ESTIMATING AND TESTING OF FUNCTIONAL DATA WITH RESTRICTIONS

A Dissertation

by

SANG HAN LEE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2007

Major Subject: Statistics

ESTIMATING AND TESTING OF FUNCTIONAL DATA WITH RESTRICTIONS

A Dissertation

by

SANG HAN LEE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Marina Vannucci
Committee Members,	Faming Liang
	Johan Lim
	Yoonsuck Choe
Head of Department,	Simon J. Sheather

August 2007

Major Subject: Statistics

ABSTRACT

Estimating and Testing of Functional Data with Restrictions. (August 2007)

Sang Han Lee, B.S., Seoul National University;

M.S., Seoul National University

Chair of Advisory Committee: Dr. Marina Vannucci

The objective of this dissertation is to develop a suitable statistical methodology for functional data analysis. Modern advanced technology allows researchers to collect samples as functional which means the ideal unit of samples is a curve. We consider each functional observation as the resulting of a digitized recoding or a realization from a stochastic process. Traditional statistical methodologies often fail to be applied to this functional data set due to the high dimensionality.

Functional hypothesis testing is the main focus of my dissertation. We suggested a testing procedure to determine the significance of two curves with order restriction. This work was motivated by a case study involving high-dimensional and high-frequency tidal volume traces from the New York State Psychiatric Institute at Columbia University. The overall goal of the study was to create a model of the clinical panic attack, as it occurs in panic disorder (PD), in normal human subjects. We proposed a new dimension reduction technique by non-negative basis matrix factorization (NBMF) and adapted a one-degree of freedom test in the context of multivariate analysis. This is important because other dimension techniques, such as principle component analysis (PCA), cannot be applied in this context due to the order restriction.

Another area that we investigated was the estimation of functions with constrained restrictions such as convexification and/or monotonicity, together with the

development of computationally efficient algorithms to solve the constrained least square problem. This study, too, has potential for applications in various fields. For example, in economics the cost function of a perfectly competitive firm must be increasing and convex, and the utility function of an economic agent must be increasing and concave. We propose an estimation method for a monotone convex function that consists of two sequential shape modification stages: (i) monotone regression via solving a constrained least square problem and (ii) convexification of the monotone regression estimate via solving an associated constrained uniform approximation problem.

To the Lord,

Thanks for this opportunity and the strength to finish this work.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Marina Vannucci, for her guidance and support throughout the study. She introduced me to this exiting area and gave me access to a motivating data set. My sincere thanks goes to committee member, Dr. Johan Lim for his insightful comment. Discussion with him enriched my research and provided deep understanding about statistics. Also, I would like to extend my thanks to Dr. Faming Liang and Dr. Yoonsuck Choe for their encouragement and kind comments. Finally, I would like to thank my family and friends for their prayer. Without your prayer, it would have been not possible.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

1.1 Motivation and problem

This work we present here was motivated by an collaboration with investigators at the New York Psychiatric Institute, at Columbia University. The overall goal of the study is to create a model of the clinical panic attack in normal human subjects, as it occurs in individuals affected by panic disorder. Sodium lactate reliably produces panic attacks in patients with panic disorder (Liebowitz et al., 1985). Normals rarely have such reactivity. A distinctive feature of lactate induced panic is a marked increase in tidal volume (Goetz et al., 1993). Klein (1993) suggested that the spontaneous panic attack may be due to a hypersensitive alarm system for detecting signals of impending suffocation, such as rising levels of CO2 or brain lactate. The endogenous opioid system is an important central regulator of respiratory drive. An exogenous opioid, such as morphine, blunts sensitivity to CO2 (Fleetham et al., 1980). Conversely, naloxone, an opioid receptor antagonist, increases the ventilatory response to hypercapnic hypoxia in normal human controls (Akiyama et al., 1993). Naloxone pretreatment may make normal individuals (who putatively have an intact opioid system) vulnerable to the marked anxiogenic and respiratory effects of lactate. In a pilot study Sinha and Klein (2005) found that lactate after naloxone, administered to normals, produced a marked increase in tidal volume that exceeded previous results from infusing only lactate. Surprisingly, lactate, despite producing a

---

The format and style follow that of *Journal of the American Statistical Association*.

metabolic alkalosis, is a tidal volume stimulant, as has been shown in both normal humans and rats.

1.1.1 Experimental study

A randomized study with normal subjects was designed to test the investigators' hypothesis. Subjects, healthy normal male and female adult volunteers, not affected by any psychiatric or significant illness, were randomized to three groups. They received either naloxone followed by lactate or saline followed by lactate or naloxone followed by saline. The hypothesis was that subjects receiving the naloxone-lactate sequence will have greater increases in tidal volume during the lactate phase than subjects in the other two groups. The naloxone-saline sequence should have lesser effects than the saline-lactate sequence. The randomization was unequal (3:3:1), with smallest number of subjects in the saline-lactate group, since prior experience with this sequence in normal subjects produced relatively minor increments in tidal volume. Establishing a lack of naloxone-saline effect was considered crucial. Respiratory and other physiological measurements were taken during the experiment, together with qualitative information measured via questionnaires and interviews.

The experiment on each individual consisted of four phases:

Phase I (baseline): Approximately 30 minutes. The subject has sensors and intravenous lines placed within 5 minutes while supine. This phase provides baseline measurements for each subject. Patency is maintained by slow saline drip, slowly increased to normal flow prior phase II. All infusion adjustments are made without the subject's knowledge. Personnel and subjects are blind to infusion contents. All randomized infusion sequences are set up in advance by the Research Pharmacist who maintains a secret subject listing.

Phase II (first infusion): Approximately 20 minutes. Subjects receive either naloxone

over approximately the first 3 to 5 minutes, within the saline flow, or just stay on saline.

Phase III (second infusion): Approximately 20 minutes. Subjects who received naloxone at the first infusion are switched to either saline or lactate, and those that received only saline at the first infusion are switched to lactate. The infusion of the experimental component in the saline flow lasts approximately 20 minutes.

Phase IV (recovery): Approximately 120 minutes. The subject remains supine, with minimal saline flow. This period allows clinical observation as well as exploration of possible prolonged effects.

The measuring and data recording device was the lifeShirt, (Wilhelm et al., 2003), a garment recently developed with embedded inductive plethysmography sensors for continuous ambulatory monitoring of respiration and other physiological functions.

1.2 Pre-processing of the data

1.2.1 Baseline adjustment and data thinning

Each subject has a different Vt baseline. We therefore performed baseline adjustment by calculating the median Vt for each subject during phase I. We then considered three ways to adjust for baseline effect: (a) subtracting the median from the Vt trace of each subject; (b) dividing the Vt trace by the median; (c) taking $\log_e$ of (b). Results from the statistical analysis we performed did not show any particular sensitivity to these different procedures for baseline adjustment. Here we report analyses performed using method (a).

Data are massive (see Figure 1). During the experiment tidal volume measurements were automatically saved 50 times per second. We reduced dimension by

considering traces obtained taking one every k-th data points. We examined plots of several reduced traces to make sure we were preserving important features of the data and decided on $k = 25$ as a safe choice. This gave us two measurements per second. In our analysis we considered data spanning over two separate time windows, covering first and second infusion, respectively. For the first infusion, clinical experience with naloxone suggests quite a fast onset, however duration of any respiratory effect is not well known. In order to cover possible prolonged effects we therefore chose a time window of approximately 8.5 minutes that covered the infusion in spans up to 2 minutes after the end of the infusion. As for the second quick onset of effect during phase III. We therefore chose a window of approximately 17 minutes before the end of the infusion.

1.2.2 Noise removal by wavelet

A smoothing procedure was necessary in order to reveal the breathing patterns of interest to the investigators. The method we investigated uses wavelet decompositions to filter out high-to-medium frequency components of the data that are unrelated to the breathing frequencies, i.e. they constitute irrelevant information. We give here a very brief description of the method. Wavelets have been extremely successful as a tool for the analysis and synthesis of discrete data. Fast algorithms, such as the discrete wavelet transform (DWT) and its undecimated version (MODWT), allow the decomposing of a curve into a set of wavelet coefficients that efficiently describe global and local features of the curve, Mallat (1989) and Percival and Walden (2002). By applying inverse transformations to coefficients at single scales one can essentially extract components that characterize the original curve at different scales (or frequency intervals). The sum of all components would give back the original curve. In a wavelet decomposition with J levels, component j roughly refers to the frequency

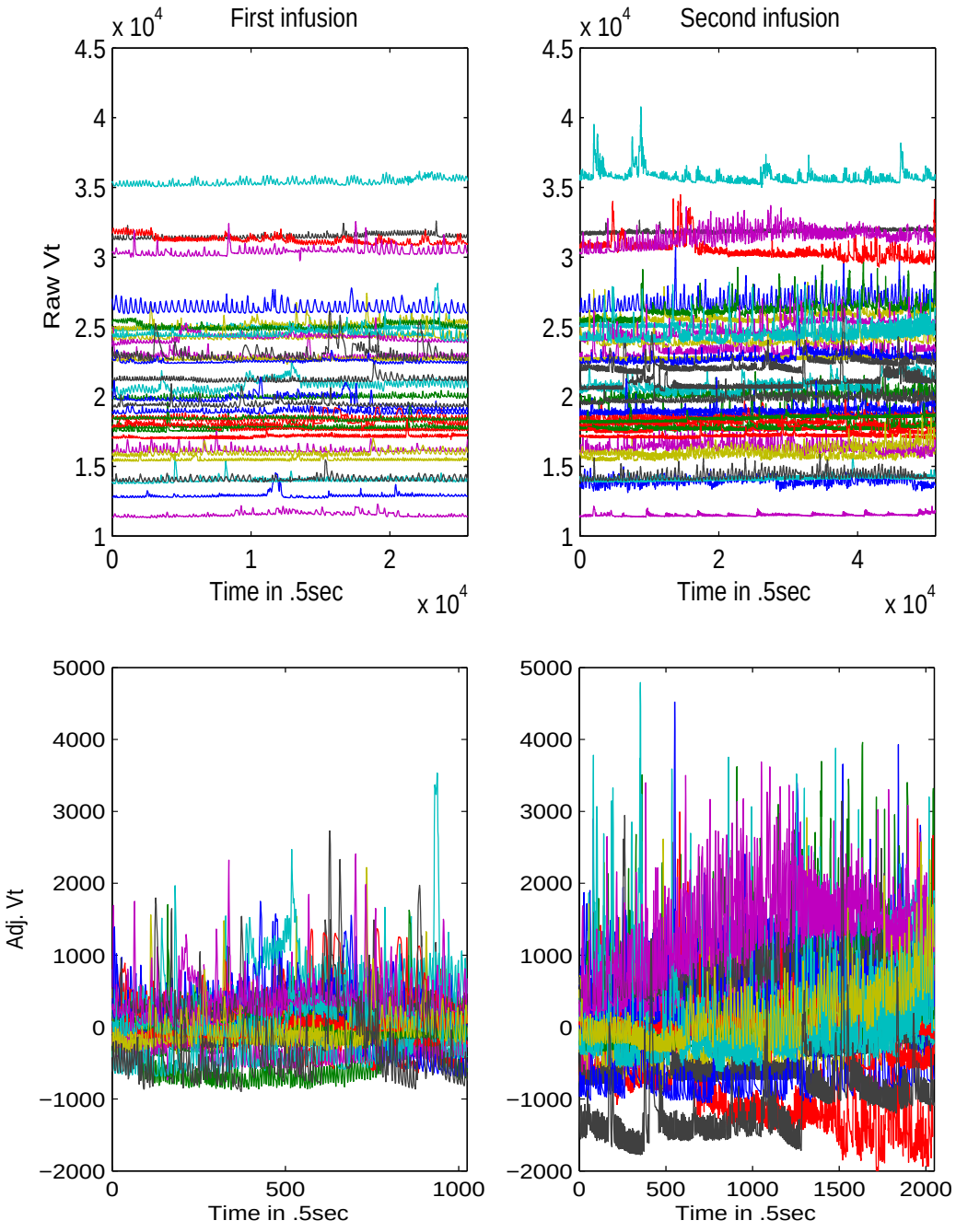

Figure 1: NL-SL-NS: Raw traces (upper) and baseline adjusted (lower). Right and left panels correspond to first and second infusion, respectively.

interval $[\frac{1}{2^{J+1}\Delta t}, \frac{1}{2^{J}\Delta t}]$, for j=1,...,J, while the component at the last level of the transform captures the "trend" of the curve, i.e. the frequency interval $[0, \frac{1}{2^{J+1}\Delta t}]$. In our

application $\Delta t = .5$sec. Considering that an average breath takes 3-5sec, we decided to analyze the Vt traces after the subtraction of components at 4sec and less. We also extracted the trends of the data (see Figure 2).

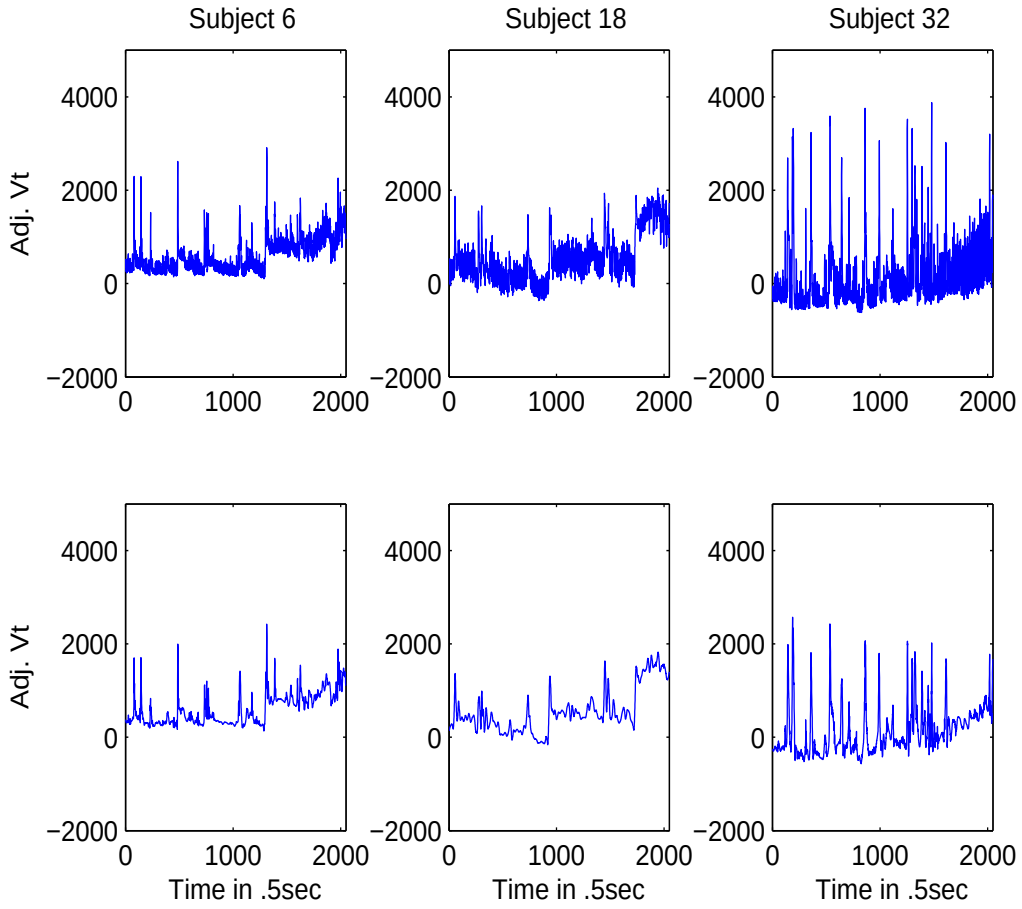

Figure 2: NL-SL-NS: Vt curves for 3 subjects during second infusion (upper) and their reconstructions after DWT with denoising (lower).

CHAPTER II

FUNCTIONAL ANALYSIS ON PANIC DATA

2.1 Introduction

A simple low power pre-planned interim analysis used t tests to compare each group's second infusion's average tidal volume. It showed the expected significant lesser effect of naloxone-saline vs both naloxone-lactate and saline-lactate. This indicated a lack of specific naloxone tidal volume effect, thus making the naloxone-saline group superfluous. However, the expected lesser effect of saline-lactate compared to naloxone-lactate was not evident in this preliminary analysis. This founding initially seemed due to an unexpectedly large tidal volume increment in the saline -lactate group. It was possible that these subjects, if given naloxone-lactate would have an even larger increment. To check this we recall the saline-lactate and naloxone-lactate subjects and administer the respective contrasting infusion.

This non blind, but objectively recorded, tidal volume crossover was reassuring since in every case the naloxone-lactate second infusion exceeded the saline-lactate infusion. This also emphasized the utility of a cross-over design.

To proceed our new, cross-over design was aimed to further resolve the differential naloxone interaction with lactate issue: each subject receives 2 infusions counterbalanced between naloxone-lactate and saline-lactate, under randomized double blind conditions. The subjects were either entirely new or subjects who had previously received naloxone-saline in the initial between group design.

2.2 Methods

2.2.1 Functional ANOVA

We looked at a set of hypothesis testing procedures adapted to functional data.

One such procedure was functional ANOVA (FANOVA), the "functionalized" version of a standard one way ANOVA, Ramsay and Silverman (1997, page 139), where, given a set of curves, an ANOVA-like test is applied at each time point. We briefly recall the procedure: For subject i and treatment l at a fixed time point t we write a fixed effect ANOVA model as

$$y_{il}(t) = \mu(t) + \alpha_l(t) + \epsilon_{il}(t),\ l = 1, \cdots, L;\ i = 1, \cdots, n_l;\ \sum_{l=1}^{L} n_l = n, \tag{2.1}$$

where $\epsilon_{il}(t)$ are independent $N(0, \sigma^2)$ errors. We again slightly abuse notation by assuming that now $\mathbf{y}_i$ indicates a reconstructed curve after wavelet denoising. When testing in model (2.1) we have

$$H_0 : \alpha_l(t) = 0,\ l = 1, \cdots, L, \tag{2.2}$$

versus the general alternative. To ensure identifiability it is standard to impose the constraint that $\sum_l n_l \alpha_l(t) = 0,\ \forall t$. Following Ramsay and Silverman (1997) we have that

$$F(t) = \frac{MSTr(t)}{MSE(t)} = \frac{SSTr(t)/(L-1)}{SSE(t)/(n-L)} \tag{2.3}$$

with $SST(t) = \sum_{i,l}[y_{il}(t) - \bar{y}_{\cdot\cdot}(t)]^2$, $SSTr(t) = \sum_l n_l[\bar{y}_{\cdot l}(t) - \bar{y}_{\cdot\cdot}(t)]^2$ and $SSE(t) = \sum_{i,l}[y_{il}(t) - \bar{y}_{\cdot l}(t)]^2$ is distributed as a non-central $F_{L-1,n-L}\left(\frac{\sum_l n_l \alpha_l^2(t)}{\sigma^2}\right)$.

For our analyses we also adapted to functional data a test procedure with order restriction suggested by Silvapulle and Sen (2005). In the context of a one-way ANOVA, in order to test $H_0 : \mu_1 = \mu_2 = \mu_3$ against $H_1 : \mu_1 \leq \mu_2 \leq \mu_3$ and $\{\mu_1, \mu_2, \mu_3\}$ not all equal, a test statistic is obtained by modifying the F-statistic as

$$\bar{F} = \{RSS(H_0) - RSS(H_1)\}/S^2 \tag{2.4}$$

where

$$RSS(H_0) = \inf_{H_0} \sum\sum (y_{il} - \mu_l)^2 = \sum\sum (y_{il} - \bar{y}_{\cdot l})^2 \ ,$$

$$RSS(H_1) = \min_{H_1} \sum\sum (y_{il} - \mu_l)^2 = \sum\sum (y_{il} - \tilde{\mu}_l)^2,$$

$$S^2 = \upsilon^{-1} \sum\sum (y_{il} - \bar{y}_{\cdot l})^2, \ \upsilon = n_1 + ... + n_L - L$$

and $(\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3)$ is the point at which the sum of squares $\sum\sum (y_{ij} - \mu_i)^2$ is minimized subject to the constraint in H_1. This constrained minimization problem can be solved using efficient computer algorithms. Silvapulle and Sen report the null distribution of $\bar{F}$ for the calculation of the p-value.

2.2.2 Permutation test

Our particular interest in designing this test was whether, over a given time lag, the difference between the area under two mean curves was positive. In order to do that we defined

$$\Delta = \int_{t \in T} (\mu_1(t) - \mu_2(t))\, dt$$

where T is the time lag under consideration and μ_1, μ_2 are the mean curves for N+L and S+L. Our hypothesis was

$$H_0 : \Delta = 0 \ \text{ versus } \ H_1 : \Delta > 0$$

and a possible test statistic is

$$\hat{\Delta} = \sum_i (\hat{\mu}_1(t_i) - \hat{\mu}_2(t_i)) \tag{2.5}$$

with $\hat{\mu}_l(t_i) = \bar{y_{\cdot}}_l(t_i)$ for $l = 1, 2$ and $\bar{y}_{\cdot l}(t_i)$ the sample mean of group l at time t_i. In order to calculate the p-value for the test we considered 1,000 permutations of the data and looked at the number of times that the value of the test statistic was greater than the observed value.

2.3 Results

2.3.1 NL-SL-NS

The above testing procedures were applied to the reconstructed curves, after denoising, during first and second infusion.

As a very simple preliminary analysis we looked at t-test comparisons among the mean tidal volumes during the two infusions for the three intervention groups, "N+S", "N+L" and "S+L". For each patient we computed the difference between the mean Vt during first and second infusion. We then applied the t-test procedure. Tables 1 and 2 report summary measures, p-values and confidence intervals. Results indicate that group N+S appears to be significantly different from groups N+L and S+L, while N+L is not significantly different from S+L. This result suggests a separation between lactate and non-lactate groups.

Table 1: NL-SL-NS: Data summary: Group means and std's of average Vt's during first and second infusions (columns 3-6) and means and std's of differences of mean Vt during first and second infusions (columns 7-8).

Group	n	mean-1st	std-1st	mean-2nd	std-2nd	mean-diff	std-diff
N+S	15	34.496	175.544	-33.973	403.282	-68.469	344.148
N+L	14	90.401	251.369	497.954	448.657	407.552	444.376
S+L	5	-63.948	189.591	281.793	181.533	345.741	97.904

Table 2: NL-SL-NS: Results from t-test.

Groups	p-value	Confidence Interval
N+S vs N+L	0.004	(-781.824, -170.212)
N+L vs S+L	0.632	(-206.766, 330.388)
N+S vs S+L	0.001	(-622.390, -206.030)

As a further investigation we looked into functional ANOVA (FANOVA). With

respect to the simple t-test analysis above, the FANOVA gives the additional information on where the actual differences among the three treatment groups occur in time. Figure 3 shows the results from the pairwise comparisons. We notice that there is no significant difference among the three groups during the first infusion. Groups N+S and N+L differ significantly during the second infusion, with a considerably large difference occurring near the end of this time period. Group S+L, on the other hand, is not well separated from the other two during the second infusion: N+L and S+L are significantly different only for a couple of time points, and only small differences occur between groups N+S and S+L.

The plot of the treatment means revealed a striking feature of the Vt curves, that is mean curves for the two lactate groups are fairly flat during the first infusion while they show a marked increase during the second infusion, unlike the mean curve of group N+S which is fairly flat over the time lags of both first and second infusion, see Figure 4. Investigators had an ordered means hypothesis going into the study, according to which $N + S \leq S + L \leq N + L$. In order to test this hypothesis we adapted to functional data a testing procedure with alternative hypothesis given by that order restriction, see Methods. Figure 5 is a graphical display of the result of the test. The test indicates no difference during the first infusion. During the second infusion, instead, the p-value falls below .05 after the first few minutes, indicating that the restricted alternative hypothesis is strongly supported by the data.

To summarize, the test procedure with an order restriction has suggested a clear ordering among the mean treatments during second infusion. The plot of means and the FANOVA had previously indicated a significant difference among groups N+S and N+L. A distinctive feature is a steady increase of the mean Vt for those who received lactate in the second infusion, N+L or S+L. However, no test procedure has been able to successfully discriminate between N+L and S+L, although the test with

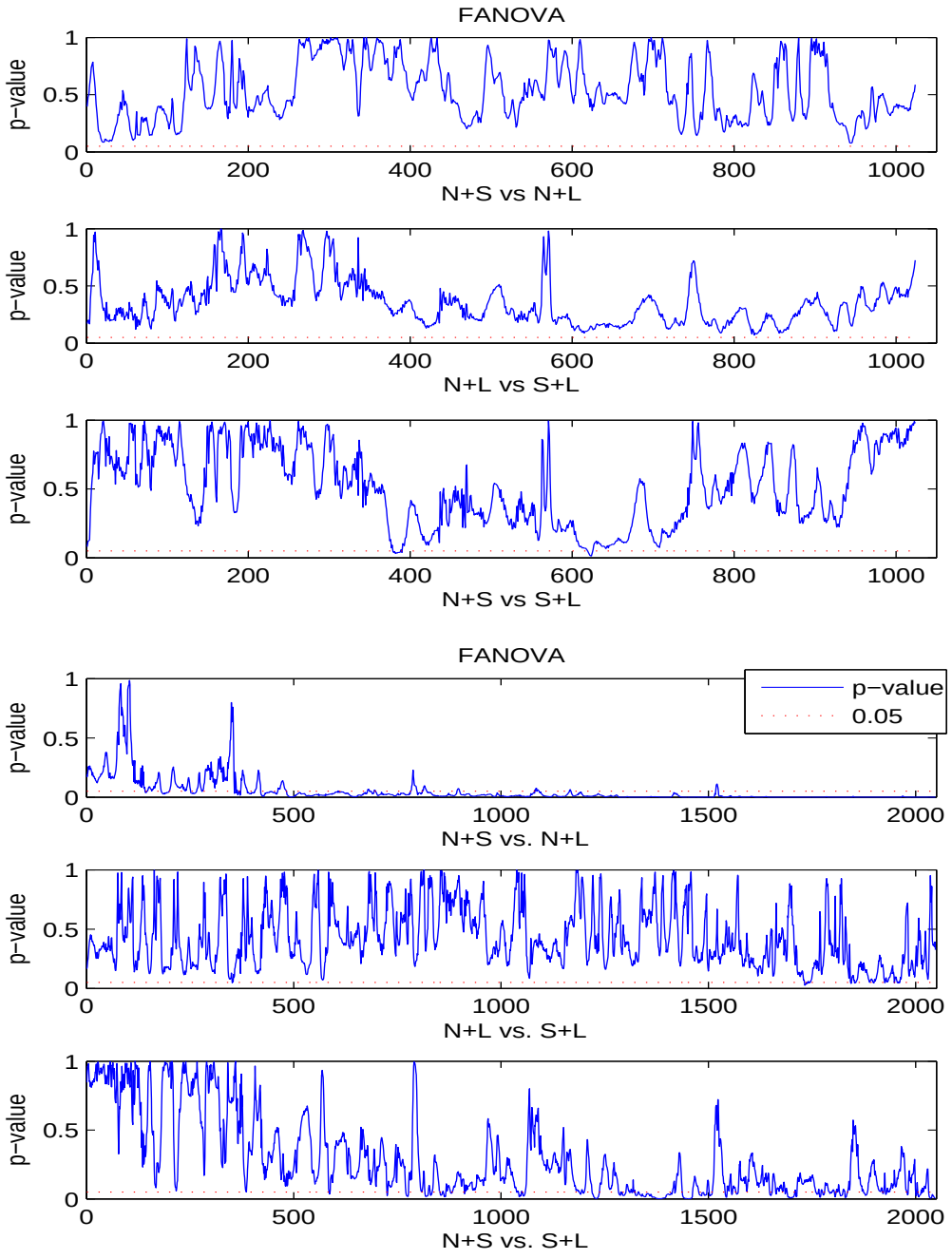

Figure 3: NL-SL-NS: FANOVA on the denoised data for first (upper 3 plots) and second (lower 3) infusion.

order restriction has indicated that the mean Vt for S+L is always smaller than or equal to the mean Vt for the N+L group.

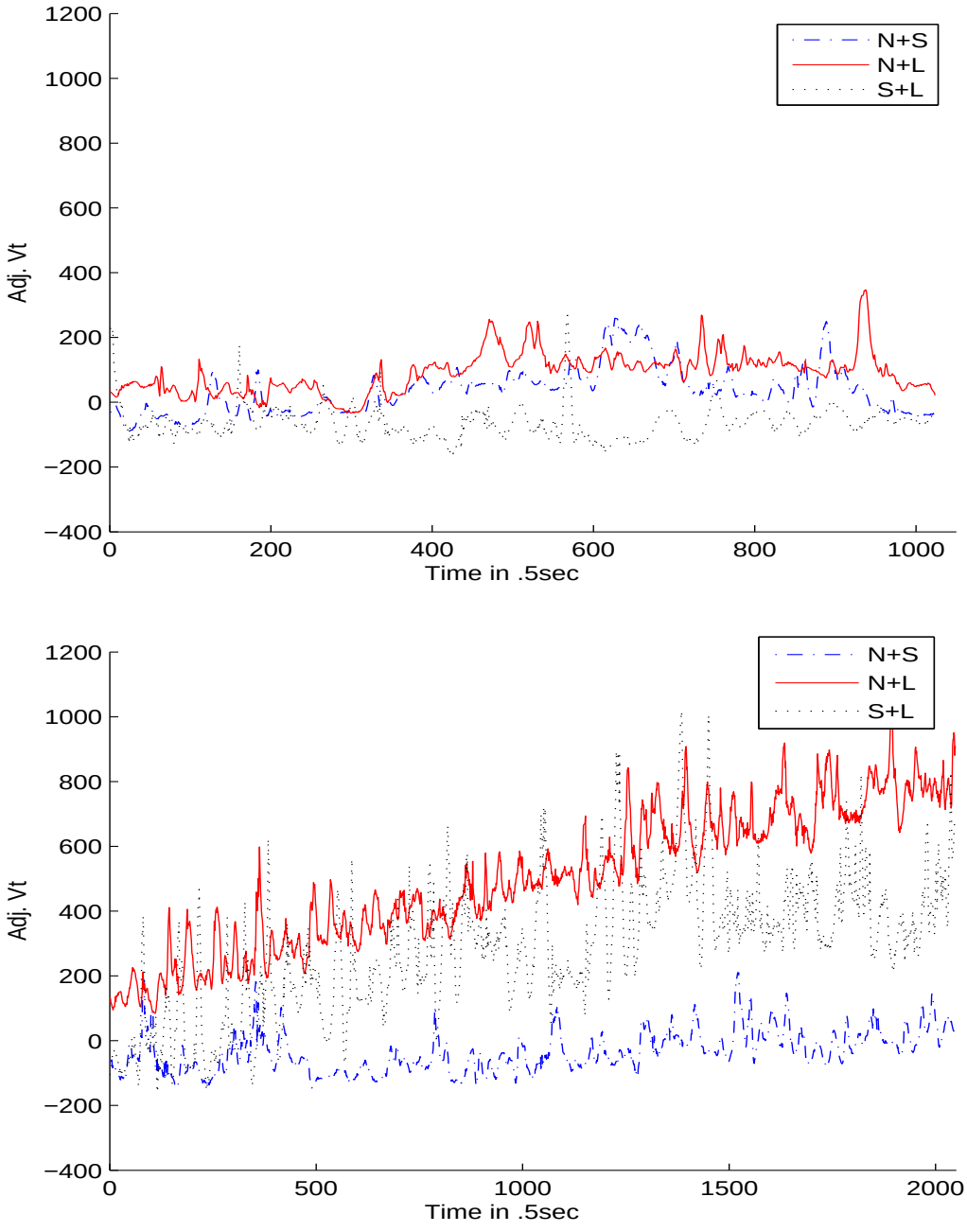

Figure 4: NL-SL-NS: Group mean curves for first (upper) and second (lower) infusion.

Looking again at the plot of means in Figure 4 we noticed that the increase of mean Vt for S+L during the second infusion seems to come to a stop in the

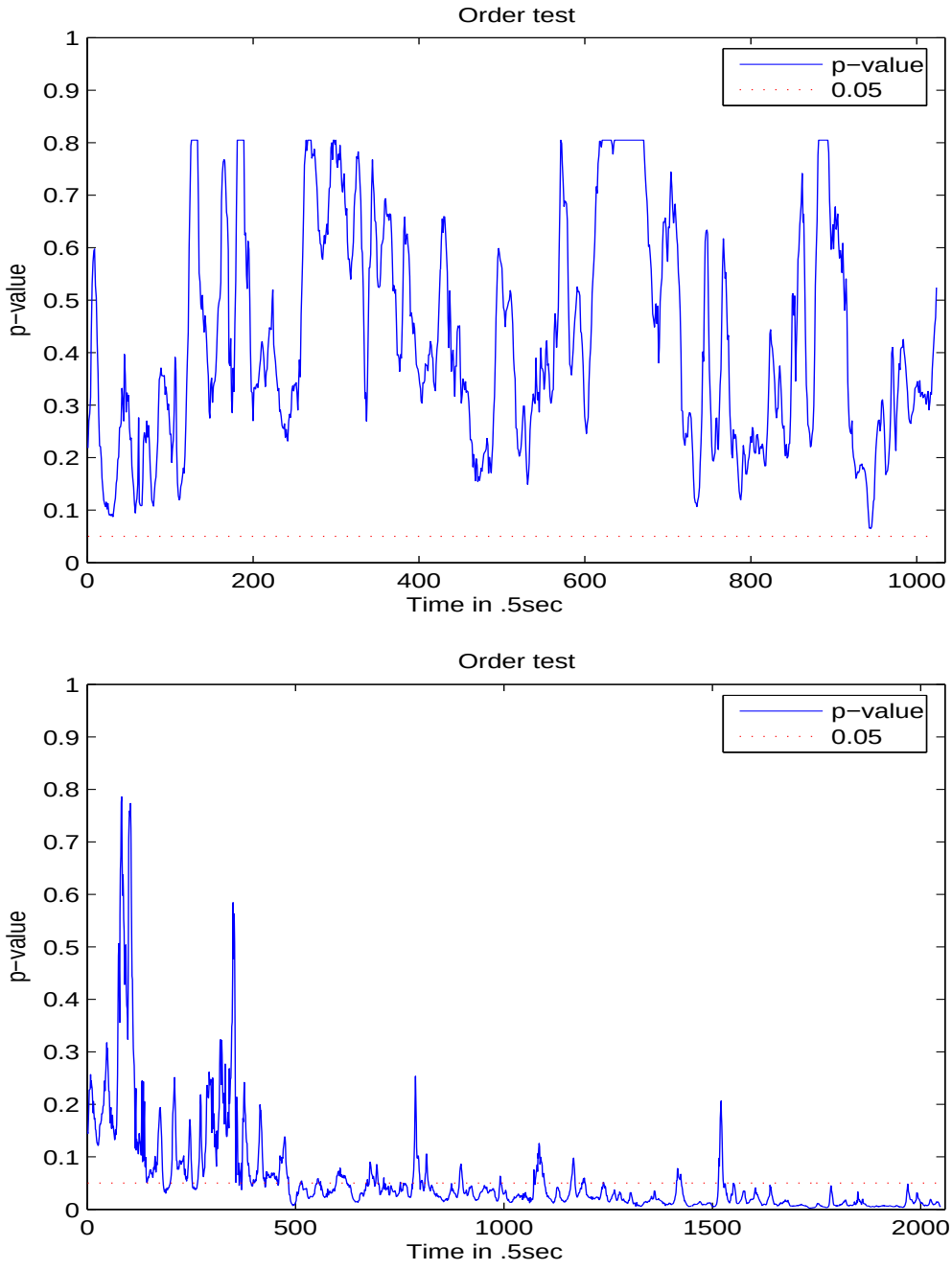

Figure 5: NL-SL-NS: p-values, test with order restriction for first (upper) and second (lower) infusion.

last few minutes, while it keeps increasing for the N+L group. In order to test this possible difference between the two lactate groups, we looked into designing a

nonparametric test specific for the data at hand, see Methods. Our particular interest was in testing whether, over a given time lag, the difference between the area under the two mean lactate curves was positive. We performed the test over several time lags and found that the null hypothesis was rejected when considering the last minutes of the second infusion, in particular the calculated p-value was .05 when considering the last 10 minutes of second infusion and .01 when considering the last 7 minutes. This provided statistical evidence of a longer lactate induced increment in tidal volume when preceded by naloxone.

2.3.2 NL versus SL - unpaired data

At the end of the follow-up study we had a total of 37 NL and 28 SL subjects. Ignoring the within subject correlation should result in a conservative error. We performed a global t-test, functional tests and the permutation test. For these analyses we focused on the 17 minutes of second infusion. We considered baseline adjusted and denoised data. Figure 6 shows the mean Vt traces. Table 3 reports data summary: We took the average of baseline-adjusted Vt for each subject and calculated the mean and standard deviation of these average Vt's. For the global t-test (see table 4) we took the average of baseline-adjusted Vt for each subject and conducted t-test on these averages. P-value for null hypothesis of $NL = SL$ against alternative hypothesis of $NL > SL$ is slightly less than 0.10.

Table 3: NL-SL: Data summary: Group means of average Vt

group	n	mean	std
N+L	37	488.0972	460.6292
S+L	28	347.7279	401.6689

For FANOVA (see figure 7) we conducted pointwise anova procedure (FANOVA)

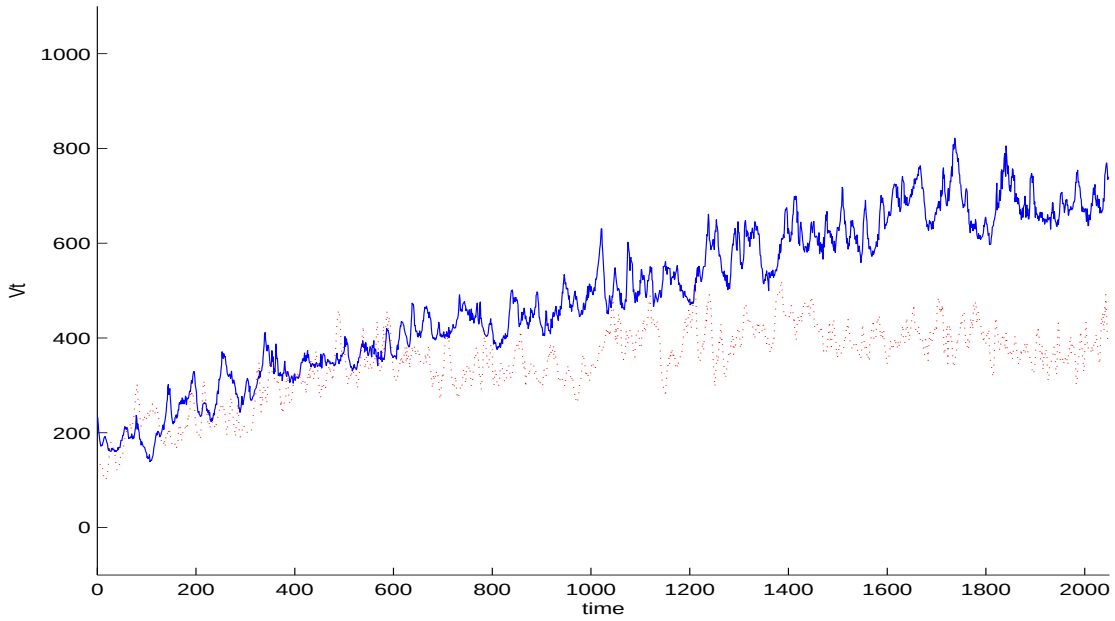

Figure 6: NL-SL: Mean curves, baseline adjusted and denoised data

Table 4: NL-SL: Results of t-test

Group	p-value	C.I.
N+L vs S+L	0.0977	(-73.0929 341.3107)

on baseline-adjusted and denoised Vt. Note that the null hypothesis is NL=SL and the alternative hypothesis is NL $\neq$ SL. Although the p-value's are wiggling, overall p-value's are decreasing over time. Noticeably p-value's falls below .05 in the last few minutes. For the permutation test (see table 5) we used baseline-adjusted and denoised Vt, $H_0 : NL = SL \;\; vs \;\; H_1 : NL > SL$. Note that p-value is decreasing for shorter time lags, for example, p-value for last 17 minutes is about 0.1073, but for last 11 minutes about 0.0597. After this time, p-value's are less than 0.05.

Table 5: NL-SL: Permutation test

time-lag	17:00	15:00	13:00	11:00	9:00	7:00	5:00	3:00
p-value	0.1080	0.0773	0.0693	0.0597	0.0473	0.0313	0.0210	0.0170

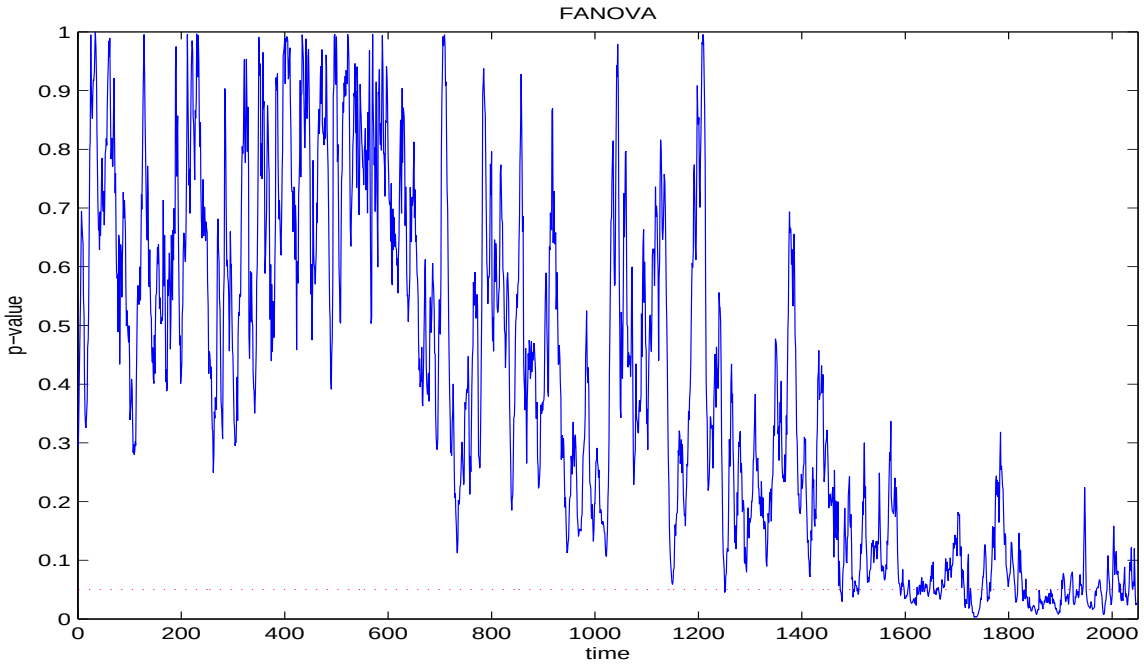

Figure 7: NL-SL: FANOVA on baseline-adjusted and denoised data

2.3.3 NL versus SL - paired data

We also looked at the paired data on the 25 cross-over subjects. Here we used differences of the data (data=NL-SL for each subject). The global t-test: $H_1 : diff > 0$ resulted in a p-value= 0.3657 and was therefore not rejected. A pointwise t-test: $H_1 : diff(t) > 0$, showed p-values decreasing over time (see figure 9). The mean curve of the difference increases over time (see figure 8). The permutation test: $H_1 : diff(\Delta t) > 0$ was significant over the whole 17 minutes time-lag at level 0.05 (see table 6).

Table 6: Paired NL-SL: p-value for permutation test on differences

time-lag	17	15	13	11	9	7	5
p-value	0.0400	0.0270	0.0250	0.0190	0.0260	0.0200	0.0117

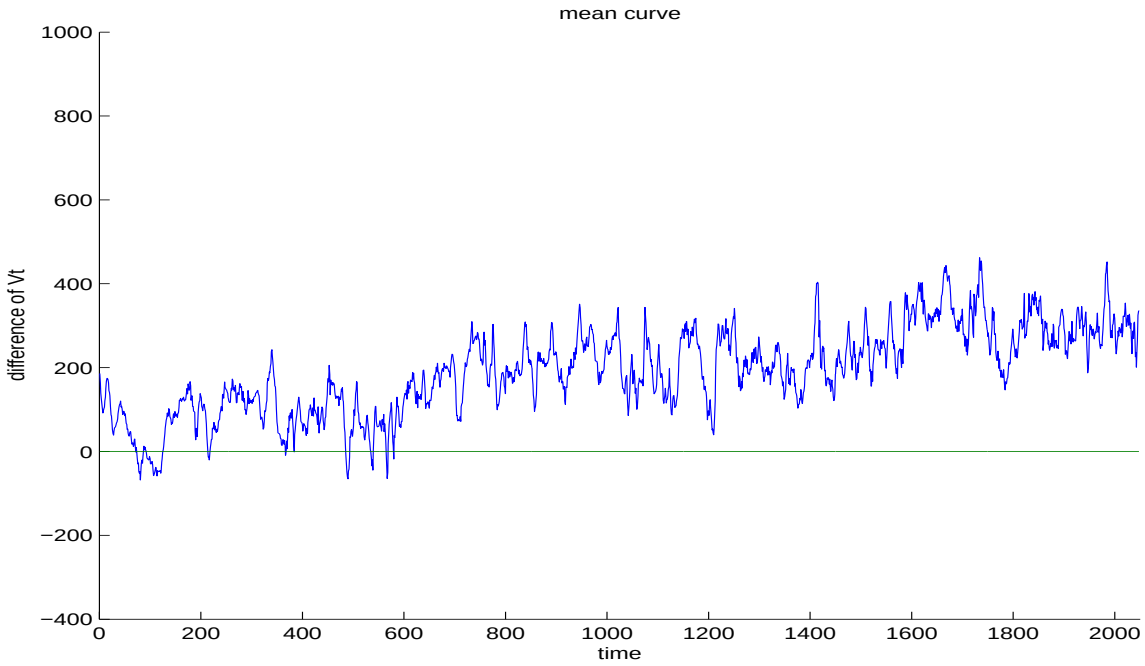

Figure 8: Paired NL-SL: Mean curve of differences, horizontal line is 0

2.3.4 Test on smoothed components

The analysis of the paired data may be affected by the lack of "registration" of the Vt curves. Curve registration (Ramsey and Silverman) is a process according to which curves are "calibrated" across time, i.e., aligned with respect to some common feature or characteristic. Since different subjects have very different breathing cycles, registration of Vt traces is not trivial. As an additional analysis we have looked into the analysis of features of the data extracted by wavelet decompositions. Essentially, using the wavelet decomposition we have filtered out high frequency components (longitudinal variance) that are unrelated to the breathing frequencies, i.e. they

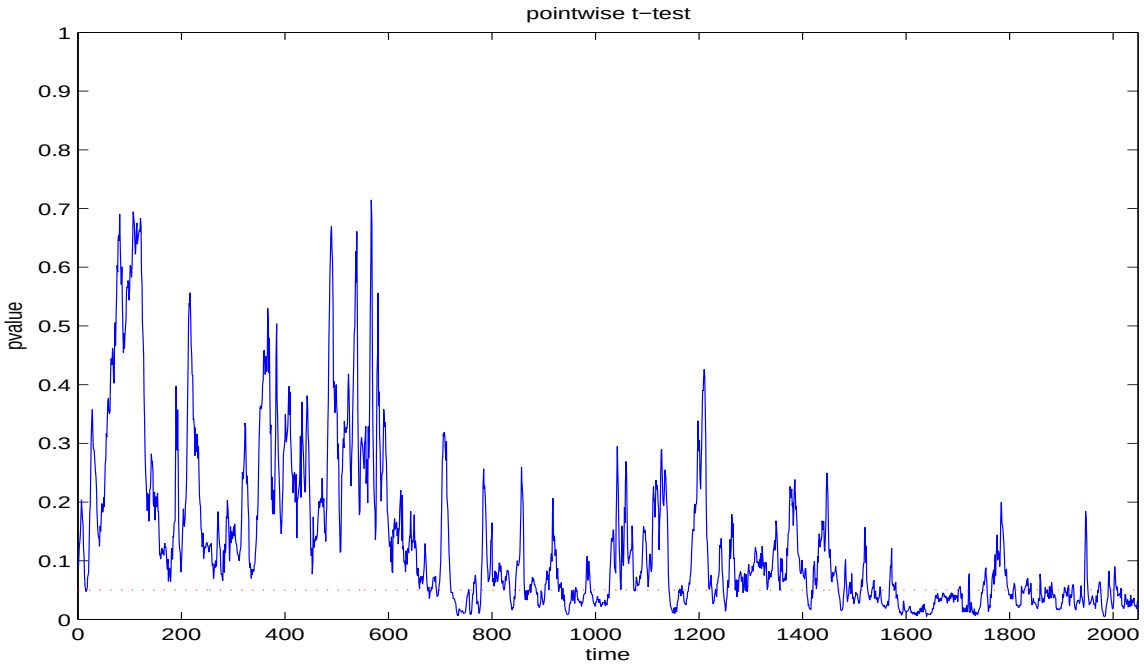

Figure 9: Paired NL-SL: Pointwise t-test on differences

constitute irrelevant information (see Section Pre-processing of the data in Chapter I). Considering that an average breath takes 4-6sec, we have analyzed Vt traces after subtraction of components at 4sec and less. We have also looked at the trend component. This drastic smoothing of the data should lessen the effect of the lack of registration.

For the 25 cross-over subjects, Figures 11 and 13 show p-values for the pointwise t-test: $H_1 : diff(t) > 0$ that decrease over time. Mean curves of differences are plotted in Figures 10 and 12. Permutation tests: $H_1 : diff(\Delta t) > 0$ were significant over the whole 17 minutes time-lag at level 0.05 (see tables 7 and 8). These results confirm previous findings. In particular, notice how the pointwise test on the trend of the data indicates that the two lactate groups are significantly different on the last 10-11 minutes of the infusion, which confirms the findings of the NL-SL-NS study.

Table 7: Paired NL-SL, extracted component: p-value for permutation test on differences

time-lag	17	15	13	11	9	7	5
p-value	0.0397	0.0283	0.0240	0.0210	0.0253	0.0197	0.0160

Table 8: Paired NL-SL, trend: p-value for permutation test on differences

time-lag	17	15	13	11	9	7	5
p-value	0.0400	0.0357	0.0220	0.0220	0.0187	0.0177	0.0160

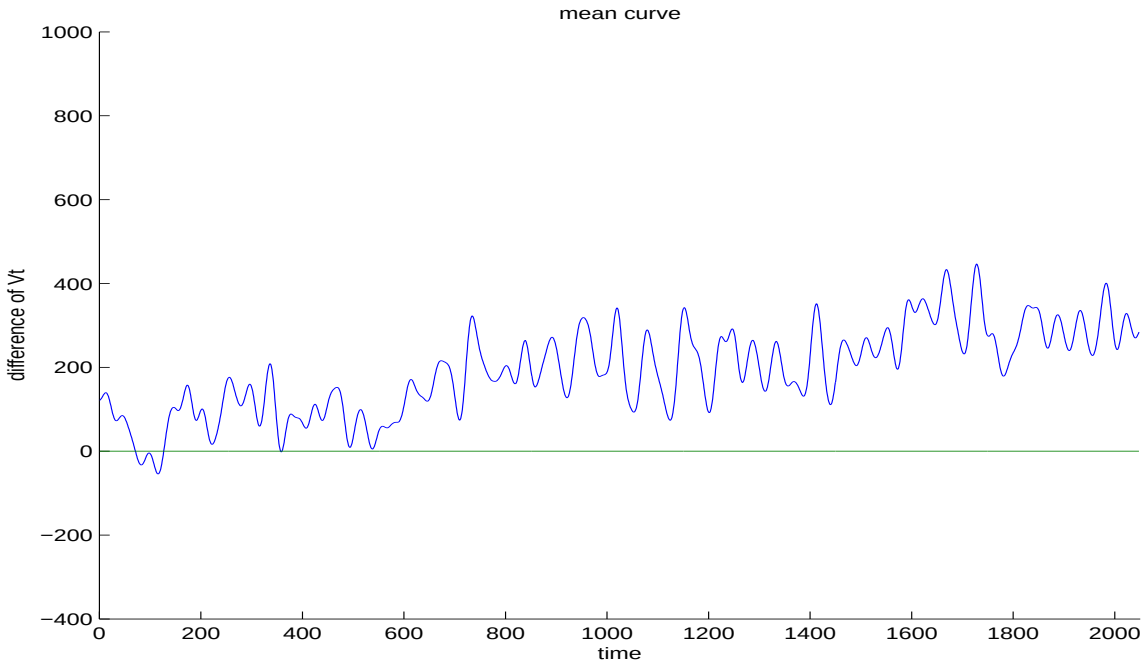

Figure 10: Paired NL-SL, extracted component: Mean curve of differences, horizontal line is 0

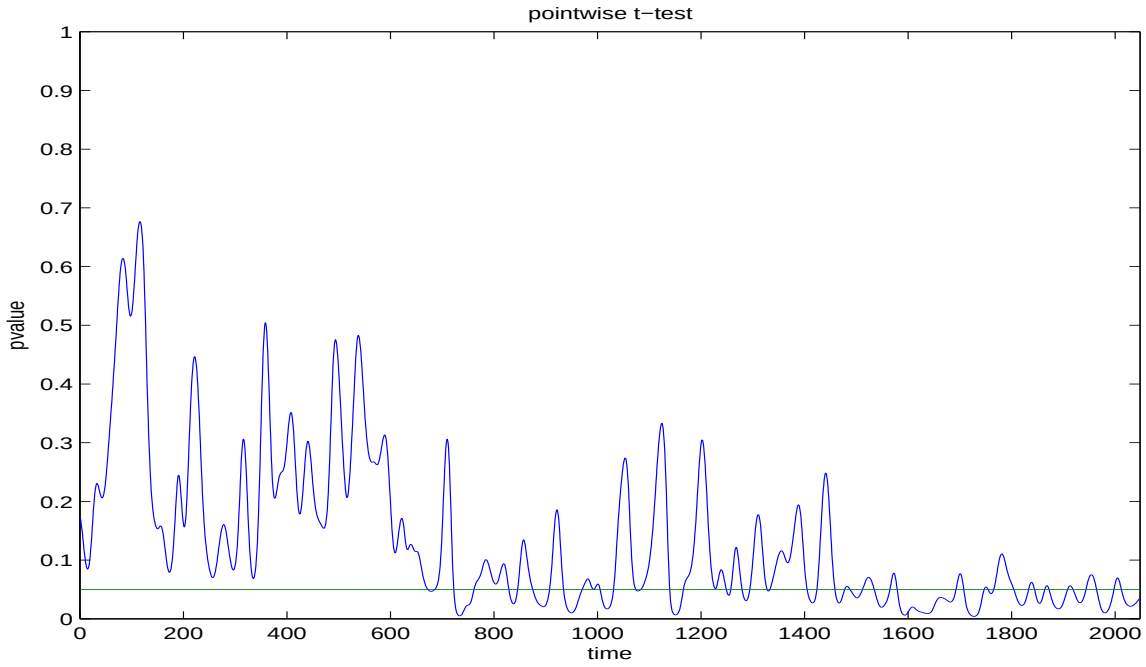

Figure 11: Paired NL-SL, extracted component: Pointwise t-test on differences

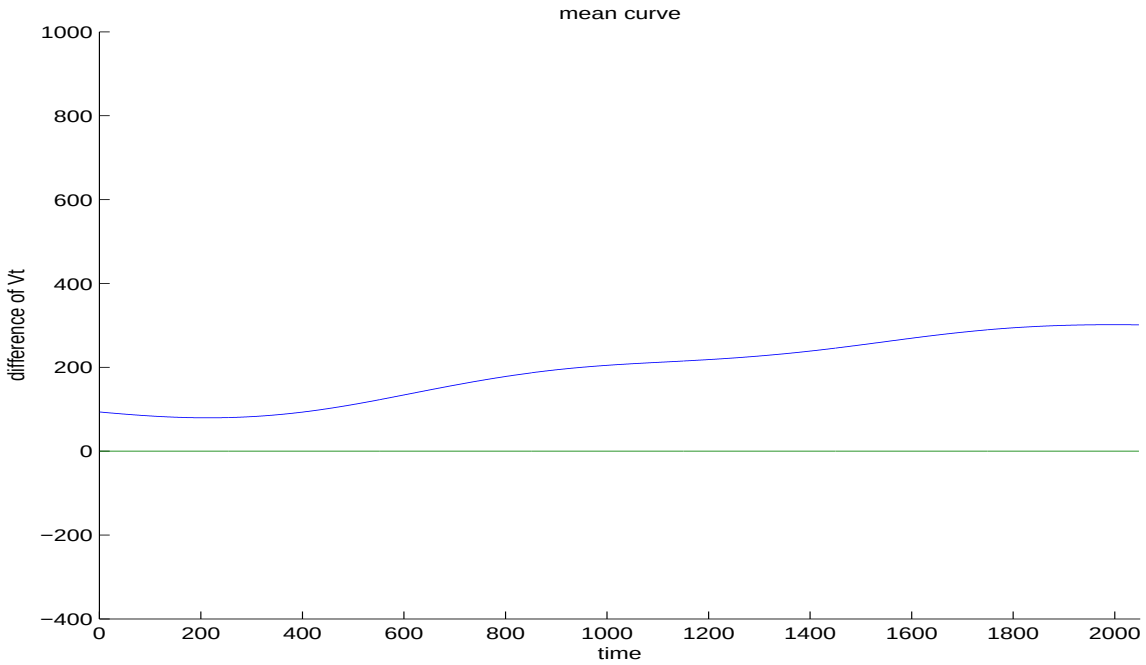

Figure 12: Paired NL-SL, extracted trend: Mean curve of differences, horizontal line is 0

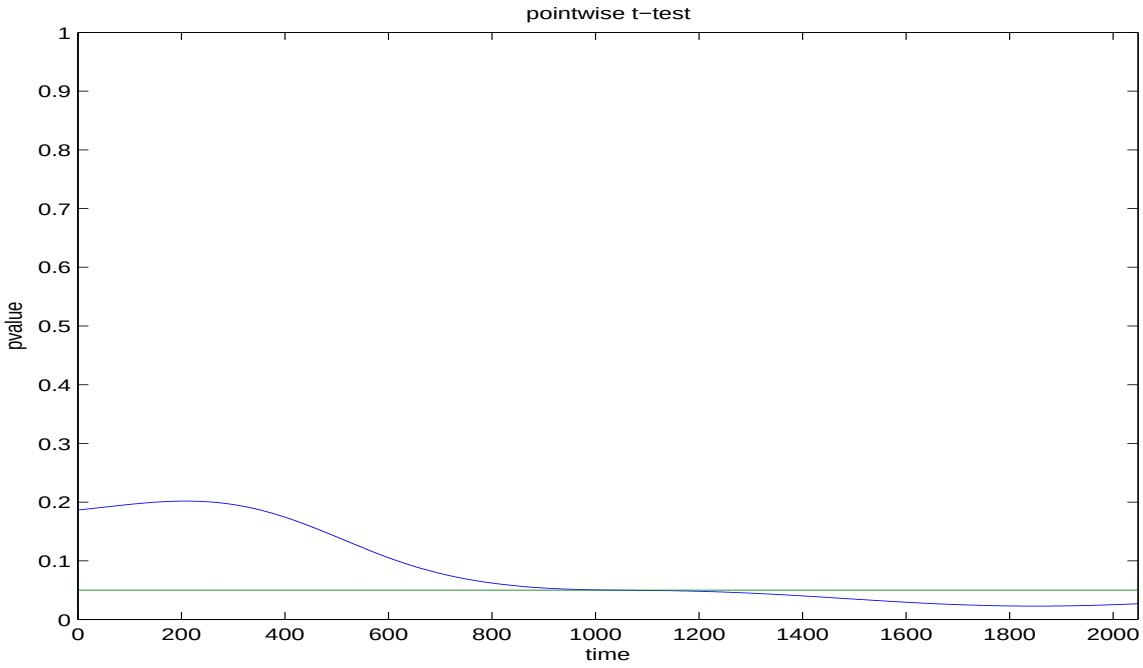

Figure 13: Paired NL-SL, extracted trend: Pointwise t-test on differences

CHAPTER III

ORDER-PRESERVING DIMENSION REDUCTION TEST FOR THE DOMINANCE OF TWO MEAN CURVES

3.1 Introduction

High-dimensional functional data have become prominent in a number of medical and biological fields. There the units of observation are curves and the observed data consist of sets of curves, often sampled on a fine grid. More and more attention among researchers is now devoted to the development of appropriate statistical methodologies suitable for the analysis of such data. Our interest here is to test the order in mean between two sets of curves. To be specific, let $f_1(t)$ and $f_2(t)$ indicate the mean curves of the two sets of curves. We are interested in testing whether $f_1(t) \geq f_2(t)$ for every t. Our work is motivated by a study that looks at high-dimensional, high-frequency measurements of tidal volume, i.e. the volume of gas exchanged during each ventilated breath, on a number of individuals subject to interventions that may induce panic attacks.

There are several approaches to deal with the testing problem at hand. A first naive approach is to compute the average value of each curve and then apply a one-side t-test. However, this global test completely ignores the point-wise nature of the data. A second naive approach is a point-wise t-test. With respect to the simple global t-test, point-wise t-tests give the additional information on where the significance occurs. This procedure, however, is sub-optimal and leads to large type I errors due to multiple testing. One can expect that an overall test that combines all these point-wise comparisons, such as a nonparametric procedure, would perform better. Fan (**?**) and other researchers, see Serban and Wasserman (**?**) and references therein,

suggested nonparametric methods for curve testing problems based on representations of the curves that use basis functions, such as wavelets or Fourier bases. These methods, however, do not consider any order constraints.

The testing problem we are considering has also been addressed in the context of multivariate statistics. O'Brien (**?**) suggested test procedures that take into account heterogeneity over t. Pocock and Tsiatis (**?**) considered the extension of the O'Brein's test to survival times subject to censoring. Follmann (**?**) proposed a modified Hotelling's T statistic and computed its asymptotic distribution. Tang and Geller (**?**) suggested the use of an approximate likelihood ratio test. All these approaches require a consistent covariance matrix estimate, which is typically not readily available with high-dimensional functional data.

We propose a new procedure to test the order in mean of two sets of regularly observed curves. The key idea of the suggested procedure relies on preserving the order in mean while reducing the dimension of the data. We do this by projecting the observed data matrix onto a space of low rank matrices which are represented as a product of a coefficient matrix and a positively constrained basis matrix. Here, the positively constrained basis matrix preserves the order between two curves, that is, if one curve is larger than the other one, then the coefficient vectors preserve the same ordering. Thus, once we find a low dimensional representation of the data matrix, we can then apply multivariate testing procedures to the coefficient vectors.

The order preserving matrix factorization we adopt, denoted by non-negative basis matrix factorization (NBMF), minimizes the Frobenius norm between the observed data matrix and a pre-specified lower rank matrix which imposes positive constraints to the basis vectors. In this sense, our procedure is comparable to other dimension reduction procedures such as principal component analysis or the non-negative matrix factorization of Lee and Seung (**?**). However, Principal component analysis does

not impose any constraint, while the non-negative matrix factorization assumes positiveness of both coefficients and basis vectors. NBMF, on the other hand, is not convex and no algorithm can guarantee the convergence to a global minimum. We propose an iterative procedure that converges to a local minimum. We also provide a probabilistic view of the adopted dimension reduction procedure that offers a good understanding of the distributional properties of the coefficient vectors and the residuals.

Here we look at data arising from an experiment where measurements of tidal volume are taken on a number of individuals subject to interventions that may induce panic attacks. Details of the study and the interventions are given in the Introduction. Prior to the study investigators had an ordered mean hypothesis of the type $f_1(t) \geq f_2(t)$ with groups 1 and 2 defined by two different interventions. Figure 14 presents the mean curves for the two groups and confirms the intuition of the investigators. Our task was to design a test to statistically validate this finding. Conclusions from the results of our analysis have suggested novel findings to the investigators.

The remainder of the chapter is organized as follows. In Section 3.2 we describe the non-negative basis matrix factorization method, provide an algorithm for its implementation and apply the procedure to testing the order between two high-dimensional mean curves. A short review of the multivariate testing procedure we use is also given. In Section 3.3 we illustrate the suggested procedures on a simulation study and in Section 3.4 we present results from the case-study example that motivated our work. Section 3.5 concludes the chapter.

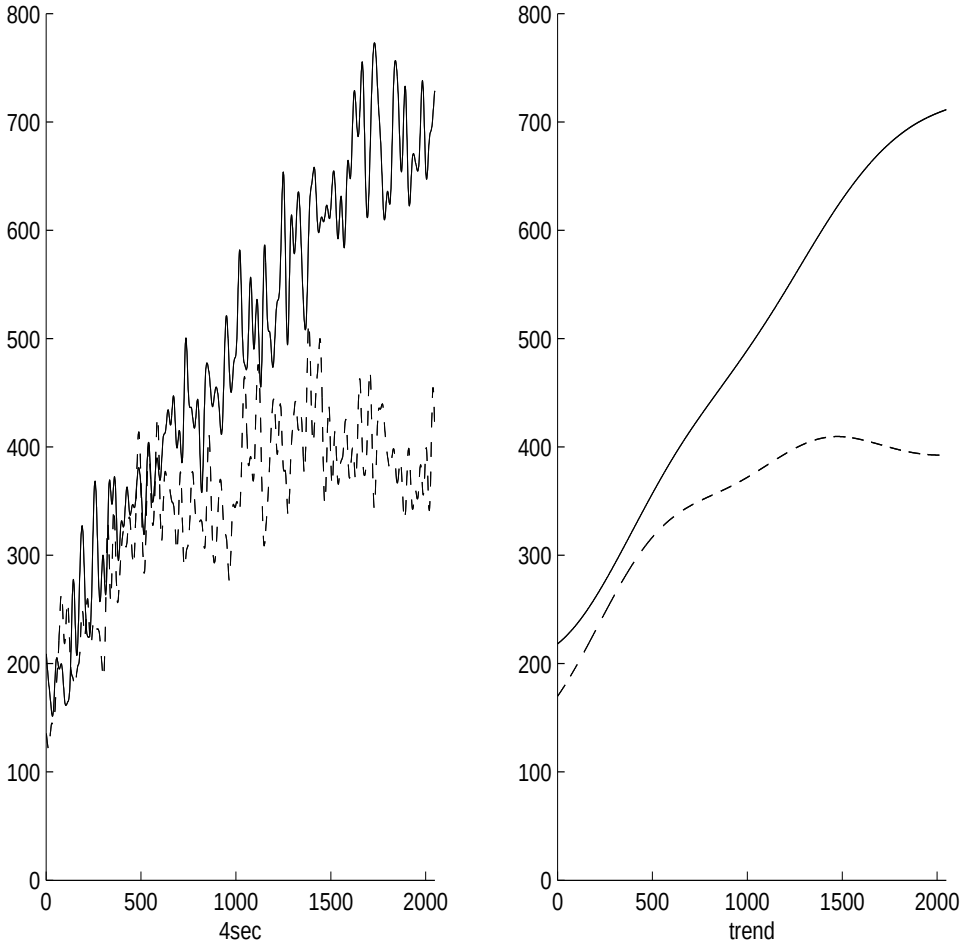

Figure 14: Mean curves for N+L (line) and S+L (dotted line) after pre-processing. The left panel is for traces after subtraction of components at 4sec and less, the right panel is for the trends.

3.2 Methods

This Section focuses on the problem of testing the order between two mean curves, f_1 and f_2, using sampled curves. To be specific, let

$$\mathbf{Y}^{(1)} = \begin{pmatrix} Y_1^{(1)} \\ Y_2^{(1)} \\ \vdots \\ Y_{n_1}^{(1)} \end{pmatrix} = \begin{pmatrix} y_{11}^{(1)}, & y_{12}^{(1)}, & \ldots, & y_{1p}^{(1)} \\ y_{21}^{(1)}, & y_{22}^{(1)}, & \ldots, & y_{2p}^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n_1 1}^{(1)}, & y_{n_1 2}^{(1)}, & \ldots, & y_{n_1 p}^{(1)} \end{pmatrix}$$

represent the observed n_1 curves for group 1, and $\mathbf{Y}^{(2)}$ the similarly defined $n_2 \times p$ matrix of curves for group 2. We assume that each curve $Y_j^{(i)}$, for $j = 1, 2, \ldots, n_i$, is independent of each other. Let $f_i = (f_{i1}, f_{i2}, \ldots, f_{ip})$, for $i = 1, 2$, indicate the p-dimensional mean curves for groups 1 and 2, respectively. We are interested in testing $f_1 \geq f_2$, i.e. $f_{1k} \geq f_{2k}$ for every $k = 1, \ldots, p$.

3.2.1 Order preserving dimension reduction

As pointed out in the introduction, there is no readily available method to test the order between f_1 and f_2 for large p. The key idea behind our proposal is to represent $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ via lower dimensional vectors of coefficients, so that the problem becomes tractable. In particular, in doing dimension reduction we want to preserve the order between the two mean curves, so that the same hypothesis, $f_1 \geq f_2$, can be tested in the lower dimensional representation of the data. In the reduced space we use existing multivariate testing procedures. Specifically, suppose that $\mathbf{W}^{(1)}$ (and $\mathbf{W}^{(2)}$) is a lower dimensional approximation of $\mathbf{Y}^{(1)}$ (and $\mathbf{Y}^{(2)}$) and that μ_1 (and μ_2) is its mean vector. Thus, we test the order between f_1 and f_2 by testing the order between μ_1 and μ_2.

3.2.2 Problem statement

In this section we state our dimension reduction problem precisely.

Let us consider n curves observed at p time points. Here p is typically much larger than n. We seek to do dimension reduction by finding a small number of local features of the curves, each defined as a positive linear combination of the p time points. More precisely, we find a low rank approximation to a $n \times p$ data matrix $\mathbf{Y}$ that consists of a coefficient matrix $\mathbf{W}$, $n \times r$, and non-negative basis matrix $\mathbf{H}$, $r \times p$,

by solving the following non-linear optimization problem

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{Y} - \mathbf{WH}\|_{\mathrm{F}}^2 \\ &\text{subject to} \quad \mathbf{H} \geq 0, \quad \|H_k\|_{\mathrm{F}}^2 = 1 \quad \text{for } k = 1, 2, \ldots, r, \end{aligned} \tag{3.1}$$

where H_k is the kth column vector of $\mathbf{H}$, $\|\mathbf{A}\|_{\mathrm{F}}$ is the Frobenius norm of $\mathbf{A}$, i.e., $\|\mathbf{A}\|_{\mathrm{F}} = \left(\sum_{ij} A_{ij}^2\right)^{1/2}$, and where the dimension r of $\mathbf{W}$ is a parameter to be set by the user.

This problem is analogous to other dimension reduction procedures such as principal component analysis (PCA) or the non-negative matrix factorization (**?**), in the sense that it projects the original data matrix onto a space of lower rank matrices by minimizing the Frobenius norm. PCA, for example, minimizes

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{Y} - \mathbf{WH}\|_{\mathrm{F}}^2, \\ &\text{subject to} \quad \|H_k\|_{\mathrm{F}}^2 = 1, \quad \text{for } k = 1, 2, \ldots, r, \end{aligned} \tag{3.2}$$

where the H_k's are orthogonal to each other. The non-negative matrix factorization, instead, minimizes (3.2) with the additional constraints $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$.

3.2.3 An iterative algorithm

The function to be minimized in (3.1) is convex either in $\mathbf{W}$ or $\mathbf{H}$, but not in both. Using this fact we propose the following iterative procedure to solve the least square problem and find a local minimum:

(i) Given the current estimate of $\mathbf{W}$, solve the constrained least squares problem

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{Y} - \mathbf{WH}\|_{\mathrm{F}}^2 \\ &\text{subject to} \quad \mathbf{H} \geq 0, \quad \|H_k\|_{\mathrm{F}}^2 = 1, \quad \text{for } k = 1, 2, \ldots, r, \end{aligned} \tag{3.3}$$

(ii) Given the current estimate $\mathbf{H}$, update $\mathbf{W}$ as

$$\mathbf{W} = \mathbf{YH}^T\left(\mathbf{HH}^T\right)^{-1} \tag{3.4}$$

which is the solution to the unconstrained least squares problem

$$\text{minimize} \quad \|\mathbf{Y} - \mathbf{WH}\|_{\mathrm{F}}^2 \tag{3.5}$$

Step (i) requires to solve a quadratic program (QP) with linear inequality constraints and quadratic equality constraints. We solve this step with a two-stage procedure: First, we solve the QP with linear inequalities

$$\begin{aligned} &\text{minimize} && \|\mathbf{Y} - \mathbf{WH}\|_{\mathrm{F}}^2 \\ &\text{subject to} && \mathbf{H} \geq 0, \end{aligned}$$

and then we normalize the resulting estimates $\mathbf{H}$ as

$$H_k = H_k / \|H_k\|_{\mathrm{F}}^2, \quad \text{for } k = 1, 2, \ldots, r.$$

Simple algebra can show the equivalence between the two optimization procedures, i.e. step (i) and two-stage procedure we use.

The Karush-Kuhn-Tucker (KKT) conditions of a QP with linear equality constraints is a set of linear equations which can be solved analytically (**?**). Two most common ways to solve a QP, or its KKT, conditions are the interior point method and the simplex method (Boyd and Vandenberghe, 2004 , and Van de Panne, 1974). The interior point method solves the QP with linear inequality constraints by reducing it to a sequence of linear equality constrained problems. The simplex method solves the KKT conditions by reformulating the problem into a linear programming problem. In this paper we use the MOSEK toolbox (**?**).

The proposed iterative least square procedure converges to a local minimum since each step finds a new estimate which improves the Frobenius norm. Let $\left(\mathbf{W}^{(k)}, \mathbf{H}^{(k)}\right)$ indicate the current state and $\left(\mathbf{W}^{(k+1)}, \mathbf{H}^{(k+1)}\right)$ the subsequent estimate from steps

(i) and (ii). Then,

$$\begin{aligned}
\|\mathbf{Y} - \mathbf{W}^{(k)}\mathbf{H}^{(k)}\|_{\mathrm{F}}^2 &\geq \min_{\mathbf{H}} \|\mathbf{Y} - \mathbf{W}^{(k)}\mathbf{H}\|_{\mathrm{F}}^2 \\
&= \|\mathbf{Y} - \mathbf{W}^{(k)}\mathbf{H}^{(k+1)}\|_{\mathrm{F}}^2 \\
&\geq \min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{W}\mathbf{H}^{(k+1)}\|_{\mathrm{F}}^2 \\
&= \|\mathbf{Y} - \mathbf{W}^{(k+1)}\mathbf{H}^{(k+1)}\|_{\mathrm{F}}^2.
\end{aligned}$$

3.2.4 A probabilistic view

In this section we provide a probabilistic view of NBMF that turns out to be helpful in setting practical guidelines for the choice of the dimension r of the matrix $\mathbf{H}$. We start by explaining the model we assume for the observation matrix $\mathbf{Y}$. Suppose that each column vector Y_j of the observed matrix $\mathbf{Y}$ has a mean vector f which belongs to the space spanned by the column vectors of the $r^* \times p$ matrix $\mathbf{H}^*$. Suppose that the Y_j's have covariance matrix Ω for every $j = 1, 2, \ldots, n$ and that they can be represented as

$$\mathbf{Y} = \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix} \mathbf{H}^* + \mathbf{E}, \tag{3.6}$$

with μ, a $1 \times p$ vector, and where each column vector of $\mathbf{E}$ has mean vector 0 and covariance matrix Ω.

Let us now consider the choice of r in the case of the NBMF method. Suppose we mistakenly choose a higher dimension r than the true dimension r^*. Let $\mathbf{H}$ be the basis matrix constructed by adding to $\mathbf{H}^*$ the extra basis vectors $H_{r^*+1}, \ldots, H_r$, and $\mathbf{H}^{\perp}$ be the matrix of basis vectors $H_{r+1}, \ldots, H_p$, orthogonal to those of $\mathbf{H}$. We re-write (3.6) as

$$\mathbf{Y} = \eta\mathbf{H} + \mathbf{E}_1\mathbf{H} + \mathbf{E}_2\mathbf{H}^{\perp}, \tag{3.7}$$

where

$$\eta = \begin{pmatrix} \mu & 0 \\ \mu & 0 \\ \vdots & \vdots \\ \mu & 0 \end{pmatrix}$$

The optimization problem (3.1) becomes

$$\begin{aligned} \left\|\mathbf{Y} - \mathbf{WH}\right\|_{\mathrm{F}}^2 &= \left\|(\eta + \mathbf{E}_1 - \mathbf{W})\mathbf{H} + \mathbf{E}_2\mathbf{H}^{\perp}\right\|_{\mathrm{F}}^2 \\ &= \left\|(\eta + \mathbf{E}_1 - \mathbf{W})\mathbf{H}\right\|_{\mathrm{F}}^2 + \left\|\mathbf{E}_2\mathbf{H}^{\perp}\right\|_{\mathrm{F}}^2. \end{aligned}$$

because of the orthogonality between $\mathbf{H}$ and $\mathbf{H}^{\perp}$.

Thus, the solution to the problem, given $\mathbf{H}$, is

$$\mathbf{W} = \eta + \mathbf{E}_1, \tag{3.8}$$

and we have

$$\begin{aligned} \min_{\mathbf{W}} \left\|\mathbf{Y} - \mathbf{WH}\right\|_{\mathrm{F}}^2 &= \left\|\mathbf{E}_2\mathbf{H}^{\perp}\right\|_{\mathrm{F}}^2 && (3.9) \\ &\approx n\left\|\mathbf{H}^{\perp T}\Omega\mathbf{H}^{\perp}\right\| && (3.10) \\ &= n\sum_{k=r+1}^{p}\left\|H_k^T\Omega H_k\right\| && (3.11) \end{aligned}$$

where $\|A\| = \sum_{ij}|A_{ij}|$.

This follows from the fact that the variance of a column vector of $\mathbf{E}_2\mathbf{H}^{\perp}$ is $\mathbf{H}^{\perp T}\Omega\mathbf{H}^{\perp}$, because of the orthogonality between $\mathbf{H}$ and $\mathbf{H}^{\perp}$ and the relationship

$$\mathbf{E} = \mathbf{E}_1\mathbf{H} + \mathbf{E}_2\mathbf{H}^{\perp},$$

and from the orthogonality among $H_{r+1}, \ldots, H_p$. Note that the Frobenius norm of $\left\|\mathbf{E}_2\mathbf{H}^{\perp}\right\|_{\mathrm{F}}^2$ is defined as the sum of squares of each component in the matrix. Hence, it is reasonable to expect it to be close to $n\left\|\mathbf{H}^{\perp T}\Omega\mathbf{H}^{\perp}\right\|$.

A difficulty arises from the fact that the covariance matrix Ω is unknown. We assume Ω to be isotropic, in the sense that it is invariant to the orthonormal rotation of the axes. With this assumption, the following statistics, which we denote by MSE in the remainder of the chapter,

$$\begin{aligned} \min_{\mathbf{W}} \left\|\mathbf{Y}-\mathbf{W}\mathbf{H}\right\|_{\mathrm{F}}^{2} \Big/ \big\{(p-r)n\big\} &= \left\|\mathbf{E}_2\mathbf{H}^{\perp}\right\|_{\mathrm{F}}^{2} \Big/ \big\{(p-r)n\big\} \\ &\approx \sum_{k=r+1}^{p} \left\|H_k^T \Omega H_k\right\| \Big/ (p-r) \end{aligned} \tag{3.12}$$

is expected to be constant for every $r \geq r^*$. In the special case $\Omega = \sigma^2\mathbf{I}$, we can expect the MSE to be approximately equal to σ^2 for $r \geq r^*$. Thus, if the MSE does not markedly decrease after a certain r, we can choose r as the dimension of the reduced space.

3.2.5 Testing the order between two mean curves

In order to use the order preserving dimension reduction method in our testing problem we apply a common basis matrix $\mathbf{H}$ to both data matrices, $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$, and find lower dimensional approximations. To be specific, we set

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{pmatrix} \tag{3.13}$$

and find $r-$rank approximations with positive basis vectors of the type

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{pmatrix} \approx \begin{pmatrix} \mathbf{W}^{(1)} \\ \mathbf{W}^{(2)} \end{pmatrix} \mathbf{H}. \tag{3.14}$$

We can now apply a multivariate statistical testing procedure to the lower dimensional approximations of the data. Testing the order between the mean curves f_1 and f_2 is equivalent to testing the order between the mean vectors of their lower

dimensional approximations, μ_1 and μ_2. From (3.14) it is clear that the following relation between f_1 and f_2, and μ_1 and μ_2 holds:

$$f_1 = \mu_1 \mathbf{H} \quad \text{and} \quad f_2 = \mu_2 \mathbf{H}. \tag{3.15}$$

That is,

$$\begin{aligned} f_1 \geq f_2 \quad &\Leftrightarrow \quad \mu_1 \mathbf{H} - \mu_2 \mathbf{H} \geq 0 \\ &\Leftrightarrow \quad \big(\mu_1 - \mu_2\big)\mathbf{H} \geq 0 \\ &\Leftrightarrow \quad \big(\mu_1 - \mu_2\big)\mathbf{H}\mathbf{H}^T \geq 0 \\ &\Leftrightarrow \quad \mu_1 - \mu_2 \geq 0 \end{aligned}$$

and we are left to test

$$\mathrm{H}_0 : \mu_1 = \mu_2 \quad \text{against} \quad \mathrm{H}_1 : \mu_1 \geq \mu_2$$

based on the lower dimensional approximation $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$.

3.2.6 Follmann's multivariate test

We use the Follmann's multivariate procedure to test the order between μ_1 and μ_2. Other procedures, such as the approximate likelihood ratio test by Tang and Geller (**?**), may be used here to test the hypothesis. The Follmann's test has a good power for the alternative hypothesis of a positive mean vector. The test rejects if a quadratic form of the sample mean vector exceeds its 2α critical value and the sum of the elements of the mean vector exceeds zero. This test is shown to have type I error rate equal to α for both cases of known and unknown covariance matrix. It can also be shown that (3.16) converges in distribution to a half chi-square distributed random variable with p degrees of freedom.

To test a one-sided alternative hypothesis (**?**) suggested the use of the following modified Hotellings' T-statistics

$$\begin{aligned} \mathrm{T} &= \big(\overline{\mathbf{W}}^{(1)} - \overline{\mathbf{W}}^{(2)}\big)^T \big(n_1^{-1}\mathbf{S}_1 + n_2^{-1}\mathbf{S}_2\big)^{-1} \big(\overline{\mathbf{W}}^{(1)} - \overline{\mathbf{W}}^{(2)}\big) \\ &\quad \times \mathrm{I}\Big(\sum_{k=1}^{p} (\overline{\mathbf{W}}_k^{(1)} - \overline{\mathbf{W}}_k^{(2)}) > 0\Big), \end{aligned} \tag{3.16}$$

where

$$\begin{aligned} \mathbf{S}_i &= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \big(\mathbf{W}_j^{(i)} - \overline{\mathbf{W}}^{(i)}\big)\big(\mathbf{W}_j^{(i)} - \overline{\mathbf{W}}^{(i)}\big)^T, \\ \overline{\mathbf{W}}^{(i)} &= \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{W}_j^{(i)}, \end{aligned}$$

for $i = 1, 2$, with $\overline{W}_k^{(i)}$ the $k-$th component of $\overline{\mathbf{W}}^{(i)}$ and $\mathrm{I}(\cdot)$ the indicator function.

3.3 Numerical examples

We consider three numerical examples to show the performance of the NBMF method. The first example illustrates the property of the MSE defined in Section 3.2.4, while the second example shows its accuracy in estimating the true lower dimensional representation. Finally, we investigate the power and coverage of the NBMF method.

3.3.1 Selection of r

Our first example illustrates how the MSE (3.12) decreases for $r \leq r^*$ and is constant for every $r \geq r^*$. Here, r^* is the true dimension of the lower dimensional representation of the data, while r is the dimension pre-specified by the user. This example confirms that the MSE provides a good reference for choosing an appropriate r.

The numerical study is set up as follows. We fix the true column rank r^* of $\mathbf{H}$ as 5, the dimension p of each curve as 10 and 100, and the number of subjects, n, as 20. We randomly generate a coefficient matrix $\mathbf{W}$ of dimensions 20×5 and a basis

matrix $\mathbf{H}$ of dimensions 5×10 and 5×100 using uniform distributions within the interval $[-10, 10]$ and $[0, 10]$, respectively. We also randomly generate an error matrix $\mathbf{E}$ using independent normal distributions with mean 0 and variance 1 and construct an observation matrix as

$$\mathbf{Y} = \mathbf{WH} + \mathbf{E}.$$

With this setting, we generate 100 observational matrices. We solve the NBMF problem with each of the generated observational matrices, denoting the solution for the kth matrix as $\widehat{\mathbf{W}}_r^{(k)}$ and $\widehat{\mathbf{H}}_r^{(k)}$, for $r = 1, 2, \ldots, 9$. Figure 15 shows the box-plots of the MSE's computed for each $r = 1, 2, \ldots, 9$. The MSE values clearly decrease as r increases towards r^* while they stay constant for $r \geq r^*$, as expected.

3.3.2 Accuracy of NBMF

The second simulation study shows the accuracy of the NBMF method in finding basis vectors in a reduced space. The numerical study is designed as follows. We fix $r^* = 1$ and assume it is correctly chosen; thus, we also fix $r = 1$. We fix $\mathbf{H}$ as a randomly selected 1×10 vector with $\|\mathbf{H}\|_{\mathrm{F}}^2 = 1$. We generate 100 matrices $\mathbf{W}$ and $\mathbf{E}$ as described in Section 3.3.1. We construct 100 observation matrices as $\mathbf{Y} = \mathbf{WH} + \mathbf{E}$ and solve the NBMF problem for each one of these. Table 9 reports the true $\mathbf{H}$, means and standard deviations (std) of the estimated basis vectors, followed by the first five estimates.

3.3.3 The power of NBMF method

Finally, a small simulation study has been conducted to investigate the performance of NBMF method, i.e., power and test size. Two true mean functions are chosen based on panic data as the below,

$$f_1(t_k) = 224 + 0.26t_k + u_1(t_k), \;\; f_2(t_k) = 200 + 0.3t_k - 0.0001t_k^2 + u_2(t_k) \qquad (3.17)$$

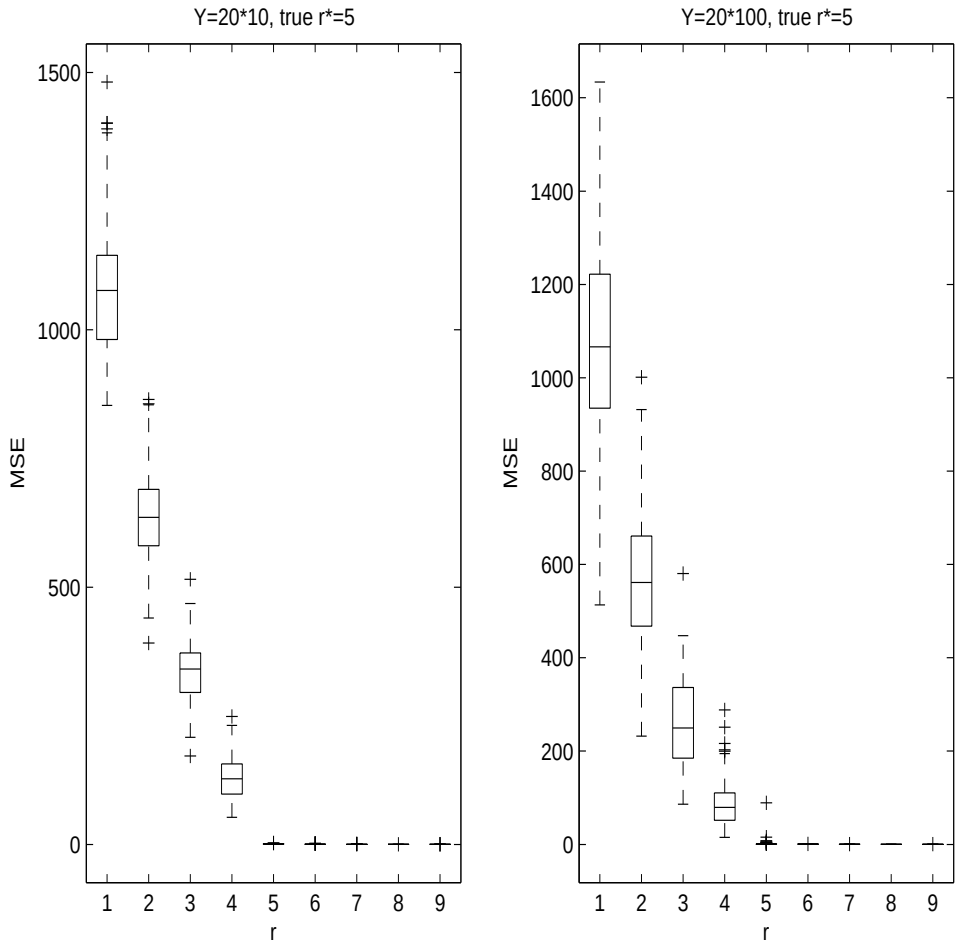

Figure 15: Box-plots of MSE's for different values of r, based on 100 generated data sets. The left panel is the boxplot for p=10 and the right panel is for p=100.

where $u_1(t) = 20\sqrt{t/2^{10}}\sin(50\pi/(t/2^{10}+0.05))$, $u_2(t) = 10\sin(\pi t/2^4)$ for $k = 1, 2, ..., p$. Doppler function and sine function are added to two main functions, respectively to resemble panic data (see figure 16). Since a multivariate statistical testing procedure will be applied to the lower dimensional approximations of the data, we find the lower dimensional approximations, $\mu_1,\ \mu_2$, $1 \times r$ vectors, of true functions $f_1,\ f_2$ as well as the lower dimension basis $\mathbf{H}$, $r \times p$ matrix, by NBMF described in Section 3.2.1.

To get the sample curves, first, sample coefficients W_{ij} in the lower dimension space are randomly generated from a multivariate normal distribution with mean μ_i

Table 9: The first column is the true basis vector $\mathbf{H}$, the second column is the mean of 100 estimates, the third column is the standard deviation of 100 estimates, and the other columns are the first 5 estimates $\widehat{\mathbf{H}}$.

true H	**mean**	**std**	$\widehat{\mathbf{H}}1$	$\widehat{\mathbf{H}}2$	$\widehat{\mathbf{H}}3$	$\widehat{\mathbf{H}}4$	$\widehat{\mathbf{H}}5$
0.0234	0.0229	0.0038	0.0262	0.0226	0.0267	0.0169	0.0266
0.0459	0.0460	0.0041	0.0505	0.0531	0.0421	0.0397	0.0468
0.0694	0.0694	0.0038	0.0715	0.0726	0.0732	0.0654	0.0689
0.0871	0.0873	0.0042	0.0840	0.0908	0.0897	0.0958	0.0905
0.1337	0.1316	0.0032	0.1328	0.1318	0.1332	0.1356	0.1329
0.1347	0.1363	0.0035	0.1330	0.1370	0.1363	0.1375	0.1382
0.2777	0.2770	0.0041	0.2781	0.2788	0.2850	0.2827	0.2743
0.5361	0.5350	0.0026	0.5357	0.5280	0.5326	0.5359	0.5324
0.5378	0.5388	0.0017	0.5408	0.5365	0.5375	0.5369	0.5391
0.5433	0.5437	0.0026	0.5408	0.5501	0.5421	0.5404	0.5460

and covariance $\sigma_i^{*2}\mathbf{I}_{r\times r},\ j = 1, 2, ..., n_i$ for each $i = 1, 2$ where $\mathbf{I}_{r\times r}$ is the identity matrix with rank r. Then, $\mathbf{WH}$ provides n sample curves where $n = n_1 + n_2$ and $\mathbf{W} = (W_{ij}),\ n \times r$ matrix. At last, we add to $\mathbf{WH}$ noise term $\mathbf{E}$, $n \times p$ matrix, generated from a normal with mean 0 and variance σ^2.

One can expect that the power of any multivariate testing procedure would decrease as σ^{*2} increases. Intuitively, bigger σ^{*2} makes two groups harder to be differentiated. σ^2 would be expected to make same effect on the performance of the test as σ^{*2} does. Most crucial impact on the performance of the test would be the difference between two mean functions. Therefore, we consider various $\sigma^{*2},\ \sigma^2$ and Δ^* for the performance of NBMF method where $\Delta^* = \frac{1}{T}\int_{t\in\mathcal{T}}(f_1(t) - f_2(t))^2 dt$, i.e., a standardized L_2 norm, T is the Euclidian measurement of a compact set of $\mathcal{T}$.

Analytic form of the power of NBMF test is not available. Hence, we calculate the power and test size in the empirical way described below.

1. Generate random sample curves under the null or alternative hypothesis for

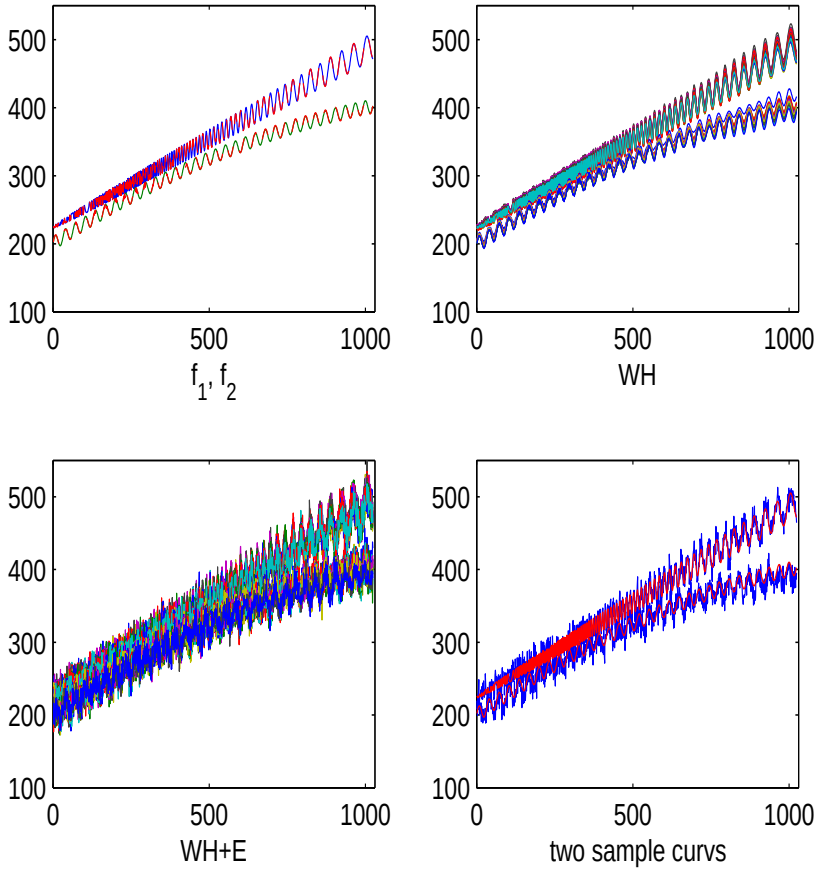

Figure 16: Sample curves in model (1); the left upper is two true population curves superimposed by W^*H, the right upper is sample curves of WH, the right lower is sample curves of $WH+E$, the left lower is two sample curves of $WH+E$ superimposed by true mean curves (red)

fixed Δ^* and SNR.

2. Run the NBMF test on the sample data from (1).

3. Repeat step (1)-(2), m times.

4. Calculate the portion of rejections out of m.

In this example, we have two mean functions and generate 25 sample curves from each

one at $p = 1024$ observed time points, $t_k \in [1, 2, ..., 1024]$, total sample size $n = 50$ with $m = 1,000$ replications as described above. We find the lower dimensional approximation at $r = 2$ for the computational convenience. In fact, r could be from 1 to 24. Reconstructed curves by $\mathbf{W}^*\mathbf{H}$ fit the true almost perfectly where $\mathbf{W}^* = (\mu_1^T, \mu_2^T)^T$ (see the left upper panel in figure 16). Hence, we are satisfied with $r = 2$. $\sigma^{*2} = 10^2$ or 10^4 are selected based on $\mu_1 = (11238, 519),\ \mu_2 = (11022, -740)$. We tried two σ^2's according to SNR =1 or 1/3 where SNR is the ratio of the standard deviation of signal to noise.

Under the null hypothesis the test size is 0.0510 at SNR=1 and 0.0630 at SNR=1/3 both with $\sigma^{*2} = 10^2$ while with $\sigma^{*2} = 10^4$, the test size of NBMF test is 0.021 at SNR=1 and 0.0580 at SNR=1/3. This makes sense since it is becoming harder to detect the difference as σ^{*2} or σ^2 increase. In other words, test should not reject the null hypothesis at $\sigma^{*2} = 10^4$ more often than $\sigma^{*2} = 10^2$. To vary the Δ^*, we set $f_1(t) = f_2(t)$ for $t \in [t_k, ..., 1024]$ or $[1, ..., t_k]$. lag is the length of interval of $f_1 \neq f_2$. Tables 10 and 11 report the power over various Δ^*, SNR at $\sigma^{*2} = 10^2$ or 10^4. Notice that at smaller Δ^* and bigger σ^{*2} the power of test is still good. For the comparison, we run the pointwise one-side t-test on the simulated data also. Figure 17 shows the p-value by pointwise t-test on data of lag=10. Pointwise t-test hardly says the difference. But, NBMF test clearly declare the overall difference.

Table 10: Power at $\sigma^{*2} = 10^2$.

lag (from the right)	10	20	30	40
SNR Δ^*	70	170	248	305
1	1.0	1.0	1.0	1.0
1/3	0.690	0.9090	0.999	1.0

Table 11: Power at $\sigma^{*2} = 10^4$.

lag (from the left)	50	75	100
SNR Δ^*	24	40	55
1	0.801	0.998	1.0
1/3	0.468	0.7290	1.0

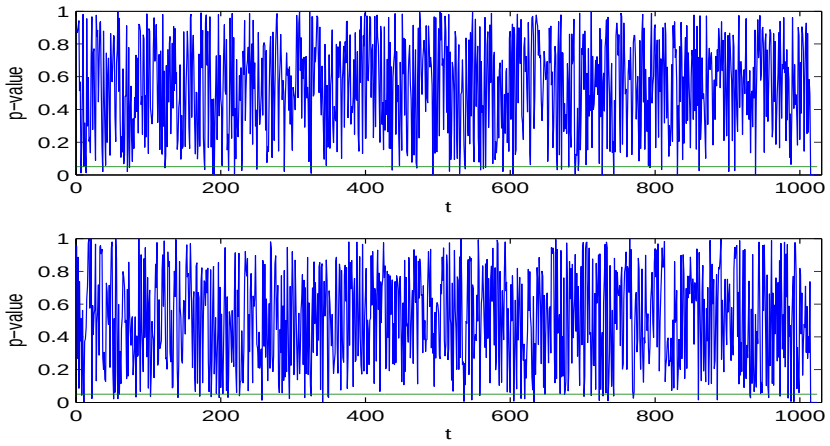

Figure 17: Pointwise t-test in lag=10; the upper is the figure at $SNR = 1$, the lower at $SNR = 1/3$

3.4 Application

Here we analyze data from an experiment that looks at high-dimensional, high-frequency measurements of tidal volume on a number of individuals subject to interventions that may induce panic attacks. Prior to the study investigators had an ordered mean hypothesis of the type $f_1(t) \geq f_2(t)$ with groups 1 and 2 defined by two different interventions.

3.4.1 Tidal volume (Vt) traces

We applied our testing procedure to the data spanning over a time window covering the lactate infusion. Based on their previous experience with lactate infusions

investigators do not expect a quick onset of effect. We therefore chose a window of approximately 17 minutes before the end of the infusion.

Figure 18 displays some of the data. The first row shows two sample Vt curves after baseline adjustment. The trace in the left panel is an N+L sample, the one in the right panel an S+L. The second row shows the same two curves after substraction of components at 4sec and less and the third row shows the corresponding trends. Figure 14 displays the mean curves for N+L (line) and S+L (dotted line). The left panel is for traces after subtraction of components at 4sec and less, the right panel is for the trends. Working with the trends has the advantage of essentially avoiding complications with registration issues (**?**). Curve registration is a process according to which curves are "calibrated" across time, i.e., aligned with respect to some common feature or characteristic. Effective registration procedures for respiratory flows and tidal volume measurements, however, are not trivial. In our study subjects have very different breathing cycles and basically do not exhibit common features.

3.4.2 Results

The investigators' claim can be formalized as an hypothesis testing problem with alternative hypothesis of the type $(N + L)(t) \geq (S + L)(t)$ for every t. We have a total of 65 curves with 2048 observed points, 37 curves belong to the N+L group and 28 to the S+L group. We can test the hypothesis on the smoothed curves and also on the trends. A very small r, i.e., a large reduction, can be used for very smooth curves such as the trend data. In our analysis, we used $r = 4$ for the trend data and $r = 25$ for the smoothed data after subtraction of components at 4sec and less. Figure 19 shows original and approximated curves by **WH** for three N+L and three S+L subjects.

In order to start our iterative procedure, **W** was randomly generated by the

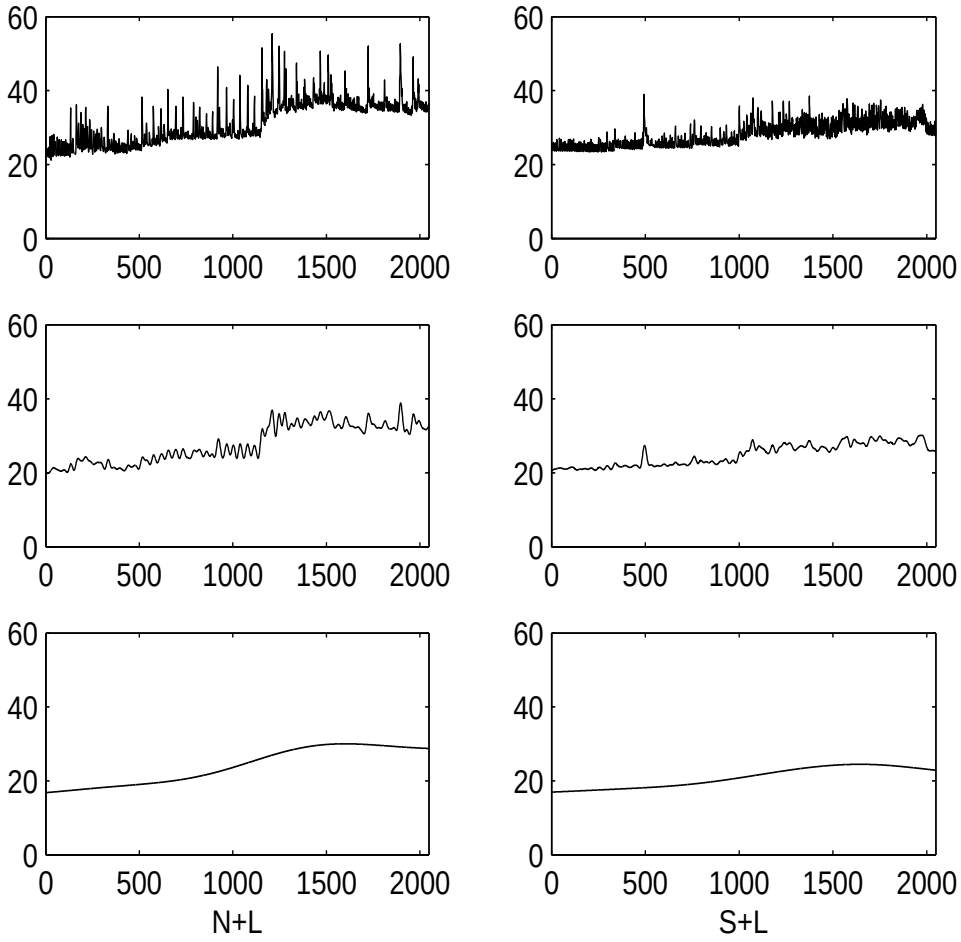

Figure 18: The first row shows two sample VT curves after baseline adjustment. The trace in the left panel is an N+L sample, the one in the right panel an S+L. The second row shows the same two curves after substraction of components at 4sec and less and the third row shows the corresponding trends.

uniform distribution in the interval [-100,100]. Convergence was achieved in only 10 iterations. We fist applied the test to the entire time-lag of the data. The test was significant, see Table 12, confirming the intuition of the investigators that subjects receiving the naloxone-lactate sequence have greater increases in tidal volume. Since investigators were also interested in an indication of the time at which the significance occurs, we applied the testing procedure to shorter time-lags of the data. Results are summarized in table 12 and seem to indicate that the dominance of the N+L mean

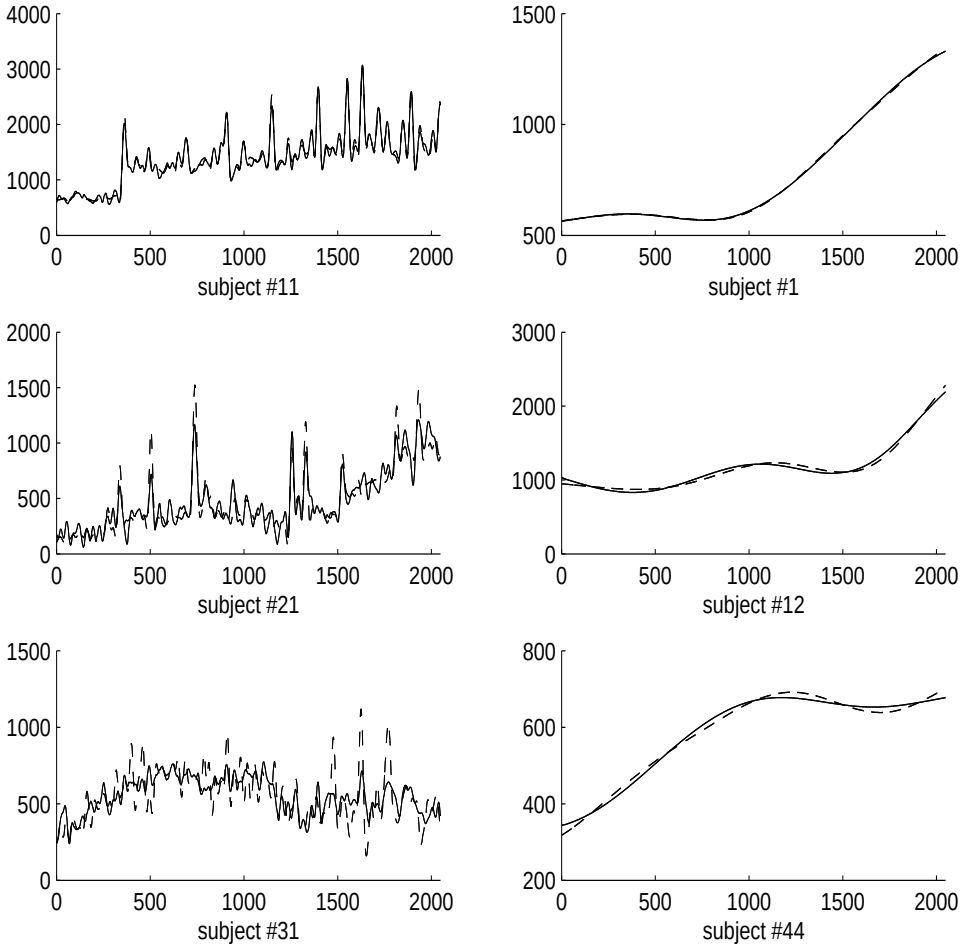

Figure 19: Original (line) and approximated (dotted line) traces for 6 randomly selected subjects. We used $r = 25,\ 4$, for smoothed data and trends, respectively. The left panel is for traces after subtraction of components at 4sec and less, the right panel is for the trends

curve over the S+L one becomes more and more significant starting at approximately 10 minutes before the end of the infusion.

Table 12: P-values of test results with smoothed data and trends. The half chisquare distribution of the Follmann's test statistics for large sample was used. P-values from a permutation test using the Follmann's test statistics are reported in parentheses.

time-lag	smoothed	trend
first 5 min	0.1435(0.1795)	0.4172(0.4295)
first 10 min	0.0564(0.0405)	0.3641(0.3710)
first 15 min	0.0031(0.0000)	0.0738(0.0815)
last 5 min	0.0180(0.0070)	0.0457(0.0445)
last 10 min	0.0082(0.0010)	0.0481(0.0615)
whole time	0.0030(0.0000)	0.0483(0.0535)

CHAPTER IV

ESTIMATING MONOTONE CONVEX FUNCTIONS VIA SEQUENTIAL SHAPE MODIFICATION

This Chapter proposes an estimation method for a monotone convex function that consists of two sequential shape modification stages: (i) monotone regression via solving a constrained least square problem and (ii) convexification of the monotone regression estimate via solving an associated constrained uniform approximation problem. This method is faster than the conventional constrained least squares (LS) method by about two orders of magnitude. Moreover, we show that, under an appropriate smoothness condition, the uniform convergence rate achieved by the proposed method is nearly comparable to the best achievable rate for a nonparametric estimate which ignores the shape constraint. Simulation studies show that the uniform error achieved by the proposed method is comparable to that achieved by the constrained LS method.

4.1 Introduction

Consider the regression model

$$Y_i = f(X_i) + \sigma\epsilon_i, \tag{4.1}$$

where f is a monotone convex function from an interval $\Omega \subseteq \mathbb{R}$ into $\mathbb{R}$, ϵ_i are independent and identically distributed (i.i.d.) random variables with zero mean and unit variance, and $\sigma > 0$. This paper concerns the problem of estimating of f, using the samples $(x_1, y_1), \ldots, (x_n, y_n)$ from this model. Monotone convex (concave) function estimation problems arise in several disciplines including economics (Aït Sahalia and Duarte, 2003; Matzkin, 1994) and medical studies (Lloyd, 2002).

One can extend the constrained least squares method, which is quite effective in monotone function estimation or isotonic regression, to (4.1). (See, e.g., Barlow, Bartholomew, Bremner and Brunk (1972) and Robertson, Wright and Dykstra (1988) for more on isotonic regression.) The extension requires one to solve an infinite-dimensional optimization problem of the form

$$\begin{aligned} &\text{minimize} \quad \textstyle\sum_{i=1}^{n} \left| y_i - \widehat{f}(x_i) \right|^2 \\ &\text{subject to} \quad \widehat{f} \in \mathcal{F}_{\mathrm{mc}}, \end{aligned}$$

where $\mathcal{F}_{\mathrm{mc}}$ is the class of all monotone convex functions from Ω into $\mathbb{R}$. Unlike isotonic regression, this problem does not has an analytic solution, and is difficult to solve even approximately. Moreover, related theoretical issues such as convergence rate have not been understood well.

One can apply the shape constrained spline estimation method, described in Mammen, Marron, Turlach, and Wand (2001), to (4.1). This method assumes that the function to be estimated is at least twice differentiable. In this paper, we do not assume such a smoothness constraint on f.

The main purpose of this Chapter is to propose a computationally efficient method for estimating a monotone convex function. This method is faster than the conventional constrained LS method by about two orders of magnitude. Moreover, we show that, under an appropriate smoothness condition, the uniform convergence rate achieved by the proposed method is nearly comparable to the previously known best uniform convergence rate of a nonparametric estimate which is not necessarily monotone convex. Simulation studies show that the uniform error achieved by the proposed sequential method is comparable to that achieved by the constrained least squares method. In these aspects, the proposed method seems to give a better compromise between computational efficiency and accuracy than the constrained LS method.

4.2 Sequential shape modification

This Section describes a method for estimating the monotone convex function f from the samples $\{(x_i, y_i)\}_{i=1}^n$.

- Monotonization. Solve the constrained LS problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n \big(y_i - f(x_i)\big)^2 \\ \text{subject to} & f \in \mathcal{F}_{\text{mon}}, \end{array} \tag{4.2}$$

  in which $\mathcal{F}_{\text{mon}}$ is the set of all monotone functions from Ω into $\mathbb{R}$.

- Convexification. Solve the constrained uniform approximation problem

$$\begin{array}{ll} \text{minimize} & \sup_{x\in\Omega} \big|\widehat{f}(x) - \tilde{f}(x)\big| \\ \text{subject to} & \tilde{f} \in \mathcal{F}_{\text{con}}, \end{array} \tag{4.3}$$

  in which $\widehat{f}$ is a solution to (4.2) and $\mathcal{F}_{\text{con}}$ is the set of all convex functions from Ω into $\mathbb{R}$.

Both optimization problems can be solved in $O(n \log n)$ operations, as will be seen soon. The overall computational complexity of the method described above is therefore $O(n \log n)$. The method is therefore faster than the constrained LS method applied to estimating f in (4.1) which requires $O(n^3)$ flops for each Newton iteration.

4.2.1 Monotonization

The first-stage problem is called isotonic or monotone regression, dating back to 1950s (Brunk, 1955; Brunk, 1958). A standard result in isotonic regression is that the solution to the minimization problem (4.2) is piecewise linear, and is given as the slope of the greatest convex minorant of the cumulative sum diagram of the points $\big(x_k,\ \sum_{i=1}^k y_i\big)$, $k = 1, 2, \ldots, n$. The slope is characterized as

$$\widehat{f}(x_k) = \max_{i\le k} \min_{j\ge k} P\big(i,j\big) = \min_{j\le k} \max_{i\ge k} P\big(i,j\big) = \min_{j\ge k} \max_{i\le k} P\big(i,j\big) \tag{4.4}$$

where

$$P(i,j) = \frac{1}{j-i+1}\sum_{k=i}^{j} y_k.$$

The solution $\widehat{f}$ can be characterized by the pooled adjacent violators algorithm with $O(n)$ operations; see, e.g., Barlow et al. (1972).

4.2.2 Convexification

To solve the convexification problem (4.3), we start by noticing that the solution $\widehat{f}$ to (4.2) is piecewise linear and continuous with n break points, say $x_1, \ldots, x_n$. Its convex envelope $\widehat{f}_{\mathbf{env}}$ is defined by the unique piecewise linear and continuous function that satisfies the interpolation condition:

$$\widehat{f}_{\mathbf{env}}(u_i) = z_i, \quad i = 1, \ldots, r,$$

where $\{(u_i, z_i) \mid i = 1, \ldots, r\}$ is the set of the vertices of the lower convex hull of the break points of $\widehat{f}$. The convex envelope can be found by the convex hull algorithm in $O(n \log n)$ operations; see, e.g., De Berg, Schwarzkopf, Van Kreveld and Overmars (2000) for more on the convex hull algorithm.

From the convex envelope, we can find a solution to the convexification problem (4.3), using the following algorithm.

Convexification algorithm

1. Find the lower envelope $\widehat{f}_{\mathbf{env}}$ of $\widehat{f}$.
2. Find the points x_l and x_r that satisfy

$$\widehat{f}_{\mathbf{env}}(x_l) + \|\widehat{f} - \widehat{f}_{\mathbf{env}}\|_{\infty,\Omega}/2 = f(x_l), \quad \widehat{f}_{\mathbf{env}}(x) + \|\widehat{f} - \widehat{f}_{\mathbf{env}}\|_{\infty,\Omega}/2 > \widehat{f}(x),\ x \le x_l,$$
$$\widehat{f}_{\mathbf{env}}(x_r) + \|\widehat{f} - \widehat{f}_{\mathbf{env}}\|_{\infty,\Omega}/2 = \widehat{f}(x_r), \quad \widehat{f}_{\mathbf{env}}(x) + \|\widehat{f} - \widehat{f}_{\mathbf{env}}\|_{\infty,\Omega}/2 > \widehat{f}(x),\ x \ge x_r.$$

3. Of the break points of $\widehat{f}_{\mathbf{env}}$, find the left adjacent point $\tilde{x}_l$ of x_l and the right adjacent point of $\tilde{x}_r$.

4. Construct the function $f^\star : \Omega \to \mathbb{R}$ by

$$f^\star(x) = \begin{cases} \max\{s_l(x), \widehat{f}_{\mathbf{env}}(x)\} & \text{if } x \le x_l, \\ \widehat{f}_{\mathbf{env}}(x) + \|\widehat{f} - \widehat{f}_{\mathbf{env}}\|_{\infty,\Omega}/2 & \text{if } x_l \le x \le x_r, \\ \max\{s_r(x), \widehat{f}_{\mathbf{env}}(x)\} & \text{if } x \ge x_r, \end{cases}$$

in which s_l is the affine function that passes through the two points

$$(x_l, \widehat{f}_{\mathbf{env}}(x_l) + \|\widehat{f} - \widehat{f}_{\mathbf{env}}\|_{\infty,\Omega}/2), \qquad (\tilde{x}_l, \widehat{f}_{\mathbf{env}}(\tilde{x}_l) + \|\widehat{f} - \widehat{f}_{\mathbf{env}}\|_{\infty,\Omega}/2)$$

and s_r is the affine function that passes through the two points

$$(x_r, \widehat{f}_{\mathbf{env}}(x_r) + \|\widehat{f} - \widehat{f}_{\mathbf{env}}\|_{\infty,\Omega}/2), \qquad (\tilde{x}_r, \widehat{f}_{\mathbf{env}}(\tilde{x}_r) + \|\widehat{f} - \widehat{f}_{\mathbf{env}}\|_{\infty,\Omega}/2).$$

The function $f^\star$ generated by the algorithm above, which is our estimate of the true function f in (4.1), is piecewise linear and completely characterized in $O(n \log n)$ operations, since we can find the break points of $\widehat{f}_{\mathbf{env}}$ in $O(n \log n)$ operations. The reader is referred to Kim and Lim (2006) for more on the algorithm described above.

Another important property of $f^\star$ is given in the following lemma.

Lemma 1. *Let $f^\star$ be the function generated by the boundary correction algorithm described above. Then,*

$$\sup_{x\in\Omega} |f^\star(x) - f(x)| \le \sup_{x\in\Omega} |\widehat{f}(x) - f(x)|.$$

The proof is given in Kim and Lim (2006).

4.3 Uniform convergence rate

To support theoretically the method described above, we analyze the uniform convergence rate achieved by the method described above. The analysis is based on Lemma 1. This lemma tells us that the uniform approximation error of the second stage estimate at the second stage is always smaller than or equal to that of the first

stage estimate. As a result, the monotone convex estimate converges uniformly to the true function at least as fast as the monotone regression estimate ignoring the convexity constraint. We therefore focus on establishing the uniform convergence rate of the first-stage estimate, i.e., the monotone regression estimate.

So far, many researchers have paid attention to convergence rate analysis of monotone regression estimates. Pointwise consistency of the monotone regression estimate was proved by Hanson, Pledger, and Wright (1973), Hanson and Pledger (1976), Mukerjee (1988), and Mammen (1991). The convergence rate at a fixed point was shown to be $n^{-1/3}$ by Brunk (1958) and Wright (1981). Durot (2002) recently proved the $\mathcal{L}_1$ convergence of the estimate beyond the pointwise convergence. However, the uniform convergence rate of the monotone estimate, which is our main focus, has received little interest in previous literature.

To carry out uniform convergence rate analysis for isotonic regression, we need to introduce some technical assumptions. We assume without loss of generality that Ω is $[0,1]$ and let Ω_δ denote the interval $[\delta, 1-\delta]$ with $\delta \in (0, 0.5)$. Suppose that the first derivative of the true function f is bounded and positive:

$$0 < a < \sup_{x \in \Omega_\delta} f'(x) < b$$

where a and b are positive numbers. Let $\mathcal{P}_n = \{\mathrm{I}_{1,n}, \mathrm{I}_{2,n}, \ldots, \mathrm{I}_{k_n,n}\}$ be a partition of the interval Ω_δ. Let the length of the interval $\mathrm{I}_{k,n}$ be an order of $n^{-1/3}$ and let $\mathbf{c}_k$ be the center of the kth interval. We further assume that $\mathbb{P}\big(X \in \mathrm{I}_{k,n}\big) \approx n^{-1/3}$, i.e., k_n is taken as $k_n \approx n^{1/3}$.

Theorem 4.3.1. Suppose that the true function f has positive bounded derivative on Ω; i.e., $a < \inf_{x\in\Omega} f'(x) \le \sup_{x\in\Omega} f'(x) < b$ for some positive a and b. Let $\widehat{f}_n$ be the monotone regression estimate of f and $\mathbf{T}_n$ be its uniform error defined by

$$\mathbf{T}_n = \sup_{x\in\Omega} \Big| \widehat{f}_n(x) - f(x) \Big|. \tag{4.5}$$

Suppose ϵ in (4.1) satisfy the moment conditions for Bernstein's inequality: (i) $\mathrm{E}\big(\epsilon\big) = 0$, and (ii), for some $c > 0$, $\mathrm{E}\big|\epsilon\big|^k \leq vk!c^{k-2}/2$ for all $k \geq 2$ and a constant v. Then,

(i) $\big(n^{1/3}\big/\log n\big)\mathbf{T}_n$ converges to 0 almost surely, and

(ii) for any $\epsilon > 0$, $\mathbb{P}\big(n^{1/3}\ \mathbf{T}_n > \epsilon\big) > 0$.

Proof. We start with recalling a sub-additive decomposition of $\mathbf{T}_n = \sup_{x\in\Omega_\delta} \big|\widehat{f}_n(x) - f(x)\big|$:

$$\mathbf{T}_n \leq \max_{k=2}^{n} \Big\{ \big|\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\big|, \big|\widehat{f}_n(\mathbf{c}_{k-1}) - f(\mathbf{c}_{k-1})\big| \Big\} + \max_{k=2}^{n} \Big| f(\mathbf{c}_k) - f(\mathbf{c}_{k-1}) \Big|. \quad (4.6)$$

Therefore, it suffices to show that

$$\frac{n^{1/3}}{\log n} \max_{k=1}^{n} \big|\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\big| \quad \text{and} \quad \frac{n^{1/3}}{\log n} \max_{k=2}^{n} \big|f(\mathbf{c}_k) - f(\mathbf{c}_{k-1})\big|$$

converge to 0 almost surely.

We define, in Section 4.3, $\mathbf{c}_k$ as the center of the kth interval in a partition of the sample space Ω_δ. Then, for every $x \in (\mathbf{c}_{k-1},\ \mathbf{c}_k]$,

$$\widehat{f}_n(\mathbf{c}_{k-1}) \leq \widehat{f}_n(x) \leq \widehat{f}_n(\mathbf{c}_k), \qquad f(\mathbf{c}_{k-1}) \leq f(x) \leq f(\mathbf{c}_k),$$

and

$$\widehat{f}_n(\mathbf{c}_{k-1}) - f(\mathbf{c}_k) \leq \widehat{f}_n(x) - f(x) \leq \widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_{k-1}).$$

Thus,

$$\begin{aligned} \widehat{f}_n(\mathbf{c}_{k-1}) - f(\mathbf{c}_{k-1}) - (f(\mathbf{c}_k) - f(\mathbf{c}_{k-1})) &\leq \widehat{f}_n(x) - f(x) \\ &\leq \widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k) + (f(\mathbf{c}_k) - f(\mathbf{c}_{k-1})), \end{aligned}$$

which implies that

$$\Big|\widehat{f}_n(x) - f(x) \leq \max\Big\{ \big|\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\big|, \big|\widehat{f}_n(\mathbf{c}_{k-1}) - f(\mathbf{c}_{k-1})\big| \Big\} + \Big|f(\mathbf{c}_k) - f(\mathbf{c}_{k-1})\Big|,$$

for every $x \in [\mathbf{c}_{k-1}, \mathbf{c}_k)$. Putting these inequalities together, we have

$$\begin{aligned}
&\sup_{x\in\Omega_\delta} \Big|\widehat{f}_n(x) - f(x)\Big| \\
&\quad\leq \max_{k=2}^{n} \Big\{ \big|\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\big|, \big|\widehat{f}_n(\mathbf{c}_{k-1}) - f(\mathbf{c}_{k-1})\big| \Big\} + \max_{k=2}^{n} \Big| f(\mathbf{c}_k) - f(\mathbf{c}_{k-1})\Big|.
\end{aligned}$$

Now we show that both terms in (4.6) converges to 0 almost surely. From the Borel-Cantelli lemma (Billingsley, 1995), we know that to show the almost sure convergence of the first term $\big(n^{1/3}/\log n\big)\ \max_k \big|\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\big|$ to 0, it suffices to prove that, for some constant $C > 0$,

$$\mathbb{P}\Big(\max_{k=1}^{k_n} n^{1/3} \Big|\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\Big| > C \log n, \text{ infinitely often} \Big) = 0. \tag{4.7}$$

The second term $\big(n^{1/3}/\log n\big)\ \max_k \big|f(\mathbf{c}_k) - f(\mathbf{c}_{k-1})\big|$ is a deterministic sequence, and it converges to 0 since $f'(x)$ is bounded on Ω.

We finish the proof of the claim by showing (4.7). Note that this is equivalent to showing that both

$$\sum_{k=1}^{k_n} \mathbb{P}\Big(n^{1/3}\big(\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\big) > C \log n \Big) \tag{4.8}$$

and

$$\sum_{k=1}^{k_n} \mathbb{P}\Big(n^{1/3}\big(\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\big) < -C \log n \Big) \tag{4.9}$$

have finite sums, since

$$\mathbb{P}\big(\max_{k=1}^{k_n} n^{1/3} \Big|\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\Big| > C \log n \big) \leq \sum_{k=1}^{k_n} \mathbb{P}\big(n^{1/3} \big|\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\big| > C \log n \big). \tag{4.10}$$

In the sequel, we show that (4.8) has a finite sum with respect to n. The proof of (4.9) is very similar to (4.8) and is omitted.

Let $N_n(a, b)$ be the number of the data points whose x values are in the interval $[a, b]$. Let u be a point such that

$$\mathbf{c}_k \leq u \leq \mathbf{c}_{k+1}, \quad N_n(\mathbf{c}_k, u) \approx \frac{n^{2/3}}{\log n}, \quad \text{and} \quad f(u) - f(\mathbf{c}_k) \approx \frac{n^{-1/3}}{\log n}.$$

Note that such u exists since $\mathbf{c}_{k+1} - \mathbf{c}_k \approx n^{-1/3}$ and $\mathbb{P}\big(X \in [\mathbf{c}_k, \mathbf{c}_{k+1}]\big) \approx n^{-1/3}$. Then,

$$\begin{aligned}
\widehat{f}_n(\mathbf{c}_k) &= \max_{\alpha \le \mathbf{c}_k} \min_{\mathbf{c}_k \le \beta} \frac{1}{N_n(\alpha,\beta)} \sum_{\{i:\alpha \le x_i \le \beta\}} y_i \\
&\le \max_{\alpha \le \mathbf{c}_k} \frac{1}{N_n(\alpha,u)} \sum_{\{i:\alpha \le x_i \le u\}} y_i \\
&= f(\mathbf{c}_j) + \big(f(u) - f(\mathbf{c}_j)\big) + \max_{\alpha \le \mathbf{c}_j} \frac{1}{N_n(\alpha,u)} \sum_{\{i:\alpha \le x_i \le u\}} \big(y_i - f(x_i)\big),
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\bigg(f(\mathbf{c}_k) - \widehat{f}_n(\mathbf{c}_k) > \frac{C\log n}{n^{1/3}}\bigg) \\
&\quad\le \mathbb{P}\left(\max_{\alpha \le \mathbf{c}_k} \frac{\sum_{\{i:\alpha \le x_i \le \beta\}} \big(y_i - f(x_i)\big)}{\sqrt{N_n(\alpha,u)}} > \frac{C\log n}{n^{1/3}} \sqrt{N_n(\alpha,u)}\right) \\
&\quad\le \sum_{\alpha \le \mathbf{c}_k} \mathbb{P}\left(\frac{\sum_{\{i:\alpha \le x_i \le \beta\}} \big(y_i - f(x_i)\big)}{\sqrt{N_n(\alpha,u)}} > \frac{C\log n}{n^{1/3}} \sqrt{N_n(\alpha,u)}\right). \qquad (4.11)
\end{aligned}$$

Here, the last term (4.11) is bounded using the Bernstein's inequality (p.855 Shorack and Wellner (1986)) with the moment conditions in the theorem. To be specific, for some K, we have

$$\begin{aligned}
&\sum_{\alpha \le \mathbf{c}_k} \mathbb{P}\left(\frac{\sum_{\{i:\alpha \le x_i \le \beta\}} \big(y_i - f(x_i)\big)}{\sqrt{N_n(\alpha,u)}} > \frac{C\log n}{n^{1/3}} \sqrt{N_n(\alpha,u)}\right) \\
&\quad\le\ n \exp\bigg(-K\frac{(\log n)^2}{n^{2/3}} N_n(\alpha,u)\bigg) \\
&\quad\le\ n\exp\big(-K\log n\big). \qquad (4.12)
\end{aligned}$$

Therefore,

$$\sum_{n=1}^{\infty}\sum_{j=1}^{k_n} \mathbb{P}\Big(\widehat{f}_n(\mathbf{c}_j) - f(\mathbf{c}_j) > \frac{C\log n}{n^{1/3}}\Big) \le n^2 \exp\Big(-K\log n\Big) \le \exp\Big(-K'\log n\Big)$$

for sufficiently large C (or equivalently sufficiently large K and K').

Proof of (ii)

Now we show the second part of the theorem that $\mathbb{P}\big(n^{1/3}\ \mathbf{T}_n > \epsilon\big) > 0$. To see this, first note that

$$\begin{aligned}\mathbb{P}\Big(n^{1/3}\ \mathbf{T}_n > \epsilon\Big) &\geq \mathbb{P}\Big(\max_{k=1}^{k_n}\ n^{1/3}\ \big\{\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\big\} > \epsilon\Big) \\ &\geq \mathbb{P}\Big(n^{1/3}\ \big(\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\big) > \epsilon\Big). \qquad (4.13)\end{aligned}$$

It is shown by Wright (1981) that

$$n^{1/3}\Big(\widehat{f}_n(\mathbf{c}_k) - f(\mathbf{c}_k)\Big) \text{ converges in distribution to } \mathbf{b}_k^n\mathbf{Z}, \qquad (4.14)$$

in which b_k^n is a bounded function of σ^2 and $f(\mathbf{c}_k)$, and $\mathbf{Z}$ is the slope of the greatest convex minorant of $\frac{1}{2}\big(W(t) + t^2\big)$ with the two-sided Brownian motion $W(t)$. In particular, the asymptotic distribution of $\mathbf{Z}$ is shown to be

$$f_Z(z) \sim K|z| \exp\Big(-\frac{2}{3}|z|^2 + 2^{-1/3}a_1|z|\Big), \quad \text{as} \quad |z| \to \infty, \qquad (4.15)$$

in which $a_1 \approx -2.3381$ (see Chernoff (1964) and Groeneboom and Wellner (2002)). Thus, the right-hand side of the last inequality in (4.13) is bounded away from 0, that is, the assertion of this theorem holds. □

The preceding theorem along with Lemma 1 implies that a lower bound on the convergence rate of the estimate via the proposed sequential method is $O_p\big(n^{-1/3}\log n\big)$. This rate is slightly slower than $O_p\big(n^{-1/3}(\log n)^{1/3}\big)$, the previously known best convergence rate which a nonparametric estimation method can achieve (Stone, 1982; p. 244 in Rao, 1983). We should emphasize that the nonparametric estimate is not necessarily monotone convex. We conclude that when the first derivative of the true function f is bounded, the uniform convergence rate achieved by the sequential estimation method is nearly comparable to the previously known best uniform convergence rate of a nonparametric estimate.

4.4 Numerical examples

We carry out a simulation study to examine the finite sample performance of the proposed sequential method and compare it with the constrained LS. In the simulation study, the design points $x_1, \ldots, x_n$ are taken as $x_i = i/n$ and the response data y_i are generated from the model (4.1). We consider three true functions: $f(x) = 1$ (neither strictly monotone nor strictly convex), $f(x) = x$ (strictly monotone but not strictly convex), $f(x) = x^2$ (strictly monotone and convex) on $\Omega = [0, 1]$. Two error distributions with mean 0 and variance σ^2 are considered for ϵ_i: the Gaussian and the double exponential distribution. The double exponential distribution has the form $f(x) = \big(1/2\sigma\big) \exp\big(- |x|/\sigma\big)$. For each pair of the true function and the error distribution, we generate 100 data sets and apply the proposed sequential method and the constrained LS to these data sets. We solve the constrained LS using the MOSEK software package (MOSEK ApS, 2002). In each of 100 data sets, we compute the uniform error, $\max_{i=1}^n \big|\widehat{f}_n(x_i) - f\big(x_i\big)\big|$, and the integrated mean squared error (IMSE), $\big(1/n\big) \sum_{i=1}^n \big|\widehat{f}_n(x_i) - f\big(x_i\big)\big|^2$.

Tables 1 and 2 summarize the simulation results for $n = 20$ and 50 for three noise level $\sigma = 0.1, 0.2$ and 0.3. The tables show that, in most cases, the performance of the proposed method is comparable to that of the constrained LS. Moreover, the proposed method is much faster than the constrained LS solved that relies on an efficient quadratic programming solver.

Table 13: Summary of simulation results for $n = 20$.

$f(x) = 1$					
		sequential estimation		constrained LS	
σ	error distribution	uniform error	IMSE	uniform error	IMSE
0.1	Gaussian	0.0921	0.0016	0.0653	0.0009
	double exponential	0.1266	0.0031	0.0856	0.0016
0.2	Gaussian	0.1964	0.0067	0.1331	0.0031
	double exponential	0.2284	0.0111	0.1438	0.0059
0.3	Gaussian	0.2473	0.0126	0.1665	0.0072
	double exponential	0.4113	0.0269	0.2916	0.0130
$f(x) = x$					
		sequential estimation		constrained LS	
σ	error distribution	uniform error	IMSE	uniform error	IMSE
0.1	Gaussian	0.1459	0.0022	0.2403	0.0338
	double exponential	0.1756	0.0040	0.2482	0.0349
0.2	Gaussian	0.2667	0.0075	0.2524	0.0341
	double exponential	0.2943	0.0124	0.2604	0.0377
0.3	Gaussian	0.3207	0.0145	0.2858	0.0414
	double exponential	0.4939	0.0290	0.3852	0.0489
$f(x) = x^2$					
		sequential estimation		Constrained LS	
σ	error distribution	uniform error	IMSE	uniform error	IMSE
0.1	Gaussian	0.1267	0.0032	0.1624	0.0110
	double exponential	0.1538	0.0052	0.1810	0.0117
0.2	Gaussian	0.2338	0.0083	0.2267	0.0155
	double exponential	0.2529	0.0134	0.2418	0.0186
0.3	Gaussian	0.2893	0.0161	0.2723	0.0209
	double exponential	0.4645	0.0308	0.3903	0.0302

Table 14: Summary of simulation results for $n = 50$.

$f(x) = 1$					
		sequential estimation		constrained LS	
σ	error distribution	uniform error	IMSE	uniform error	IMSE
	Gaussian	0.0735	0.0012	0.0446	0.0002
0.1	double exponential	0.1077	0.0027	0.0027	0.0005
	Gaussian	0.1870	0.0051	0.1417	0.0014
0.2	double exponential	0.2191	0.0125	0.1418	0.0029
	Gaussian	0.2546	0.0130	0.1757	0.0029
0.3	double exponential	0.3097	0.0228	0.2183	0.0064
$f(x) = x$					
		sequential estimation		constrained LS	
σ	error distribution	uniform error	IMSE	uniform error	IMSE
	Gaussian	0.1636	0.0047	0.2406	0.0316
0.1	double exponential	0.2030	0.0076	0.2459	0.0324
	Gaussian	0.2881	0.0124	0.2580	0.0327
0.2	double exponential	0.3319	0.0221	0.2610	0.0315
	Gaussian	0.3578	0.0240	0.2810	0.034
0.3	double exponential	0.4402	0.0370	0.3239	0.0387
$f(x) = x^2$					
		sequential estimation		constrained LS	
σ	error distribution	uniform error	IMSE	uniform error	IMSE
	Gaussian	0.1464	0.0032	0.1542	0.0096
0.1	double exponential	0.1925	0.0054	0.1709	0.0098
	Gaussian	0.2685	0.0093	0.2125	0.0113
0.2	double exponential	0.3221	0.0181	0.2352	0.0149
	Gaussian	0.3430	0.0192	0.2525	0.0142
0.3	double exponential	0.4254	0.0319	0.3067	0.0197

CHAPTER V

CONCLUSIONS

We have proposed a dimension reduction procedure to test the significance of whether a mean curve dominates another one. The key idea of the suggested method relies on preserving the order in mean while reducing the dimension of the data. We have made use of a novel dimension reduction procedure that preserves the order between the two curves. We have then applied a multivariate testing procedure to the coefficient vectors that represent the data matrix in a lower dimension. In a addition, we have proposed an iterative algorithm to solve the projection problem.

Our work was motivated by a study that looks at high-dimensional, high-frequency measurements of tidal volume on a number of individuals subject to interventions that may induce panic attacks. Our results have confirmed the hypothesis of the investigators according to which subjects receiving sodium lactate after naloxone have greater increases in tidal volume than subjects that do not receive the prior infusion of naloxone.

The initial hypothesis was that impairing normal subjects' endogenous opioidergic system by naloxone (N) should make them vulnerable to the panicogenic effects of subsequent lactate (L). The ultimate goal is to prove that an opioidergic dysfunction may be the pathophysiological mechanism underlying panic disorder.

For the initial study on NL, SL and NS, functional ANOVA-type testing procedures applied to the Vt traces during first and second infusion have led to the discovery that the lactate group N+L, in particular, is significantly different from N+S during the time of second infusion. Also, a steady increase of mean tidal volume was observed during both lactate infusions while it was absent in the N+S group.

A test with an order restriction showed that the mean N+L curve is always higher than or equal to the mean S+L curve during lactate. Finally, a permutation test revealed a significant difference among the two lactate groups during the last minutes of the second infusion.

The follow-up study on SL and NL confirmed the steady increase of mean tidal volume during lactate infusion. The permutation test on all 38+27 subjects confirmed that there is a significant difference among the two lactate groups during the last minutes of the second infusion. This difference was observed also in a cross-over study where tests were carried out on the differences of Vt. Our conclusions are that during the lactate infusion there is a significant effect over time which manifests itself as a steady increase of the tidal volume. In addition, when preceded by naloxone this differential increase lasts significantly longer in time. If these findings are replicated and extended, the next step would be to see if this naloxone-lactate interaction could be blocked by specific anti-panic agents in a controlled study.

In Chapter IV, we have described a computationally efficient two-stage procedure for estimating monotone convex functions, based on $\mathcal{L}_2$ monotonization and uniform convexification. The monotonization problem at the first stage has an analytic solution that can be characterized in $O(n)$ operations by the pool adjacent violators algorithm. The convexification problem at the second stage can be solved in $O(n \log n)$ operations by means of the convex hull algorithm. This computational merit is the main motivation of using the uniform norm instead of the $\mathcal{L}_2$-norm in convexifying the first-stage estimate. The proposed method is much faster than the constrained LS and performs as well as the constrained LS.

The method is similar in spirit to the two-stage estimation method described in Kim and Lim (2006), which can handle general shape constraints. The method consists of nonparametric function estimation without taking into account the shape

constraint and shape modification of the nonparametric estimate by solving a constrained uniform approximation problem. The uniform convergence rate of this general method is determined by the first-stage nonparametric one. On the other hand, that of the method proposed in this paper is nearly comparable to that of the best first-stage nonparametric estimation method.

REFERENCES

Aït-Sahalia, Y. and Duarte, J. (2003). "Nonparametric Option Pricing under Shape Restrictions." *Journal of Econometrics*, 116, 9–47.

Akiyama, Y., Nishimura, M., Kobayashi, S., Yoshioka, A., Yamamoto, M., Miyamoto, K., and Kawakami, Y. (1993). "Effects of Naloxone on the Sensation of Dyspnea during Acute Respiratory Stress in Normal Adults." *Journal of Applied Physiology*, 74, 590–595.

Barlow, R., Bartholomew, D., Bremner, J., and Brunk, H. (1972). *Statistical Inference under Order Restrictions*. New York: John Wiley and Sons.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge, UK: Cambridge Unviersity Press.

Brunk, H. (1955). "Maximum Likelihood Estimates of Monotone Parameters." *The Annals of Mathematical Statistics*, 26, 607–616.

— (1958). "On the Estimation of Parameters Restricted by Inequalities." *The Annals of Mathematical Statistics*, 29, 437–454.

De Berg, M., Schwarzkopf, O., Van Kreveld, M., and Overmars, M. (2000). *Computational Geometry: Algorithms and Applications*. New York: Springer-Verlag.

Durot, C. (2002). "Sharp Asymptotics for Isotonic Regression." *Probability Theory and Related Fields*, 122, 222–240.

Fan, J. (1996). "Tests of Significance Based on Wavelet Thresholding and Neyman's Truncation." *Journal of American Statistical Association*, 91, 674–699.

Fleetham, J., Clarke, H., Dhingra, S., Chernick, V., and Anthiosen, N. (1980). "Endogenous Opiates and Chemical Control of Breathing in Humans." *Am Rev Respiratory Disease*, 121, 1045–1049.

Follmann, D. (1996). "A Simple Multivariate Test for One-sided Alternatives." *Journal of American Statistical Association*, 91, 854–861.

Goetz, R., Klein, D., Gully, D., Kahn, J., Liebowitz, M., Fyer, A., and Gorman, J. (1993). "Panic Attacks during Placebo Procedures in the Laboratory: Phsyiology and Symptomatology." *Arch Gen Psychiatry*, 50, 280–285.

Hanson, D. and Pledger, G. (1976). "Consistency in Concave Regression." *Annals of Statistics*, 4, 1038–1050.

Hanson, D., Pledger, G., and Wright, I. (1973). "On Consistency in Monotonic Regression." *Annals of Statistics*, 1, 401–421.

Kim, S. and Lim, J. (2006). "Uniform Approximation and Estimation of Shape Restricted Functions, Preprint." Available from www.stanford.edu/∼sjkim/Papers/shape_con_est.pdf.

Klein, D. (1993). "False Suffocation Alarms, Spontaneous Panics and Related Conditions: An Integrative Hypothesis." *Arch Gen Psychiatry*, 50, 306–317.

Lee, D. and Seung, H. (1999). "Learning the Parts of Objects by Non-negative Matrix Factorization." *Nature*, 401, 788–793.

Liebowitz, M., Gorman, J., Fyer, A., Levitt, M., Dillon, D., Levy, G., Appleby, I., Anderson, S., Palij, M., Davies, S., and Klein, D. (1985). "Lactate Provocation of Panic Attacks: Clinical and Behavioral Findings." *Arch Gen Psychiatry*, 41, 764–770.

Lloyd, C. (2002). "Estimation of a Convex ROC Curve." *Statistics and Probability Letters*, 59, 99–111.

Mallat, S. (1989). "Multiresolution Approximations and Wavelet Orthonormal Bases of $L_2(R)$." *Transactions of the American Mathematical Society*, 351(1), 69–87.

Mammen, E. (1991). "Nonparametric Regression under Qualitative Smoothness Assumption." *Annals of Statistics*, 19, 741–759.

Mammen, E., Marron, J., Turlach, B., and Wand, M. (2001). "A General Framework for Constrained Smoothing." *Statistical Science*, 16, 232–248.

Matzkin, R. (1994). "Restrictions of Economic Theory in Nonparametric Methods." In *Handbook of Econometrics*, eds. R, Engle and D, McFadden. Amsterdam, Netherlands: North-Holland.

MOSEK-ApS (2002). *The MOSEK Optimization Tools Version 2.5. User's Manual and Reference*. Available from www.mosek.com.

Mukerjee, H. (1988). "Monotone Nonparametric Regression." *Annals of Statistics*, 16, 741–750.

O'Brien, P. (1984). "Procedures for Comparing Samples with Multiple Endpoints." *Biometrics*, 40, 1079–1087.

Percival, D. and Walden, A. (2002). *Wavelet Methods for Time Series Analysis*. Cambridge, UK: Cambridge University Press.

Pocock, S., Geller, N., and Tsiatis, A. (1987). "The Analysis of Multiple Endpoints in Clinical Trials." *Biometrics*, 43, 487–498.

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. New York: Springer.

Rao, P. (1983). *Nonparametric Functional Estimation*. New York: Academic Press.

Robertson, T., Wright, F., and Dykstra, R. (1988). *Order Restricted Statistical Inference*. New York: John Wiley and Sons.

Serban, N. and Wasserman, L. (2005). "CATS: Clustering after Transformation and Smoothing." *Journal of American Statistical Association*, 100, 990–999.

Silvapulle, M. and Sen, P. (2005). *Constrained Statistical Inference*. Hoboken, NJ: Wiley.

Sinha, S., Goetz, R., and Klein, D. (2005). "Physiological and Behavioral Effects of Naloxone and Lactate in Normals with Relevance to the Pathophysiology of Panic Disorder." *Psychiatry Research*, in press.

Stone, C. (1982). "Optimal Global Rates of Convergence for Nonparametric Regression." *Annals of Statistics*, 10, 1040–1053.

Tang, D.-I., Gnecco, C., and Geller, N. (1989). "An Approximate Likelihood Ratio Test for a Normal Mean Vector with Nonnegative Components with Application to Clinical Trials." *Biometrika*, 76, 577–583.

Van de Panne, C. (1974). *Methods for Linear and Quadratic Programming*. Amsterdam, Netherlands: North-Holland Publishing Company.

Wilhelm, F., Roth, W., and Sackner, M. (2003). "The LifeShirt: An Advanced System for Ambulatory Measurement of Respiratory and Cardiac Function." *Behav Modif*, 27(5), 671–691.

Wright, I. (1981). "The Asymptotic Behavior of Monotone Regression Estimates." *Annals of Statistics*, 9, 443–448.

VITA

Sang Han Lee, son of Jung-Woo Lee and Inn-Sook Yoon, was born in Masan, Korea. He received a Bachelor of Science degree in statistics from Seoul National University in Seoul, Korea in 1999. He received a Master of Science degree in statistics from the same university under the direction of Dr. Woo-Chul Kim in 2001. He continued his studies in statistics under the direction of Dr. Marina Vannucci and received a Doctor of Philosophy degree in statistics from Texas A&M University in College Station, Texas, in August 2007. His address is Department of Statistics, Texas A& M University, 3143 TAMU, College Station, TX 77843.