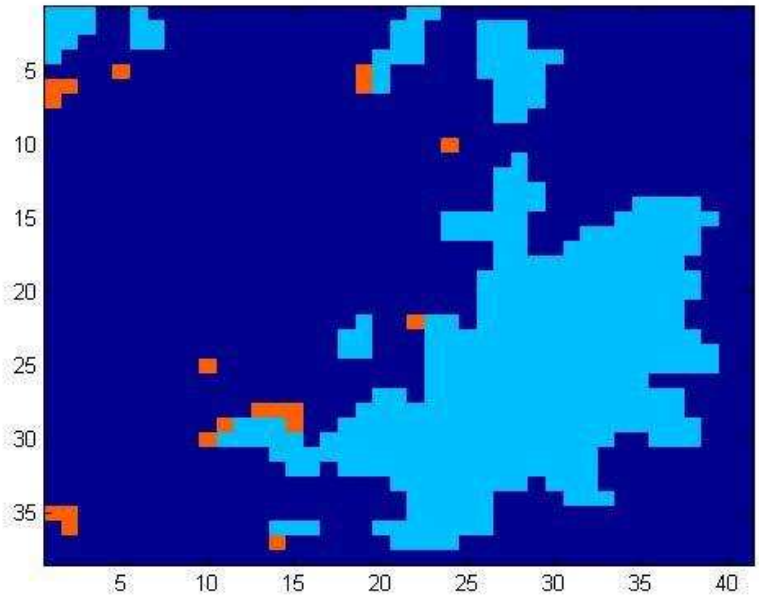


Figure 6.1: *An example of I-map based on FDR ratio mapping algorithm. The red blocks indicate the locations that were falsely discovered significant. The light blue blocks are the true significant improved regions.*

6.3 Backfitting Algorithm for Semiparametric Regression

In Chapter 4, we proposed a semiparametric regression model for spatial-temporal data in the NMES study. The semiparametric regression can be viewed as a standard linear mixed model when using general radial splines to expand the zero-mean random field. The log likelihood of the fixed parameters for model (4.10) is

$$l(\boldsymbol{\beta},\boldsymbol{\theta}) = \log L(\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \frac{n}{2}\log(2\pi), \tag{6.1}$$

where $\mathbf{V} = \sigma_\varepsilon^2\mathbf{I} + \sigma_u^2\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T$ and $\boldsymbol{\theta} = (\sigma_\varepsilon^2, \sigma_u^2)$.

For fixed $\boldsymbol{\theta}$, taking the derivative of the log-likelihood with respect to $\boldsymbol{\beta}$, we find the estimate of $\boldsymbol{\beta}$ as the solution of

$$(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}, \tag{6.2}$$

which is the well-known *generalized least-squares* (GLS) formula.

To estimate the random effects, consider the log-likelihood of all the parameters. First, note that

$$L(\boldsymbol{\beta},\boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{u})f(\mathbf{u}).$$

From the mixed model specification, the conditional distribution of $\mathbf{y}$ given $\mathbf{b}$ is normal with mean $\mathrm{E}(\mathbf{y}|\mathbf{u}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ and variance $\sigma_\varepsilon^2\mathbf{I}$. The random effects $\mathbf{u}$ is normal with mean zero and variance $\sigma_u^2\mathbf{I}$. Hence,

$$\begin{aligned} & l(\boldsymbol{\beta},\boldsymbol{\theta}) = \log L(\boldsymbol{\beta},\boldsymbol{\theta}) \\ =& -\frac{1}{2}\log|\sigma_\varepsilon^2\mathbf{I}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\widetilde{\mathbf{Z}}\mathbf{u})^T(\sigma_\varepsilon^{-2}\mathbf{I})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\widetilde{\mathbf{Z}}\mathbf{u}) \\ & -\frac{1}{2}\log|\sigma_u^2\mathbf{I}| - \frac{1}{2}\mathbf{u}^T(\sigma_u^{-2}\mathbf{I})\mathbf{u} - \frac{n}{2}\log(2\pi). \end{aligned} \tag{6.3}$$

Given the fixed parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$, the estimate of $\mathbf{u}$ is the solution of

$$\big(\sigma_\varepsilon^{-2}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}} + \sigma_u^{-2}\mathbf{I}\big)\mathbf{u} = \sigma_\varepsilon^{-2}\widetilde{\mathbf{Z}}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{6.4}$$

This estimate is also known as the *best linear unbiased predictor* (BLUP). In practice the unknown fixed parameters are replaced by their estimates through the *profile log-likelihood* of $\mathbf{V}$.

$$\begin{aligned} l_p(\mathbf{V}) = \log L(\mathbf{V}) = &-\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\mathbf{y}^T\mathbf{V}^{-1}\Big[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\Big]\mathbf{y} \\ &- \frac{n}{2}\log(2\pi). \end{aligned} \tag{6.5}$$

That is, the log-liklihood above is obtained by plugging the GLS estimate of $\boldsymbol{\beta}$ into the log-likelihood function (6.1).

Note that our model consists of two components, the fixed effect and nonparametric random field. It is interesting to estimate the $\boldsymbol{\beta}$ and $\mathbf{u}$ simultaneously. The derivative of the log-likelihood of all the parameters (6.3) with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial\, l(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \sigma_\varepsilon^{-2}\mathbf{X}^T\big(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \widetilde{\mathbf{Z}}\mathbf{b}\big). \tag{6.6}$$

Setting (6.6) equal to zero and then combining the equation (6.4) (the derivative of the log-likelihood of all the parameters (6.3) with respect to $\mathbf{u}$ and setting equal to zero), we obtain the mixed model equations.

$$\begin{bmatrix} \sigma_\varepsilon^{-2}\mathbf{X}^T\mathbf{X} & \sigma_\varepsilon^{-2}\mathbf{X}^T\widetilde{\mathbf{Z}} \\ \sigma_\varepsilon^{-2}\widetilde{\mathbf{Z}}^T\mathbf{X} & \sigma_\varepsilon^{-2}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}} + \sigma_u^{-2}\mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \sigma_\varepsilon^{-2}\mathbf{X}^T\mathbf{y} \\ \sigma_\varepsilon^{-2}\widetilde{\mathbf{Z}}^T\mathbf{y} \end{bmatrix} \tag{6.7}$$

The estimates we compute from this simultaneous equation (6.7) are exactly the same as those from (6.2) and (6.4). This suggests fitting the model by a iterative backfitting algorithm to fixed effect component and nonparametric random field component. Our algorithm is similar to the *backfitting algorithm*

in the additive model (Hastie and Tibshirani, 1990; Hastie et al., 2001) which is a general algorithm that enables one to fit an additive model using any regression-type fitting mechanism.

Algorithm 6.3.1. *Backfitting algorithm for the semiparametric model*

1). Given the initial estimate $\hat{\boldsymbol{\beta}}$ *by the ordinary least squares estimate*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

2). Compute the estimate $\hat{\mathbf{u}}$ *from a nonparametric (spline smoothing) model*

$$\mathbf{y}^* = \widetilde{\mathbf{Z}}\mathbf{u} + \varepsilon,$$

where $\mathbf{y}^* = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. *The estimate of* $\hat{\mathbf{u}}$ *is still the BLUP by the formula as before,*

$$\left(\sigma_\varepsilon^{-2}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}} + \sigma_u^{-2}\mathbf{I}\right)\mathbf{u} = \sigma_\varepsilon^{-2}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{y}}.$$

3). Re-estimate $\hat{\boldsymbol{\beta}}$ *from a fixed effect model*

$$\mathbf{y}^{**} = \mathbf{X}\boldsymbol{\beta} + \varepsilon,$$

where $\mathbf{y}^{**} = \mathbf{y} - \widetilde{\mathbf{Z}}\hat{\mathbf{u}}$.

4). Iterate step's 2 and 3 until convergence.

As a next step, we propose to study the efficiency of the backfitting algorithm and compare it to the convenient mixed model fitting. Moreover, we use the AIC of the full log-likelihood as the criterion for our smoothing parameter in Chapter 5. AIC is justified from a model prediction perspective. It is designed to choose the model with the lowest predictive log-likelihood and is related to cross-validation and Mallows C_p. In the context of smoothing, AIC has been used mostly for selecting the smoothing parameter (see chapter 3). However, for complex semiparametric models, model selection is still in its infancy. The backfitting algorithm also inspires us to study the marginal AIC of the model as the criterion for the smoothing parameter.

6.4 Extensions of Measurement Error Problems

In Chapter 5 we successfully developed new fast non-Fourier estimators for the densities and regression functions in measurement error models. There are many interesting and challenging problems open to our research.

Variance Reduction Estimators

It is observed that $\hat{F}_-(x)$ has smaller variance in the left tail of F while $\hat{F}+(x)$ has small variance in the right tail. Sun et al. (2002) consider a form of combined estimator of F,

$$\hat{F}^*(x) = [1 - p(x)]\hat{F}_-(x) + p(x)\hat{F}_+(x)$$

where p is a distribution function. Two typical choices of p are an *ad hoc* choice $e^x/(1 + e^x)$ and one that minimizes the variance of $\hat{F}^*(x)$. This estimator is asymptotically unbiased and normally distributed. It has better performance in practice than $\hat{F}_-(x)$ and $\hat{F}_+(x)$.

The simulation of Chapter 5 suggests that our 3U deconvolving estimators and non-Fourier kernel estimators did not perform well in the tails. We propose to study the combined estimators following the same idea above with the expectation of better performance. Since our 3U deconvolving estimators are inspired from the principle of random number generation, we will be able to develop error-corrected estimators for any arbitrary distribution.

Nonparametric Estimation of ARAM and GARCH-processes

GARCH, *generalized autoregressive conditional heteroscedasticity* process (Bollerslev, 1986; Engle, 1982), is a popular stochastic process which has been used fairly successfully in modeling time series in finance. As a basis for analyzing

the risk of financial investments the GARCH model has been frequently used to modeling asset price volatility over time.

GARCH processes are closely related to *autoregressive moving average* (ARMA) processes. If we square a GARCH(1,1) process then we get an ARMA(1,1) process. Therefore, as an intermediate step towards GARCH processes, we study the nonparametric estimation for ARMA models, which is closely related to the regression with errors-in-variables we studied in Chapter 5. A linear ARMA(1,1) model with mean w is given by

$$X_{t+1} = w + aX_t + b\varepsilon_t + \varepsilon_{t+1} \tag{6.8}$$

where ε_t is zero-mean white noise. So, the nonparametric generalization of this model is,

$$X_{t+1} = f(X_t, \varepsilon_t) + \varepsilon_{t+1} \tag{6.9}$$

for some unknown function $f(x, u)$ which is monotone in the second argument u. If f does not depend on the second argument, (6.9) reduces to a nonparametric autoregression of order 1:

$$X_{t+1} = f(X_t) + e_{t+1}. \tag{6.10}$$

The autoregression function $f(x)$ under this model (6.10) can be estimated by common kernel estimates or local polynomials (Fan and Yao, 2003). However, in the general case of (6.9) we have the problem of estimating a function of "unobservable" variables. As f depends also on the observable time series X_t, the basic idea of constructing a nonparametric estimate of $f(x, u)$ is to combine a common kernel smoothing in the first variable x with a deconvolution smoothing in the second variable u. We plan to study the further properties of nonparametric SWAP estimators of ARMA and GARCH processes under this general setting.

PLM and GPLM with Measurement Errors

Recall that in Chapter 5 we studied a semiparametric model for the spatial-temporal data,

$$Y = \mathbf{X}\boldsymbol{\beta} + Z(\mathbf{s}) + \varepsilon$$

where $Z(\mathbf{s})$ is a random field. This model has a strong relationship with a *partial linear model* (PLM) (Engle et al., 1986; Speckman, 1988) which consists of two additive components, a linear and a nonparametric part.

$$Y = \mathbf{X}\boldsymbol{\beta} + g(\mathbf{T}) + \varepsilon, \tag{6.11}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a finite dimensional parameter and $g(\cdot)$ is any type of smooth function. Here we assume a decomposition of the explanatory variables into two vectors, $\mathbf{X}$ and $\mathbf{T}$. There is a straightforward generalization of this model to the case with a known link function $L(\cdot)$. This semiparametric extension of the *generalized linear model* (GLM)

$$E(Y|\mathbf{X}, \mathbf{T}) = L(\mathbf{X}\boldsymbol{\beta} + g(\mathbf{T})) \tag{6.12}$$

is denoted as a *generalized partial linear model* (GPLM) (Härdle et al., 1998; Severini and Staniswalis, 1994).

PLM and GPLM have received a considerable amount of research in the past two decades. One reason is that it is much more flexible than the standard linear model since it combines both parametric and nonparametric components. Another reason is that it allows easier interpretation of the effect of each variable compared to a completely nonparametric regression.

The typical estimate of model (6.11) is based on the profile likelihood. Consider a simple case of the smoothing part where T is one-dimensional. Then a kernel estimate of g is

$$g_j(T_j) = \frac{\sum_{i=1}^n K_h(T_i - T_j)(Y_i - \mathbf{X}_i^T\boldsymbol{\beta})}{\sum_{i=1}^n K_h(T_i - T_j)}. \tag{6.13}$$

If we define a smoother matrix $\mathbf{S}$ by its elements

$$S_{ij} = \frac{K_h(T_i - T_j)}{\sum_{i=1}^n K_h(T_i - T_j)},$$

then (6.13) has a matrix form

$$g = \mathbf{S}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

The *Speckman estimators* (Speckman, 1988) of PLM are

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{Y}}, \\ \hat{g} &= \mathbf{S}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \end{aligned}$$

where $\widetilde{\mathbf{X}} = (\mathbf{I} - \mathbf{S})\mathbf{X}$ and $\widetilde{\mathbf{Y}} = (\mathbf{I} - \mathbf{S})\mathbf{Y}$.

It is of interest to study PLM and GPLM with error-in-variables. More specifically, we are interested in models when the nonlinear variable is measured in error, *i.e.* when T is one-dimensional

$$\begin{aligned} E(Y|\mathbf{X}, T) &= \mathbf{X}\boldsymbol{\beta} + g(T) \\ Z &= T + U \end{aligned} \tag{6.14}$$

where U is a measurement error.

The model we studied in Section 2 of Chapter 5 is just a special case of the model 6.14, where $\beta = 0$ and T is observed with uniform measurement error $U \sim \mathcal{U}(-\theta, \theta)$. Let us denote

$$m_{ni}^*(t) = \frac{\widetilde{K}_h(t - Z_i)}{\sum_{i=1}^n \widetilde{K}_h(t - Z_i)},$$

where $\widetilde{K}_h(t) = \widetilde{K}(t/h)/h$ and $\widetilde{K}$ is defined by (5.25).

The error-corrected estimators of PLM with nonlinear variable-in-error are

$$\begin{aligned} \hat{\boldsymbol{\beta}}^* &= (\widetilde{\mathbf{X}}^{*T}\widetilde{\mathbf{X}}^*)^{-1}\widetilde{\mathbf{X}}^{*T}\widetilde{\mathbf{Y}} \\ \hat{g}^*(t) &= \sum_{j=1}^n m_{nj}^*(t)(Y_i - \widetilde{\mathbf{X}}_i^*\hat{\boldsymbol{\beta}}) \end{aligned} \tag{6.15}$$

where $\widetilde{\mathbf{Y}} = (\widetilde{Y}_1, \cdots, \widetilde{Y}_n)^T$ with

$$\widetilde{Y}_i = Y_i - \sum_{j=1}^{n} m^*_{nj}(Z_i) Y_j,$$

and $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{X}}_1, \cdots, \widetilde{\mathbf{X}}_n)^T$ with

$$\widetilde{\mathbf{X}}_i = \mathbf{X}_i - \sum_{j=1}^{n} m^*_{nj}(Z_i) X_j.$$

We intend to study the asymptotics of the estimators 6.15. Moreover, the backfitting algorithm we proposed in Section 6.3 will also be examined in this case.

Appendix

A.1 Consistency of Midline Regression

Consider the case of simple linear regression of the midline defined in section 2.3.1,

$$Y_i = \beta_0 + \beta_1 z_i + \epsilon_i,$$

where the horizontal axis $z_i = 1, 2, \cdots, n$ and the ϵ_i are independent with mean 0 and variance σ^2. If we define

$$\omega_i = \frac{z_i - \bar{z}}{\sum_{j=1}^{n}(z_j - \bar{z})^2} \quad \text{and} \quad v_i = \frac{1}{n} - \bar{z}\omega_i,$$

then the least square estimators of β_0 and β_1 are:

$$\hat{\beta}_{0n} = \sum_{i=1}^{n} v_i Y_i \quad \text{and} \quad \hat{\beta}_{1n} = \sum_{i=1}^{n} \omega_i Y_i$$

respectively. Since $EY_i = \beta_0 + \beta_1 z_i$, we have

$$\mathrm{E}\hat{\beta}_{0n} = \beta_0 + \beta_1 \bar{z} - \beta_0 \bar{z} \sum_{i=1}^{n} \omega_i - \beta_1 \bar{z} \sum_{i=1}^{n} \omega_i z_i$$

and

$$\mathrm{E}\hat{\beta}_{1n} = \beta_0 \bar{\omega} - \beta_1 \sum_{i=1}^{n} \omega_i z_i.$$

Note that $\sum_{i=1}^{n} \omega_i = 0$ and $\sum_{i=1}^{n} \omega_i z_i = 1$. Therefore, $\mathrm{E}\hat{\beta}_{0n} = \beta_0$ and $\mathrm{E}\hat{\beta}_{1n} = \beta_1$ which is to say that $\hat{\beta}_{on}$ and $\hat{\beta}_{1n}$ are unbiased. A sufficient condition for the consistency of $E\hat{\beta}_{0n}$ and $E\hat{\beta}_{1n}$ is that their variance tends to zero as $n \to \infty$. Since $\mathrm{Var}Y_i = \sigma^2$, we have

$$\mathrm{Var}\hat{\beta}_{0n} = \sigma^2 \sum_{i=1}^{\infty} v_i^2 \quad \text{and} \quad \mathrm{Var}\hat{\beta}_{1n} = \sigma^2 \sum_{i=1}^{\infty} \omega_i^2.$$

with $\sum_{i=1}^{n} \omega_i^2 = \{\sum_{i=1}^{n} (z_i - \bar{z})^2\}^{-1}$. These expressions simplify to

$$\mathrm{Var}\hat{\beta}_{0n} = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{z}^2}{\sum_{j=1}^{n} (z_j - \bar{z})^2},$$

and

$$\mathrm{Var}\hat{\beta}_{1n} = \frac{\sigma^2}{\sum_{j=1}^{n} (z_j - \bar{z})^2}.$$

Therefore, $\hat{\beta}_{0n}$ and $\hat{\beta}_{1n}$ are consistent because

$$\frac{\bar{z}^2}{\sum_{j=1}^{n} (z_j - \bar{z})^2} = \frac{(n+1)^2/4}{n(n+1)(2n+1)/6 - (n+1)^2/4} \longrightarrow 0$$

and

$$\frac{1}{\sum_{j=1}^{n} (z_j - \bar{z})^2} = \frac{1}{n(n+1)(2n+1)/6 - (n+1)^2/4} \longrightarrow 0.$$

Here we use the fact $z_i = 1, \cdots, n$.

A.2 Explicit Formulae for Bivariate Local Estimators

In the image application we are interested in a two-dimensional smoothing problem. The explicit formulae of bivariate local estimators are useful in order to reduce the computational loading. Following the notation of Section 3.2, the local regression model with a bivariate covariate becomes

$$Y_i = m(X_{1i}, X_{2i}) + \epsilon_i,$$

where $m(\cdot,\cdot)$ is unknown. A suitably smooth function m can be approximated in a neighborhood of a point $\mathbf{x} = (x_1, x_2)$ by a bivariate local polynomial.

A local linear approximation is

$$m(x_1, x_2) \approx b_0 + b_1(x_1 - X_1) + b_2(x_2 - X_2).$$

A local quadratic approximation is

$$\begin{aligned} m(x_1, x_2) \approx & b_0 + b_1(x_1 - X_1) + b_2(x_2 - X_2) \\ & + \frac{b_3}{2}(x_1 - X_1)^2 + \frac{b_4}{2}(x_2 - X_2)^2 + b_5(x_1 - X_1)(x_2 - X_2). \end{aligned}$$

The local coefficients are estimated by solving the weighted least squares problems (3.4) and (3.4). Here $m(x_1, x_2)$ is the first component of the local coefficient, $\hat{b}_0$.

In order to derive explicit formulae for bivariate local linear estimators, consider the sums

$$S_{pq} = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})(X_{1i} - x_1)^p (X_{2i} - x_2)^q$$

$$Z_{pq} = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})(X_{1i} - x_1)^p (X_{2i} - x_2)^q Y_i$$

where $p, q = 0, 1, 2, \cdots$. Then for the local linear estimate we can write

$$\hat{m}_{\mathbf{H}}(x_1, x_2) = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} S_{00} & S_{10} & S_{01} \\ S_{10} & S_{20} & S_{11} \\ S_{01} & S_{11} & S_{02} \end{bmatrix}^{-1} \begin{bmatrix} Z_{00} \\ Z_{10} \\ Z_{01} \end{bmatrix} \tag{A.1}$$

We are able to fit the explicit formula for the estimated regression function on one line,

$$\hat{m}_{\mathbf{H}}(x_1, x_2) = \frac{(S_{20}S_{02} - S_{11}^2)T_{00} + (S_{10}S_{11} - S_{01}S_{20})T_{01} + (S_{01}S_{11} - S_{02}S_{10})T_{10}}{2S_{01}S_{10}S_{11} - S_{02}S_{10}^2 - S_{00}S_{11}^2 - S_{01}^2 S_{20} + S_{00}S_{02}S_{20}} \tag{A.2}$$

Then by (3.11) the CV bandwidth selector function for the bivariate local linear estimator is,

$$\mathrm{CV}(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)\}^2 w(\mathbf{X}_i)\cdot \left\{1 - \frac{(S_{20}S_{02} - S_{11}^2)K_{\mathbf{H}}(0)}{S_{00}(S_{20}S_{02} - S_{11}^2) + S_{10}(S_{01}S_{11} - S_{10}S_{02}) + S_{01}(S_{10}S_{11} - S_{01}S_{20})}\right\}^{-2} \tag{A.3}$$

For the local quadratic estimate we can write

$$\hat{m}_{\mathbf{H}}(x_1, x_2) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} S_{00} & S_{10} & S_{01} & S_{20} & S_{02} & S_{11} \\ S_{10} & S_{20} & S_{11} & S_{30} & S_{12} & S_{21} \\ S_{01} & S_{11} & S_{02} & S_{21} & S_{03} & S_{12} \\ S_{20} & S_{30} & S_{21} & S_{40} & S_{22} & S_{31} \\ S_{02} & S_{12} & S_{03} & S_{22} & S_{04} & S_{13} \\ S_{11} & S_{21} & S_{12} & S_{31} & S_{13} & S_{22} \end{bmatrix}^{-1} \begin{bmatrix} Z_{00} \\ Z_{10} \\ Z_{01} \\ Z_{20} \\ Z_{02} \\ Z_{11} \end{bmatrix} \tag{A.4}$$

Formulae (A.2), (A.3) and (A.4) are very useful in terms of reducing the numerical burden.

A.3 Gaussian Random Fields

Gaussian random fields have been applied in a large number of fields to a diverse range of ends, and very many deep theoretical analyses of various properties are available. Adler (1981) gives a good introduction to the area. For modeling spatial data we refer to the monograph by Cressie (1993). Here we give a few basic definitions related to our semiparametric model.

Definition A.3.1 (Gaussian random field)**.** $\{Z(\mathbf{s}) : \mathbf{s} \in \mathbf{D}\}$ *is a Gaussian random field with mean function* $\mu(\mathbf{s}), \mathbf{s} \in \mathbf{D}$ *and covariance function* $C(\mathbf{h})$ *if for every finite collection of sites,*

$$\mathbf{s}_1, \mathbf{s}_2 \cdots, \mathbf{s}_n$$

the vector

$$Z = \begin{bmatrix} Z(\mathbf{s}_1) \\ Z(\mathbf{s}_2) \\ \vdots \\ Z(\mathbf{s}_n) \end{bmatrix}$$

is multivariate normally distributed (i.e. $Z \sim N(\mu, \Sigma)$*) with mean vector*

$$\mu = E(Z) = \begin{bmatrix} \mu(\mathbf{s}_1) \\ \mu(\mathbf{s}_2) \\ \vdots \\ \mu(\mathbf{s}_n) \end{bmatrix}$$

and variance-covariance matrix

$$\Sigma = var(Z) = \begin{bmatrix} \sigma^2 & C(\mathbf{s}_1 - \mathbf{s}_2) & \cdots & C(\mathbf{s}_1 - \mathbf{s}_n) \\ C(\mathbf{s}_2 - \mathbf{s}_1) & \sigma^2 & \cdots & C(\mathbf{s}_2 - \mathbf{s}_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(\mathbf{s}_n - \mathbf{s}_1) & C(\mathbf{s}_n - \mathbf{s}_2) & \cdots & \sigma^2 \end{bmatrix}.$$

Consider a random field $\{Z(\mathbf{s}) = \mu + \varepsilon(\mathbf{s})\colon\ \mathbf{s} \in \mathbf{D}\}$, where μ is the population mean, the error function $\varepsilon(\mathbf{s})$ is a zero-mean random function of the spatial location $\mathbf{s}$. Next we define a stationary Gaussian random field.

Definition A.3.2 (Second-order stationary)**.** *The random field* $\{Z(\mathbf{s}) : \mathbf{s} \in \mathbf{D}\}$ *is second-order stationary if assumptions (1)* $\sim$ *(3) are satisfied.*

Assumption 1. *The errors have mean zero, i.e.,* $E\{\varepsilon(\mathbf{s})\} = 0, \mathbf{s} \in \mathbf{D}$*. Then* $E\{Z(\mathbf{s})\} = \mu, \mathbf{s} \in \mathbf{D}$.

Assumption 2. *Homoscedasticity, i.e.,* $var\{\varepsilon(\mathbf{s})\} = \sigma^2, \mathbf{s} \in \mathbf{D}$*, does not depend on spatial locations* $\mathbf{s} \in \mathbf{D}$*. Then* $var\{Z(\mathbf{s})\} = \sigma^2, \mathbf{s} \in \mathbf{D}$.

Assumption 3. *The covariance function*

$$C(\mathbf{s} - \mathbf{u}) = cov\{\varepsilon(\mathbf{s}), \varepsilon(\mathbf{u})\}; \mathbf{s}, \mathbf{u} \in \mathbf{D}$$

only depends on the difference in locations (distance and direction) of the pair of sites $\mathbf{s}, \mathbf{u} \in \mathbf{D}$. *Then*

$$C(\mathbf{s} - \mathbf{u}) = cov\{Z(\mathbf{s}), Z(\mathbf{u})\}; \mathbf{s}, \mathbf{u} \in \mathbf{D}.$$

Definition A.3.3 (Isotropic)**.** *The random field* $\{Z(\mathbf{s}) : \mathbf{s} \in \mathbf{D}\}$ *is isotropic if assumption (4) is satisfied.*

Assumption 4. *The covariance function*

$$C(\|\mathbf{s} - \mathbf{u}\|) = cov\{\varepsilon(\mathbf{s}), \varepsilon(\mathbf{u})\}; \mathbf{s}, \mathbf{u} \in \mathbf{D}$$

depends on the distance $\|\mathbf{s} - \mathbf{u}\|$ *between the sites* $\mathbf{s}, \mathbf{u} \in \mathbf{D}$.

Bibliography

Adler, R. J. (1981). *Geometry of Random Fields*. John Wiley and Sons, New York.

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22:203–217.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple hypothesis testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston.

Besl, P. J. and McKay, N. D. (1992). A method for registration of 3D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:239–256.

Bogie, K. M. and Triolo, R. J. (2003). The effects of regular use of neuromuscular electrical stimulation on tissue health. *Journal of Rehabilitation Research and Development*, 40:469–475.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31:307–327.

Carroll, R. J. and Hall, P. (1989). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Associations*, 83:1184–1186.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. Chapman Hall, New York.

Cator, E. A. (2001). Deconvolution with arbitrarily smooth kernels. *Statistics and Probability Letters*, 54:205–214.

Chu, C. K. and Marron, J. S. (1991). Choosing a kernel regression estimator (with discussion). *Statistical Science*, 6:404–436.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610.

Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., and Marchal, G. (1995). Automated multi-modality image registration based on information theory. *Information Processing in Medical Imaging 1995*, eds: Bizais et al., Kluwer Academic, Dordrecht:263–274.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403.

Cressie, N. (1993). *Statistics for Spatial Data (revised edition)*. John Wiley and Sons, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Efromovich, S. (1997). Density estimation for the case of supersmooth measurement error. *Journal of the American Statistical Association*, 92:526–535.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica*, 50:987–1008.

Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81:310–320.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19:1257–1272.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87:998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21:196–216.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1997). Local polynomial regression: Optimal kernels and asymptotic minimax effciency. *Annals of the Institute of Statistical Mathematics*, 49:79–99.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20:2008–2036.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B*, 57:371–394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Application: Theory and Methodologies*. Chapman and Hall, New York.

Fan, J. and Truong, Y. (1993). Nonparametric regression with errors in variables. *The Annals of Statistics*, 21:1900–1925.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag, New York.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman and Hall, New York.

Fitzpatrick, J. M., Hill, D. L. G., and Maurer Jr., C. R. (2000). Image registration, Chapter 8. *Handbook of Medical Imaging, Volume 2: Medical Image Processing and Analysis*, eds: Fitzpatrick, J.M. and Sonka, M., SPIE Press, Bellingham, Washington:447–513.

Friston, K. J. (2004). Introduction: experimental design and statistical parametric mapping. *Human brain function, 2nd Edition*, eds: Frackowiak et al., Academic Press, London.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., and Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, 2:189–210.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons, New York.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data hypothesis (with discussion). *Test*, 12:1–77.

Gentle, J. E. (2003). *Random Number Generation and Monte Carlo Methods, Second Edition*. Springer-Verlag, New York.

Gerlot, P. and Bizais, Y. (1988). Image registration: A review and a strategy for medical applications. *Information Processing in Medical Imaging 1987*, eds: Graaf, C.N.de and Viergever, M.A., Plenum Press, New York:81–89.

Groeneboom, P. and Jongbloed, G. (2003). Density estimation in the uniform deconvolution model. *Statistica Neerlandica*, 57:136–157.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer-Verlag, New York.

Hall, P., Sheather, S. J., Jones, M. C., and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 78:263–269.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, New York.

Härdle, W. (1991). *Smoothing Techniques with Implementation in S*. Springer-Verlag, New York.

Härdle, W., Hall, W., and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameter selectors from their optimum. *Journal of the American Statistical Association*, 83:86–101.

Härdle, W., Mammen, E., and Müller, M. (1998). Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, 93:1461–1474.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Hill, D. L. G., Batchelor, P. G., Holden, M., and Hawkes, D. J. (2001). Medical image registration. *Physics in Medicine and Biology*, 46:1–45.

Hochberg, Y. and Tamhane, A. (1987). *Multiple comparison procedures*. Wiley, New York.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman and Hall, New York.

Hurvich, C. M., Simonoff, J. S., and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society, Series B, Methodological*, 60:271293.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91:401–407.

Kendall, M. G., Stuart, A., Ord, J. K., Arnold, S. F., and O'Hagan, A. (1998). *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory, 6th edition*. Edward Arnold, London.

Loader, C. (1999). *Local Regression and Likelihood*. Springer-Verlag, New York.

Maintz, B. A. and Viergever, M. A. (1998). A survey of medical image registration. *Medica Image Analysis*, 2:1–36.

Maurer, C. R. and Fitzpatrick, J. M. (1993). A review of medical image registration. *Interactive Imageguided Neurosurgery*, ed. Maciunas, R. J., American Association of Neurological Surgeons, Parkridge, IL:17–44.

Mendelsohn, J. and Rice, J. (1982). Deconvolution of microfluorometric histograms with b-splines. *Journal of the American Statistical Association*, 77:748–753.

O'Sullivan, F. and Roy Choudhury, K. (2001). An analysis of the role of positivity and mixture model constraints in poisson deconvolution problems. *Journal of Computational and Graphical Statistics*, 10:673–696.

Pelizzari, C. A., Chen, G. T., Spelbring, D. R., Weichselbaum, R. R., and Chen, C. T. (1989). Accurate three-dimensional registration of CT, PET, and/or MR images of the brain. *Journal of Computer Assisted Tomography*, 13:20–26.

Ripley, B. D. (1987). *Stochastic Simulation*. Wiley, New York.

Rosenfeld, A. and Kak, A. C. (1982). *Digital Picture Processing,* Vol. I and II. Academic Press, Orlando.

Roy Choudhury, K. (1998). *Additive Mixture Models for Multichannel Image Data*. Ph.D. Dissertation, University of Washington, Seattle.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22:1346–1370.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

Schimek, M. G. (2000). *Smoothing and Regression: Approaches, Computation, and Application*. John Wiley & Sons, New York.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York.

Seeger, M. (2004). Gaussian processes for machine learning. *International Journal of Neural Systems*, 14:1–38.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of the American Statistical Association*, 82:605–610.

Severini, T. A. and Staniswalis, J. G. (1994). Quasilikelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89:501–511.

Shannon, C. E. (1948a). The mathematical theory of communication (part 1). *Bell System Techinical Journal*, (reprint available from http:// www.lucent.com) 27:379–423.

Shannon, C. E. (1948b). The mathematical theory of communication (part 2). *Bell System Techinical Journal*, (reprint available from http:// www.lucent.com) 27:623–656.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53:683–690.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag, New York.

Snyder, D. L., Schulz, T. J., and O'Sullivan, J. A. (1992). Deblurring subject to nonnegativity constraints. *IEEE Transactions on Signal Processing*, 40:1143–1150.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, 50:413–436.

Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, 21:169–184.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 5:595–620.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8:1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053.

Studholme, C., Hill, D. L. G., and Hawkes, D. J. (1995). Multiresolution voxel similarity measures for MRPET registration. *Information Processing in Medical Imaging 1995*, eds. Bizais, Y. et al., Kluwer Academic, Dordrecht:287–298.

Studholme, C., Hill, D. L. G., and Hawkes, D. J. (1996). Automated 3D registration of MR and CT images of the head. *Medical Image Analysis*, 1:163–175.

Studholme, C., Hill, D. L. G., and Hawkes, D. J. (1997). Automated 3D registration of MR and PET brain images by multi-resolution optimization of voxel similarity measures. *Medical Physics*, 24:25–35.

Studholme, C., Hill, D. L. G., and Hawkes, D. J. (1999). An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32:71–86.

Sun, J. (1991). Significance levels in exploratory projection pursuit. *Biometrika*, 78:759–769.

Sun, J. (1993). Tail probabilities of the maxima of gaussian random fields. *The Annals of Probability*, 21:34–71.

Sun, J. (1995). Iterative estimates for a smoothing parameter. *Statistics and Probability letters*, 24:347–356.

Sun, J. (2001). Multiple comparisons for a large number of parameters. *Biometrical Journal*, 43:627–643.

Sun, J. and Feuerverger, A. (2002). Measurement errors: connections and solutions. *Workshop on Developments and Challenges in Mixture Models, Bump Hunting and Measurement Error Models 2002*.

Sun, J. and Loader, C. (1994). Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics*, 22:1328–1345.

Sun, J., Loader, C., and McCormick, W. P. (2000). Confidence bands in generalized linear models. *Annals of Statistics*, 28:429–460.

Sun, J., Morrison, H., Harding, P., and Woodroofe, M. (2002). Density and mixture estimation from data with measurement errors. Revision invited by *Journal of the American Statistical Association*.

Sun, J. and Woodroofe, M. B. (1996). Adaptive smoothing for a penalized npmle of a non-increasing density. *Journal of Statistical Planning and Inference*, 52:143–159.

van den Elsen, P. A., Pol, E. J. D., and Viergever, M. A. (1993). Medical image matching - a review with classification. *IEEE Engineering in Medicine and Biology*, 12:26–39.

Vardi, Y. and Lee, D. (1993). From image deblurring to optimal investment: maximum likelihood solutions for positive linear inverse problems. *Journal of the Royal Statistical Society: Series B*, 55:569–612.

Wahba, G. (1990). *Spline Models for Observational Data*. Regional Conference Series in Applied Mathematics 59, SIAM, Philadelphia.

Wand, M. P. (1998). Finite sample performance of deconvolving density estimators. *Statistics and Probability Letters*, 37:131139.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing, Vol. 60 of Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

Woods, R. P., Cherry, S. R., and Mazziotta, J. C. (1992). Rapid automated algorithm for aligning and reslicing pet images. *Journal of Computer Assisted Tomography*, 16:620–633.

Woods, R. P., Grafton, S. T., Holmes, C. J., Cherry, S. R., and Mazziotta, J. C. (1998a). Automated image registration: I. General methods and intrasubject, intramodality validation. *Journal of Computer Assisted Tomography*, 22:139–152.

Woods, R. P., Grafton, S. T., Holmes, C. J., Cherry, S. R., and Mazziotta, J. C. (1998b). Automated image registration: II. Intersubject validation of linear and nonlinear models. *Journal of Computer Assisted Tomography*, 22:153–165.

Woods, R. P., Mazziotta, J. C., and Cherry, S. R. (1993). Mri-pet registration with automated algorithm. *Journal of Computer Assisted Tomography*, 17:536–546.

Yang, L. and Tschernig, R. (1999). Multivariate bandwidth selection for local linear regression. *Journal of the Royal Statistical Society: Series B*, 61:793–815.

Yarkony, G. M. and Heinemann, A. W. (1995). Management of pressure ulcers after spinal cord injury. *Spinal cord injury: Clinical Outcomes from the Model Systems*, Gaithersburg, MD: Aspen Publishers:100–119.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82:171–196.

Zhang, C. H. (1990). Fourier methods for estimating mixing densities and distributions. *Journal of Statistical Planning and Inference*, 18:806–831.

Zhang, Z. (2005). *Multiple Hypothesis Testing for Finite and Infinite Number of Hypotheses*. Ph.D. Dissertation, Case Western Reserve University, Cleveland.