On the Statistical Analysis of Functional Data Arising from Designed Experiments

by

Monica Sirski

A Thesis submitted to the Faculty of Graduate Studies
of the University of Manitoba
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Department of Statistics
University of Manitoba
Winnipeg, Manitoba

Abstract

We investigate various methods for testing whether two groups of curves are statistically significantly different, with the motivation to apply the techniques to the analysis of data arising from designed experiments. We propose a set of tests based on pairwise differences between individual curves. Our objective is to compare the power and robustness of a variety of tests, including a collection of permutation tests, a test based on the functional principal components scores, the adaptive Neyman test and the functional F test. We illustrate the application of these tests in the context of a designed 2^4 factorial experiment with a case study using data provided by NASA. We apply the methods for comparing curves to this factorial data by dividing the data into two groups by each effect (A, B, ..., ABCD) in turn. We carry out a large simulation study investigating the power of the tests in detecting contamination, location, and shift effects on unimodal and monotone curves. We conclude that the permutation test using the mean of the pairwise differences in L_1 norm has the best overall power performance and is a robust test statistic applicable in a wide variety of situations. The advantage of using a permutation test is that it is an exact, distribution-free test that performs well overall when applied to functional data. This test may be extended to more than two groups by constructing test statistics based on averages of pairwise differences between curves from the different groups and, as such, is an important building-block for larger experiments and more complex designs.

Acknowledgments

If it takes a village to raise a child, it also takes a network of support to complete a Ph.D. I would first like to thank my advisor Dr. John Brewster, for his patience, guidance, and support over the years. I would also like to thank the other members of my committee: Dr. Alexandre LeBlanc, Dr. Lisa Lix and Dr. Robert McLeod, for encouraging me and answering questions and also for challenging me when needed, thereby making my dissertation a better work. Thank you also to my external examiner, Dr. G. Geoffrey Vining, for helpful feedback regarding this dissertation. Thanks to several of you for travelling to my Ph.D. oral defense.

Thank you to Dr. Peter Parker, of the United States' National Aeronautics and Space Administration (NASA), who provided the data that inspired this research in the first place. Thank you also to Dave Gabrielson and Alan Forbes for their help with the many subtleties of Xgrid.

Thanks also go to my parents, especially my mom, who have looked after my children for many, many, many hours as I escaped to the library to work. And to my husband, words cannot express how important you have been to me over the years. I could not have accomplished all of this without your support.

Funding for this project has been provided by: a PGS D from the Natural Sciences and Engineering Research Council of Canada (NSERC), a University of Manitoba Graduate Fellowship, Dr. John Brewster and Dr. Robert McLeod through NSERC Discovery Grants, and the University of Manitoba Department of Statistics. I am very grateful for all of the support I have received.

Dedication

This work is dedicated to my family.

Nothing in the world is worth having or worth doing unless it means effort, pain, difficulty . . . I have never in my life envied a human being who led an easy life. I have envied a great many people who led difficult lives and led them well.

— Theodore Roosevelt

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Data where the response is a curve is becoming more prevalent with the advent of sophisticated data collection techniques and increased computing power. Such data are often called *functional data*. Corresponding to the increased availability of functional data are questions regarding the analysis of data of this form. This dissertation is a comprehensive exploration of various approaches to analyzing functional data, focusing on methods for testing for significant differences between two groups of curves. The emphasis is on data generated from designed experiments, although the techniques used are in general more broadly applicable.

We have several goals in writing this dissertation. Our first goal is a comprehensive examination and synthesis of the current status of the literature in areas dealing with data that are curves. Second is the development of new tests appropriate for testing whether two groups of curves are different. Third is a simulation study which is the systematic comparison between various testing procedures, including the new tests we propose. Our final goal is the application of the testing procedures to a set of real data whereby we develop a strategy for analyzing functional data arising from 2^k and 2^{k-p} designed experiments, using both cross-sectional and functional approaches.

The problem of comparing two groups of curves is a fundamental question at the heart of more sophisticated techniques such as analysis of variance and regression. Comparing two groups is also the foundation of the analysis of data arising from 2^k factorial experiments or 2^{k-p} fractional factorial experiments, serving as building blocks for tests used in larger experiments and more complex designs. Thus, although the research in this dissertation focusses on the properties of test procedures for comparing two groups of curves, we also illustrate how the procedures can be applied in analyzing data from a 2^k experiment.

In the traditional case, where the response is scalar, we compare data points from two groups using a parametric approach, such as a t-test, or a nonparametric approach, such as a rank-sum test — or, with more than two groups, parametric or nonparametric analysis of variance techniques. These established techniques are included in many introductory statistics textbooks.

When the responses become curves, the situation becomes more complicated. This situation arises when a series of points is recorded over time, for example. There are several approaches to the analysis of such data. It is possible to analyze these data as a series of points and carry out a multivariate analysis or longitudinal data analysis. However, using multivariate methods may lead to issues if the dimension of the data is high, for example, if we have many points collected over time. As discussed in Faraway [1997, p. 256], multivariate statistics “become dominated by variation in unimportant directions” as the dimension of the data grows. Longitudinal data analysis techniques are designed for quite sparse data and the model structures are potentially quite restrictive. Another technique, the focus of this research, is to smooth the data (using a B-spline, for example) and treat the series of points as a curve. In this dissertation we limit ourselves to the case where the data are smooth curves. We do not concern ourselves with issues related to smoothing, the number of points (or the locations of points) at which observations have been made, or the possibility of missing data; these are problems to be addressed at a different time.

Denote the jth observation from the ith group as $Y_{ij}(t)$. We use t because these data are often collected over time, in which case the index is a natural one. This is not always the case, however, as in the data set that inspired this work. This data, displayed in Figure 1.1, was provided by Peter Parker of the United States' National Aeronautics and Space Administration (NASA). The NASA data were collected on a fairly coarse grid; in Figure 1.1, the discrete data are simply joined by straight lines to form "curves." The data in tabular form is located in Appendix A. This data arose from a full 2^4 screening computer experiment that was conducted by NASA to examine the effects and importance of four factors: Tower Length (A), Tower Diameter (B), Tip Fineness Ratio (C), and Tip Shape (D), on the response, Drag. The experiment was run at 10 different Mach numbers, in unequally-spaced intervals from Mach 0.7 to 4.0 — so, in this case, t represents Mach number, not time. Here, for each combination of factor levels there are 10 responses, one for each of the 10 Mach numbers. The question is how to test for significant effects using these curves. As a starting point, consider testing for a factor B effect by separating the curves into two groups based on the level of B. These curves are

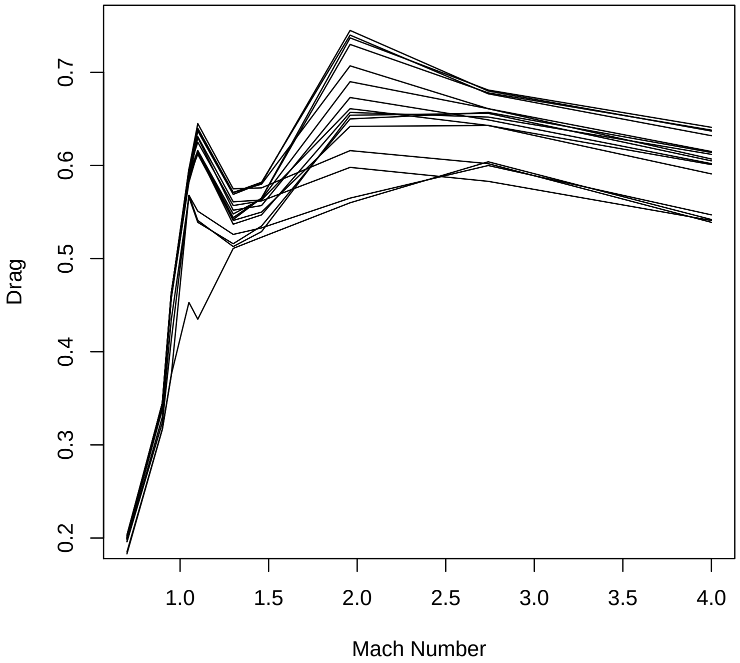

Figure 1.1: Response data as collected, a series of points.

plotted in Figure 1.2, with different line types indicating the two levels of factor B, Tower Diameter. This time, we have smoothed the curves using splines. This is the type of data that we focus on in this dissertation.

There are a number of approaches to comparing two sets of curves. One option is to run tests at each recorded time point along the curve. This is a fairly naïve solution to the problem, as it does not account for the functional nature of the data, and it suffers from a scale problem: for example, a data set with data recorded at 1000 time points would require carrying out 1000 t-tests and will suffer from the multiple comparisons problem. Instead, we could use procedures that treat the data as curves. A few parametric approaches have been developed, such as the functional F test [Shen and Xu, 2007]. Alternatively, some tests use dimension reduction to express the curves in a smaller number of coefficients, as in the adaptive Neyman

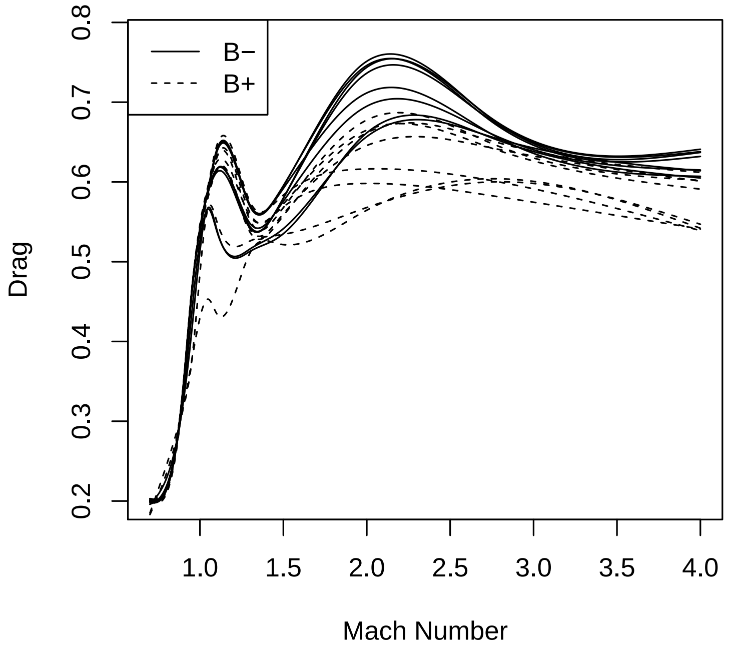

Figure 1.2: Response data as curves, grouped by factor B.

test [Fan, 1996a, Fan and Lin, 1998] which tests the coefficients of the Fourier transformation, or the method proposed by Sturino et al. [2010], which compares the functional principal components scores.

Another method is the permutation test, also called the randomization test, that dates back to the early years of statistics. This method was proposed by Pitman [1937a] (see also Pitman [1937b]). The permutation test is distribution-free; it uses the data to assess the probability that the observed result (or something more extreme) would be obtained under the null hypothesis. The drawback of the technique is that carrying out a permutation test is very calculation intensive and therefore has traditionally been limited in application to small data sets. Happily, with the increasing computing power that has yielded functional data has also come the ability to run permutation tests on larger and larger data sets, and as a result permutation tests have enjoyed a bit of a renaissance of late. Sturino et al. [2010] consider permutation tests based on summary curves (such as the mean curves) in the context of functional data, and Ramsay et al. [2009] use a permutation test as the basis for their t-test-like test.

In this dissertation we propose our solution to the problem of comparing two groups of curves — a collection of permutation tests that we have developed. These permutation tests are calculated in a pairwise fashion, that is, we calculate the distance between pairs of curves, one from each group, and base the test statistic on averages of these pairwise comparisons. The pairwise approach used with these new tests is in contrast to the permutation tests used by Sturino et al. [2010], which are based on comparisons of a summary measure calculated for each group. This idea of constructing test statistics based on pairwise comparisons is central to the methods that we propose. The concept of using pairwise comparisons in this context is not something that we have seen in the literature, although it does seem to be a natural thing to do. One difficulty with using tests based on summary curves is that the shapes of the summary curves may not resemble the individual curves at all. By comparing the curves at the individual level we avoid the risk of losing information about the differences between individual curves that may be masked by considering only differences between summary curves.

Having been presented with several options for testing whether two groups of curves are different, and adding our own technique as an option, we then ask: which test should we use? The answer to this question is not clear and is the motivation factor for the simulation study we conduct. The comparisons between testing procedures are limited since most articles in the literature consider only a few test statistics to compare at a time. Shen, in his thesis [Shen, 1999], used simulated data having various covariance structures to compare the power of the functional F test with a test proposed by Faraway [1997] and one discussed by Ramsay and Silverman [1997]. He found that the functional F test is rarely the most powerful test compared to the other two. In follow-up work [Shen and Faraway, 2004], the authors compared the F test to Faraway's bootstrapping test [1997], and two other tests. The functional F test has the best power of the statistics they consider but they find the F statistic to be sensitive to the error structure of the data, indicating that the functional F test may not be very robust. Simulations performed by Sturino et al. [2010] compare the permutation tests they define to the tests they developed based on principal components. The power of the tests depends on the structure of the simulation and the effect being tested. However no comparisons were made between the tests in the paper and other options.

Therefore, we see that no comprehensive comparison has been done between the many testing procedures. The lack of comparisons makes it difficult to determine which test to use under which conditions. In order to gain a better understanding of the nature of the various testing procedures, we carry out a simulation study in order to examine the behaviour of various tests, including our proposed pairwise-based testing procedures, to assess how powerful and robust the tests are under varying conditions. A robust statistic is one that is not sensitive to the testing conditions; it performs well in a variety of situations, for example, it is not sensitive to the shape of the underlying curves or the correlation structure of the data.

The rest of this dissertation is structured as follows: We begin with a synthesis of the literature in Chapter 2, briefly discussing terminology and presenting examples of areas of application where data are curves. We continue with a review of the literature relating to testing for significance between two groups of curves, and expand the discussion to the current status of the research relating to the three areas where data as curves are most prominent: profile analysis in statistical process control, functional data analysis, and longitudinal data analysis. This literature review is extensive and is presented in the hopes that by bringing together the three disciplines using data that are curves, there will be opportunities for cross-collaboration between the distinct areas, allowing them to learn from each other, thereby making each field stronger. Chapter 3 begins with the development of the pairwise permutation tests we have proposed, and continues with the presentation of the other testing procedures under consideration. We present a case study in Chapter 4, where we analyze the NASA data. We use this case study to demonstrate the development of a strategy for approaching the analysis of data that are curves. We first take a cross-sectional approach and show how these techniques can illuminate features of interest. We then move to a functional approach, applying the testing procedures discussed in the previous chapter. Here, we use the techniques designed for comparing curves to analyze the data, generated by a 2^4 factorial design. We carry out the tests exploiting the natural form of this data, by grouping the data into two groups according to the effect being tested. Following are the results of simulation studies that explore the accuracy, robustness, and power of the statistical procedures in Chapter 5. Finally, we conclude with Chapter 6, where we summarize the results, discuss the limitations of this work and outline some directions for future research.

CHAPTER

TWO

LITERATURE REVIEW

The structure of this chapter is as follows. First is an overview of the areas of application involving data where the response is a curve. Second is a review of the methods proposed to test whether two groups of curves are significantly different. The current status of the research in areas that commonly deal with data as curves is reviewed. Finally, we consider permutation tests as applied to data that are curves. We discuss how most of the permutation tests used in application are carried out using pointwise techniques and that there has been little comparison done between permutation tests and other functional data tests.

It is important to note that the term *functional data analysis* means different things to different people. For example, Cardot focuses on functional covariates with a scalar response (as examples, see Cardot et al. [2003], Cardot et al. [2004], and Cardot et al. [2007a]). The focus of this research is on the case where the response is a curve with scalar covariates.

One person who addressed the problem of how to compare two sets of curves some time ago was Rao [1958], who discusses how to compare groups of growth curves. He proposes several approaches to testing. A basic test is estimating the rate of growth and then comparing the mean rates between groups. He also defines a likelihood ratio test. Müller [2005] says "Rao (1958) developed preliminary ideas on functional principal components in applications to growth curves."

Other early papers in the field of functional data analysis include the following: Ramsay and Dalzell [1991], who divide functions into structural and residual components in an extension of linear regression to the functional case; Rice and Silverman [1991], who discuss how to estimate the mean and covariance of a set of curves; and Besse and Cardot [1996], who adapt principal components analysis for functional data.

Since these early papers, the area of functional data analysis has experienced an explosion of interest and the techniques that have been developed have been applied in a vast variety of situations. One area where data is treated as a curve is statistical process control, for example: mayonnaise product quality monitoring [Sahni et al., 2005]; wood products manufacturing [Staudhammer et al., 2005, 2007]; injection moulding, where the compression strength of foam is measured over different levels of compression in a robust design study [Nair et al., 2002]; tonnage signals in stamping, torque signals in tapping, and force signals in welding [Jin and Shi, 2001]; semiconductor manufacturing [Kang and Albin, 2000]; and automobile manufacturing [Lawless et al., 1999]. Another aspect of statistical process control is calibration, which has also used functional data techniques: Stover and Brill [1998] discuss calibration of chromatographs using ion chromatography calibration data, while Mestek et al. [1994] deal with calibration curves of the photometric determination of Fe^{3+} with sulfosalicylic acid.

The field of medicine is rich with functional data: growth curves (Gasser et al. [1985], James et al. [2000], among others); developmental toxicity studies in lab animals [Hall and Severini, 1998]; ^{14}C-folate in plasma of healthy adults [Yao et al., 2003]; periodically stimulated fetal heart rates [Ratcliffe et al., 2002b,a]; neuron firing in electrophysiological studies [Behseta et al., 2007]; monitoring gene expression using a microarray experiment [Li et al., 2002, López et al., 2004]; EEG data [Abramovich et al., 2004]; primary biliary cirrhosis studies [Müller, 2005]; tobacco-treatment clinical trials [Hall et al., 2001b, Yang et al., 2007]; density profiles of brain tissue (using chromatography) in the study of aging [Muñoz Maldonado et al., 2002]; medical imaging [Kelemen et al., 1997, Yushkevich et al., 2001]; opthamology [Loncatore et al., 1999]; and forecasting mortality and fertility rates [Hyndman and Ullah, 2007].

Finally, we mention a selection of other applications: radar range profiles [Hall et al., 2001a], the study of brightness of stars over time [Hall et al., 2000], ergonomics data studying driver movement in an automobile [Faraway, 1997, Shen and Faraway, 2004], profiles of atmospheric radioactivity [Hlubinka and Prchal, 2007], oceanology [Nerini and Ghattas, 2007], and the weight loss behaviour of garlic [Castano et al., 2006].

2.1 Background Information

The analysis of data that are curves falls under the purview of several areas of statistics, primarily profile analysis, which is a subdiscipline of statistical process control (SPC, also known as quality control), functional data analysis (FDA) and longitudinal data analysis (LDA). These three fields have developed separately, with different goals and foci, based on the nature of the data and the motivations of the disciplines.

Statistical process control has two separate goals depending on the situation, which we will briefly outline. See Woodall et al. [2004] for a more in-depth discussion. One goal is to use a group of curves to set control limits, the other goal is to use these established control limits to monitor a curve with data collected in real time and detect processes that have deviated from the control situation. The data under consideration in this field is usually fairly high frequency.

Functional data is also characterized in general by data collected regularly at a high frequency, while longitudinal data is usually more sparse and collected at irregular intervals [Rice, 2004]. Consequently, functional data analysis focuses more on dimension reduction. In addition, longitudinal data analysis is more model-based and inferential, while functional data analysis has a more exploratory and non-parametric point of view with a focus on describing the data (with principal components, smoothing, etc.). The nature of each subject is revealed by comparing the classic FDA text, Ramsay and Silverman [2005], with the classic LDA text, Diggle et al. [2002]. Happily, cross-collaboration between the two areas has begun over the past few years (see Rice [2004], Marron et al. [2004], and Müller [2005]), reducing the distinction between the fields.

2.2 Comparing Curves

At the root of these three disciplines is the issue of comparing curves. Given two or more sets of curves, how can we know if they differ? There are various ways for curves to differ — level shifts, different shapes or different peaks, for example. There has been much interest in this area, with many different approaches to the problem. We begin with a discussion of how the disciplines of FDA and LDA approach the problem of comparing two sets of curves. The field of profile analysis does not feature prominently in this section because the motivations in this area are generally not about testing whether two groups of curves are significantly different.

Interestingly, much of the literature in the field of functional data analysis is concerned with describing the data (with principal components, smoothing, etc.) as opposed to formal hypothesis testing, including assessing statistical significance between groups of curves. The texts by Ramsay and Silverman [1997, 2002, 2005] barely touch on the subject while the text by Ferraty and Vieu [2006] does not discuss hypothesis testing at all. Even when testing is done, it is often poorly structured. For example, Ramsay and Silverman [2005] use pointwise F tests to test for significant effects, completely ignoring the issue of multiple comparisons. Others have attempted to use common tests from multivariate data analysis in the functional case, which may not be advisable, as multivariate techniques break down as the number of time points increases, as discussed by Faraway [1997].

2.2.1 Comparing Curves Using Functional Data Analysis Techniques

Despite the emphasis on describing curves, researchers in the functional data area offer several methods for comparing curves. Ramsay and Silverman [2005] suggest calculating the F-test statistic at each time point, but do not address how to deal with the resulting statistics. Faraway [1997] presents a regression analysis technique for data having a functional response. This approach is affected by the number of points sampled per curve and requires bootstrapping. Chiou et al. [2004] also use functional linear regression. Cuevas et al. [2004] develop a one-way ANOVA for functional data that does not require homoskedasticity. This is an asymptotic test. Both the test by Faraway [1997] and that of Cuevas et al. [2004] require heavy use of computer simulations to carry out the procedures. Simplifying matters greatly, Shen and Faraway [2004] extend Faraway's work and propose a functional F test for nested functional linear models. This test is easy to calculate, it does not require any computer simulations, and its performance is not affected by the number of points sampled per curve. The data does not need to be smoothed to carry out this test and uses the unbiased least-squares estimator. Shen and Faraway derive the null distribution of their F test and show how it can be approximated. They compare their test to Faraway's bootstrapped test, the multivariate log-likelihood ratio test, and a method based on data smoothed with B-splines. Their simulations show that the power of the tests depends on the data. However, they argue that the F test has the benefit of avoiding the risk that the other two comparison tests have with being influenced by "unimportant directions of variation" (the two comparison tests may declare effects statistically significant that are not of practical importance) [Shen and Faraway, 2004].

The text by Ramsay et al. [2009] does address the issue of hypothesis testing (see Chapter 10). They discuss functional analysis of variance (fANOVA) to test for effects. They fit basis functions to their data to reduce the dimension of the data and use penalized least squares to estimate the coefficients of the model, usually penalizing derivatives of the functions. In terms of hypothesis testing, they use permutation tests based on the absolute value of the test statistic similar in form to the t-test statistic at each point in time. They take the maximum value of the statistic and compare it to the maximums calculated for 200 random regroupings of the data. They also propose a functional F statistic and use a permutation test based on the maximum F value over time.

Another group of tests are based on expansion techniques. Fan [1996a] develops an adaptive Neyman test and a thresholding statistic to compare two sets of curves. Fan uses the Neyman approach with the Fourier transform and the thresholding approach with wavelet transforms. Fan [1996a] compares the power of the Kolmogorov-Smirnov, Cramér-Von Mises, Anderson-Darling, and Shapiro-Wilk tests to the adaptive Neyman test and the thresholding test. He finds that the thresholding test has highest power in detecting local features, followed by the adaptive Neyman test. For global features both of Fan's proposed tests have high power compared to the other tests. Fan and Lin [1998] extend this idea for use with k groups of curves. They transform the data using a Fourier transformation and test the coefficients of the transformed data using an adaptive Neyman test. They expand their method into a functional ANOVA, which they call high dimensional ANOVA, for use with multiple sets of curves.

Of a similar nature is the work by Abramovich et al. [2004], who use wavelet decompositions in their non-adaptive and adaptive methods for testing main effects and interactions in the functional ANOVA model. They argue that their methods are asymptotically optimal non-adaptive and adaptive procedures. The carry out a simulation study to examine how their test performs against two alternatives. The results show that the tests have proper 5% rejection rate under the null hypothesis and increase satisfactorily as the alternative deviates from the null. They do not compare the power of their test to any other tests.

Others have proposed techniques based on principal components analysis, reducing the curves to a smaller set of coefficients. Sturino et al. [2010] fit principal components to the centred curves and use the principal component scores as the basis for testing. Benko et al. [2009] also use principal components in a functional extension of the common principal components concept introduced by Flury [1988]. Compared to the former test, the test by Benko et al. has a far more complicated structure and requires bootstrapping to carry out the hypothesis testing. Another approach is that of Horváth et al. [2009], who compare two sets of curves by modelling the relationship between effect and response variables with linear operators, and testing whether the operators are the same, although they deal with the case where both explanatory and response variables are functions. To accomplish this testing procedure they expand the data using functional principal components. Their test statistics have an asymptotically χ^2 distribution.

Some procedures use alternative data-reduction techniques to reduce the information in the curves into a smaller set of variables or coefficients and then use the reduced information to test hypotheses. Ferraty et al. [2007] discuss specifically the problem of how to compare two sets of curves from the perspective of cluster analysis. They use a basis expansion (as they say, the basis can be anything: “splines, Fourier functions, wavelets, functional PCA components, . . . ”) to express the data and test hypotheses based on the coefficients of the expansion. They also discuss testing hypotheses based on the derivatives of the curves rather than on the curves themselves.

Munk and Dette [1998], Dette and Derbort [2001], and Neumeyer and Dette [2003] provide methods for testing differences between groups of curves under various sample size and variance conditions. Munk and Dette [1998] develop a consistent test for two independent samples that accommodates unequal sample sizes. This test is based on the weighted L_2 distance between the two regression functions. They compare the power of their test to the procedure proposed by Delgado [1993] that is restricted to equal sample sizes and equal time points. They find that Delgado’s test performs better when the difference between the two functions is close to linear. Munk and Dette’s procedure has larger power when the difference is “more wiggly.” Dette and Derbort [2001] focus on nonparametric regression and deal with testing higher order interactions. Neumeyer and Dette [2003] discuss how to compare two regression curves by testing the mean functions where the data is unequally-spaced with unequal variances. They propose a bootstrapping procedure to carry out the test. The authors compare their test to two others and conclude that they prefer their test.

Some procedures take a Bayesian approach. Behseta and Kass [2005], Behseta et al. [2005] and Behseta et al. [2007] use an MCMC-based approach called Bayesian adaptive regression splines (BARS) as a dimension-reduction technique to fit splines to curves and test the equality of two or more functions. They are concerned with methods for very noisy functions. Behseta et al. [2007] modify multivariate ANOVA techniques, using BARS to fit curves and using likelihood ratio tests and approximating the distribution of the test statistic with a χ^2 distribution. Note that the code for running BARS is readily available in R and S.

Testing procedures have the potential to get very complicated. Hastie and Tibshirani [1993] present a varying coefficients model, where the coefficients of model parameters are allowed to change over time. Guo [2002] proposes a maximum likelihood ratio test for smoothing spline ANOVA (SS ANOVA) models. This model and test combination can be used to test for significant fixed effects in the functional model, including testing for a significant difference between two groups of curves. SS ANOVA was shown to be connected to mixed effects models by Speed [1999]. Antoniadis and Sapatinas [2007] instead suggest using wavelet decompositions as a means for testing model effects, both fixed and random. Cuesta-Albertos et al. [2007] propose a random projection model to test for goodness of fit to parametric families that can also be used to test for goodness of fit to other models (such as the Black-Scholes model). They use bootstrapping to carry out the hypothesis testing. Kuelbs and Vidyashankar [2010] develop one- and two-sample tests for high-dimensional data making extensive use of asymptotic results. Morris and Carroll [2006] develop a method for functional mixed-effects modelling using a Bayesian wavelet-based approach.

Others have taken very basic ideas and adapted them to the realm of FDA. Heckman and Zamar [2000] discuss the problem of defining a measure of similarity of shape using rank correlation and use it to divide a collection of curves into groups, and also use the measure to study monotonicity. Using this measure for testing is problematic because the estimator is not consistent. Fraiman and Muniz [2001] discuss trimmed means for functional data in order to measure the median of a group of curves. Hall et al. [2001b] point out that a permutation test is an alternative to other formal tests as it requires few assumptions, although they caution that the method becomes complicated with a complex design and requires programming competency. Muñoz Maldonado et al. [2002] use permutation tests to carry out hypothesis testing. They carry out a small power analysis limited to comparing the tests they develop in the paper. James and Sood [2006] use the permutation technique to test whether a curve or set of curves is equal to an estimated mean curve. Sturino et al. [2010] also use a permutation test to compare groups of curves but compare their method only to the principal components method they also propose in their paper. We will return to the permutation test shortly, in Section 2.4, where we provide a greater overview of this technique.

In this section we have reviewed the various approaches developed to test whether two groups of curves are significantly different. These tests vary greatly in simplicity, required computing resources, and assumptions. We discussed the situations where power comparisons between procedures have been done. We now briefly discuss how the area of longitudinal data analysis approaches the problem of comparing two sets of curves.

2.2.2 Comparing Curves Using Longitudinal Data Analysis Techniques

One may also approach the problem of comparing curves from the LDA point of view, treating the observations of individual curves as repeated measures. In random effects models, the regression coefficients are not restricted to be the same for all individuals [Diggle et al., 2002]. These models accommodate both discrete and continuous covariates and the covariates are allowed to vary over time. The difference between the random effects model approach and the model-based approach of Shen and Faraway [2004], for example, is that the former model assumes an explicit structure for the covariances among the responses. The choice of covariance structure may affect parameter estimates and their variances (see Hand and Crowder [1996], p. 73). In contrast, Shen and Faraway's model makes no such assumptions and therefore avoids the inherent risk of misspecifying the covariance structure. Computationally, random effects models use complex algorithms to approximate solutions but the code is readily available in statistical analysis software. However, increasing the dimension of the data (that is, more frequently collected data) causes problems with these models. Most LDA techniques are not equipped to handle high-dimensional data.

2.2.3 Discussion

We have discussed many procedures for testing whether two groups of curve are different. There is no shortage of options for comparing curves and we have our choice of approach to take. Some techniques are fairly complex, some are limited in scope. What is not clear is which test to choose. In the literature there is limited comparison of various testing procedures for comparing two sets of curves. Articles that compare tests and carry out simulation studies consider only a few options and have limited simulation studies.

Few comparisons have been done between longitudinal and functional approaches. One investigation by Yang et al. [2007] analyzes a set of data using both Shen and Faraway's regression structure and a linear mixed-effects model with a random intercept. The former model allows the model effects to change over time. Plotting these effects allows us to see how the parameter effects change over time, which can be enlightening. The mixed-effects model yields merely a point estimate for each parameter, which is not nearly as revealing. Yang et al. [2007] run simulations to compare the performance of the functional F test to linear mixed-effects models and Wilks' Lambda test using a Fourier transformation. They find that the power of the tests depends on the assumed covariance structure of the error process being simulated, although the functional F test has higher power than the linear mixed-effects model overall. They point out that the functional F test takes much less time to simulate than the linear mixed-effects models.

We have spent some time reviewing techniques for comparing curves. We have presented the few comparisons that have been done between procedures. These limited comparisons are the motivation for a more comprehensive examination of the power performance between testing procedures. This simulation study is presented in Chapter 5. We now expand our scope and offer a review of the literature of the fields of statistical process control, functional data analysis, and longitudinal data analysis, beyond the realm of comparing curves. This is a broader review of the major topics that arise in the three fields.

2.3 A Brief Summary of Research Using Data that are Curves

In this section we present a brief overview of the research activity in three areas: statistical process control, functional data analysis, and longitudinal data analysis, which have developed separately over the years. We hope that bringing everything together in one document sets the stage for more encompassing research that spans two or all of these fields, taking the strength of each and creating better methods applicable in wider settings. There have been some beginning steps towards this goal, with cross-over between functional and longitudinal data analysis, and this activity is summarized.

2.3.1 Quality Control

In this section, we discuss current research in statistical process control, where the analysis of curves is called profile monitoring. The curve is usually called a profile, although terminology varies. Jin and Shi [2001] call profiles “waveform signals,” while Gardner et al. [1997] refer to a signature instead of a profile in their methodology applied to equipment fault detection.

In statistical process control, Woodall et al. [2004] and Woodall [2007] are thorough reviews of the status of research in the area. The problem of dealing with data as curves arose while considering alternatives to summary (univariate) measures of quality. Ding et al. [2006] discuss the difficulties in applying a nonlinear perspective to profiles and the drawbacks of using summary statistics such as the maximum magnitude or the average value. Summarizing the data in such a way squanders the richness of the information contained in the curves. Consequently, “a monitoring system based merely on the simple statistics often suffers from a high false-alarm rate and/or a high miss-detection rate,” (Ding et al. [2006], p. 200).

In lieu of univariate measures, as a very simple alternative, the profile may be modelled by a straight line, which may be useful, if not over the whole curve, then on a region of the index variable (usually time). For example, Mahmoud and Woodall [2004] focus on linear profiles, which can be modelled using linear regression analysis. The slope and intercept can be monitored using a multivariate T^2 control chart [Stover and Brill, 1998, Kang and Albin, 2000]. Kang and Albin [2000] also suggest monitoring the residuals using exponentially weighted moving average (EWMA) and range (R) charts. They demonstrate through simulation studies that monitoring both the regression parameters and residuals work well to detect a shift in the slope, intercept (or both), and the standard deviation. Kim et al. [2003] propose modifications of the techniques of Kang and Albin [2000]. Mahmoud and Woodall [2004] use indicator variables and an F test to monitor the error variance. Jensen et al. [2007] investigate estimation methods that are robust with respect to outliers.

More sophisticated models to analyze profiles have been considered. These techniques are usually applications of methodologies borrowed from other areas of statistics, such as time series analysis and longitudinal data analysis. Staudhammer et al. [2005] incorporate autocorrelation using time-series models but conclude the models are not especially helpful in statistical process control applications. Jensen et al.

[2008] use linear mixed models, pooling information from the profiles to improve the fit of the model. These types of models include a correlation structure, allowing the model to reflect the relationship between profiles and between measurements within a profile. They report that for balanced data, the additional correlation structure does not add any benefit. For unbalanced data or when there is missing data, they find that the linear mixed model approach is preferable.

At the same time, Woodall et al. [2004] say, “Most of the current work on the topic of profile monitoring, however, still relies fundamentally on standard ideas from classical regression and multivariate analysis,” (p. 311). The field is ripe with opportunities for future research using functional techniques.

2.3.2 Current Topics in Functional and Longitudinal Data Analysis

The term functional data analysis was coined in a landmark paper by Ramsay and Dalzell [1991]. The term is not universal; searching *functional data analysis* as a keyword does not uncover all references for the analysis of data that are curves. A broad overview of the key topics in the field is found in Ramsay and Silverman's book *Functional Data Analysis, Second Edition* [Ramsay and Silverman, 2005] (the first edition was published in 1997 [Ramsay and Silverman, 1997]). A new text in the field, *The Oxford Handbook of Functional Data Analysis* [Ferraty and Romain, 2011] provides a slightly different perspective and gives a good summary of the common current topics in functional data analysis. The classic reference for longitudinal data analysis is Diggle et al. [2002].

In this section, some of the key areas of research in functional and longitudinal data analysis are discussed.

Smoothing

Smoothing is frequently the first step in a data analysis using curves. Rice [2004] discusses the different approaches to smoothing between longitudinal and functional data analysis. He says that in general, LDA smoothing procedures are model driven, with smoothing parameters determined using maximum likelihood estimation and therefore not explicitly chosen. In contrast, FDA smoothing is more likely to consider several bandwidths and choose one based on plots of the smoothing effect, or will use cross-validation, which is driven by the data instead of an imposed model structure. Rice [2004] applies a typical LDA smoothing procedure to functional data, and notes how this can provide better-fitting smooths of the individual curves. Fan and Zhang [2000] apply functional linear models to longitudinal data analysis and propose a two-step smoothing approach which is not as computationally intensive as spline and kernel methods.

Rice [2004] says, “One of the main advantages of the basis function approach is that the reduction to a finite dimensional representation allows subsequent use of a variety of standard statistical tools,” (p. 641). This technique is ripe for exploitation in applications. For instance, Hall et al. [2001a], use FDA to reduce the dimension of radar range profiles and then apply multivariate analysis techniques to analyze the data. Shi et al. [1996] convert curves into B-splines and use principal components of the covariance matrix to reduce the number of parameters in their mixed effects model. Rice and Wu [2001] also fit splines to curves to fill in sparse curves, estimate the covariance matrix using the EM algorithm and use a cross-validated log-likelihood to evaluate the fit of the model.

An issue to consider is that Hall and Van Keilegom [2007] (see also their technical report written in 2006 on the same issue, Hall and Van Keilegom [2006]) discuss how using different smoothing parameters for different curves has the potential to reduce power when conducting two-sample hypothesis testing. Based on their investigations they recommend that the same smoothing parameters be used for all curves when the curves are used in hypothesis tests. This is an interesting problem that deserves further attention.

Principal Components

Principal components analysis (PCA) is a common technique in FDA as a method for reducing the dimension of the data. The technique is primarily used as an exploratory, descriptive tool rather than for hypothesis testing. The exceptions are discussed in the section on comparing curves. Rice and Silverman [1991] and Silverman [1996] develop smoothed principal components techniques. The book by Ramsay and Silverman [2005] contains a chapter providing a good overview on using principal components to describe characteristics of the data. James et al. [2000] present a method for PCA with sparse functional data. Yao et al. [2003] propose the use of shrinkage methods to improve functional PCA. Li et al. [2002] use PCA combined with nested modelling to group genes in gene expression analysis. Jones and Rice [1992] use principal components to select “representative” curves from a large set of curves, in order to create summary displays of the data as an alternative to superimposing a large set of raw curves on top of each other, where details might be hidden. While most common in FDA, some cross-over into applying these ideas to LDA has begun. Yao et al. [2005a] develop a technique to apply functional PCA to sparse longitudinal data.

Aligning Curves

Curve registration (also called warping or alignment) is a topic of interest, at least from the FDA point of view. Consider the growth curves of children [Gasser et al., 1985]. The timing of growth spurts varies among individuals, leading to a large amount of variability in the growth curves. Aligning the growth spurts reduces this variability and helps illuminate trends in the data. Kneip and Gasser [1992]; Gasser and Kneip [1995]; Capra and Müller [1997]; Ramsay and Li [1998]; Wang and Gasser [1999]; Kneip et al. [2000]; and Liu and Müller [2003] all consider approaches to align curves; Ramsay and Silverman [2005] cover the topic in some detail.

Kneip et al. [2000, p. 28] advise some caution when registering curves, because:

> The registration problem has the potential to be ill posed. It seems clear that amplitude and phase variation can only be separated if the latter varies slowly with respect to the former. In many problems, including brain imaging, the amount of variation in amplitude will vary greatly from one part of the domain to another. For example, a slice through the middle of the brain shows high-frequency variation in various types of neuroimages in the cortical regions near the outer boundaries of the brain, but only low-frequency variation in central regions, where much of the space is taken up by the ventricular chambers. This suggests that

> local algorithms will have to vary bandwidths to adapt to the frequency of amplitude variation.

It seems one should be careful when registering curves. Curve registration may help illuminate some characteristics of the curves but caution should be used. Further investigation into combining curve registration with testing procedures is required.

Modelling Relationships

Model-based approaches are very popular in LDA (especially generalized estimating equations (GEE)), but do not appear as much in the FDA literature, most likely because of the descriptive nature of much of FDA analysis, or perhaps because good inferential techniques are only now being developed. The high-dimensional nature of most functional data may also be a limiting factor.

Yao et al. [2005b] and Müller [2005] both propose extensions of functional linear models that accept sparse, irregularly-spaced longitudinal data. Ratcliffe et al. [2002a] develop what they call functional logistic regression, using functional data analysis with a binary response, and extend it to a covariate with a repeated stimulus. Müller [2005] also discusses a logistic regression method. Shi and Choi [2011] discuss functional regression analysis for functional response variables and scalar and functional covariates with a Bayesian perspective using Gaussian process regression. Yang et al. [2007] apply a functional data analysis approach to longitudinal data that was collected fairly frequently. Yang et al. [2007] make a good point: "By applying both longitudinal and functional data analysis to the same set of data, the overall time-averaged treatment efficacy and the dynamic time-changing effects of treatment can be jointly targeted so that we may obtain multi-faceted enhanced understanding of the studied phenomena," (p. 1562). Fan and Zhang [2000] propose a two-step process to fit a functional ANOVA model to longitudinal data whereby the estimates are calculated on unsmoothed data and then those estimates are smoothed in a secondary step. They do not do hypothesis testing, instead they use pointwise confidence bands, and they require data recorded at the same time points for all responses.

Graphical Techniques

Visual techniques have been developed to better understand and view the data. Dawson et al. [1997] present two graphical techniques — draftman's display and a parallel axis system — to assist in determining the initial form of the covariance matrix.

Chaudhuri and Marron [1999] and Chaudhuri and Marron [2002] introduce SiZer (SIgnificant ZERo crossing of derivatives) analysis. Zhao et al. state, "The main goal of SiZer analysis is to understand which clusters in the data (i.e., bumps in the curves) represent important population structure, and which can be attributed to sampling variability. . . The SiZer map uses colors to indicate statistical significance of slopes, with blue for significantly increasing, red for significantly decreasing, and the intermediate color of purple when the slope is not significant. The fourth color of gray is used to indicate locations where the data are too sparse for statistical inference," [Zhao et al., 2004, p. 801]. This technique studies a range of bandwidths at the same time.

2.3.3 Cross-Disciplinary Opportunities Between Functional and Longitudinal Data Analysis

We conclude this section with a discussion on bringing the fields of functional data analysis and longitudinal data analysis together. While reading articles from FDA and LDA, it became clear that each field had very different approaches to analyzing data where the response is a curve, despite often facing similar issues. However, in more recent years there seems to have developed a willingness towards cross-disciplinary application of techniques, and an awareness of how this may afford some benefits. For example, in a special issue of *Statistica Sinica*, Marron et al. [2004, p. 615], discuss the differences between longitudinal data and functional data and call for a unifying theory between them:

> Longitudinal data and functional data are both data collected over a period of time on the same subject. They both depict the realization of a smooth underlying process at discrete time points. However, there are intrinsic differences between the two approaches, partly due to different sampling schemes. . . . The two fields have recently crossed paths due to challenges faced in each and this has lead to the beginning of fruitful interactions. From the longitudinal point of view, there is a need to pursue more flexible non- or semi-parametric frameworks that may better capture the complex data features that are present in many longitudinal studies. From the functional side, there is a need to provide techniques that work for 'sparse' data commonly encountered in longitudinal studies.

The issue of the separation between FDA and LDA (that is, the development of the two fields as separate, distinct areas) is raised by several authors (see for example Zhao et al. [2004], Rice [2004], and Yang et al. [2005]). Rice [2004] distinguishes between typical longitudinal data and typical functional data. In Rice's opinion, functional data is more likely to be collected frequently, often using automatic equipment, while longitudinal data is usually collected with bigger gaps between measurement times, and often with irregular measurement times. He points out that the aims between the two approaches also differ: in FDA the focus is on data exploration and feature discovery, while LDA is more interested in inference. He also observes that the LDA literature is more concerned with missing data. He notes that they do have many common goals, such as describing an average curve over time, dealing with noisy data when describing curves, and relating covariates to curve characteristics.

Müller discusses the non-parametric tendencies of FDA and how this is beneficial, "The prepubertal growth spurt that had almost vanished from paediatrics textbooks of the 1980s, since the parametric models that had become popular for fitting the human growth curve did not have room for such a second growth spurt — it took a non-parametric analysis to bring this growth spurt (which had been recognized in the pre-modelling era) back on the map (Gasser et al., 1985)," [Müller, 2005, pp. 224–225]. There are advantages to exploring the data in different ways.

Researchers are beginning to realize the advantages of applying FDA to longitudinal data, and vice versa. As Müller [2005] says, "Functional methods provide a variety of valuable and potentially powerful tools for longitudinal data analysis if a bridge can be built in which longitudinal data, sampled on sparse designs, can be brought under the umbrella of functional tools," (p. 225). In addition, Yang et al. [2005] point out that repeated measurements on a subject could be treated as samples from a curve, which can be smoothed and tested with FDA. "With time-dependent coefficients, functional regression analysis captures the time-varying exposure-response relationship, thus providing a simpler data structure with intuitive interpretations. A time series plot of the estimated coefficient function vividly reveals how the effect of a predictor can change along the time axis," (p. 4). They also mention that FDA is more robust compared to LDA, because it does not make as many assumptions about the correlation structure as LDA does.

In conclusion, just as the field of profile analysis stands to benefit from the application of techniques from FDA and LDA, so too do longitudinal data analysis and functional data analysis have opportunities to develop further by integrating techniques between the fields.

2.4 Permutation Tests

In this section we discuss the permutation test. We briefly mentioned this approach in the section on comparing curves and now we will discuss permutation tests in greater detail. We begin with a short history of the test, and provide some examples of applications using the permutation test with data that are curves, illustrating how most of these tests are done in a pointwise manner instead of exploiting the functional nature of the data. We conclude with a discussion of the advantages of using a permutation test.

The permutation test, sometimes called a randomization test, is a non-parametric approach to hypothesis testing that has been in use for many years, dating back to Pitman [1937a], (see also Pitman [1937b, 1938]) who references Fisher (1935). The permutation technique is enjoying a renaissance for the same reasons functional data analysis is growing: data collection has become more sophisticated, giving rise to extremely large data sets, and computing power has grown, meaning computer-intensive techniques are more doable. Texts by Edgington [1980] and Good [2006] are classics in this area but they do not discuss permutation tests with functional data. The text by Pesarin and Salmaso [2010] discusses permutation tests in great detail, covering basic two-sample (univariate) tests, multivariate testing procedures, repeated measurements (where the focus is mostly on testing with missing values), survival analysis and shape analysis, but does not deal with tests for functional data.

Permutation tests have been used with non-functional data in a wide variety of applications: hydrology [Bardsley et al., 1999]; radar detection [González-García et al., 2005, Sanz-González et al., 2007]; evolutionary biology [Griswold et al., 2008]; population genetics [Price et al., 2010]; ecology [Champley et al., 1997, Shipley, 2010]; shape discrimination [Terriberry et al., 2005]; chromatography [Muñoz Maldonado et al., 2002]; and genetic epidemiology [Fang and Wang, 2009].

Two relatively new fields that make broad use of permutation tests are human brain imaging and gene expression. In human brain imaging, functional magnetic resonance images (fMRIs) are taken of the brain at a baseline control level and under some stimulus. These images of the brain are then compared for each voxel (which is like a pixel for 3D images). Based on the sheer number of voxels and the correlation between voxels that are close together, a permutation test is often carried out instead of using a model- or distribution-based hypothesis test (Holmes et al. [1996] and Arndt et al. [1996] are early papers, see also Locascio et al. [1997], Bullmore et al. [1999] and Belmonte and Yurgelun-Todd [2001]; Nichols and Holmes [2001] have a step-by-step article). However, the basis of the test is often simply a t-test carried out at each voxel. Then the permutation test is used to adjust for the multiple testing problem and reduce the false discovery rate [Benjamini and Hochberg, 1995]. Similarly, Suckling and Bullmore [2004] analyze fMRI data from a factorial experiment with an F ratio combined with permutations. The same techniques, applying a series of pointwise tests and using permutation tests to adjust for multiple testing, are used in genetic mapping: Churchill and Doerge [1994] is an early reference, followed up by Doerge and Churchill [1996]. Gene expression analysis (also called microarray analysis) uses the same technique [Dudoit et al., 2002, Landgrebe et al., 2002, Almind and Kahn, 2004], although the treatment of comparisons can get quite sophisticated [McIntosh et al., 2004, Barry et al., 2005, Ptitsyn et al., 2006, Efron and Tibshirani, 2007, Barry et al., 2008, Distaso et al., 2008, Nettleton et al., 2008, Pounds et al., 2009]. This technique is also applied in other areas of genetics [Wu and Lin, 2006, Yap et al., 2009]. Chau et al. [2004] and Teismann et al. [2007] use permutation tests in the area of magnetoencephalography.

Essentially, then, this technique is just a series of pointwise tests. The image or microarray is not treated as a function. Other techniques in shape analysis in medical imaging that do treat the image as a function reduce data dimension by principal components [O'Connor et al., 2010] or wavelets [Bullmore et al., 2001, Şendur et al., 2005, Nain et al., 2007] and then use permutations to calculate p-values.

These pointwise techniques have crept into other fields. Cox and Lee [2008] use the same analysis structure wherein they calculate pointwise tests and then adjust the p-value using a permutation test. Castano et al. [2006] also use permutation tests to calculate p-values in a garlic storage application, using the techniques developed for brain imaging in Nichols and Holmes [2001].

However, we are more concerned with methods that treat the data as curves. Blair et al. [1994] investigated the applicability of permutation tests in place of Hotelling's T^2 test in behavioural science research, when the T^2 test is not appropriate, for example when the assumptions (such as normality) of the parametric test are violated or when there are more variables than subjects in the study. This idea dates back to Chung and Fraser [1958]. Another technique is adapting nonparametric tests for the functional case and then applying permutation tests. Bugni et al. [2009] describe a method to test goodness of fit to a specified theoretical parametric distribution using a generalization of the Cramér-von Mises test. Hall and Tajvidi [2002] test the equality of two distributions based on a test statistic similar to the Cramér-von Mises statistic and use permutation methods to determine the p-value. They report that their test is quite sensitive, detecting alternatives close to the null hypothesis. Fay and Shih [1998] (and also Shih and Fay [1999]) adapt the Mann-Whitney test and use permutation tests for empirical distribution functions that are applicable to repeated measures or censored data. Muñoz Maldonado et al. [2002] use Pearson's sample correlation coefficient to measure the similarity between two (registered) sample curves and then use permutations to calculate the p-value.

Sometimes a more model-based approach is used, fitting models and using permutation tests to determine significance. Cuevas et al. [2004] adapts the F test for use with functional data using an asymptotic test and uses bootstrapping to calculate the distribution of the test statistic under the null hypothesis. Delicado [2007] applies Cuevas et al.'s methodology, approaching the problem from an analysis of distance point of view, but uses a permutation test instead of bootstrapping. Cardot et al. [2007b] propose a functional F statistic similar to Shen and Faraway's [2004] and a test statistic based on smoothing residuals and use permutations to calculate the p-values of the tests. Cardot et al. [2004] use a permutation test approach for a model with functional covariates and a scalar response (the opposite situation to the focus here). Greenwood et al. [2010] fit penalized regression splines to wetland hydrology data and use permutation tests to test hypotheses.

Others have used a wide variety of techniques in combination with the permutation approach. James and Sood [2005] compute the residuals of the observed curve from the mean curve and permute these residuals to carry out hypothesis testing that a set of curves (or one curve) are generated from a specific shape. López-Pintado and Romo [2007] use a measure of depth to analyze the shape of a set of curves (but do not test if two sets of curves are similar or not) which they use to detect outliers. Chiou and Müller [2007] use a randomization test for their goodness-of-fit test to test the overall significance of the model parameters based on functional principal components. Castano et al. [2006] fit a functional linear model to their data and use permutation tests to test for significant model effects. Schmoyer [1994] tests for correlation using a permutation approach.

More specifically related to the problem of comparing groups of curves, one can measure the distance between curves using some sort of norm and use the norm as the basis for a permutation test. This is the approach taken by Zerbe and Walker [1977], who use a permutation technique to test whether the mean curves differ between groups over a specified interval. They use the L_2 norm to measure the distance between the mean curves. They also discuss using only a sample of permutations instead of the complete list. The test is applied to growth curves of girls. Also, Sturino et al. [2010] measure the squared distance between the mean curves of the two groups and use a permutation test to test for significance. They also suggest two other measures of differences, the median curves and the area under the mean curve of each group, as bases for permutation tests.

We see that there are a wide variety of ways to use permutation tests with functional data. Thus far we have mentioned many ways to test whether two groups of curves are significantly different from each other. Next we outline a few characteristics of permutation tests that have been discussed in the literature.

Permutation tests have some very desirable properties. Good states that, "Permutation tests correctly applied are exact, unbiased, and distribution-free," [Good, 2006, p. 73]. We do have to ensure that our hypothesis is constructed in a valid way for the permutation test to be appropriate. The only caveat is that the permutation test assumes, under the null hypothesis, that the two samples come from the same distribution [Good, 2006].

Along with the desirable properties, permutation tests also have good power. As Good [2006] states, "The power of permutation tests is quite high. For small samples, a permutation test may be the most powerful test available. For very large samples, a permutation test will be as powerful as the most powerful parametric test."

Edgington [1964] discusses the advantages of permutation tests compared with nonparametric tests:

> A desirable property of rank-order tests is freedom from the normality assumption and other parametric assumptions about the shape of the populations. The major limitation of rank-order tests is their lack of power: When parametric assumptions are met, rank-order tests are less likely than parametric tests to reject false null hypotheses.

He continues:

> Randomization tests combine the best features of parametric and nonparametric tests. They are nonparametric yet have the power of parametric tests because they use the numerical values of the scores. A randomization test drives a sampling distribution of a statistic from repeated computations of the statistic for various ways of dividing the scores. The purpose of the test determines the appropriate divisions of the score for computing the statistic. The significance of an obtained statistic is the proportion of the statistics in the sampling distribution that exceeds the obtained value.

Edgington adds, "Perhaps the most important aspect of randomization tests is the fact that, because they produce their own probability tables, they can be used to test any sort of quantitative relationship." Thus it seems that permutation tests applied to functional data may be a powerful method for comparing two groups of curves.

As we have stated, the p-value produced by the permutation test is exact if we examine all possible rearrangements. However, for large samples of curves the number of permutations becomes prohibitively large. In such cases we may consider taking a random sample of permutations and use these to calculate the p-value, which is discussed by Edgington [1969] and Good [2006]. This p-value will not be exact. Happily, Keller-McNulty and Higgins [1987] found that a permutation test based on a subset of 1600 permutations had power reasonably close to the power of the test using all permutations. Note that the simulations done by Keller-McNulty and Higgins in 1987 required the use of a supercomputer to carry out the calculations. We have access to much greater computing power today.

Alternatively one could consider other methods to increase the computational efficiency of permutation tests. Heckel et al. [1998] use a greedy algorithm to reduce the number of computations required for each subsequent permutation, recalculating only the summary measure for the elements that have changed groups. Zhou et al. [2009] and [Zhou and Wang, 2009] obtain the first four moments of the Pearson distribution series to approximate the distribution of the permutation test statistic to avoid the computational cost of the usual permutation test. This is based on something similar to what Kazi-Aoual et al. [1995] have done but for a specific application. Raz [1990] uses an approximation based on the gamma distribution, which Pitman [1937a] also did.

In conclusion, permutation tests have a number of advantages over other approaches. They are widely used in certain fields of research, however most applications are based on a series of pointwise tests and do not exploit the functional nature of the data. Since permutation tests are not very widely applied to actual curves, we have limited information regarding how they compare with other tests used in functional data analysis, nor has anyone compared the variety of tests developed in the functional data analysis area. In the literature, a proposed test is often compared with one of two other tests based on a limited number of simulations. What is missing is a systematic examination of the performance of a wide variety of tests for functional data, including permutation tests.

CHAPTER

THREE

TESTS AND NOTATION FOR COMPARING TWO GROUPS

In Chapter 2 we discussed the various testing procedures that have been developed to test whether two groups of curves are different. We also discussed the limited comparisons that have been done between the various testing procedures. We reviewed the benefits of permutation tests.

In this chapter we begin by presenting a set of new permutation tests that we have developed based on the idea of comparing the curves in a pairwise fashion. These tests have all the benefits of permutation tests as discussed in the previous chapter on page 37, that is, they are exact, nonparametric tests. In addition, unlike the permutation tests proposed by Sturino et al. [2010], the test statistic is based on comparisons between the individual curves, which may provide a more powerful test because for tests based on summary curves, the summary curve may not resemble the shapes of the individual curves, and differences between curves in the two groups may be obscured. Determining if it matters whether we use a permutation test with pairwise comparisons between curves or base the test on summary curves is one motivation for a simulation study comparing the power performance of these two approaches to comparing curves.

Based on our literature review, we know there exist many other testing procedures for comparing two sets of curves. We include some of these tests to compare with the two types of permutation tests. These tests will be applied in the case study in Chapter 4 and studied further in the simulation study in Chapter 5. We are interested in comparing different approaches to dealing with functional data and therefore include tests based on the nature of their approach to testing. The test proposed by Sturino et al. [2010] is based on data reduction, expressing the curves in terms of principal components and using the principal component scores as the basis for testing. Another reason for including this test is that it is the functional data approach that Sturino et al. [2010] take as in comparison to their permutation tests, thus, it seems prudent to include. The adaptive Neyman test proposed by Fan and Lin [Fan, 1996a, Fan and Lin, 1998] is selected as an example of a testing procedure using data expansion techniques. The functional F test developed by Shen and Faraway [Shen and Faraway, 2004, Shen and Xu, 2007] is included because it is a functional version of the familiar F test used in the non-functional situation. We also include a permutation test by Ramsay et al. [2009] because it is included in the **fda** package that we use to carry out some of our data analysis.

The purpose of this chapter is to outline the structure of all of the tests under consideration along with all relevant notation. We introduce each test in turn and conclude the chapter with a discussion regarding the relative benefits of each approach.

Consider an observed response curve $Y_{ij}(t)$ for the jth observation in the ith group, $i = 1, 2$, $j = 1, \ldots, n_i$. The index variable here is t which usually indicates time although it does not have to be. We take $t \in [0, 1]$ for simplicity, although this is not necessary.

3.1 Permutation Tests

In this section we will describe a large number of permutation tests. These permutation tests are divided into three categories: permutation tests based on calculating pairwise differences between curves, permutation tests based on calculating differences between summary curves, and a permutation test based on a t-test-like statistic.

3.1.1 Pairwise Permutation Tests

We propose a set of permutation tests based on pairwise comparisons between curves. In this approach, we begin by defining a distance measure between two curves, with one curve taken from each group. One option for the distance measure is to use the integrated absolute difference (L_1 norm) between two curves f_1 and f_2 on the interval $[0, 1]$, given by

$$\delta_1(f_1, f_2) = \int_0^1 |f_1(t) - f_2(t)| \, \mathrm{d}t \qquad \text{for } f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2.$$

There are $n_1 \times n_2$ possible pairs of curves between the two groups. Then we compute the test statistic $\overline{T}_1$, which is the mean of these integrated absolute differences,

$$\overline{T}_1 = \frac{1}{n_1 n_2} \sum_{f_1 \in \mathcal{F}_1} \sum_{f_2 \in \mathcal{F}_2} \delta_1(f_1, f_2). \tag{3.1}$$

This test statistic, $\overline{T}_1$, is a distance measure, $\overline{d}_1(\mathcal{F}_1, \mathcal{F}_2)$, between two groups of curves. Here we use the subscript $_1$ to indicate the type of norm (distance measure) being used. The type of hat on the T denotes the type of summary statistic. The value of the statistic in equation (3.1) is for one grouping of the variables. To carry out the permutation test we need to recalculate this test for all reorderings of the curves into two groups. Let $B = \frac{1}{2}\binom{n_1+n_2}{n_1}$ be the number of unique groupings of the curves into two groups. We repeat the calculations for each grouping of the curves into two groups, computing T_b each time, for $b = 1, \ldots, B$. Then we compare the value of the test statistic $\overline{T}_1$ for the original grouping of curves to those of the permuted groups. The p-value is the proportion of T_b's which are greater than or equal to the observed value of $\overline{T}_1$ (the value of the test statistic for the original groups of curves),

$$p\text{-value} = \frac{1}{B} \sum_{b=1}^{B} \mathcal{I}(T_b \geq \overline{T}_1),$$

where $\mathcal{I}(\cdot)$ is the indicator function.

This permutation test is only one approach among many. Instead of using the L_1 norm, one could use alternative measures of how different two curves are, such as the integrated squared difference (which for simplicity we will call the L_2 norm although we do not take the square root),

$$\delta_2(f_1, f_2) = \int_0^1 [f_1(t) - f_2(t)]^2 \, \mathrm{d}t \qquad \text{for} f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2,$$

or the maximum absolute difference (L_∞ norm),

$$\delta_\infty(f_1, f_2) = \sup_{t \in [0,1]} |f_1(t) - f_2(t)| \qquad \text{for} f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2.$$

Another option is to calculate the absolute difference in area under each curve,

$$\delta_A(f_1, f_2) = \left| \int_0^1 f_1(t)\mathrm{d}t - \int_0^1 f_2(t)\mathrm{d}t \right| \qquad \text{for} f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2.$$

We could also change the summary measure and compute the median, trimmed mean or another m-estimator, rather than the mean. Define the trimmed mean as follows: Let $\delta_{[i]}$ be the ith ordered pairwise difference between two curves, $\delta_{[1]} < \delta_{[2]} < \cdots \delta_{[n_1 n_2]}$. Let the integer k, where $k < \frac{1}{2} n_1 n_2$, be the number of observations trimmed from each end of the ordered statistics, calculated by $k = \ell n_1 n_2$ (where the right hand side is rounded to the nearest integer), for $\ell = 0.10, 0.20$. Then define the ℓ-trimmed mean as

$$\overline{T}_\ell = \frac{1}{(n_1 n_2 - 2k)} \sum_{k+1}^{n_1 n_2 - 2k} \delta_{[i]}(f_1, f_2)$$

Combining the various summary statistics and distance measures results in several test statistics. We use a bar ($^-$) to indicate that the mean was used as the summary statistic, tilde ($^\sim$) to indicate median, and a bar with a subscript to indicate the 10% and 20% trimmed means ($^-{}_{10}$, $^-{}_{20}$, respectively) .

These combinations result in the pairwise mean L_2 norm,

$$\overline{T}_2 = \overline{d}_2(\mathcal{F}_1, \mathcal{F}_2) = \frac{1}{n_1 n_2} \sum_{f_1 \in \mathcal{F}_1} \sum_{f_2 \in \mathcal{F}_2} \delta_2(f_1, f_2),$$

the pairwise mean L_∞ norm,

$$\overline{T}_\infty = \overline{d}_\infty(\mathcal{F}_1, \mathcal{F}_2) = \frac{1}{n_1 n_2} \sum_{f_1 \in \mathcal{F}_1} \sum_{f_2 \in \mathcal{F}_2} \delta_\infty(f_1, f_2),$$

and the pairwise mean difference in integrated area,

$$\overline{T}_A = \overline{d}_A(\mathcal{F}_1, \mathcal{F}_2) = \frac{1}{n_1 n_2} \sum_{f_1 \in \mathcal{F}_1} \sum_{f_2 \in \mathcal{F}_2} \delta_A(f_1, f_2).$$

The statistics using other summary statistics are computed similarly. The notation is laid out in Table 3.1.

3.1.2 Method without Pairwise Comparisons

Next, consider an alternative to doing pairwise comparisons, instead taking the summary statistic of the group first and then calculating the distance measure. This is the approach taken by Sturino et al. [2010], and here we describe the test statistics as defined in their paper. Define the sample mean curve for the ith group as

$$\bar{f}_i(t) = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}(t),$$

and the sample median curve as

$$\widetilde{f}_i(t) = \text{median}_{j \in n_i} \{Y_{ij}(t)\}.$$

Table 3.1: Pairwise permutation test statistics notation.

pairwise mean L_1 norm	$\overline{T}_1$
pairwise mean L_2 norm	$\overline{T}_2$
pairwise mean L_∞ norm	$\overline{T}_\infty$
pairwise mean difference in area	$\overline{T}_A$
pairwise median L_1 norm	$\widetilde{T}_1$
pairwise median L_2 norm	$\widetilde{T}_2$
pairwise median L_∞ norm	$\widetilde{T}_\infty$
pairwise median difference in area	$\widetilde{T}_A$
pairwise 10% trimmed mean L_1 norm	$\overline{T}_{10,1}$
pairwise 10% trimmed mean L_2 norm	$\overline{T}_{10,2}$
pairwise 10% trimmed mean L_∞ norm	$\overline{T}_{10,\infty}$
pairwise 10% trimmed mean difference in area	$\overline{T}_{10,A}$
pairwise 20% trimmed mean L_1 norm	$\overline{T}_{20,1}$
pairwise 20% trimmed mean L_2 norm	$\overline{T}_{20,2}$
pairwise 20% trimmed mean L_∞ norm	$\overline{T}_{20,\infty}$
pairwise 20% trimmed mean difference in area	$\overline{T}_{20,A}$

When dealing with discrete points, it is possible to calculate the summary measures at fixed time points and then smooth the resulting curve to carry out the rest of this procedure. This is the approach taken by Sturino et al. [2010]. When dealing with smooth curves we take a sample of points from each curve and use those points to calculate the sample median curve.

Sturino et al. [2010] define the squared difference between the mean curves,

$$S_2 = \int_0^1 \left[\bar{f}_{1\cdot}(t) - \bar{f}_{2\cdot}(t)\right]^2 \mathrm{d}t,$$

while using the L_1 norm with the median curves we obtain

$$S_1 = \int_0^1 \left|\widetilde{f}_{1\cdot}(t) - \widetilde{f}_{2\cdot}(t)\right| \mathrm{d}t.$$

The L_∞ norm is

$$S_\infty = \sup_{t \in [0,1]} \left| \widetilde{f}_{1\cdot}(t) - \widetilde{f}_{2\cdot}(t) \right|,$$

and the integrated area difference is

$$S_A = \left| \int_0^1 \bar{f}_{1\cdot}(t) \mathrm{d}t - \int_0^1 \bar{f}_{2\cdot}(t) \mathrm{d}t \right|.$$

To calculate p-values, the statistics are recalculated for each regrouping of curves as described in the previous section.

3.1.3 Ramsay's Permutation Test

Ramsay et al. [2009] discuss a permutation test between two groups based on the t-test. This test is included in the **fda** R package [Ramsay et al., 2011]. They begin by taking the absolute value of a t-test-type statistic at each point along the curve

$$T(t) = \frac{\left| \overline{Y}_{1\cdot}(t) - \overline{Y}_{2\cdot}(t) \right|}{\sqrt{\frac{1}{n_1} s_1^2(t) + \frac{1}{n_2} s_2^2(t)]}}.$$

The test statistic is the maximum value of $T(t)$,

$$R = \sup_{t \in [0,1]} T(t).$$

Then they use a permutation test to assess significance, by randomly reordering the curves and recalculating the test statistic with the new groups of curves. The test uses 200 random reorderings by default. By nature, then, this test is similar to the test using the L_∞ norm but scaled by variances and calculated pointwise on the mean curves instead of computing the maximum along a smoothed curve. In principle this test is done continuously but it is carried out at discrete points as necessary.

Other Test Statistics as Permutation Tests

The beauty of the permutation test is that one can use any test statistic as the measure and apply the permutation test methodology to that measure in order to obtain a p-value. We calculate the test statistic for each regrouping of the curves into two groups. Then we take the test statistics calculated for the original groupings of curves, compare them with all of the other values computed for the regroupings, and count how many are as extreme as the original test statistic. The proportion of values as extreme as the observed value gives the p-value.

In this manner we will carry out permutation test versions based on test statistics outlined in the following sections: the principal components test, adaptive Neyman test, and functional F test. We will compare the power performance of these permutation versions of the test statistics to the performance of the original forms.

3.2 Test Based on Principal Components

Sturino et al. [2010] define a test based on functional principal components. We have $Y_{ij}(t)$ as the jth observed response curve for the ith group, $i = 1, 2$, and $j = 1, \ldots, n_i$. Centre the observations by subtracting the mean curve $\overline{Y}(t) = \frac{1}{n_1+n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}(t)$, resulting in the centred data $\breve{Y}_{ij}(t) = Y_{ij}(t) - \overline{Y}(t)$. The covariance function $\sigma(s,t)$ is estimated by

$$\hat{\sigma}(s,t) = \frac{1}{n_1+n_2-1}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\breve{Y}_{ij}(s)\breve{Y}_{ij}(t).$$

Similarly to the multivariate case, we use the eigenequation

$$\int \sigma(s,t)\psi(t)\mathrm{d}t = \lambda\psi(s) \tag{3.2}$$

to solve for the eigenvalues $\lambda_1 > \lambda_2 > \ldots > 0$ and eigenfunctions $\psi_1(t), \psi_2(t), \ldots$.

Sturino et al. [2010] use the Karhunen-Loève expansion, approximating every centred functional observation $\breve{Y}_{ij}(t)$ by

$$\breve{Y}_{ij}(t) \approx \sum_{k=1}^{K}\gamma_{ijk}\psi_k(t),$$

with γ_{ijk} as the principal component scores, $\gamma_{ijk} = \int \breve{Y}_{ij}(t)\psi_k(t)\mathrm{d}t$. The γ_{ijk} are random variables with zero mean and $\mathrm{cov}(\gamma_{ijk},\gamma_{ijl}) = \mathcal{I}(l=k)\lambda_k$, where $\mathcal{I}(\cdot)$ is the indicator function. They argue that the resulting estimates $\hat{\gamma}_{ijk}$ are approximately Normal, $\hat{\gamma}_{ijk} \approx \mathrm{Normal}(\mu_{ik}, \sigma^2_{ik})$ and that they are approximately independent across i and k. With two principal components, they test $H_0 : \mu_{11} = \mu_{21}$ and $\mu_{12} = \mu_{22}$.

The test statistic for the first principal component is

$$z = \frac{\sum_{j=1}^{n_1}\hat{\gamma}_{1j1}/n_1 - \sum_{j=1}^{n_2}\hat{\gamma}_{2j1}/n_2}{\sqrt{\lambda_1/n_1 + \lambda_1/n_2}}.$$

Then the p-value is calculated, $p_1 = P(Z > |z|)$. The calculations for the second principal component follow similarly. The authors propose an alternative rejection region, rejecting H_0 if either $p_1 < \alpha\lambda_1/(\lambda_1+\lambda_2)$ or $p_2 < \alpha\lambda_2/(\lambda_1+\lambda_2)$, where p_1 is used to test the first principal component and p_2 the second one. They state that weighting the rejection regions in this way greatly increases the power of the test.

3.3 Adaptive Neyman Test

Fan [1996a] proposes an adaptive Neyman test to compare two sets of curves in an overall test. This test is further discussed in Fan and Lin [1998]. First consider a single set of curves $Y_1(t), \ldots, Y_n(t)$, an iid sample from a distribution $F(t)$. The authors begin by testing goodness of fit of the distribution $F(t)$ against a given distribution function F_0,

$$H_0 : F(t) = F_0(t) \qquad \text{versus} \qquad H_1 : F(t) \neq F_0(t).$$

The traditional method of attack for such a problem is a Kolmogorov-Smirnov test or a Cramér-Von Mises test. However, Fan [1996a] notes that one can transform this test into the equivalent form

$$H_0 : F(t) = U(t) \qquad \text{versus} \qquad H_1 : F(t) \neq U(t),$$

where $U(t)$ is the CDF of the uniform distribution over the unit interval. He then considers the Fourier transform

$$\begin{aligned} \theta_{2r-1} &= \int_0^1 \cos(2\pi r t)\mathrm{d}F(t), \\ \theta_{2r} &= \int_0^1 \sin(2\pi r t)\mathrm{d}F(t), \qquad r = 1, 2, \ldots. \end{aligned}$$

The $\hat{\theta}_r$, $r = 1, 2, \ldots,$ are the empirical Fourier coefficients. Then the test can be expressed as

$$H_0 : \theta_r = 0, \qquad r = 1, 2, \ldots \qquad \text{versus} \qquad H_1 : \text{at least one } \theta_r \neq 0. \tag{3.3}$$

Fan proceeds to express the Kolmogorov-Smirnov and Cramér-Von Mises tests in terms of the Fourier coefficients and shows that they are based on only the first few Fourier coefficients. The consequence of this phenomenon is that such distribution-based tests have limited power when testing distributions that have high-frequency components (that is, distributions having large Fourier coefficients θ_r for some large r). Another consequence is that these tests are not good at detecting local features, such as bumps. Fan sets out to find an alternative form of such tests that do not exhibit such power problems.

Fan [1996a] points out that under the null hypothesis in (3.3), the coefficients $\hat{\theta}_r$ are asymptotically independent and normally distributed: $\hat{\theta}_r \sim N(\theta_r, n^{-1})$ for $r = 1, \ldots, N$, where $N/n \to 0$. This led him to the Gaussian white noise model.

Fan [1996a] and Fan and Lin [1998] consider $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\theta}, I_n)$, an n-dimensional normal random vector. The test of interest is

$$H_o : \boldsymbol{\theta} = 0 \qquad \text{vs} \qquad H_1 : \boldsymbol{\theta} \neq 0.$$

They suggest modifying the maximum likelihood test statistic $\|\mathbf{Y}\|^2$, which considers the entire $\mathbf{Y}$, by testing only the first m components, determining m by

$$\hat{m} = \underset{m:1 \leq m \leq n}{\arg\max} \left\{ m^{-1/2} \sum_{i=1}^{m} (\mathrm{Y}_i^2 - 1) \right\}.$$

The form of the adaptive Neyman test statistic is

$$\begin{aligned} T^*_{AN} &= (\sqrt{2\hat{m}})^{-1}\sum_{i=1}^{\hat{m}}(Y_i^2-1) \\ &= \max_{1\le m\le n}\left\{(\sqrt{2m})^{-1}\sum_{i=1}^{m}(Y_i^2-1)\right\}, \end{aligned}$$

which is equivalent to rejecting H_0 when

$$T_{AN} = \sqrt{2\log\log n}T^*_{AN} - \{2\log\log n + 0.5\log\log\log n - 0.5\log(4\pi)\}.$$

The authors provide a table of simulated distribution quantiles in their paper.

To compare two sets of curves, they suggest the following process. Assume the first group of curves are a random sample from

$$Y_{1j}(t) = f_1(t) + \epsilon_j(t), \qquad t = 1, \ldots, T, \qquad j = 1, \ldots, n_1,$$

and the second group of curves are a random sample from

$$Y_{2j}(t) = f_2(t) + \epsilon'_j(t), \qquad t = 1, \ldots, T, \qquad j = 1, \ldots, n_2.$$

The errors $\epsilon_j(t)$ and $\epsilon'_j(t)$ are assumed to have mean 0. It is assumed that the time observations t are equally-spaced, in order to facilitate the Fourier transform applied later. The hypotheses of interest are

$$H_0 : f_1(t) = f_2(t) \qquad \text{vs} \qquad H_1 : f_1(t) \neq f_2(t).$$

For the case of independent errors, it is assumed for all j and t that $\epsilon_j(t) \sim N(0, \sigma_1^2(t))$ and $\epsilon'_j(t) \sim N(0, \sigma_2^2(t))$. Then, define

$$\overline{Y}_{1\cdot}(t) = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j},$$

$$\overline{Y}_{2\cdot}(t) = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j},$$

$$\hat{\sigma}_1^2(t) = \frac{1}{(n_1-1)}\sum_{j=1}^{n_1}\{Y_{1j} - \overline{Y}_{1\cdot}(t)\}^2,$$

and

$$\hat{\sigma}_2^2(t) = \frac{1}{(n_2-1)}\sum_{j=1}^{n_2}\{Y_{2j} - \overline{Y}_{2\cdot}(t)\}^2.$$

The standardized difference is

$$Z(t) = \frac{\overline{Y}_{1\cdot}(t) - \overline{Y}_{2\cdot}(t)}{\sqrt{\frac{1}{n_1}\hat{\sigma}_1^2(t) + \frac{1}{n_2}\hat{\sigma}_2^2(t)}}, \tag{3.4}$$

leading to $\mathbf{Z} = (Z(1), \ldots, Z(T))^T$. The next step is to apply the Fourier transform, obtaining the vector $\mathbf{Z}^*$. Then the adaptive Neyman test is carried out on the vector $\mathbf{Z}^*$.

For the case where $\epsilon(t)$ and $\epsilon'(t)$ are stationary error processes, Fan and Lin [1998] suggest performing Fourier transforms for each individual curve before running the testing procedure, which they argue gives approximately independent Normal errors. The rest of the adaptive Neyman test procedure can then be applied to the transformed data, with one modification — it is no longer necessary to run the Fourier transform on the vector $\mathbf{Z}$. Note that there is no explicit structure imposed on the stationary error process; it is not necessary to model the covariance structure in order to run the testing procedure.

3.4 Functional $\boldsymbol{F}$ Test

Shen and Faraway [2004] focus on developing a test for the functional regression model similar in form to the traditional F test that was not sensitive to the number of points sampled along the curve. They are motivated by a collection of ergonomics data.

3.4.1 General Case

Consider first the more general case discussed by Shen and Xu [2007], where we ignore groups and have $n = n_1 + n_2$, where $i = 1, \ldots, n$. Shen and Xu [2007] model the relationship between the functional response, with the form $Y_i(t),\ i = 1, \ldots, n,\ t \in [a, b]$, and a vector of predictors, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$. We write these functions are written in terms of time, t, which is the usual convention. The regression model is of the form:

$$Y_i(t) = \mathbf{x}_i^T \boldsymbol{\beta}(t) + \epsilon_i(t), \tag{3.5}$$

with unknown parameter functions $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^T$, where the $\epsilon_i(t)$ follow a Gaussian stochastic process with mean 0 and covariance function $\gamma(s, t)$. The errors $\epsilon_i(\cdot)$ and $\epsilon_j(\cdot)$ are assumed to be independent for $i \neq j$. Note that the covariance function is not assumed to have any specific structure.

To carry out the functional regression, Shen and Xu [2007] estimate the unknown functions $\boldsymbol{\beta}(t)$ by minimizing $\sum_{i=1}^{n} \|Y_i - \mathbf{x}_i^T \boldsymbol{\beta}\|^2$, where $\|f\| = (\int f(t)^2 \mathrm{d}t)^{1/2}$ is the L_2 norm of the function $f = f(t)$. This results in the least squares estimates $\hat{\boldsymbol{\beta}}(t) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}(t)$, where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is the $n \times p$ model matrix (as in linear regression models) and $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))^T$ is the vector of responses. The predicted responses are $\hat{Y}_i(t) = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(t)$, with residuals $\hat{\epsilon}_i(t) = Y_i(t) - \hat{Y}_i(t)$. The residual sum of squares is

$$rss = \sum_{i=1}^{n} \|\hat{\epsilon}_i\|^2 = \sum_{i=1}^{n} \int \hat{\epsilon}_i(t)^2 \mathrm{d}t. \tag{3.6}$$

Shen and Xu [2007] discuss how to compare two nested linear models ω and Ω, where ω has q parameter functions, Ω has p parameter restrictions, and ω is a subset of Ω (that is, $p > q$). The functional F test defined by Shen and Faraway [2004] is of the form

$$F = \frac{(rss_\omega - rss_\Omega)/(p-q)}{rss_\Omega/(n-p)},$$

where rss_ω and rss_Ω are residual sums of squares under models ω and Ω calculated according to equation (3.6).

Shen and Faraway [2004] derive an approximation for the distribution of this F test. They suggest using an ordinary F distribution with $df_1 = \lambda(p-q)$ and $df_2 = \lambda(n-p)$, where the degrees of freedom adjustment factor, λ, is defined as

$$\lambda = \left(\sum_{k=1}^{\infty} \lambda_k \right)^2 \Big/ \sum_{k=1}^{\infty} \lambda_k^2,$$

where the λ_i are the eigenvalues of the covariance function.

Shen and Xu [2007] discuss how to use the functional F test to perform tests and calculate the p-value for stepwise model selection, testing $\beta_k(t) \equiv 0$. To carry out this test, fit the reduced model without the kth covariate. The form of the test is

$$F_k = \frac{rss_k - rss}{rss/(n-p)},$$

where rss_k is the residual sum of squares under $\beta_k(t) \equiv 0$. More simply, Shen and Faraway [2004] showed that F_k can be calculated using

$$F_k = \frac{\|\hat{\beta}_k\|^2}{(rss/(n-p))(\mathbf{X}^T\mathbf{X})^{-1}_{kk}} = \frac{\int \hat{\beta}_k^2 \mathrm{d}t}{(rss/(n-p))(\mathbf{X}^T\mathbf{X})^{-1}_{kk}},$$

where $(\mathbf{X}^T\mathbf{X})^{-1}_{kk}$ is the kth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$, $\hat{\beta}_k(t)$ is the estimate of $\beta_k(t)$, and rss is the residual sum of squares under the full model. This form of the test is more convenient as it does not require the fitting of a new model for each of the k parameters. Under the null hypothesis, F_k is approximately distributed as an F distribution with degrees of freedom $df_1 = \lambda$ and $df_2 = \lambda(n-p)$, where λ is the degrees of freedom adjustment factor.

3.4.2 Method Using Equally-Spaced Time Points

When data are collected at evenly-spaced fixed time points, $t_1, \ldots t_M$, Shen and Xu [2007] suggest a simplification. In this case the model equation (3.5) becomes

$$Y_i(t_m) = \mathbf{x}_i^T \boldsymbol{\beta}(t_m) + \epsilon(t_m),$$

for $i = 1, \ldots, n$, $m = 1, \ldots, M$. Then instead of integration, summation is used, with $\|\hat{\epsilon}_i\|^2 = \sum_{m=1}^{M} \hat{\epsilon}_i(t_m)^2/M$ and $\|\hat{\beta}_k\|^2 = \sum_{m=1}^{M} \hat{\beta}_k(t_m)^2/M$. The covariance function $\gamma(s,t)$ is estimated by the empirical covariance matrix $\hat{\mathbf{\Sigma}} =$

$(\sum_{i=1}^{n} \hat{\epsilon}_i(t_k)\hat{\epsilon}_i(t_l)/(n-p))_{M\times M}$. The degrees of freedom adjustment factor can be estimated by

$$\hat{\lambda} = \text{trace}(\hat{\mathbf{\Sigma}})^2/\text{trace}(\hat{\mathbf{\Sigma}}^2).$$

The authors state that large degrees of freedom ($n-p \geq 30$) are desired for a good estimation of λ. They also state that if the data are collected at different time points for different i, smoothing techniques can be used to obtain fixed time points.

3.5 Discussion

In this chapter we have introduced several tests in detail. The tests presented here differ in their approach to measuring and testing differences between curves. We briefly discuss the relative benefits of the various measures.

The advantage of the pairwise permutation approach is that there is no risk that the differences between the curves will be obscured by a summary measure that does not capture the essence of any on the individual curves. These are exact tests that do not depend on a distribution. Therefore, they are more flexible and more widely applicable than many other tests. The data do not have to follow a specific structure in order for the test to work. Within the pairwise permutation framework we have offered many combinations of distance measures and summary measures. We expect that the performance of the various tests within this framework depends on the nature of the problem begin studied. For example, we would expect tests based on comparing the area under the curves to do poorly in detecting a location shift, but do well in detecting level shifts between two sets of curves.

The permutation tests proposed by Sturino et al. [2010] take the opposite order of calculating the distance and summary measures. We are curious to see whether the order matters. Because they do permutations only on the summary curves, these tests will run faster than the pairwise permutation tests. The tests proposed by Sturino et al. [2010] have the same advantages other permutation tests in that they are exact and distribution free.

As discussed when it was presented, Ramsay et al.'s [2009] permutation test is similar in spirit to the test based on the L_∞ norm, therefore we expect that it will perform similarly to the other permutation tests using that measure. This test uses only a sample of the permutations (in fact, it samples with replacement). However, we do not have any guidelines as to how large a sample to take to obtain an accurate test with adequate power.

The test introduced by Sturino et al. [2010] is based on reducing the amount of data from a set of curves into a set of principal components scores. This is a popular data-reduction technique applied to a hypothesis test in the functional setting. This test will run much faster than any permutation test, which is an attractive characteristic. One issue with this test is that we do not have any indication as to how many principal components to include, nor how sensitive the test is to using too few or too many principal components.

The adaptive Neyman test developed by Fan and Lin [1998] uses the Fourier transformation as its expansion method. Consequently, the test is going to perform well insofar as the Fourier transform gives a good representation of the curves. This is most likely the case for periodic functions. In other situations, with curves not well described by the Fourier transform, we do not expect that the test will perform well.

The functional F test is an intriguing adaptation of the scalar case to functional data. Again we have the same issues that arise with the adaptive Neyman test, because this test assumes a specific model structure to the data, and if this model does not accurately describe the data our inferences will be affected. An issue we will have is that we are dealing with small samples and this test needs approximately 30 degrees of freedom to accurately calculate the degrees of freedom adjustment factor. With the small sample sizes used in this dissertation we will be unable to accurately calculate this value and this may adversely affect the performance of this test.

CHAPTER

FOUR

MOTIVATING EXAMPLE

In this chapter we use a set of data generated from a 2^k factorial experiment and explore several approaches for analyzing the data. We carry out analyses from two points of view. In the first approach, we ignore the functional nature of the data and base the analysis on a series of F tests at points along the curve. We discuss some graphical techniques that illuminate the behaviour of the effects. We have not seen this strategy outlined anywhere in the literature with respect to the analysis of functional data. Then, we apply the tests designed for comparing curves, outlined in Chapter 3, to the data, testing each effect by dividing the data into two groups based on the effect. We introduce a graphical technique that displays the interaction effects over the index variable. We offer some comments on the non-functional and functional approaches but leave a detailed discussion of the latter to the next chapter, following the simulation studies.

4.1 Description of NASA Experiment

As part of the next generation of space exploration, NASA is developing new spacecraft to transport humans and cargo into space. The Ares I Crew Launch Vehicle (CLV) is one component of the program which has since been cancelled [NASA, 2012]. One segment of the development of the CLV is an analysis of the aerodynamic performance of the launch abort system.

A full 2^4 screening computer experiment was conducted by NASA to examine the effects and importance of four factors: Tower Length (A), Tower Diameter (B), Tip Fineness Ratio (C), and Tip Shape (D), on the response, Drag. The experiment was run at 10 different Mach numbers, in unequally-spaced intervals from Mach 0.7 to 4.0. Thus, for each combination of factor levels there are 10 responses, one for each of the 10 Mach numbers. These points can be interpolated, resulting in a curve as the response. The data, provided by Peter Parker of NASA, are included in a table in Appendix A and are plotted as points joined by straight lines in Figure 4.1.

The purpose of this chapter is to use the NASA data as a case study and apply some of the techniques discussed in Chapter 3. All of the analysis is carried out using the R statistical software package [R Development Core Team, 2011].

4.2 Cross-Sectional Analysis

As a very simplistic method one can fit a model at a chosen Mach number and take this model as a representation of the entire process over all Mach numbers. This approach is only applicable if the system does not change over Mach number. A

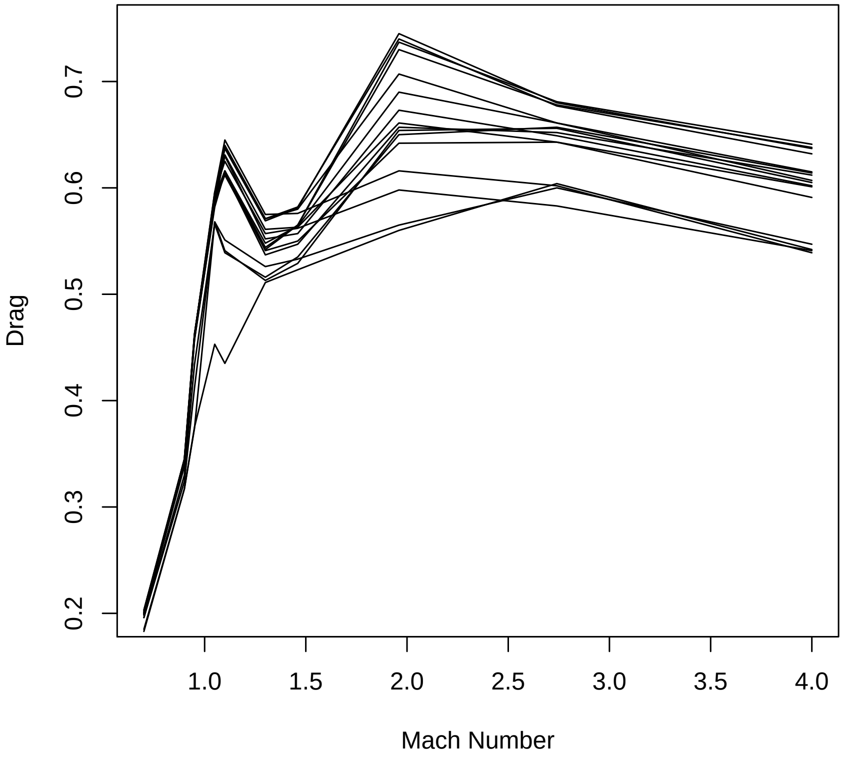

Figure 4.1: Response data plotted as points joined by straight lines.

glance at the response curves displayed in Figure 4.1 shows that drag clearly changes as Mach changes, and therefore this approach is not appropriate in this situation.

4.2.1 Pointwise Analysis at Each Mach Number

A better approach is to conduct a separate regression analysis at each Mach number. This approach is only applicable if the data are collected at the same Mach numbers for each response. If this is not the case, it is possible to smooth the data and sample at the same Mach number for each curve. However, if undertaking this process one might just be tempted to treat the responses as curves, as we will do in the next section.

In any case, here regressions were run at each Mach number, in each case reducing the model to the significant effects in a step-wise fashion. A summary of these regressions with the significant effects, their F statistics, and corresponding p-values can be found in Table 4.1. Note that Int refers to the intercept of the models.

In general, the results are fairly consistent. The effects C, D, and CD appear in most of the models. The main effect B also appears quite consistently. Other interactions, such as BC and BD, seem to come and go, while the rest of the effects do not appear in most models. Mach level 6 seems to be a special case, as the model contains far more significant effects there than at other levels.

4.2.2 Graphical Approaches

We can also present these results in graphical form. The goal of these plots is to illustrate which effects are large at each Mach level, and to indicate the size of the

	Mach Level									
Factor	1	2	3	4	5	6	7	8	9	10
Int	0.198	0.337	0.445	0.577	0.599	0.546	0.557	0.664	0.645	0.598
	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
A						-0.007	-0.004	-0.005		0.003
						< 0.01	< 0.01	< 0.01		< 0.01
B	-0.001	-0.002				-0.001	-0.006	-0.043	-0.023	-0.026
	0.07	0.03				0.02	< 0.01	< 0.01	< 0.01	< 0.01
C	0.003	0.006	0.016	0.010	0.025	0.006	0.007	0.034	0.017	0.022
	< 0.01	< 0.01	< 0.01	0.17	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
D	-0.002	-0.005	-0.016	-0.014	-0.029	-0.013	-0.011	-0.009		0.002
	< 0.01	< 0.01	< 0.01	0.08	< 0.01	< 0.01	< 0.01	< 0.01		< 0.01
AB						0.002				
						< 0.01				
AC						-0.004	-0.003			
						< 0.01	0.05			
AD						0.004	0.003			
						< 0.01	0.03			
BC	0.002					-0.003	-0.004		0.007	0.008
	< 0.01					< 0.01	0.01		< 0.01	< 0.01
BD	-0.002					-0.001				0.002
	< 0.01					0.02				0.01
CD	0.002	0.004	0.015	0.014	0.028	0.011	0.0010	0.014		0.002
	< 0.01	< 0.01	< 0.01	0.08	< 0.01	< 0.01	< 0.01	< 0.01		0.02
ABC										
ABD						0.001				
						0.02				
ACD						-0.003				
						< 0.01				
BCD	0.002									
	< 0.01									

Table 4.1: Pointwise results: estimated effects with p-values below.

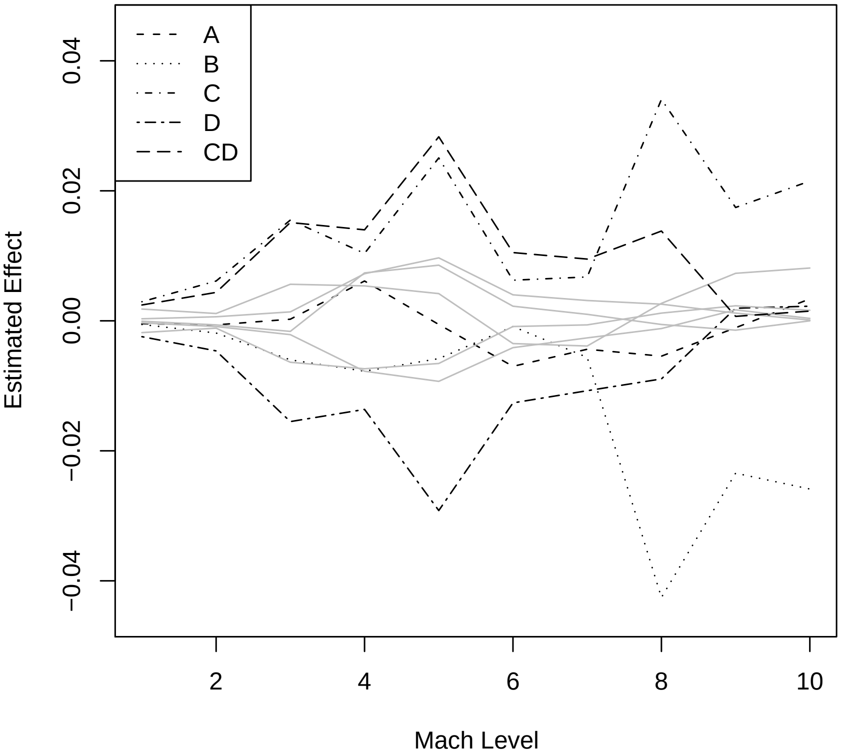

Figure 4.2: Estimated effects for each factor at each Mach level.

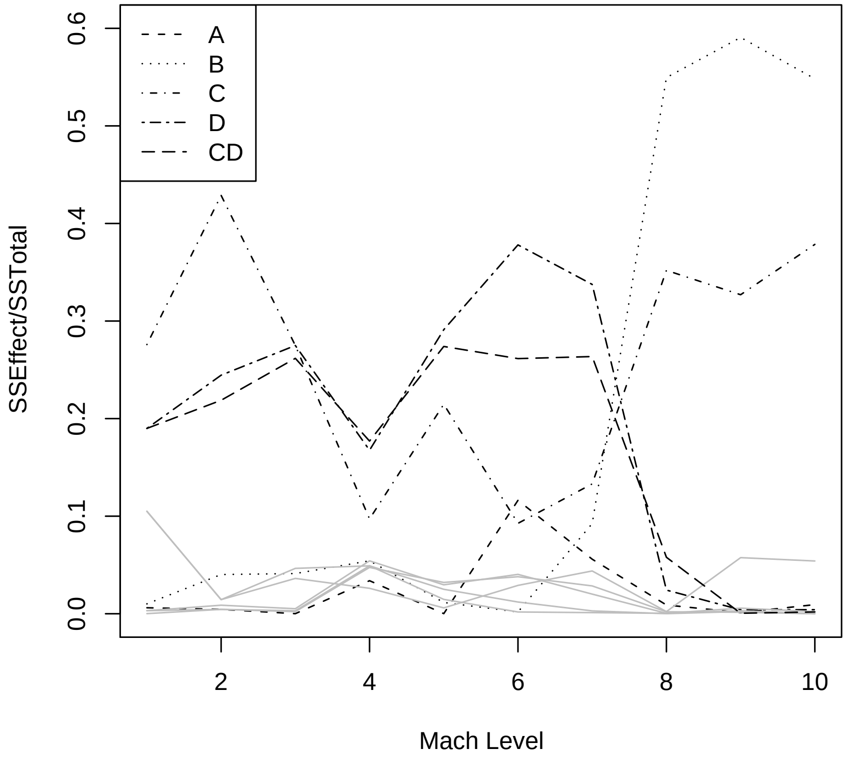

Figure 4.3: Ratio of effect sum of squares to total sum of squares.

effects as Mach level changes. To produce these plots separate regressions were run at each Mach level, with all models containing the four main effects and six two-way interactions. The magnitude of the estimated effect at each Mach level is plotted in Figure 4.2, while Figure 4.3 displays the ratio of the sum of squares of the effect to the total sum of squares. The largest effects are highlighted with different line types. Other effects are included as grey lines.

These graphs are quite enlightening. The main effect A does not explain much of the variation in the models. Its highest percentage of explained variation is approximately 10% at the sixth Mach level. This demonstrates that A is not an important factor in the models. In contrast, effect C has consistently high percentages of explained variation, although the effect seems to explain less of the variation at the middle Mach levels, and more at the lower and higher extremes. Effects B and D are intriguing. Effect B is not important at the lower Mach levels, but increases in importance at the seventh Mach level and explains much of the variation at levels eight through ten. On the other hand, effect D shows the reverse pattern: it is more important at the lower Mach levels and explains very little of the variation at Mach levels eight through ten. This graph also reveals a great deal about the interactions, namely that the CD interaction is the only interaction term of influence. It is interesting to note how the CD effect drops off at the highest three Mach levels.

Figure 4.4 is a plot of the residuals generated by running regressions at each Mach level. There is a great deal more variation in the residuals at Mach levels 3, 4, and 5, corresponding to Mach 0.95, 1.05, and 1.10. This suggests that the models do not fit as well at these Mach levels. From an engineering point of view, this variation is explainable due to the changes in aerodynamic phenomena associated

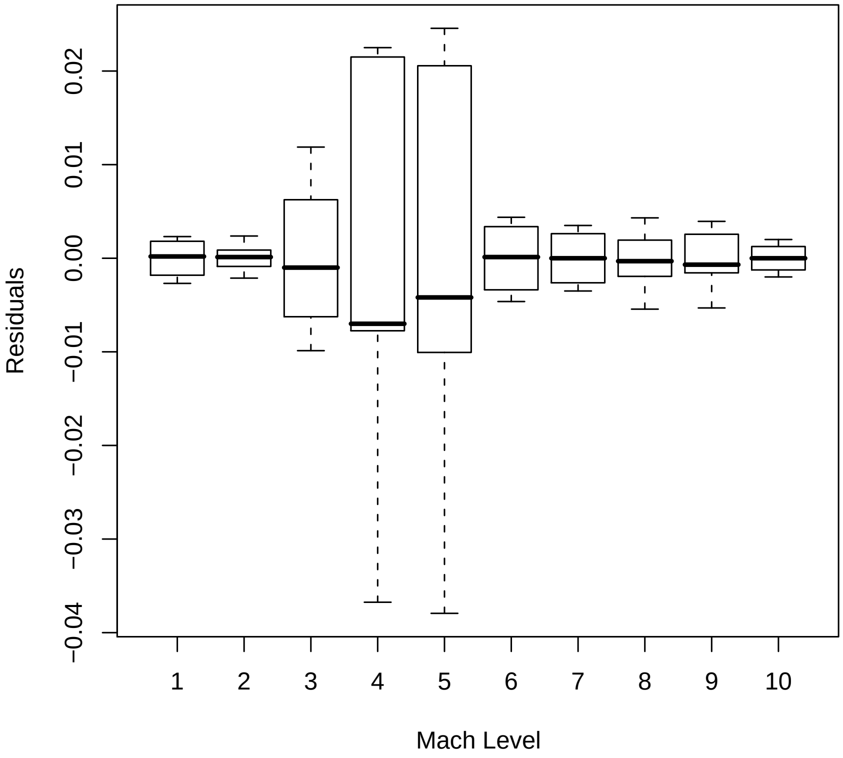

Figure 4.4: Residual boxplots.

with the transition from subsonic to supersonic speed, and may indicate the need for a more complex model (personal communication with Peter Parker).

The cross-sectional analysis is insightful, but it does not account for the fact that the responses are curves and there are some issues with treating the responses as individual sets of points. The first issue is that these sets of points are not independent, but rather are correlated along the response. We would like to account for this dependence in the analysis. Another issue is that we do not get an overall picture of the significance of the effects, instead we have ten separate groupings of significant variables, not all of which are the same. Certainly the separate tests are informative, but they do not summarize the overall situation. To fully exploit the functional characteristics of the data, we need to consider other approaches. We begin in the next section.

4.3 Smoothing

Our data set consists of a series of points, with points collected over the range of a variable, in our case Mach number, but which is often time. How do we convert these points into curves? The answer is smoothing. We fit splines to the data points using the **interpSpline** function in the splines R package [R Development Core Team, 2011]. The splines fit a smooth curve to the points; this curve may then be used as the basis for testing. If necessary, we are able to sample points from the curves. We may sample only a few points or many, depending on our requirements. In this manner it is possible to obtain equally-spaced points even when the data itself has unequally-spaced points.

One sees, then, why those who deal with functional data as functions do not concern themselves about missing data along a curve. We literally smooth away the problem, generating curves and producing the points we require, if necessary.

The smoothed NASA curves are plotted in Figure 4.5. These are the curves we will use in the rest of this chapter.

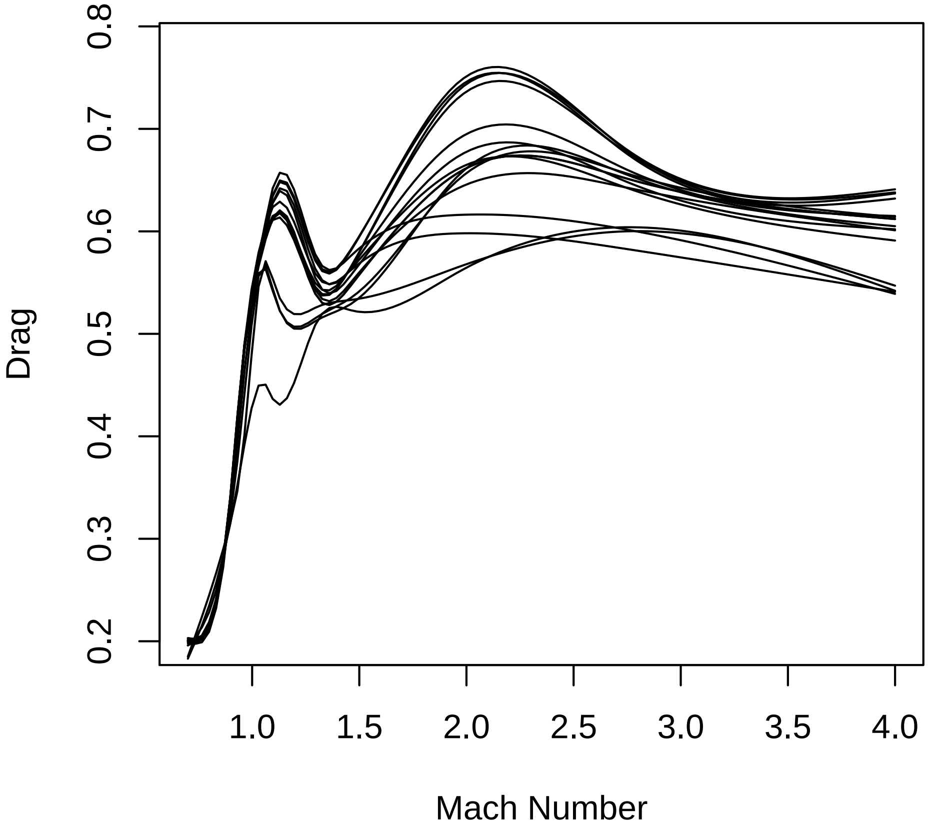

Figure 4.5: Smoothed response curves.

4.4 Analysis with Curves as Functions

We have 16 curves and 15 effects. We can test for the significance of an effect by separating the data into two groups of eight curves based on the high and low values of the effect of interest. Then we apply the techniques for comparing curves on the data as two groups. We continue in this manner for each effect in turn. By carrying out this procedure we obtain assessments of significance for each effect. In this section we apply each testing procedure and present the results.

4.4.1 Permutation Tests

The idea of a permutation test was introduced in Section 2.4. We will use two of the permutation tests here as examples, the mean of the pairwise integrated squared differences between curves and the mean of the pairwise integrated absolute differences between the curves.

Mean of Squared Differences

Consider the permutation test based on the mean of the pairwise integrated squared differences between the curves. We calculate the value of this statistic for each effect. These effect sizes are plotted in a dot chart in Figure 4.6 and the p-values calculated using the permutation test are provided on the right hand side of the plot for each effect. From this plot it is easy to see that effects D and CD have the largest calculated pairwise integrated squared difference and these effects correspond to significant p-values of 0.016 and 0.015, respectively. Effect C is the next largest; the p-value in this case is 0.476, which is not significant. However, since there is a

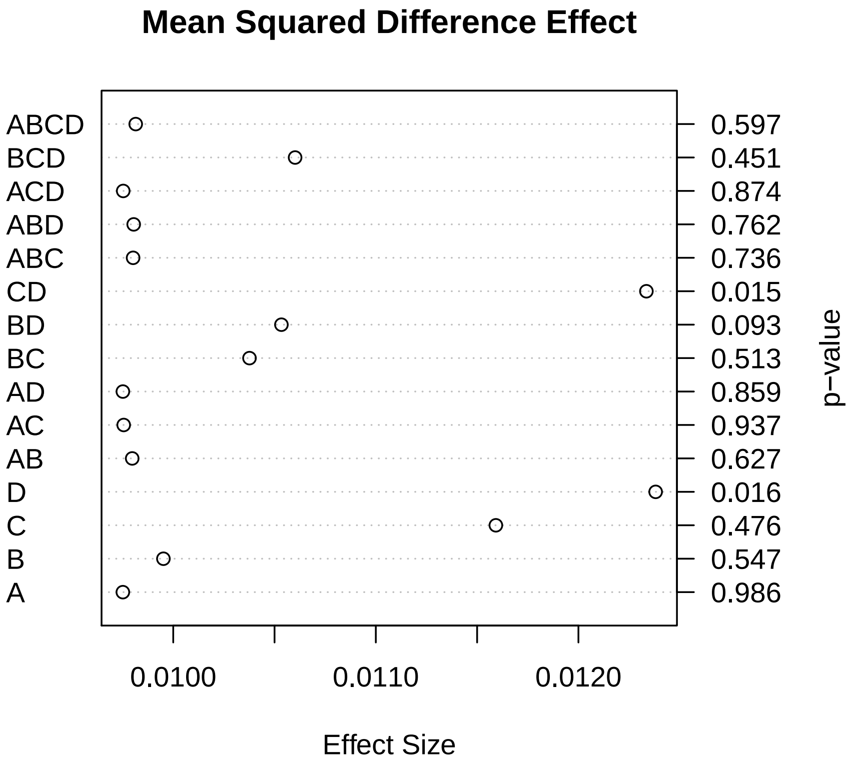

Figure 4.6: Dot plot and *p*-values of NASA experiment effects, calculated using pairwise integrated squared differences.

significant CD interaction the effect C should be considered important because of the hierarchy principle.

Mean of Absolute Differences

Consider also the mean of the pairwise integrated absolute differences between the curves. The results here are consistent with the mean pairwise integrated squared differences; the effect C has a larger test statistic but still tests as not significant.

Recall also that we may apply the methodology of the permutation test to any test statistic we choose. We do so in the next section.

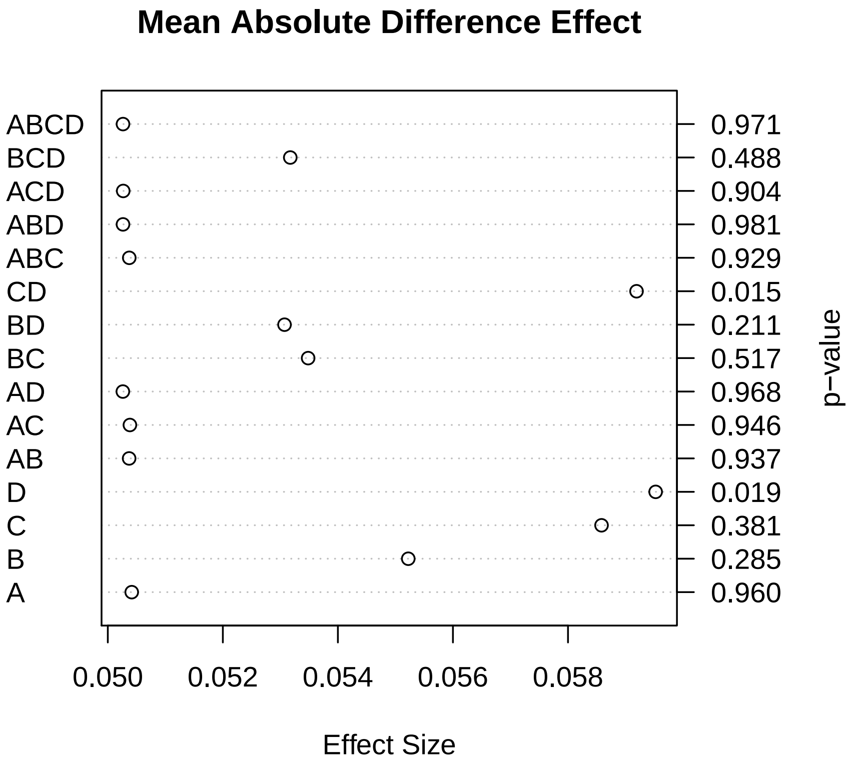

Figure 4.7: Dot plot and p-values of NASA experiment effects, calculated using pairwise integrated absolute differences.

4.4.2 A Functional Response Model

Shen and Xu [2007] propose a functional F test that accounts for the functional nature of the responses. This test is similar in spirit to the usual F test, and it is used to test nested designs as discussed on page 56.

The data are not equally spaced, so there are two options: 1) to use the data as it is and apply the F test method using integrations instead of the method using summations, or 2) to smooth the data and choose equally-spaced points from the smooths. We choose the latter option since we are smoothing the curves for other purposes as well. One question of interest is how much of an effect smoothing has on the power of this F test. This is a point for further study.

We smooth the data and sample 32 equally-spaced points from the smooths. These points are used to assess signficance in two ways. The first method proceeds as outlined in Section 3.4, using the parametric assumptions for the test, that is, using the F distribution to calculate the p-values. The second method uses the functional F statistic as the basis for a permutation test, as discussed in Section 3.1.3. We call the former the *parametric* version of the functional F test, and the latter the *permutation* version of the functional F test. This will be the convention for our terminology with the adaptive Neyman test and the test based on principal components as well.

The F statistics and p-values for both the functional F tests and permutation tests are displayed in Figure 4.8. From the dot chart we see that B, C, D and CD have the largest F test statistics. According to the parametric test p-values, all four effects are significant with small p-values. The permutation test p-values show the

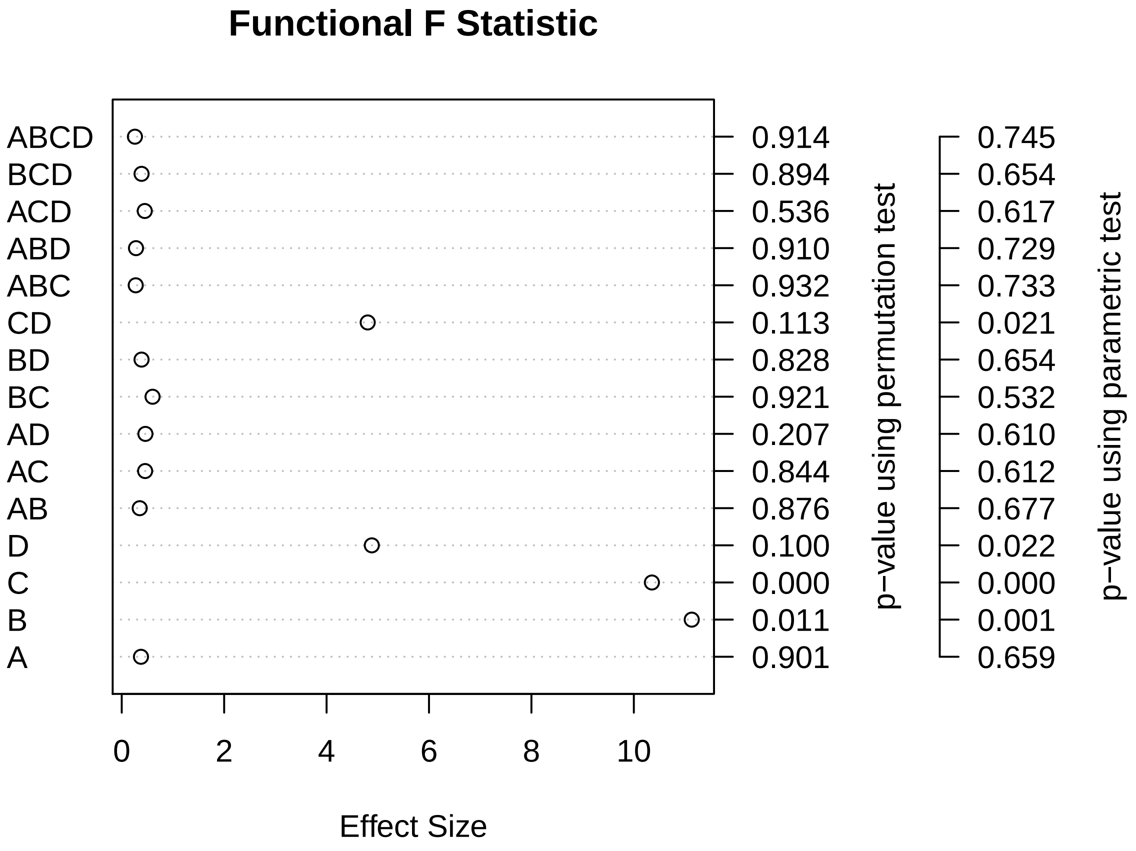

Figure 4.8: Dot plot and p-values of NASA experiment effects, calculated using functional F test, with p-values calculated using the parametric normal assumption and the permutation test technique.

three main effects B, C, and D as significant at $\alpha = 10\%$ but the CD interaction p-value is 0.113.

A final point is that the degrees of freedom were small for this experiment, $n - p = 14$. Shen and Xu recommend having $n - p \geq 30$.

4.4.3 Adaptive Neyman Test

The adaptive Neyman test uses the Fourier transform, which requires data collected at equally-spaced points, thus, the first step is to smooth the data. The data are smoothed in the same manner carried out for the functional F test, using the function **interpSpline** [R Development Core Team, 2011], which fits splines to the points. The code used for the adaptive Neyman test was written by Fan and is available on his website [Fan, 1996b].

We then sample 32 points from the curves, apply the Fourier transform, and calculate the adaptive Neyman test statistics for each effect. Again, we may use the adaptive Neyman test statistic for each effect as the basis for a permutation test. The values of the test statistics are plotted and permutation test p-values are displayed in Figure 4.9. This plot also contains a vertical reference line indicating the critical value of the adaptive Neyman test statistic at $\alpha = 10\%$ for 30 points, which is the closest value to 32 available in the table. Here we have an interesting dilemma. On the one hand, based on the critical value for the test statistic, all effects except AD, ABD, ACD and ABCD are significant. On the other hand, according to the permutation test p-values none of the effects are significant. Note that the effect sizes are so large that they are highly significant according to the parametric test regardless of the level of α chosen.

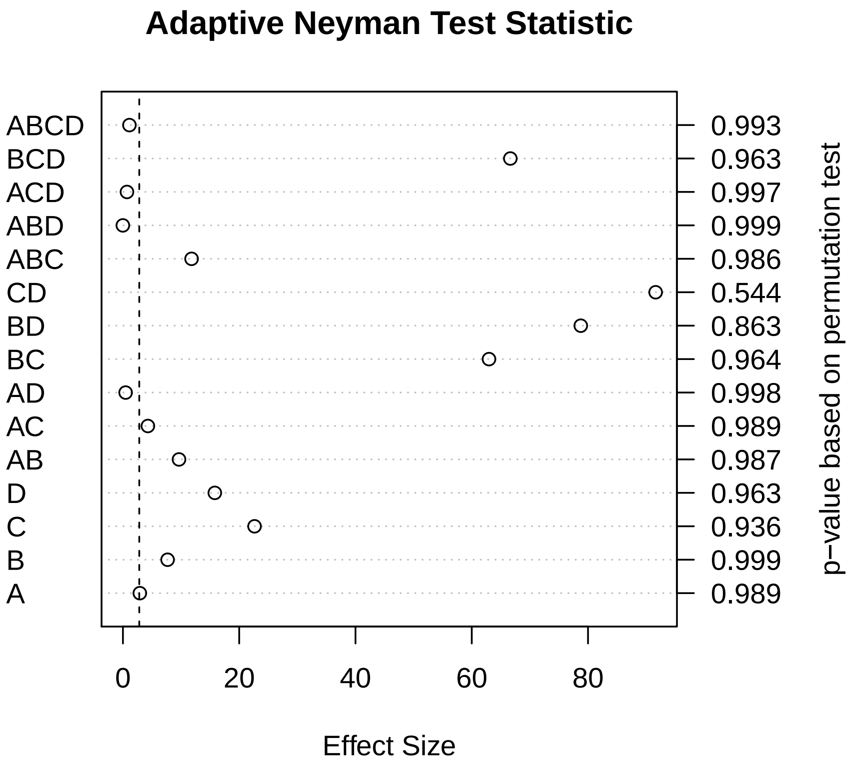

Figure 4.9: Dot plot and p-values of NASA experiment effects, calculated using the adaptive Neyman test. The dashed vertical line is the critical value of the parametric version of the test for $n = 30$, with $\alpha = 0.10$. The p-values are calculated using the permutation version of the test.

4.4.4 Test Based on Principal Components

To carry out the test based on principal components, we first need the curves to be recognized as functional data objects by R. We begin by defining a basis for the curves, using the create.bspline.basis function in the **fda** package [Ramsay et al., 2011]. Then we fit the basis to the points, without doing any additional smoothing than we have already done to the curves. This step gives us a functional data object. We fit two principal components to the functional data object. The first principal component accounts for most of the variation (82%). The second principal component accounts for 15% of the total variation. We use the eigenvalues and eigenfunctions of the principal components to calculate the test statistics.

The test using the normal distribution combines the results of the tests on the first and second principal components (PCs) and rejects the null hypothesis if either the test on the first PC is rejected or the test on the second PC is rejected. The test statistics and p-values for each component are displayed in Figure 4.10. The p-values are calculated both using the normal distribution assumption and the permutation test procedure. The tests on the first principal component reveal B and C to have the largest test statistics, followed to a much lesser extent by CD. According to the parametric test, significant effects are B and C, with p-values of 0.02 and 0.04. The results are similar using the permutation test with the same effects having small p-values.

Examining the results for the second principal component, we see the largest effects are D and CD, followed by effect B. The p-values calculated using the parametric assumptions yield D and perhaps CD as significant effects, with p-values

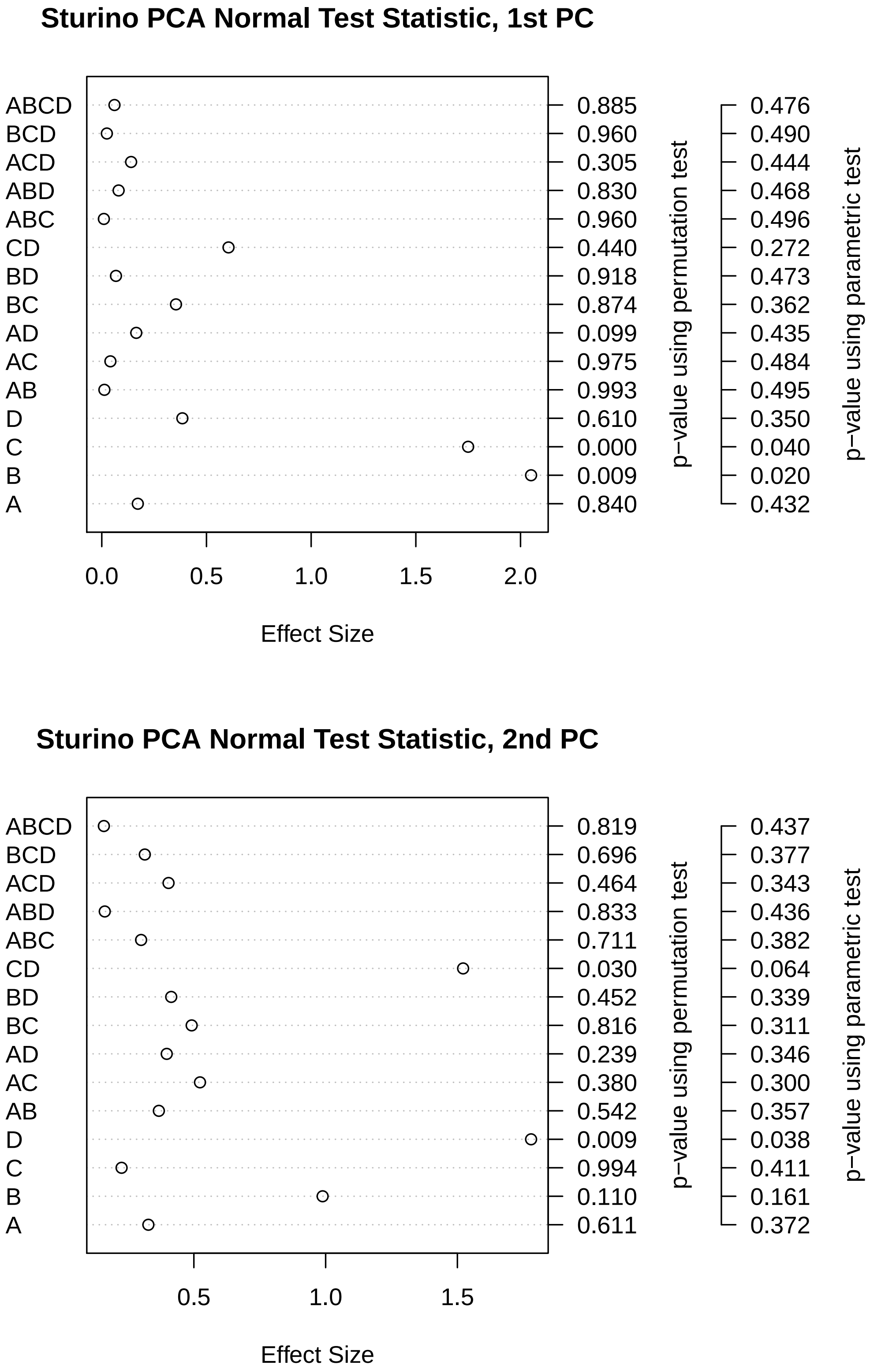

Figure 4.10: Dot plot and p-values of NASA experiment effects, calculated using the principal components test, with p-values calculated using the parametric normal assumption and the permutation test technique.

of 0.038 and 0.064, respectively. The permutation test has the same conclusions, albeit with smaller p-values of 0.009 and 0.030, respectively.

Recall that we reject the null hypothesis if either the first or second component's p-values are $< \alpha/2 = 0.05$ for a 10% level test. Thus, the significant effects are B, C, and D based on the parametric p-values and B, C, D and CD based on the permutation test p-values.

In their article, Sturino et al. [2010] suggest using alternative rejection regions which weight the rejection regions based on the size of the eigenvalues. That is, instead of using the same rejection region for each principal component, we calculate a different region for each PC, based on that PC's eigenvalue. For an $\alpha = 0.10$ level test the rejection regions are $p_1 = 0.08423402$ and $p_2 = 0.01576598$. Comparing these cut-off values to the p-values found in Figure 4.10 based on the parametric version of the test, we have effects B and C significant by the test based on the first principal component and nothing significant based on the test using the second principal component. The permutation p-values indicate that B and C are significant according to the first principal component and D is significant according to the second principal component.

Alternatively we could combine the tests and use the χ^2 test. The test statistics and p-values calculated using both parametric and permutation approaches are shown in Figure 4.11. Effects B, C, D, and CD have the largest test statistics in this case. The significant effect according to the parametric test is B with a p-value of 0.075. The p-values calculated using the permutation test indicate that B (p-value < 0.001), D (0.029), and CD (0.077) are significant at $\alpha = 0.10$.

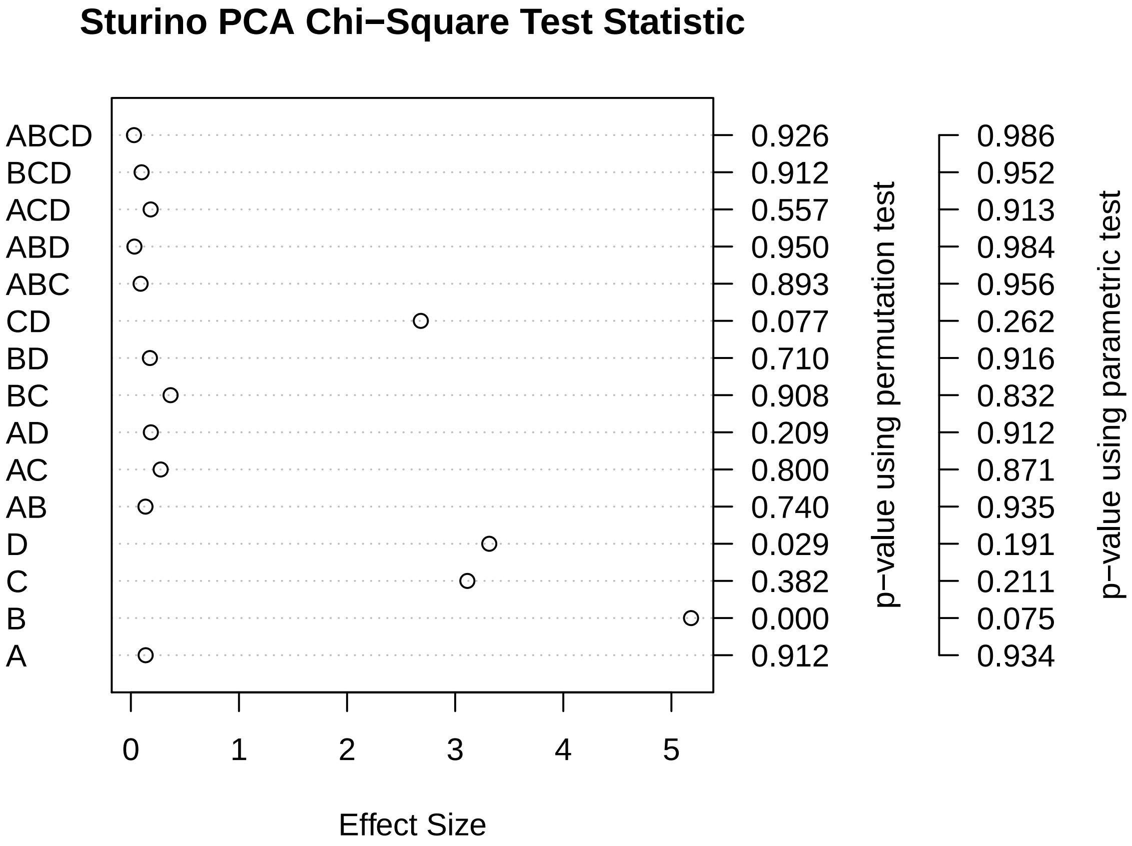

Figure 4.11: Dot plot and p-values of NASA experiment effects, calculated using principal components test using χ^2 option, with p-values calculated using the parametric normal assumption and the permutation test technique.

4.4.5 Summary

The results of these tests are summarized in Table 4.2, which lists the significant effects according to each test and the p-values where appropriate.

Table 4.2: Significant effects according to each test.

Test	Significant Effects (p-value)
Mean pairwise squared difference	D (0.016), CD (0.015)
Mean pairwise absolute difference	D (0.019), CD (0.015)
Functional F test, parametric	B (0.001), C ($<$ 0.001), D (0.022), CD (0.021)
Functional F test, permutation	B (0.011), C ($<$ 0.001), D (0.100)
Adaptive Neyman test, parametric	A, B, C, D, AB, AC, BC, BC, CD, ABC, BCD
Adaptive Neyman test, permutation	(none significant)
PC test, normal, parametric	B, C, D
PC test, normal, permutation	B, C, D, CD
PC test, alt. cut-offs, parametric	B, C
PC test, alt. cut-offs, permutation	B, C, D
PC test, χ^2, parametric	B (0.075)
PC test, χ^2, permutation	B ($<$ 0.001), D (0.029), CD (0.077)

The conclusions differ based on the test used, although there are some consistent patterns. Effects B, C, and D appear regularly, and the CD interaction is the most common interaction showing significance. Note that the pairwise permutation tests detect the CD interaction while several others do not. Not all the tests agree and we would like to know more about the nature of the robustness of the tests under consideration.

4.5 Examining Interactions

In the progression of analysis, it became clear that an interaction effect exists in the data. The CD interaction consistently appears as a significant effect according to several of the testing approaches. It was of interest to explore this effect further.

There is very little discussion in the literature about interaction effects with functional regression. The book by Ramsay and Silverman [2005] does not discuss interactions, nor do Shen and Xu [2007] in the residual analysis techniques they develop. Nair et al. [2002] discuss control-by-noise interactions in robust design studies but do not specifically examine the relationship between two interacting variables in functional regression.

The NASA data plotted in terms of high/low C and D effects are found in Figure 4.12. Another way of viewing the CD interaction effect is to examine the mean curves of each factor combination. These curves, calculated pointwise along a very fine grid of points sampled from the smoothed curves, are plotted in Figure 4.13. The curves have a different overall shape for the C-D+ factor combination, most obvious at the first smaller peak which has a different characteristic at these levels. We also see that the top peaks are lower for C- regardless of the level of D. Note also in this plot that the height of the second peak in the response appears to be related to the level of factor B. When B is high this peak is not as high as it is when B is set to the low level.

Let us consider a way of plotting the interactions as Mach number changes. The starting point is the usual interaction plot constructed in design of experiments. Consider the first Mach level, $y = 0.7$. Referring to the data on page 132, the CD interaction effects are tabled as

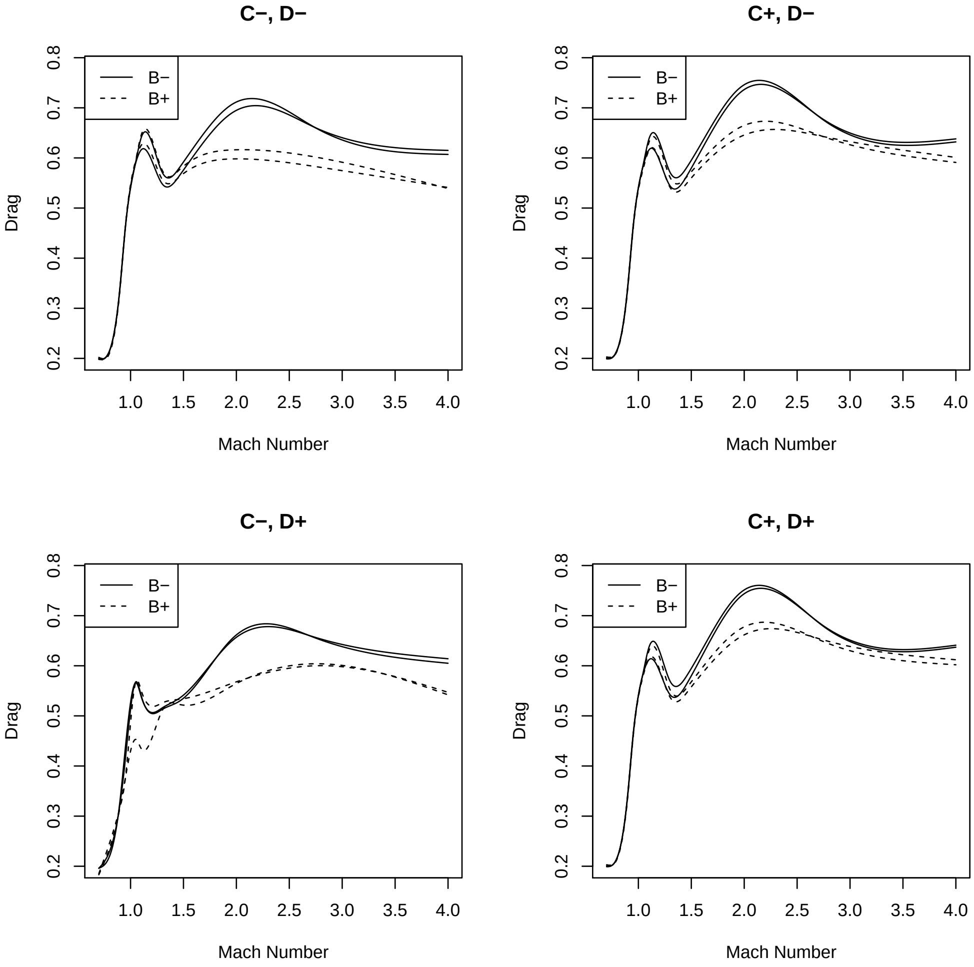

Figure 4.12: Response curves separated by high and low levels of factors C and D, with different line types indicating the level of factor B.

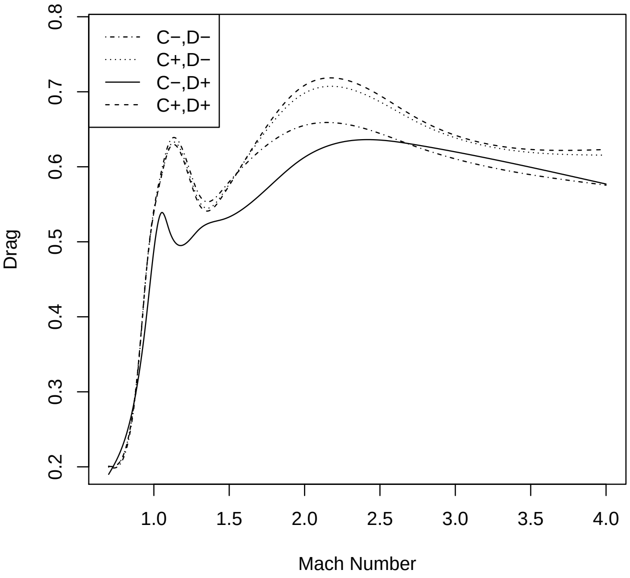

Figure 4.13: Mean response curves, mean calculated over high and low levels of factors C and D.

	D-	D+
C-	0.199	0.196
	0.198	0.196
	0.202	0.183
	0.200	0.185
C+	0.200	0.200
	0.199	0.199
	0.203	0.203
	0.201	0.201

.

The average effects are calculated to be

	D-	D+
C-	0.19975	0.19000
C+	0.20075	0.20075

.

Normally, plotting these numbers is simple because there is only one response level. However, as there are data for ten Mach numbers, this process is repeated at each level of the response. The resulting average effects are plotted on one graph to simultaneously illustrate the interaction effects for all levels. See Figure 4.14.

This plot illustrates many aspects of the interaction effects. Note how the average effects mimic the pattern of the response curves. The interactions are more pronounced at the middle Mach levels than at either tail. At Mach levels 9 and 10, the lines are parallel, indicating that there is no interaction between C and D at these levels. This result is consistent with Figure 4.2, the plot of estimated effects, where the CD interaction effect tailed off at the higher Mach numbers.

In terms of the effect on the response, this plot shows that the effect of D is negligible when C is at its high level. In contrast, D has a large effect when C is at the low level. When C is low, moving from D low to D high reduces the average Drag. The combination of C low and D high produces the least amount of Drag. This outcome is not indicated in any of the other graphs generated in this paper, but may be important from a design point of view.

This graph is a new tool for use with designed experiments having a functional response. It demonstrates the interaction effects in a clear manner and greatly aids in their interpretation. It does not require equally-spaced points; if necessary one may smooth the data to obtain the desired points. It does require curves to be discretized in some way.

4.6 Discussion

We have considered cross-sectional and functional approaches to analyzing a set of functional data. Let us summarize our impressions of each approach.

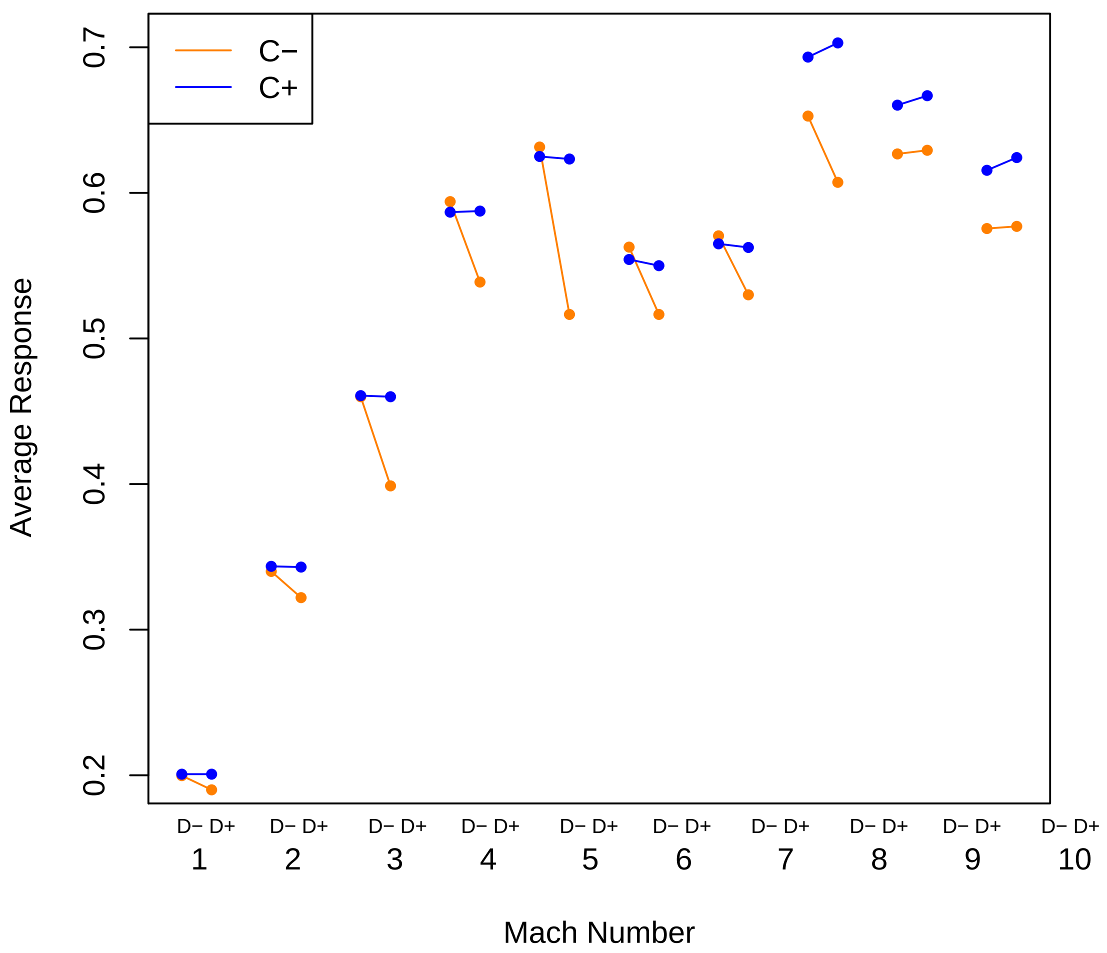

Figure 4.14: CD interaction effects for each Mach level.

4.6.1 The Cross-Sectional Approach

Advantages:

- The pointwise approach is simple and easy to carry out.
- It uses familiar tests.
- Examining the significant effects at each point gives us a good picture of the behaviour of the curve overall and indicates the important effects.
- The effect plot, ratio of effects plot, and boxplots are valuable and informative and give an excellent illustration as to how the effects vary in strength over time.
- The interaction plot allows us to examine the nature of the interactions over Mach number in a way not possible using the functional point of view.

Disadvantages:

- We do not get an overall assessment of significance.
- It is awkward and labour-intensive in that we have to build a model at each point along the curves. If we needed to do a cross-sectional analysis at a large number of points this could get tedious.
- We have some concerns regarding smoothing and its effect on the power of the tests.
- It does not treat the data in its true form, a curve, and does not account for the relationship between points along the curve at each analysis point.

Overall we feel this approach has merit. We find it informative and feel that it reveals a great deal about the nature of the effects in the data. An interesting note is that the points at which to carry out analyses are up to the user. We could potentially choose to carry out the analysis cross-sectionally at points of interest along the curves (for example, at peaks or along a steep climb or fall on the curves). This could give us a picture of interesting features without having to carry out an abundance of tests along data collected at finely-spaced time points. Then the cross-sectional analysis becomes an exploratory technique and formal testing procedures could be carried out using functional testing procedures.

4.6.2 The Functional Approach

With the variety of testing procedures used here, the advantages and disadvantages depend on the test under consideration. We will discuss each test in turn.

Pairwise Permutation Tests

Advantages:

- These tests are simple to understand and calculate.
- No distributional assumptions are needed.

Disadvantages:

- Some programming sophistication is required.
- The time to compute the test may be lengthy. It took 14 minutes to calculate all of the NASA effects based on 16 curves, which is not a large sample size.

Functional F test

Advantages:

- The form of the test is familiar.
- Calculations are quick.
- It is relatively simple to carry out the procedure.

Disadvantages:

- The distributional assumptions may not be true.
- The test is sensitive to the error structure [Shen and Faraway, 2004].

Adaptive Neyman Test

Advantages:

- It is simple to carry out with the code provided.

Disadvantages:

- The test depends on Fourier transform and assumptions which may or may not be appropriate for the data.
- Complex argument is used to develop the test statistic.
- No simple p-value calculation.

Test based on Principal Components

Advantages:

- It is easy to fit functional principal components to the data and carry out the test.

Disadvantages:

- The test as structured treats each principal component separately. It is not clear how many principal components to fit to the data, nor do we know if the procedure is sensitive to having not enough or too many principal components.

The choice between these tests is not clear at this point. Further investigation is needed, which is the basis for the simulation study in the next chapter.

4.7 Conclusions

In this chapter we explored various methods to analyze data with a functional response. The first approach fit separate least-squares regressions at each Mach level. The estimated effects and sums of squares were plotted to illustrate which effects were significant at each Mach level. Boxplots were plotted to show changes in variation across Mach levels.

We then applied a number of testing procedures designed for comparing two groups of curves to a set of data from a 2^4 experiment, including pairwise permutation tests, the functional F test, the adaptive Neyman test, and tests based on principal components. The results differed between the tests but for the most part gave a fairly consistent message about the significant effects.

Having applied these tests, we would now like to know how robust the parametric tests are to deviations from assumptions. How do the powers of the tests compare? Is there a “best” test that has good power and performs well in a variety of situations? Which permutation test is preferred? Does it matter whether the permutation tests are carried out pairwise or on summary curves? These are the questions that will be addressed in the simulation study in the next chapter.

Finally, we attempted to illustrate interaction effects using some graphical techniques. By plotting the interaction effect between two variables in series across all levels of the response, it became simple to characterize the nature of the interaction. This technique is a useful additional tool in the analysis of functional data generated by a designed experiment.

CHAPTER

FIVE

SIMULATION STUDY AND RESULTS

In this chapter we discuss the structure and results of the simulation study. We discuss how we generate curves for the simulations and how we use these curves in simulations. Then we discuss the results of the simulations, examining the power of the test statistics and comparing and contrasting the results over the varying conditions. All simulations are carried out in R [R Development Core Team, 2011]. We wish to make two points very clear regarding the purpose and structure of these simulations. One point of consideration to remember is that our goal is merely to generate a set of smooth curves on which to carry out some statistical tests. We do not require the data to follow a specific model or structure in any way. All we require is the ability to change the shape of the curves to account for the changing alternative hypothesis. Another point is that in developing our process for generating curves, we are in no way attempting to mimic the shape of the NASA curves used in the previous chapter. While the NASA data is a very interesting data set to analyze, we are not concerned with capturing the spirit of this data in the curves we generate. The NASA data set is fairly sparse and requires smoothing while our focus here is on creating a set of smooth curves. That is, we assume that the smooth curves generated through the process are our data. Our objective is to examine the analysis of data where the data are smooth curves.

5.1 Making Curves

The first step in carrying out a simulation study is to generate curves. One possibility is to start with a baseline curve and perturb it in some way. We will explain this multi-step process in this section, and then provide the notation. We restrict the data, without loss of generality, to the interval $[0, 1]$. We generate two types of curves, unimodal and monotone. We begin by describing how we generate unimodal curves.

5.1.1 Unimodal Curves

First, we need to specify a baseline curve. We use a Beta probability density function. This curve is not viewed as a density function but rather just “a curve.” A Beta density is a convenient baseline because it is flexible; it has two shape parameters on the interval $[0, 1]$, allowing level shifts to the curve as well as the ability to change the shape and peak of the curve. Notationally we refer to the baseline parameters as α_1° and β_1°. There is a $_1$ subscript because this is the first group of curves. Later we will specify a second baseline curve with α_2° and β_2° to generate the second group of curves in order to test hypotheses.

Once the baseline curve is defined, the second step is to generate a set of n_1 curves around this baseline curve by perturbing the parameters of the baseline with multiplicative lognormal noise. This results in n_1 Beta curves, with parameters α_{1i} and β_{1i}, $i = 1, 2, \ldots n_1$. [At this stage, noise variability parameters σ^2_{1a} and σ^2_{1b} must be specified.]

These curves are still smooth, since they are simply Beta densities. The third step in the process introduces randomness around the curves, using multiplicative autoregressive errors. [To do this, we need to specify the autoregressive coefficient, τ_1, the error variability parameter, $\sigma^2_{\omega_1}$, and J, the latter giving us $J + 1$ equally-spaced points on [0,1] at which the errors will be introduced.]

Now that we have a set of points that outline n_1 curves following noisy paths, the final step is to fit splines to these points, giving us n_1 smooth curves. In our case we fit splines to the points using the **interpSpline** command in R to obtain the smooths. These curves are our data.

More formally, having specified $\alpha^\circ_1, \beta^\circ_1, \sigma^2_{1a}, \sigma^2_{1b}, \tau_1, \sigma^2_{\omega_1}$ and J, the curves in the first group are given by

$$Y_{1i}(x) = \text{spline}[(x_0, y_{1i0}), (x_1, y_{1i1}), (x_2, y_{1i2}), \ldots, (x_J, y_{1iJ})](x),$$

for $0 \leq x \leq 1$, $i = 1, 2, \ldots, n_1$, where

$$x_j = \frac{j}{J}, \qquad j = 0, 1, \ldots, J,$$

and, for $i = 1, \ldots, n_1$ and $j = 1, \ldots, J$,

$$y_{1ij} = e^{\eta_{1ij}}[\text{dBeta}(x_j, \alpha_{1i}, \beta_{1i})],$$

where

$$\alpha_{1i} = e^{a_{1i}}\alpha^\circ_1, \qquad \beta_{1i} = e^{b_{1i}}\beta^\circ_1,$$

$$a_{1i} \sim N(0, \sigma^2_{1a}), \qquad i = 1, \ldots, n_1,$$

$$b_{1i} \sim N(0, \sigma^2_{1b}), \qquad i = 1, \ldots, n_1,$$

and

$$\begin{aligned} \eta_{1i0} &= \omega_{1i0}, \\ \eta_{1ij} &= \tau_1 \eta_{1i(j-1)} + \omega_{1ij}, \qquad i = 1, \ldots, n_1,\ j = 1, \ldots, J, \end{aligned}$$

$$\omega_{1ij} \sim N(0, \sigma^2_{\omega_1}), \qquad i = 1, \ldots, n_1,\ j = 0, \ldots, J,$$

with all of the a_{1i}'s, b_{1i}'s and ω_{1ij}'s being independent.

Figure 5.1 on page 98 shows (a) the baseline curve, (b) six Beta curves with random α_{1i}'s and β_{1i}'s, (c) the six curves with a layer of autoregressive error added, and (d) the spline fits, which are considered the observed data. The parameter values for this set of curves are: $\alpha^\circ_1 = 2$, $\beta^\circ_1 = 5$, $J = 7$, $\sigma^2_{1a} = \sigma^2_{1b} = \sigma^2_{\omega_1} = 0.09$, and $\tau_1 = 0.7$.

The second group of curves is the comparison group which is generated under the alternative hypothesis conditions. One feature we may want to add to the second group is a mixture of another curve, that is, a second bump, as contamination. This mixture could be generated by anything, but for now let's simply add a proportion p of another Beta density defined by $\alpha^{*\circ}$ and $\beta^{*\circ}$, where these parameter values are generated similarly to the error process used for the main curve's α° and β°, obtaining a collection of α^*_i's and β^*_i's.

The curves in the second group are generated from

$$Y_{2i}(x) = \text{spline}[(x_0, y_{2i0}), (x_1, y_{2i1}), (x_2, y_{2i2}), \ldots, (x_J, y_{2iJ})](x),$$

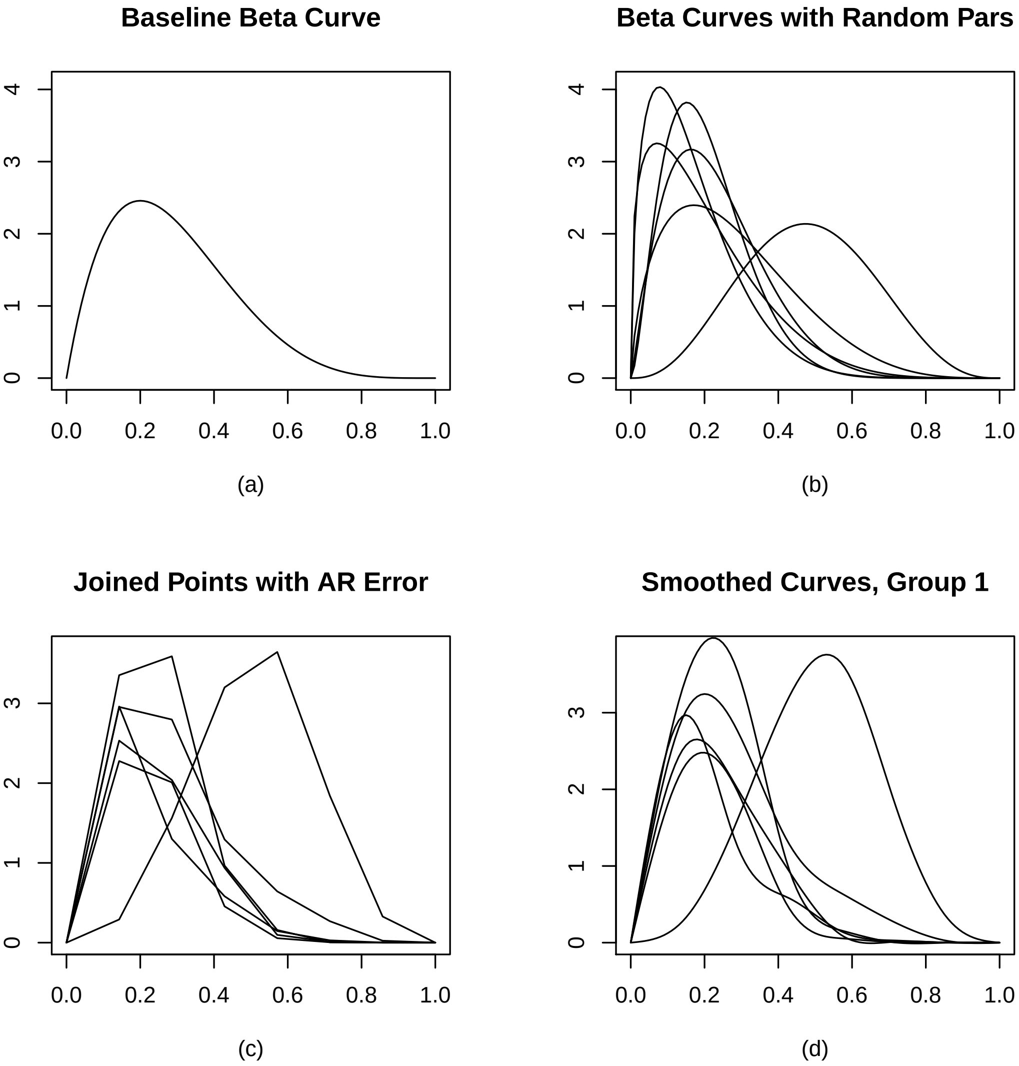

Figure 5.1: Generation of noisy curves for group 1: (a) is the baseline curve, (b) six Beta curves with random α_{1i}'s and β_{1i}'s, (c) the six curves with a layer of autoregressive error, and (d) the spline fits.

for $0 \leq x \leq 1$, $i = 1, 2, \ldots, n_2$, where

$$x_j = \frac{j}{J}, \qquad j = 0, 1, \ldots, J.$$

For $i = 1, \ldots, n_2$ and $j = 1, \ldots, J$,

$$Y_{2ij} = e^{\eta_{2ij}}[(1-p)\text{dBeta}(x_j, \alpha_{2i}, \beta_{2i}) + p\text{dBeta}(x_j, \alpha_i^*, \beta_i^*)]$$

where

$$\alpha_{2i} = e^{a_{2i}}\alpha_2^\circ, \qquad \beta_{2i} = e^{b_{2i}}\beta_2^\circ,$$

$$\alpha_i^* = e^{a_i^*}\alpha^{*\circ}, \qquad \beta_i^* = e^{b_i^*}\beta^{*\circ},$$

$$a_{2i} \sim N(0, \sigma_{2a}^2), \qquad i = 1, \ldots, n_2,$$

$$b_{2i} \sim N(0, \sigma_{2b}^2), \qquad i = 1, \ldots, n_2,$$

$$a_i^* \sim N(0, \sigma_{a^*}^2), \qquad i = 1, \ldots, n_2,$$

$$b_i^* \sim N(0, \sigma_{b^*}^2), \qquad i = 1, \ldots, n_2,$$

(these are all independent). The autoregressive error term is added using the same procedure used for the first group of curves.

To illustrate, to generate a second group of six curves, the entire curve generation process is repeated with the same parameter values, but adding in a $p = 0.1$ mixture of $\alpha^{*\circ} = 4$, $\beta^{*\circ} = 8$. These curves are displayed in Figure 5.2.

To generate the data for simulations, the baseline is fixed, we simulate new curves from it, evaluate a number of measures, and calculate test statistics. The output from the simulations are smooth curves. We consider these curves to be the data. That is, for inference we will not be making use of the model that we know produced the curves. We certainly could, but that is not the focus of this research.

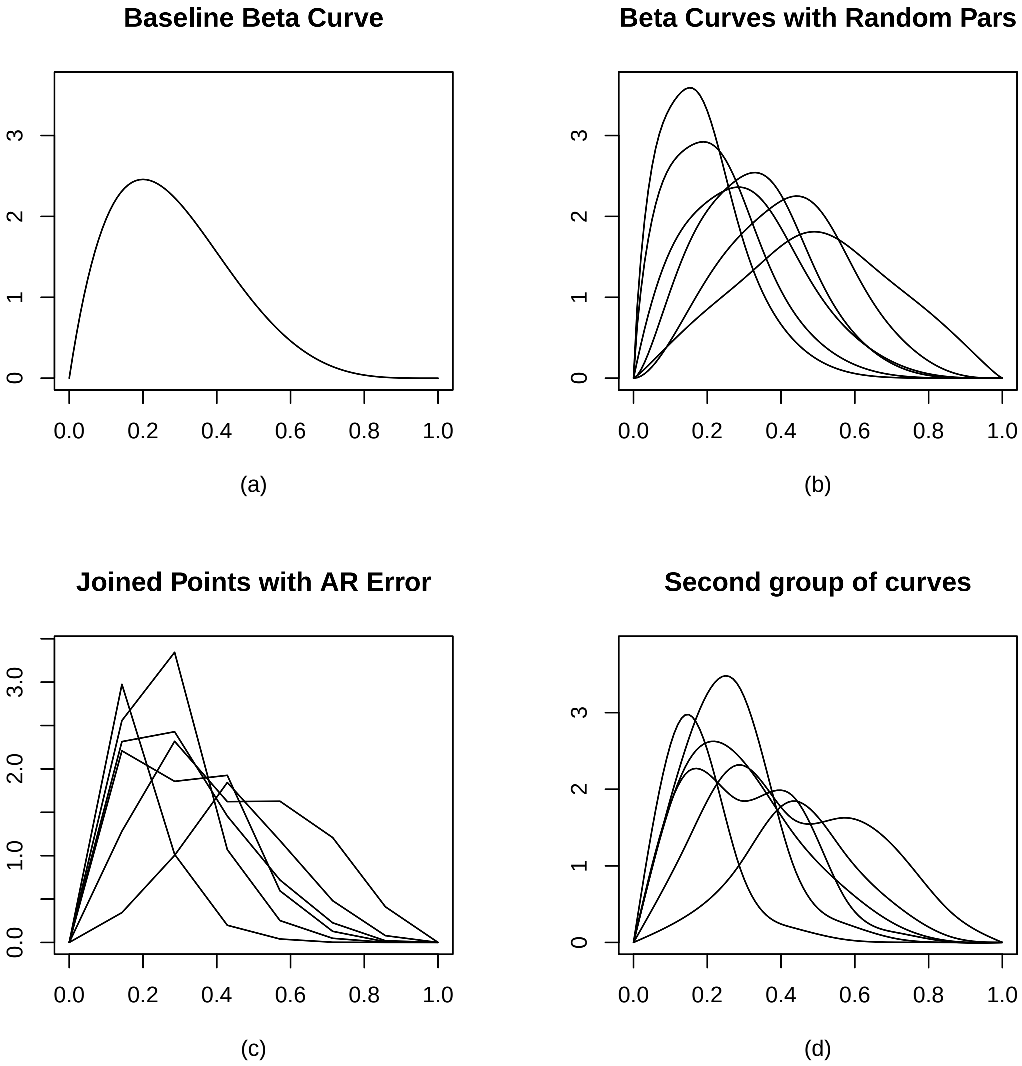

Figure 5.2: Generation of noisy curves for group 2: (a) is the baseline curve, (b) six Beta curves with random α_{2i}'s and β_{2i}'s, (c) the six curves with a layer of autoregressive error, and (d) the spline fits.

5.1.2 Simulation Study

We have discussed how to obtain the data and some methods for testing whether or not two groups of curves are different. Now we will discuss how to compare the methods. The objective is to compare the performance of the various procedures discussed in Chapter 3 in a variety of contexts. We carry out a simulation study that systematically changes the experimental conditions and allows us to assess the performance of the tests as conditions change. The construction of the curve generation developed in the last section is complex but flexible and it can be simplified (by making all of the variances identical, for example).

We set the parameters in the second group in terms of the first group's values, by setting

$$\alpha_2^\circ = \Delta_L \Delta_S \alpha_1^\circ, \qquad \beta_2^\circ = \Delta_S \beta_1^\circ.$$

Then we change the parameters in the simulation by changing the Δ's. We call Δ_L the location parameter because changing it moves the mean of the curve along the x axis. We call Δ_S the shift parameter because it changes the height of the peak of the curve.

The values of the variables are set to: $\alpha_1^\circ = 2$, $\beta_1^\circ = 5$, $\alpha^{*\circ} = 5$, $\beta^{*\circ} = 5$, $\sigma_{1a} = \sigma_{1b} = \sigma_{a^*} = \sigma_{b^*} = \sigma_{\omega_1} = \sigma_{\omega_2} = 0.2$, and $\tau_1 = 0.8$. We set $J = 10$ and the number of curves per group is $n_i = 5,\ i = 1, 2$. Then we generate groups of curves by changing the parameters according to:

$$p \in (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 0.95),$$

$$\Delta_S \in (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0),$$

$$\Delta_L \in (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0),$$

and keeping the rest of the parameters constant at their original values.

Note that the combination $p = 0,\ \Delta_S = 1.0,\ \Delta_L = 1.0$ generates the null distribution. We change one variable at a time. We investigated changing the parameters using a factorial structure in a pilot study but the results proved difficult to interpret and present; it was challenging to detect differences and examine the results. The change one variable at a time structure required 20 simulation runs. We ran 1000 simulations at each variable combination. This number of simulations is a balance between computer resources and accurate capture of the results. We are confident that this is a large enough number based on the results of our pilot study where we found that the permutation tests, which are exact, unbiased tests, had appropriate size under the null distribution. In addition, we find the power curves are smooth, indicating that we have enough simulations to overcome random error and see the overall large-sample trends. Another confirmation that 1000 simulations is large enough is we found that the power plots are consistent between the pilot study we conducted and the larger full study. Both studies used 1000 simulations, and both exhibited the same patterns. Simulations were run on the Statistics Department's network of computers using Xgrid, which is a parallel processing software. Each set of 1000 simulations took approximately 10 hours. Running the 20 simulations in parallel means the entire simulation experiment ran in half a day.

5.1.3 Monotone Curves

We adapted the curve-generating structure specified in 5.1.1 to generate monotone curves. We changed the underlying curves to Beta cumulative density functions.

The rest of the process progressed as described to generate unimodal curves.

The values of the variables for this set of simulations are set to: $\alpha_1^\circ = 2$, $\beta_1^\circ = 5$, $\alpha^{*\circ} = 5$, $\beta^{*\circ} = 5$, $\sigma_{1a} = \sigma_{1b} = \sigma_{a^*} = \sigma_{b^*} = 0.2$, $\sigma_{\omega_1} = \sigma_{\omega_2} = 0.05$, and $\tau_1 = 0.995$. We set $J = 10$ and the number of curves per group is $n_i = 5,\ i = 1, 2$.

This set of simulations uses the same parameter values as in the unimodal case:

$$p \in (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 0.95),$$

$$\Delta_S \in (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0),$$

$$\Delta_L \in (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0),$$

keeping the rest of the parameters constant at their original values. Again, we ran 1000 simulations at each variable combination.

5.2 Simulation Results

In this section we present the results of the simulations. We compare the power performance of the various statistics as p, Δ_S, and Δ_L change, and compare and contrast the results in the unimodal and monotone cases. We investigate many questions in turn. We compare power between distance measures and summary measures. We investigate whether the order of calculating the summary measure and the distance measure matters. We compare the various tests based on principal components and compare parametric- and permutation-based versions of the tests.

Then we examine more closely the behaviour of some of the statistics. We check if the p-values are relatively uniformly distributed over the null hypothesis. We also examine the distribution of the p-values as p, Δ_S, and Δ_L change. We conclude with a discussion and summary of the results.

5.2.1 Comparing Power

Comparing Distance Measures

Let us first examine the power plots for some of the pairwise permutation statistics, comparing the four distance measures. We keep the mean as the summary method and study the power as each variable, p, Δ_S, and Δ_L, changes. The power plots of the statistics based on $\overline{T}_1$, $\overline{T}_2$, $\overline{T}_\infty$ and $\overline{T}_A$ by the three variables are shown in Figure 5.3. The first column of this figure contains power plots of the test statistics under unimodal curves and the second column contains power plots under monotone curves. In each plot only one parameter is changing at a time, the others are held at their baseline values. For example, in the top row of plots, when plotting power as p changes, $\Delta_S = 1.0$ and $\Delta_L = 1.0$.

This figure tells an interesting story. Note that all of these test statistics have accurate behaviour under the null hypothesis, with rejection of the null approximately 10% of the time when $\alpha = 0.10$. One characteristic of note is that the difference in area ($\overline{T}_A$) does a poor job of detecting differences between two groups of unimodal curves regardless of the effect under scrutiny. This is not surprising because the unimodal curves as we designed them are based on probability density functions and as such have an integrated area consistently close to one. On the other hand, with monotone curves $\overline{T}_A$ does detect differences in p and Δ_L, although not as well as the other measures. Another message is that the three norm measures do a very similar job of detecting differences for p over both univariate and monotone curves. The biggest difference is in Δ_S where the L_1 norm distinguishes itself as best method of detecting differences in a shift. $\overline{T}_\infty$ has slightly lower power in the unimodal case

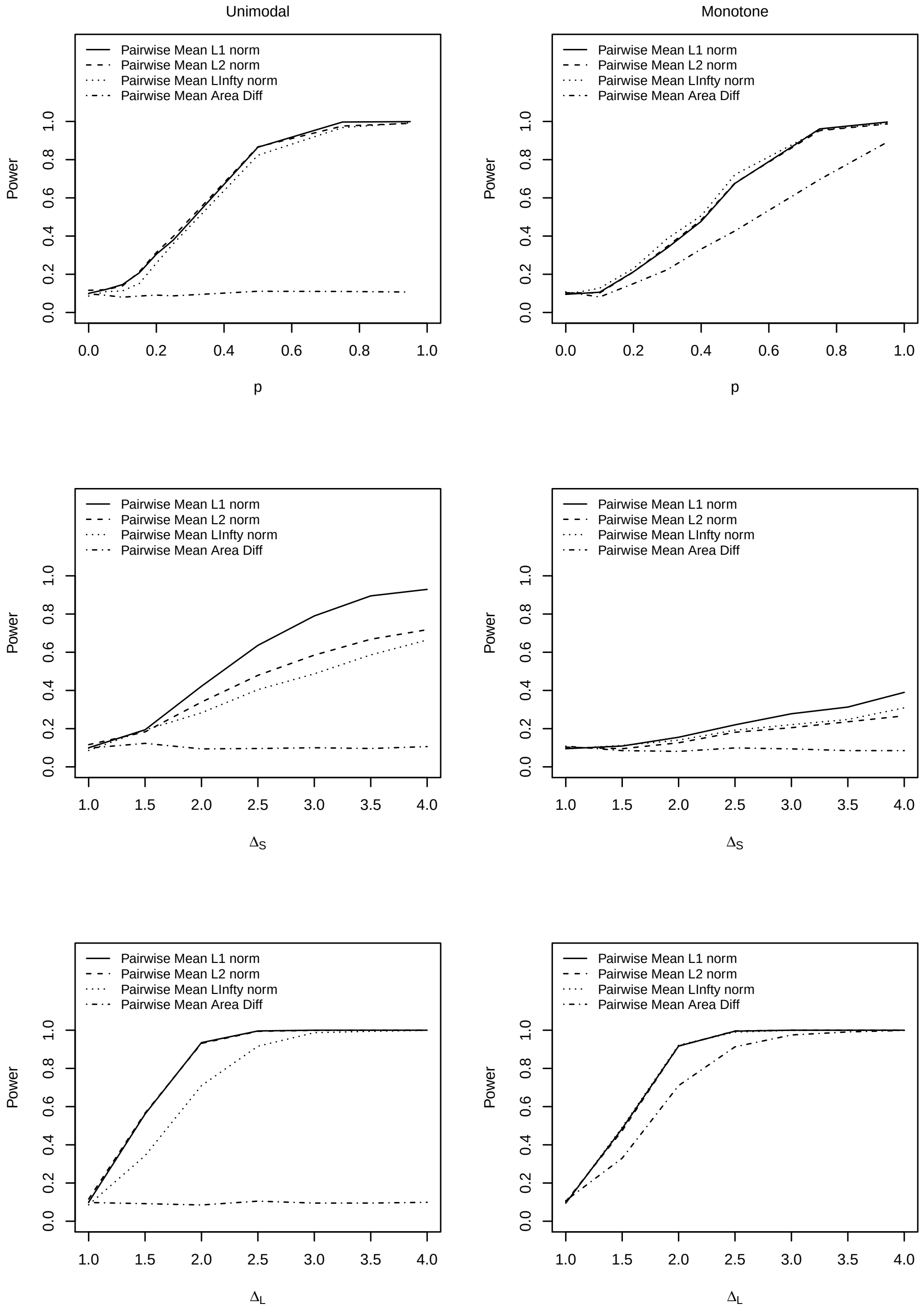

Figure 5.3: Power plots comparing norms, with mean as summary statistic.

but has slightly higher power in the monotone case. This test statistic does not do as well as the tests based on the L_1 and L_2 norms in detecting differences in Δ_L in the unimodal case. Thus, overall, based on the mean, the pairwise L_1 norm is the best measure of differences between groups of curves under a variety of situations.

Comparing Summary Measures

Now let us examine the behaviour of the pairwise permutation tests based on the L_1 norm using different summary measures. Let us compare the power curves of the mean ($\overline{T}_1$), median ($\widetilde{T}_1$), 10% trimmed mean ($\overline{T}_{10,1}$) and 20% trimmed mean ($\overline{T}_{20,1}$), which are overlaid in Figure 5.4. In these sets of curves we consistently see that the pairwise mean L_1 norm has the highest power over all variables for both unimodal and monotone curves, while the median consistently has the lowest power.

In contrast, the power curves of the pairwise permutation tests using the L_2 norm are plotted in Figure 5.5. Using the L_2 norm, in most cases there is very little difference in performance between the three summary statistics that use means. Interestingly, the trimmed 20% mean L_2 norm is the best at detecting a shift in unimodal curves. Comparatively, the median does not perform as well overall. None of the measures are good at detecting a shift in the monotone case over the range of values considered in the simulation.

Comparing the Order of Calculating Summary Measure

We focus mostly on pairwise comparisons of distance measures, but we also consider test statistics where we calculate the summary statistic first and then compute the

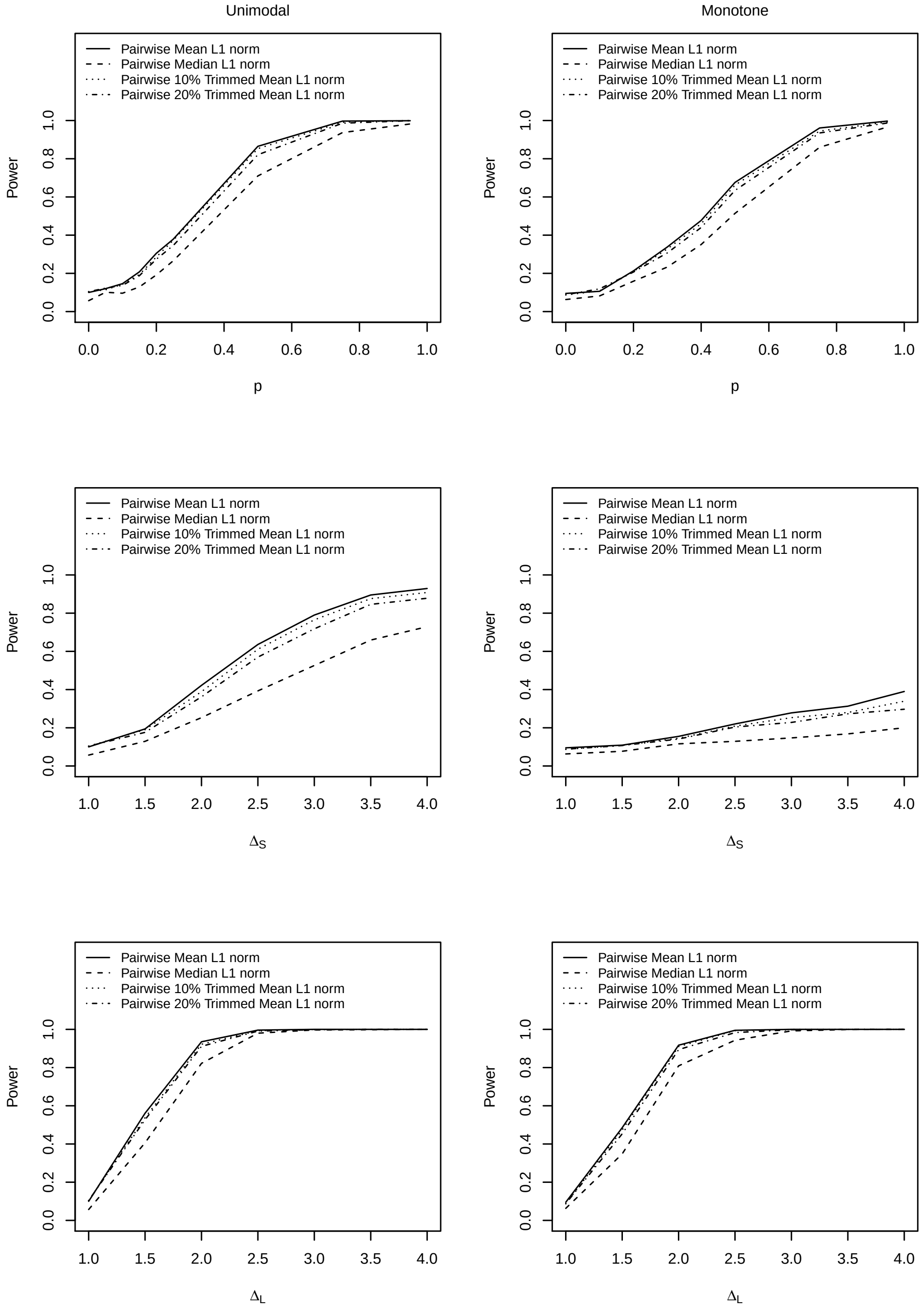

Figure 5.4: Power plots for pairwise comparisons with various summary methods, using L_1 norm as distance measure.

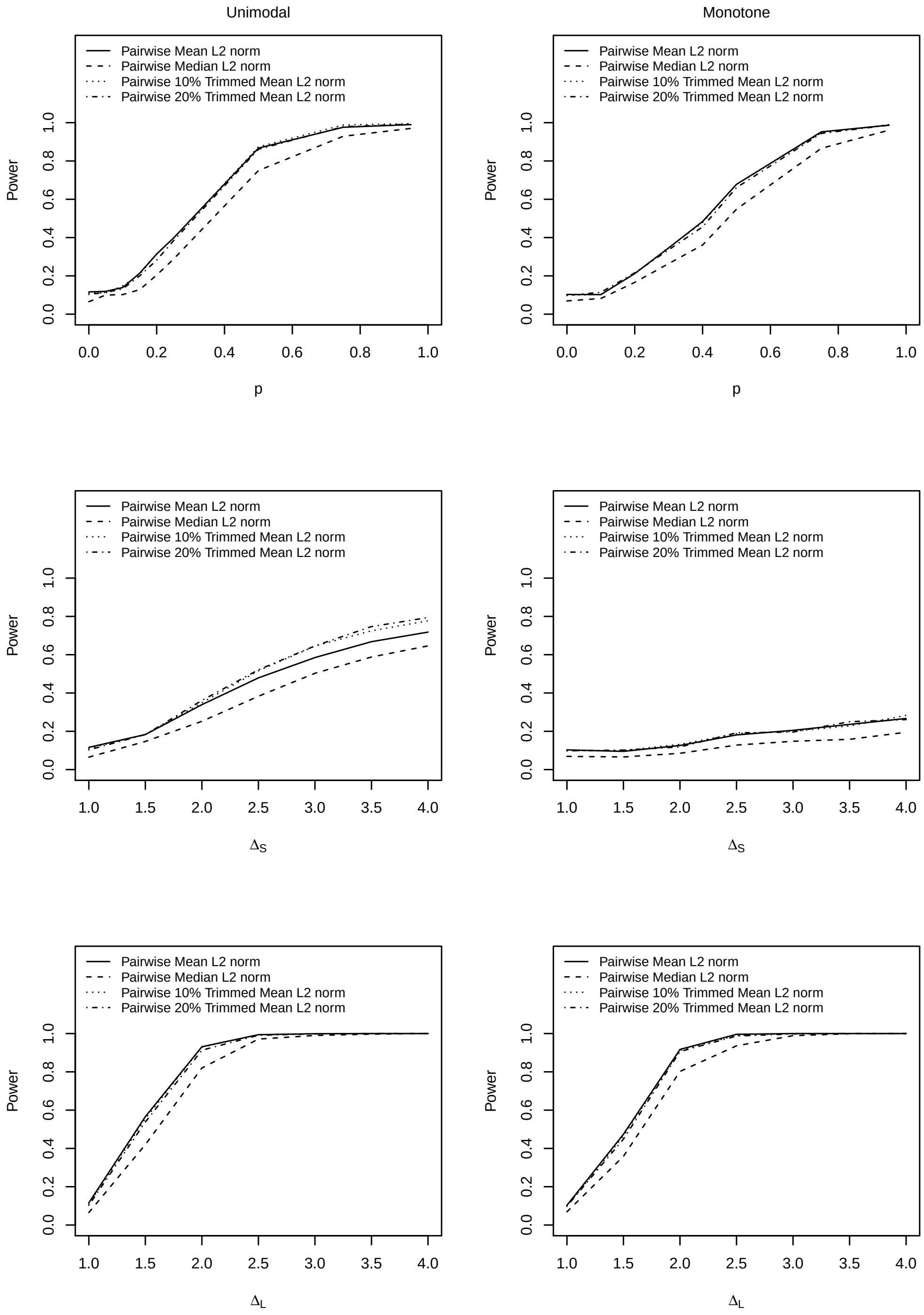

Figure 5.5: Power plots for pairwise comparisons with various summary methods, using L_2 norm as distance measure.

distance measure. The power curves for these statistics, S_1, S_2, S_∞ and S_A and are found in Figure 5.6. We also include the Ramsay test in this figure. The results here are not clear. The L_1 norm performs well for detecting a shift in the unimodal case. The L_2 norm has the highest power in the other cases. We also see that switching the order does not help the performance of the test using the difference in area in the unimodal case, and that overall the statistics do not do well detecting a shift in the monotone case. The Ramsay test does not perform well overall, with low power curves throughout.

Now let us compare the performance of two of the pairwise permutation tests, the pairwise mean L_1 norm and the pairwise mean L_2 norm, with two of the non-pairwise permutation tests, those based on the L_1 and L_2 means. The power curves for these tests are plotted in Figure 5.7. We see that the order in which we calculate the distance and summary measures matters in some cases but not in others. Both methods work equally well for detecting p and Δ_L in the unimodal case. Note that in these power plots, we often see the pattern that the non-pairwise L_1 norm performs slightly worse than the other methods being considered here. In detecting differences in Δ_S, the pairwise mean L_1 norm is the clear winner. Therefore, we see that in some cases, there is no difference between the top pairwise and non-pairwise procedures. However, in other cases the order in which we calculate the distance and summary measures does matter. The non-pairwise approach is never better overall than the pairwise approach.

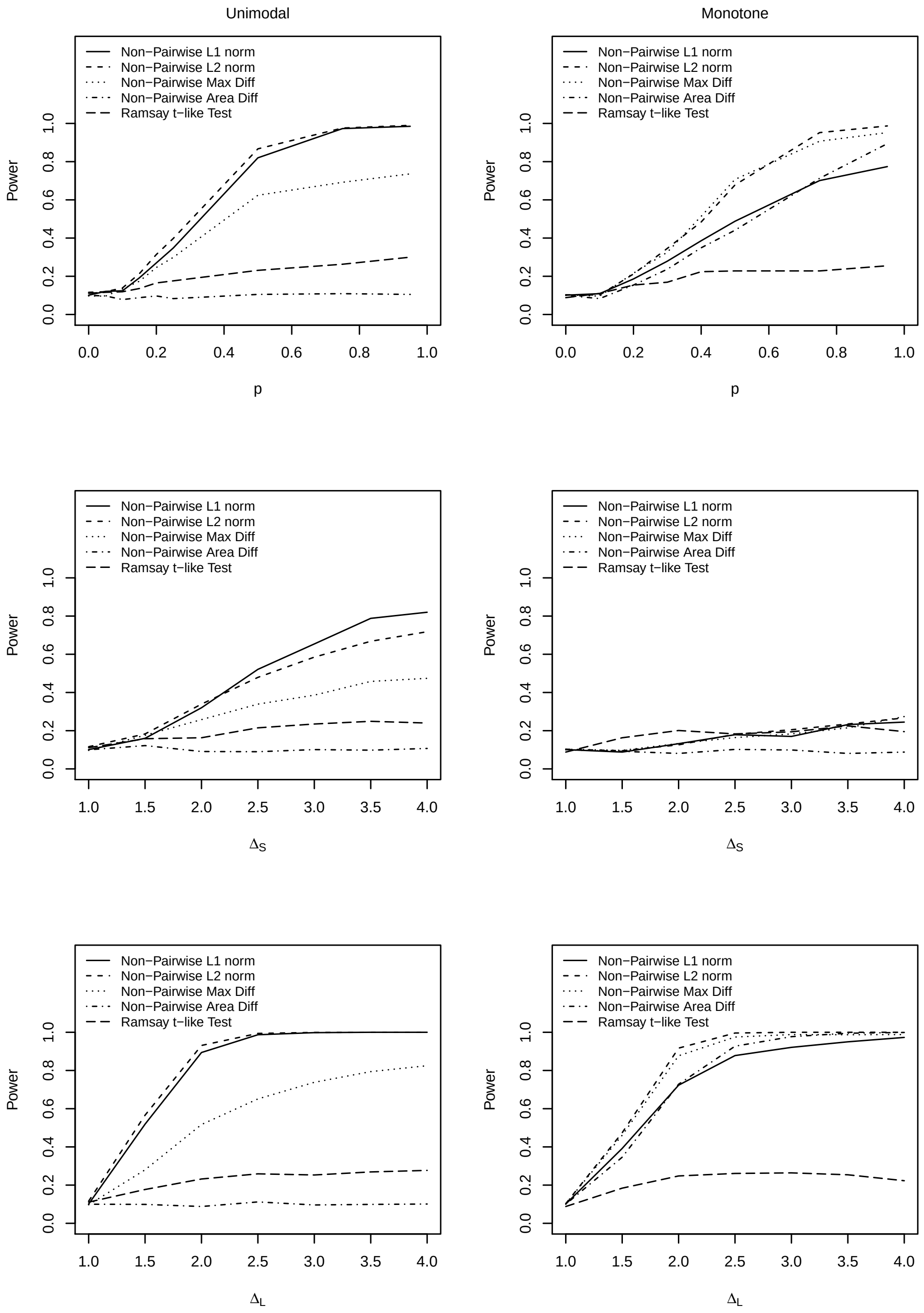

Figure 5.6: Power plots for comparisons with summary statistic computed first, then distance measure.

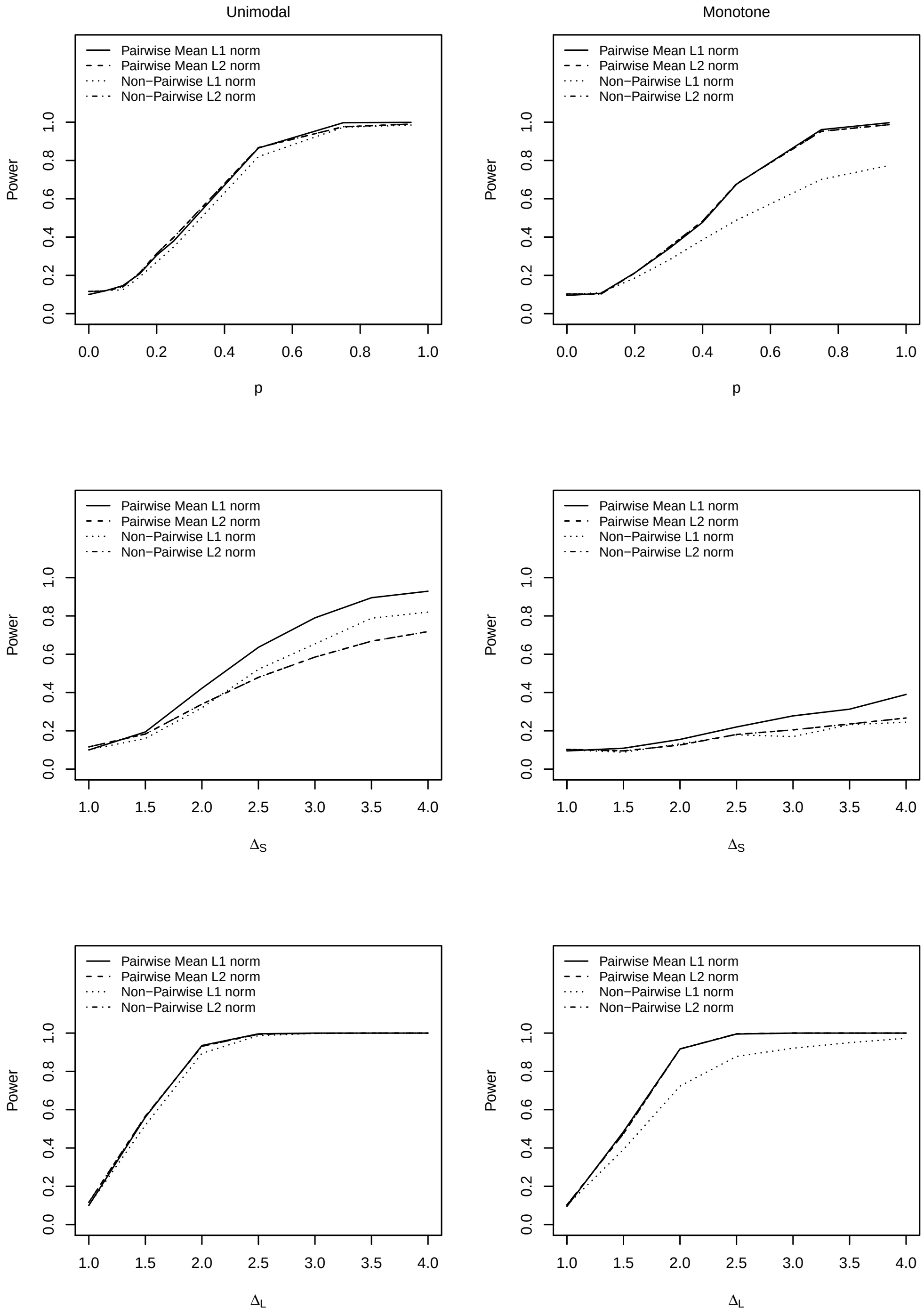

Figure 5.7: Power plots comparing order of calculating summary measure (that is, comparing pairwise and non-pairwise tests).

Comparing Tests Based on Principal Components

Shifting attention from permutation tests for a moment, consider the test statistics based on principal components as developed by Sturino et al. [2010]. The authors discuss a test based on the normal distribution and one defined using alternative rejection regions. We also consider the permutation version of the test based on the normal distribution as well as a χ^2 approach combining the two principal components' statistics. The power curves of these tests are plotted in Figure 5.8. One feature of note in these sets of curves is that the method using the normal distribution assumption rejects the null hypothesis when the null is true about 25% of the time. This is not a promising characteristic for a test statistic. Happily, the other versions of the test are better behaved. As p changes in the unimodal case, the method using the alternative cut-offs is slightly better overall than the χ^2 version, and both have higher power than the permutation version of the test. This is also the case over Δ_L. The reverse is true for monotone curves, where the χ^2 version edges out the method using the alternative cut-offs over both p and Δ_L. Changing the nature of the curves has a slight effect on the performance of the test statistics although the difference between the alternative cut-offs and χ^2 versions is very small over p and Δ_L.

Over Δ_S we see a much more gradual slope in the power curves, but still steadily increasing in the unimodal case. Again we see the issue the normal test has at the null hypothesis. The other three tests perform reasonably similarly, but it seems that the χ^2 version does the best job of detecting changes in shift in the unimodal case. None of the tests distinguish themselves in detecting shifts in the monotone case.

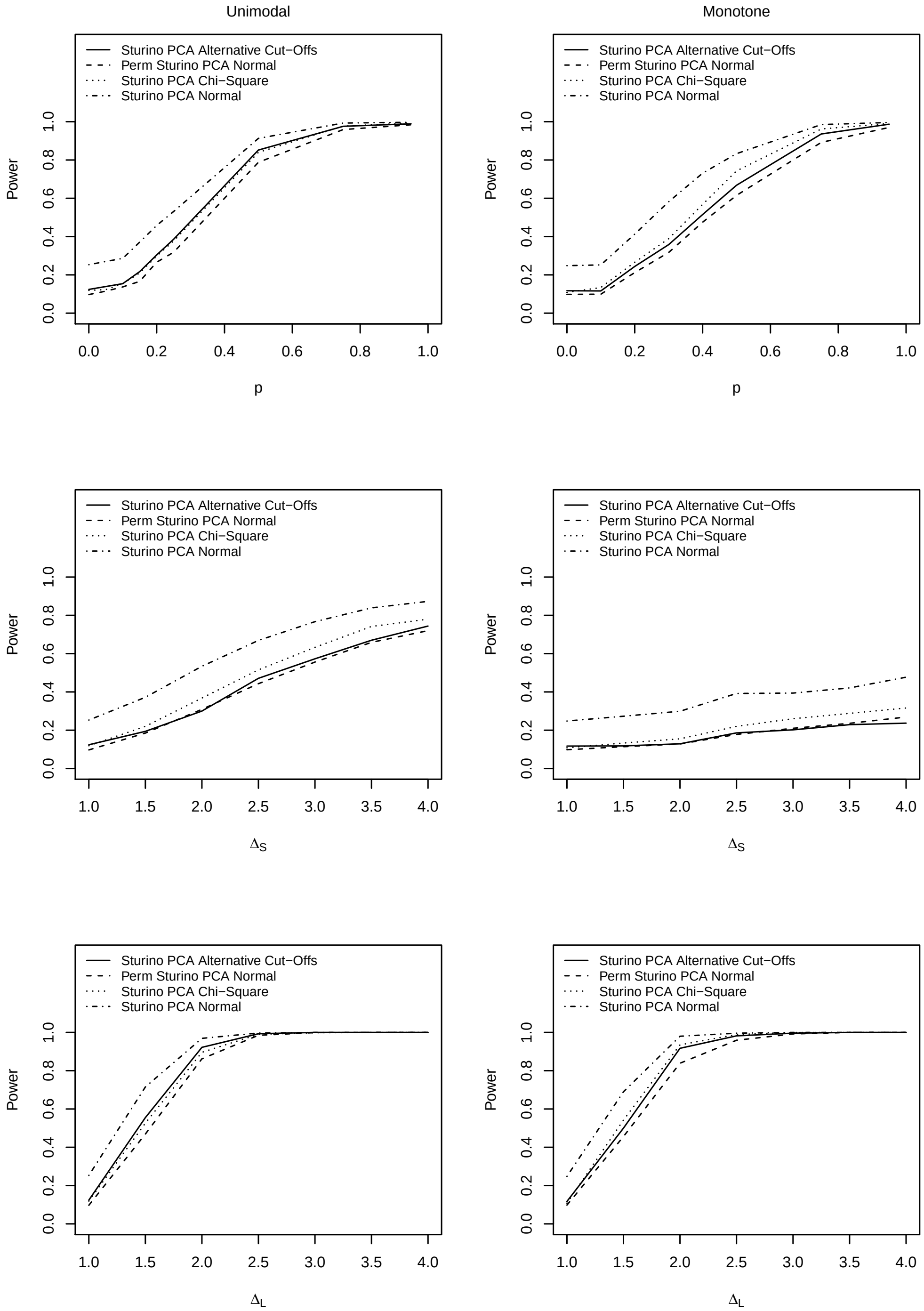

Figure 5.8: Power plots for Sturino PCA methods.

Note that the permutation test uses the same test statistic as the method using the normal distribution assumption. We see that the distribution-free permutation test has a much more accurate size, $\sim 10\%$, than the test using distribution theory. Overall, excluding the test statistic using the normal distribution, the Sturino test statistics based on principal components do a fairly good job of detecting differences between two groups of curves. The test with highest power depends on the type of effect under consideration but overall the χ^2 test seems the best candidate.

Comparing Parametric- and Permutation-Based Versions of Tests

Let us now compare the behaviour of some of the parametric tests and their permutation-based counterparts. Power curves are plotted in Figure 5.9 for the Shen-Xu F test, the adaptive Neyman test, and the Sturino PCA tests for both parametric and permutation versions. We see that the Shen-Xu F test rejects the null hypothesis too frequently, about 30% of the time, when the null hypothesis is true, but that the power curve of the permutation version of the test behaves well. The adaptive Neyman test does not perform admirably under these conditions; neither the test statistic as designed nor the permutation version have any strength as tests compared to the other tests' power performances, although it is interesting to note that the permutation version of the test has better power in most cases than the test as designed. The adaptive Neyman test is especially poor in detecting differences in the monotone case. Considering that this test is based on Fourier basis expansions this behaviour is not unexpected; the nature of a Fourier expansion does not lend itself well to describing a monotone function. This indicates that this test would most likely do best if the underlying curves were periodic in nature. Overall,

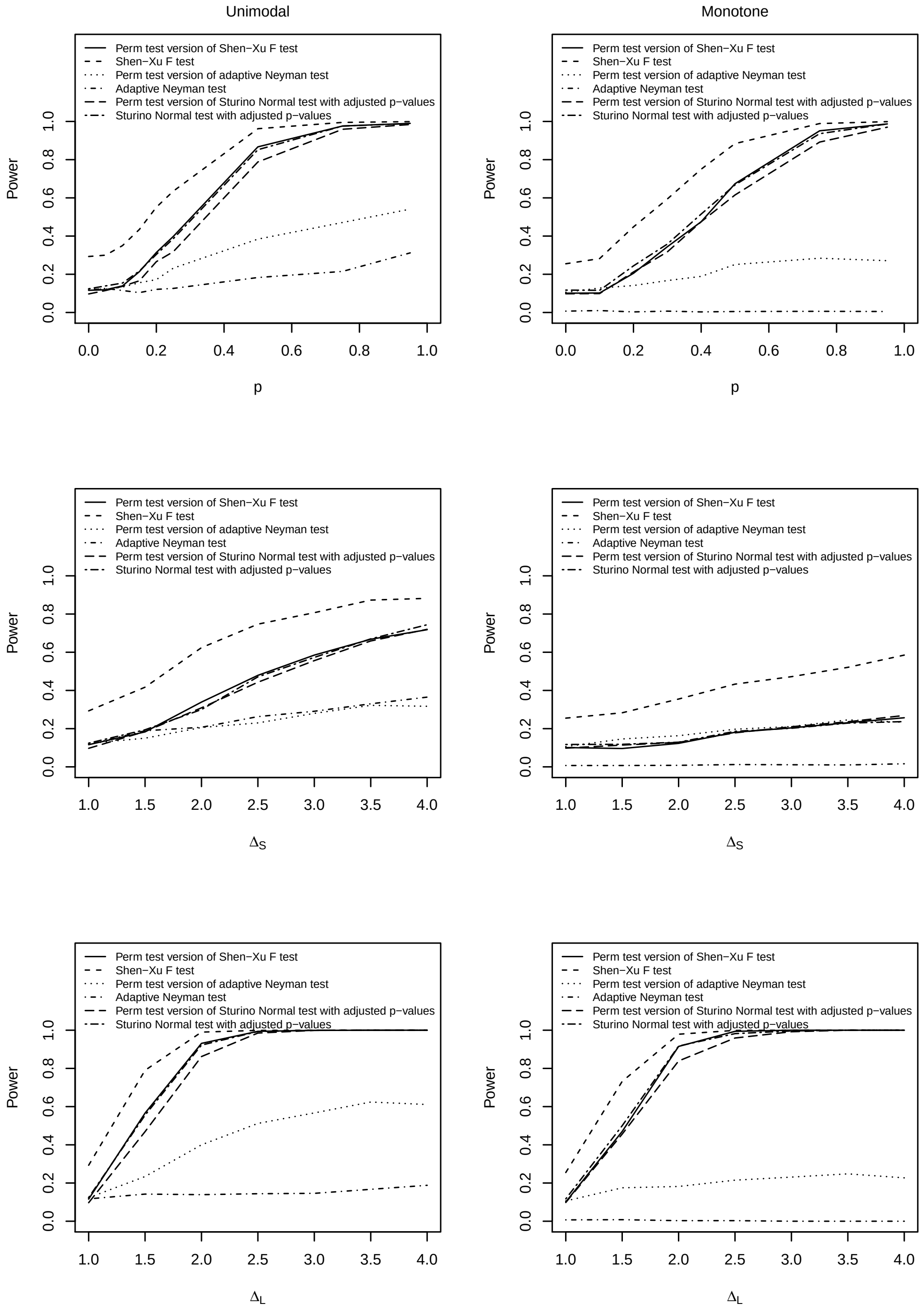

Figure 5.9: Power plots for parametric and permutation-based methods.

it seems that the best performance here is given by the permutation version of the F test.

Top Performers

We have seen several plots, now let us compare the top power performers out of all candidates. The power curves for the measures with the highest power that we have reviewed thus far are plotted in Figure 5.10. The pairwise L_1 norm performs very well overall. It is consistently at the top amongst the power curves. The next best statistic is the pairwise 20% trimmed mean. This is followed by the Sturino PCA statistic with alternative cut-offs, which has slightly higher power at lower levels of p in the monotone case but with the reverse pattern for larger values of p. Following these tests in the rankings are the Sturino PCA method with alternative cut-offs and the permutation version of the Shen-Xu F test. These two tests are competitive with the pairwise L_1 norm for detecting differences in p and Δ_L but do not do as well detecting differences in Δ_S.

5.2.2 Examining Probability of Rejection

The power curves tell part of the story as far as the performance of statistical tests go, and we will return to them, but now let us examine the behaviour of some of the measures in a slightly different way.

Norm Based on Absolute Difference

First consider the plots in Figure 5.11. Here we have a histogram of the p-values at the null hypothesis, side-by-side boxplots of the p-values over the range of p, and

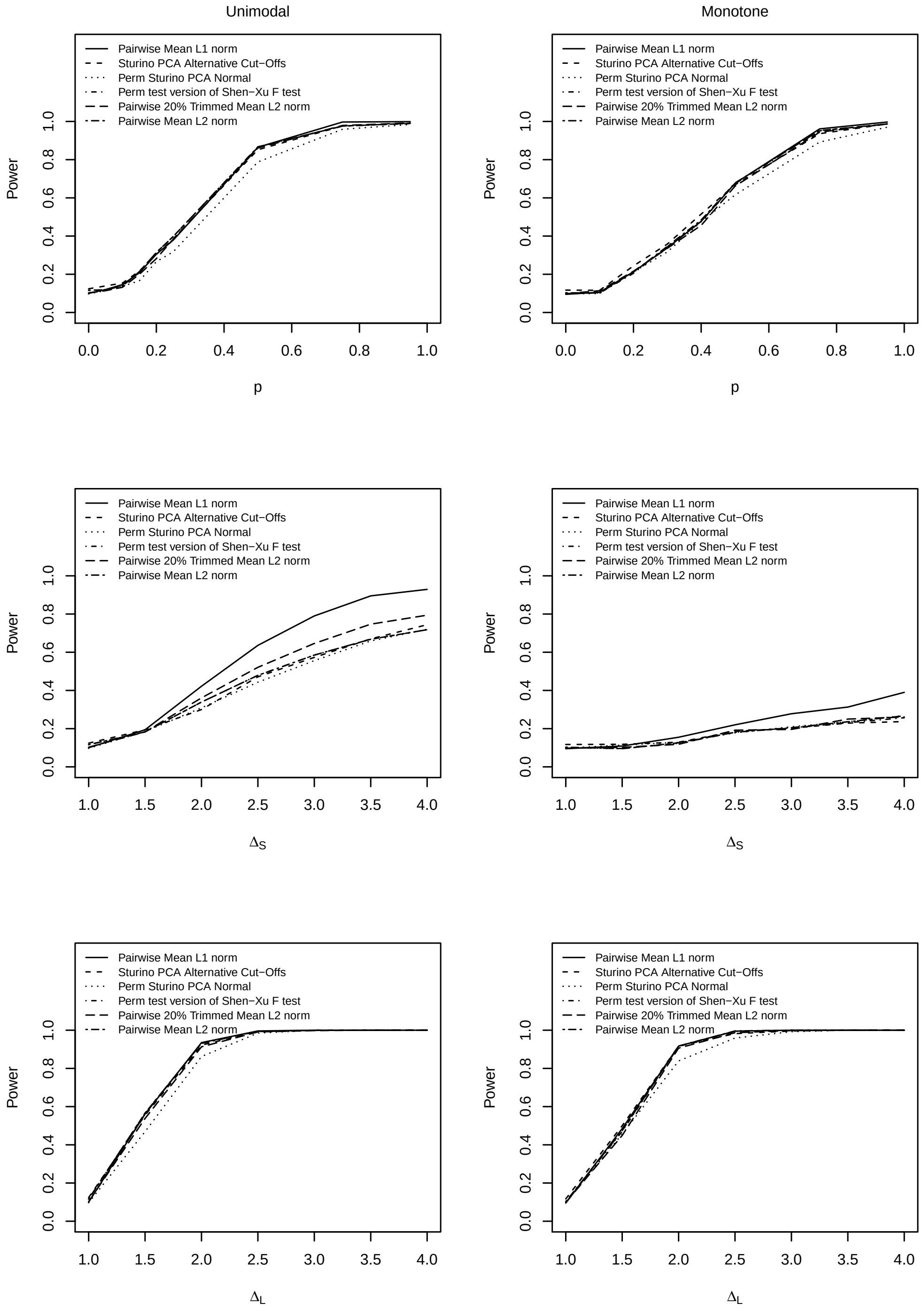

Figure 5.10: Top power performers.

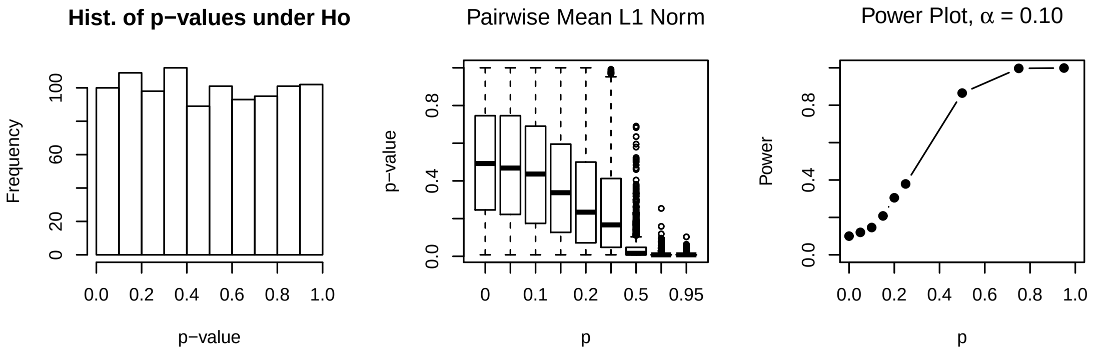

Figure 5.11: Histogram, boxplots, and power plot for pairwise mean L_1 norm permutation test on unimodal curves, plotting over p.

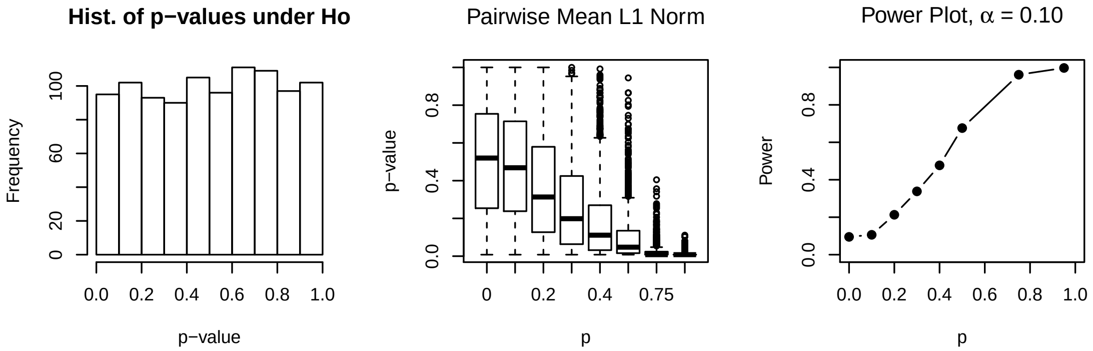

Figure 5.12: Histogram, boxplots, and power plot for pairwise mean L_1 norm permutation test on monotone curves, plotting over p.

the power curve over p, all for the pairwise mean L_1 norm in the unimodal case. The corresponding curves over Δ_S and Δ_L are included in the Appendix, in Figures B.1 and B.2 on page 133.

The histogram looks fairly flat over the null distribution. Under the null, we would expect the distribution of p-values to be fairly even, because when the null is true we would expect 5% of the observations to be rejected under an $\alpha = 5\%$ test, 10% of the observations to be rejected under at 10% test, et cetera. By this

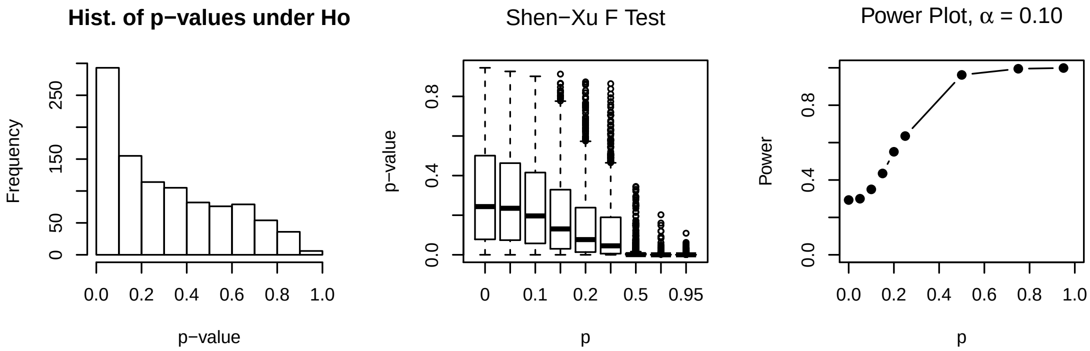

Figure 5.13: Histogram, boxplots, and power plot for parametric Shen-Xu F test on unimodal curves, plotting over p.

argument the histogram of p-values should be fairly evenly distributed when a test has accurate size. We do note the consideration that it is not possible to obtain a true zero p-value since for a permutation test the minimum value is 1/(number of combinations). Thus, this flat histogram is in line what we would expect under the null distribution. The p-values get smaller as p increases, as illustrated in the boxplots in the middle cell. This is consistent with what we would expect. And finally, in the third cell, again we see the power curve for the variable, exhibiting the same pattern we have repeatedly seen. The results are similar for the monotone case, as illustrated in Figure 5.12. The plots over Δ_S and Δ_L are included the Appendix, see Figures B.3 and B.4 on page 134.

Shen-Xu F test

The same plots for the Shen-Xu parametric F test are shown in Figure 5.13. Here we see a histogram behaving badly; it is certainly not reasonably flat over $(0, 1)$. The p-values plotted in the boxplots in the middle cell do decrease as p increases but based on the histogram and the power plot with its over-active null hypothesis

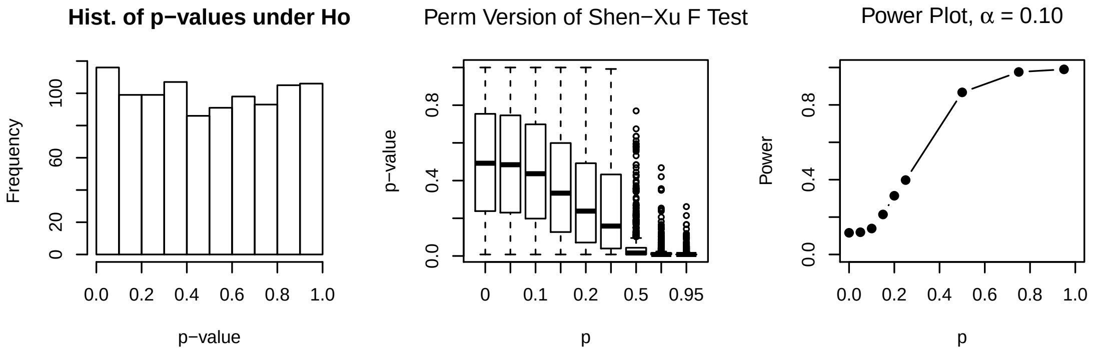

Figure 5.14: Histogram, boxplots, and power plot for permutation version of Shen-Xu F test on unimodal curves, plotting over p.

rejection, this statistic is not behaving as an ideal test statistic. Contrast this with the behaviour of the permutation version of the F test, shown in Figure 5.14. In these plots we see a nice, fairly flat histogram, p-values that decrease as Δ_S increases, and a power curve with an increasing trend and the proper rejection probability under the null hypothesis. The parametric version of the F test does not perform well under the experimental conditions we have created, but the test statistic can be used with the permutation technique to get a more accurate testing mechanism. Plots for both the parametric and permutation versions over Δ_S and Δ_L and all plots for the monotone case are included in Appendix B beginning on page 135.

Norm based on Squared Difference

Let us examine the behaviour of the test statistics using the L_2 norm with various summary measures, shown in Figure 5.15 for the unimodal case and Figure 5.16 for the monotone case. Other plots can be found in Appendix B beginning on page 139. There is little difference between the plots arising from the mean measures.

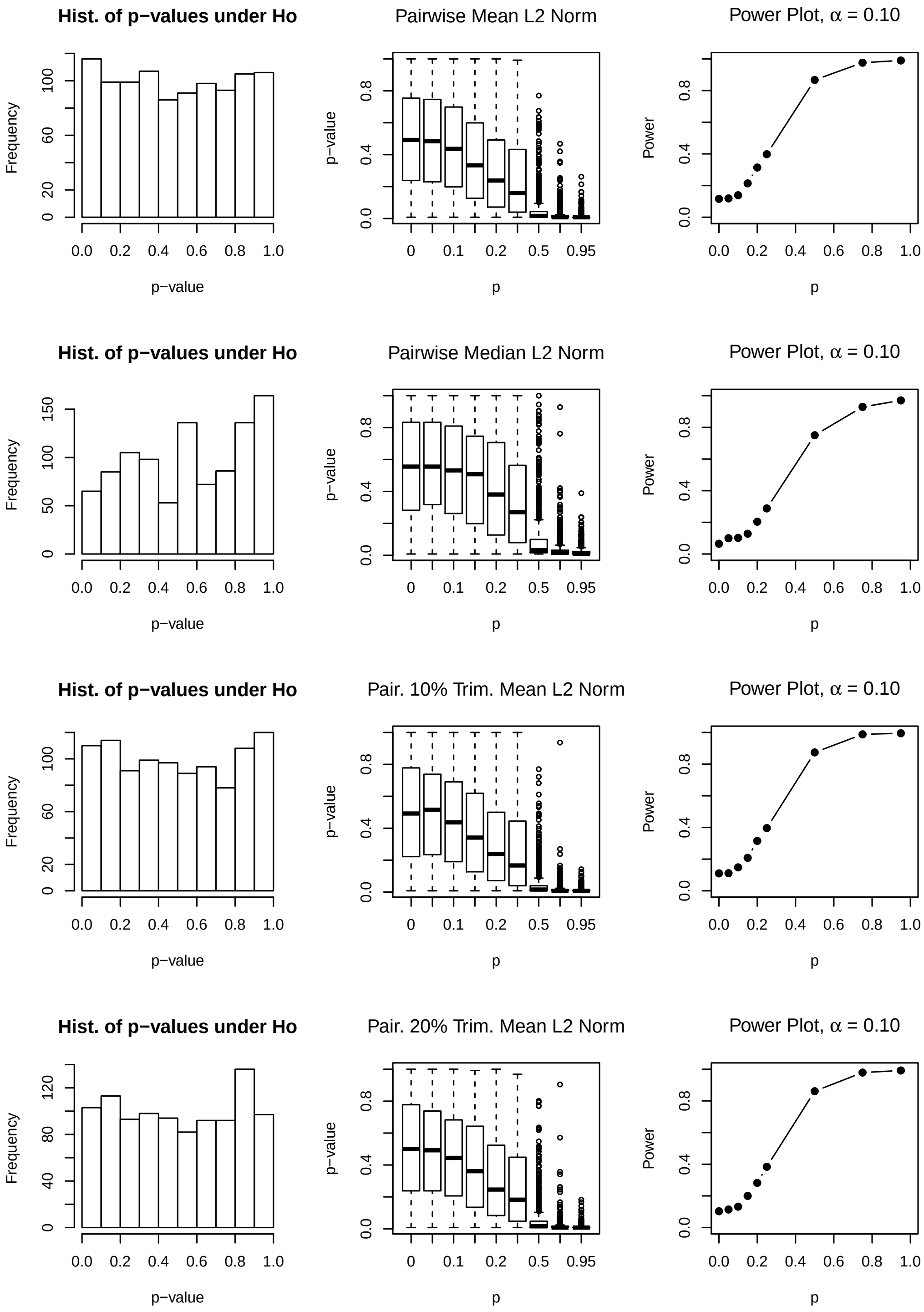

Figure 5.15: Histogram, boxplots, and power plots for L_2 norms with varying summary statistics on unimodal curves, plotting over p

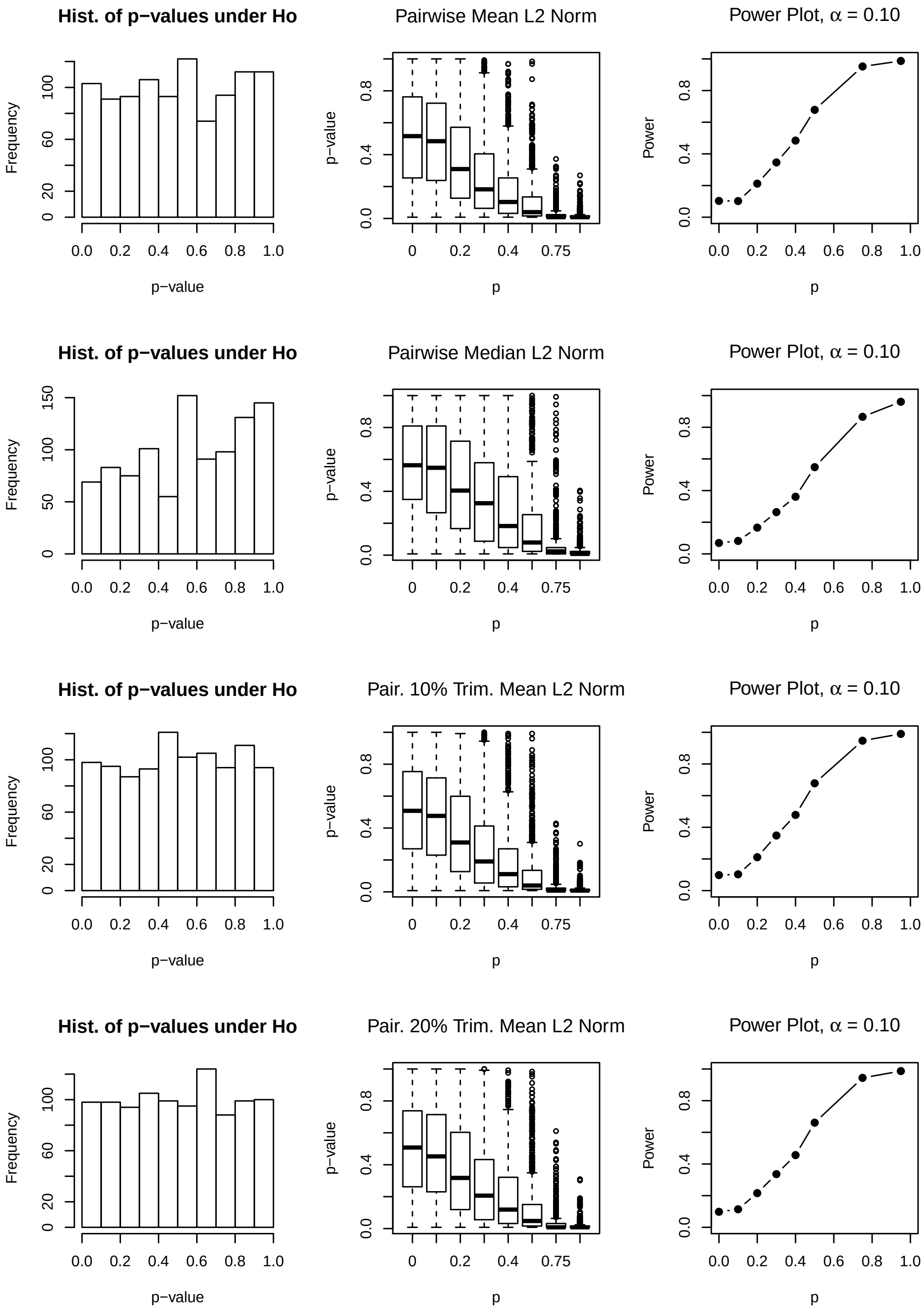

Figure 5.16: Histogram, boxplots, and power plots for L_2 norms with varying summary statistics on monotone curves, plotting over p

The difference lies in the median. Recall that the median L_2 norm had lower power than the other L_2 norm statistics. We see more evidence against the use of the median summary measure in the histogram, which is certainly not flat, while the histograms for the other test statistics are relatively evenly distributed over $(0, 1)$. The side-by-side boxplots for all of the test statistics follow the expected decreasing pattern as p increases. Results are similar for the other cases, see Appendix B for the plots.

5.3 Discussion

At the end of the functional data analysis section of the NASA case study, we briefly discussed some of the advantages and disadvantages of each of the tests we considered. Now, with more information regarding the comparative power of the tests, we consider each test in turn.

We considered a large number of pairwise permutation tests. The version with the best power performance was the test based on the mean of the pairwise L_1 norms. This test has all the advantages of a permutation test. It is simple to understand, it is easily carried out once the programming code is complete, and there are no distributional assumptions. The disadvantages are that it may take a long time to compute for large sample sizes and it requires some programming sophistication to code.

The non-pairwise permutation tests did not perform as well as the pairwise tests did. These tests have the same advantages as the pairwise tests: they are simple to understand, easily carried out once the programming code is complete, and there are no distributional assumptions. The disadvantages are that it may take a long time to compute for larger sample sizes, some programming sophistication is required, and the power isn't as good as the test based on the L_1 norm. We feel that if one is going to spend the time to carry out a permutation test, one might as well do the full pairwise test because it has higher power.

The permutation test proposed by Ramsay et al. [2009] did not perform well. As we found that the permutation tests using L_∞ norm-based measures had poor power compared to the other norms, this is not so surprising.

The parametric version of the functional F test had a poor performance. The test rejected the null hypothesis too often when the null hypothesis was true. The good performance of the permutation version of the test indicates that the test statistic measures the effect well but that the distributional assumptions are not met here. This test has been found to be sensitive to the error structure. The error structure of the curves in the simulations is not a Gaussian stochastic process and this may have contributed to the poor performance. This test is based on assuming a specific model structure to the data, if the data do not follow this structure then the test will perform well. However, considering that the permutation version using the test statistic as a basis for a permutation test performed well, this suggests that the issue relates to the parametric assumptions. The test statistic does in fact do well in detecting differences between groups of curves, as shown by the good performance of the permutation version of the test. This suggests that the issue is with the translation to the parametric test parameters. Contributing to this problem may be the small sample size in this case. Shen and Xu recommend having enough observations (curves) to allow for 30 degrees of freedom after parameters are accounted for. In this situation we are nowhere near 30 degrees of freedom.

This may at least partly explain why the parametric approach performs so poorly. Another possibility is that we made an error in the calculation of this test statistic, perhaps in a denominator division or in the calculation of λ — that is, in an area that affects the translation to the parametric assumptions. This is entirely possible but if a mistake was made, it has not been found in many uses of the code. One advantage of the functional F test is that it does not require intensive calculations.

The adaptive Neyman test did not do well in either its original or permutation-based form. This result indicates that this test statistic does not do a good job of capturing the differences between two groups of curves in our simulations. Since this test is based on Fourier transformations, perhaps these transformations did not adequately capture the features we were studying. This test may do better with periodic data.

The tests based on principal components were much more satisfactory. They had good power. The permutation test was outperformed by the distribution-based tests, indicating that the distribution theory was of some benefit in this case. The test based on alternative cut-off values is moderately easy to apply once one has figured out basis functions and principal components in the functional realm. It is quick to calculate and has good power performance.

In conclusion, our simulations suggest that the pairwise permutation test based on the L_1 norm is the best test to use, with our second choice being the principal components method using alternative cut-offs. Both of these tests perform admirably in a wide variety of situations. The permutation test is desirable for several reasons. It performs well in a variety of conditions, detecting various differences between two groups of curves. It is exact, unbiased, and model free. Many other testing procedures depend on a model, and these types of tests run into problems when the model is mis-specified. Consider for example the functional F test. The permutation version of this test performs more accurately than the parametric version. The drawback is that it does take some time to calculate. A faster test is the principal components method using alternative cut-offs. However, we do not have information on how sensitive this test is to having too few or too many principal components.

5.4 Summary

We have examined the behaviour of many test statistics under varying conditions. These simulations suggest:

- The L_1 norm is the best distance measure.
- The mean is the best summary measure.
- Taking the distance measure before the summary measure yields a more powerful test.
- Methods using principal components perform well using alternative cut-offs or the χ^2 assumption.
- In general, the permutation versions of the parametric tests perform better than the parametric versions.

We conducted simulation studies using both smaller and larger standard deviations and the results were similar. The overall patterns did not change.

Based on the results of the simulation studies, we recommend using the pairwise permutation test based on the L_1 norm to test for differences between two sets of curves. Our second choice is the principal components method using alternative cut-offs.

CHAPTER

SIX

CONCLUSIONS

We have explored methods for comparing two sets of curves. We began with an extensive literature review which summarizes the current status of research in areas concerned with data that are curves. We proposed a set of tests using pairwise comparisons between curves. We reviewed a number of tests, both parametric and nonparametric. We conducted a case study based on a set of data from NASA, demonstrating various approaches to analyzing the data. In the case study we applied techniques for comparing two groups of curves to a factorial experiment by grouping the data into two curves by each effect. Not all approaches exploited the functional nature of the data. We introduced a graphical technique to examine interaction effects for functional data. Finally, we carried out a series of simulation studies exploring the power performance and robustness of a large list of test statistics under varying conditions.

We have seen that overall, permutation tests are good tests to use for conducting tests on functional data. We found that the test based on the mean of the pairwise

L_1 norms performed very well in detecting different effects in both the unimodal and monotone cases. While a parametric test may be more convenient and faster to calculate, the permutation test is more robust and is applicable in a wider variety of settings.

We find that applying techniques for comparing curves to functional data generated by a factorial design is a useful approach to analyzing such data.

We do believe that the cross-sectional analysis method has some merit. The graphical methods are very revealing and informative regarding the changing nature of the effects over time. Carrying out a cross-sectional analysis on a set of functional data will illuminate features of interest and allow us to examine interaction effects. This information could be used as an exploratory analysis before formal testing procedures or considered a complement to the overall perspective provided by tests using the entire curve at once.

There are some limitations to the research. While we conducted an extensive simulation study it is not possible to examine every possible set of circumstances (curve shapes, effects to detect, error structure, amount of noise, number of points sampled along the curve, etc.). There are many other power simulation studies that could be carried out. Further investigation is needed towards the missing data problem, which is so important in the field of longitudinal data analysis, and the impact that smoothing away missing data has on the power of the various tests.

Permutation tests have two main limitations. They become prohibitively large to carry out for even medium-sized samples. In this situation it is possible to take a sample of permutations but little research has been done into the effects of doing so on power. Permutation test also require a very specific hypothesis testing structure.

The null hypothesis must be stated in a way that under the null hypothesis the two samples come from the same distribution [Good, 2006].

We carried out multiple testing procedures on the NASA data but did not adjust for multiple comparisons. However, in our framework, with 2^k and 2^{k-p} experiments, the purpose of such studies is generally on identifying the important factors, that is, on screening. It is not common in screening experiments to account for multiple comparisons. We feel that the approach taken in this dissertation illuminates the key factors in the NASA data and therefore that we achieved what we set out to do.

Another limitation is that our sample sizes were small. The large sample case may be very different.

We propose some future directions of research:

- Extending pairwise permutation tests to comparing more than two groups. For three groups of curves, we could take one curve from each group and compute a distance measure between each pair of curves: 1 vs 2, 2 vs 3, and 1 vs 3. We then calculate the average, obtaining a measure of the average distance between pairs of curves. We repeat this process for all sets of three curves among the three groups, and end by computing an overall measure of average distance between pairs of curves. Then we regroup the curves and repeat the computations to carry out the permutation test.

- Detecting outliers as part of model diagnostics. Could permutation tests be used to compare one specified curve to all others as an outlier detection method? These could be of use in model-building and testing and perhaps even in profile analysis.

- Investigating the effects of smoothing on power.

- Investigating the effects on power of taking a subset of the permutations instead of using every single regrouping of the curves into two groups.

- We would have liked to include a method based on wavelet decompositions but our curves were much too smooth to carry out a wavelet method to great satisfaction. Carrying out a simulation study on much rougher curves and exploring how wavelet methods perform is an area for future research.

- Investigating the application of permutation tests treating data as curves in other settings.

- Further simulations. Other covariance functions for the error could of course be examined in separate power analyses. There are many options to consider. One we have mentioned is the AR(1) structure. A compound symmetry structure has constant variance and constant covariance. Or we could follow in the spirit of Shen [1999], who used five different covariance functions in his simulations, not all of which have a formal structure: a constant variance process, a unimodal process, two random processes and a fluctuating variance process.

APPENDIX

A

NASA DATA TABLE

Effects				Response at each Mach Number									
A	B	C	D	0.7	0.9	0.95	1.05	1.1	1.3	1.46	1.96	2.74	4.0
-1	-1	-1	-1	0.199	0.342	0.460	0.593	0.640	0.571	0.580	0.707	0.661	0.607
1	-1	-1	-1	0.198	0.340	0.458	0.591	0.616	0.548	0.564	0.690	0.661	0.615
-1	1	-1	-1	0.202	0.340	0.461	0.596	0.645	0.575	0.576	0.616	0.602	0.539
1	1	-1	-1	0.200	0.338	0.461	0.596	0.625	0.557	0.562	0.598	0.583	0.541
-1	-1	1	-1	0.200	0.345	0.462	0.591	0.638	0.571	0.582	0.740	0.677	0.632
1	-1	1	-1	0.199	0.343	0.460	0.588	0.616	0.544	0.565	0.730	0.678	0.638
-1	1	1	-1	0.203	0.345	0.463	0.586	0.631	0.561	0.563	0.661	0.643	0.591
1	1	1	-1	0.201	0.341	0.458	0.582	0.615	0.541	0.550	0.642	0.643	0.601
-1	-1	-1	1	0.196	0.324	0.412	0.568	0.541	0.513	0.529	0.654	0.656	0.605
1	-1	-1	1	0.196	0.330	0.434	0.566	0.539	0.516	0.535	0.650	0.657	0.614
-1	1	-1	1	0.183	0.317	0.375	0.453	0.435	0.511	0.523	0.560	0.604	0.542
1	1	-1	1	0.185	0.317	0.374	0.568	0.551	0.526	0.533	0.565	0.600	0.547
-1	-1	1	1	0.200	0.345	0.462	0.591	0.637	0.569	0.581	0.745	0.680	0.637
1	-1	1	1	0.199	0.343	0.459	0.588	0.612	0.542	0.565	0.737	0.681	0.641
-1	1	1	1	0.203	0.344	0.462	0.587	0.630	0.552	0.557	0.673	0.649	0.602
1	1	1	1	0.201	0.340	0.457	0.584	0.614	0.537	0.547	0.657	0.652	0.612

Table A.1: NASA data

APPENDIX

B

ADDITIONAL FIGURES

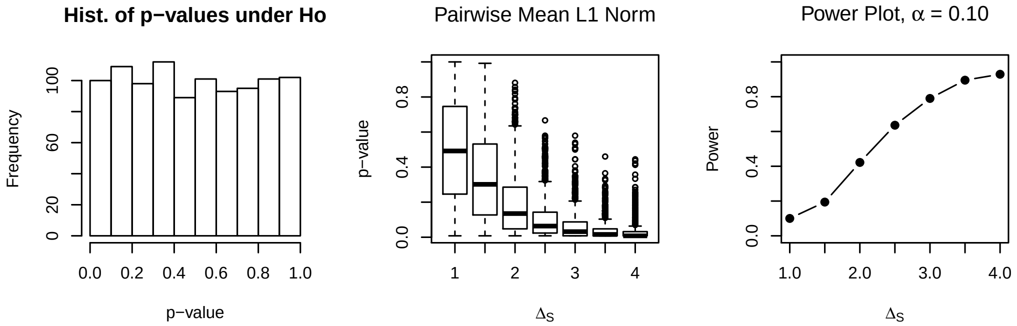

Figure B.1: Histogram, boxplots, and power plot for pairwise mean L_1 norm permutation test on unimodal curves, plotting over Δ_S.

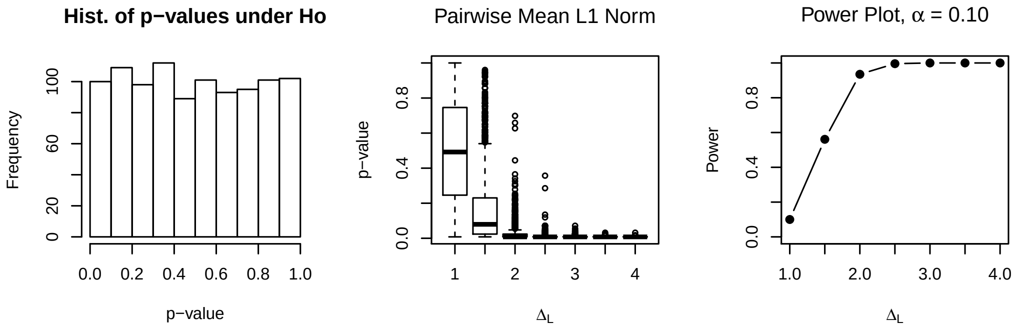

Figure B.2: Histogram, boxplots, and power plot for pairwise mean L_1 norm permutation test on unimodal curves, plotting over Δ_L.

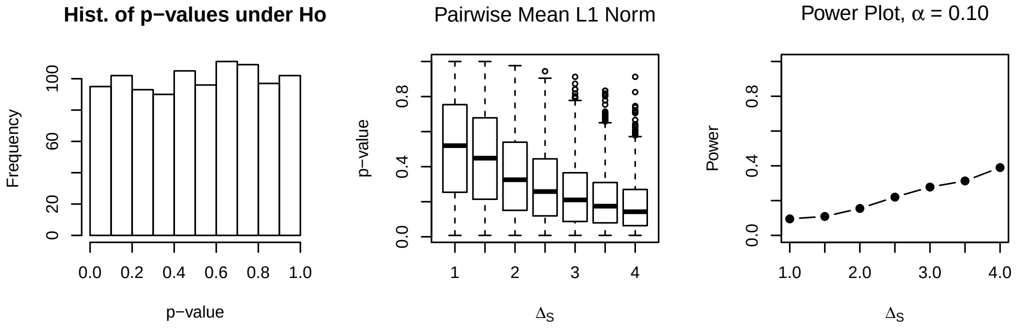

Figure B.3: Histogram, boxplots, and power plot for pairwise mean L_1 norm permutation test on monotone curves, plotting over Δ_S.

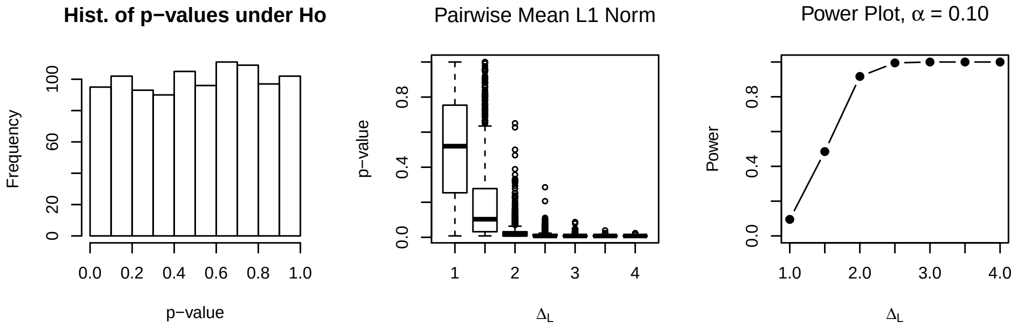

Figure B.4: Histogram, boxplots, and power plot for pairwise mean L_1 norm permutation test on monotone curves, plotting over Δ_L.

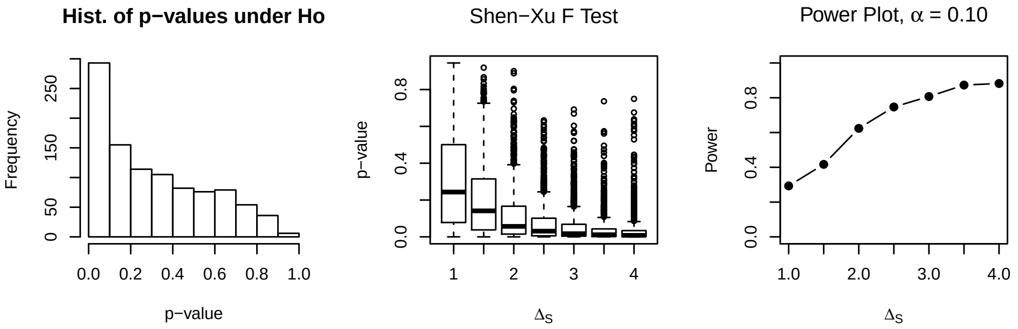

Figure B.5: Histogram, boxplots, and power plot for parametric Shen-Xu F test on unimodal curves, plotting over Δ_S.

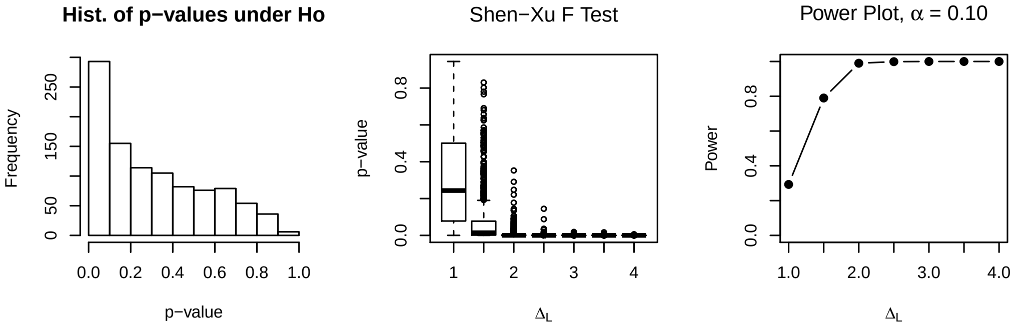

Figure B.6: Histogram, boxplots, and power plot for parametric Shen-Xu F test on unimodal curves, plotting over Δ_L.

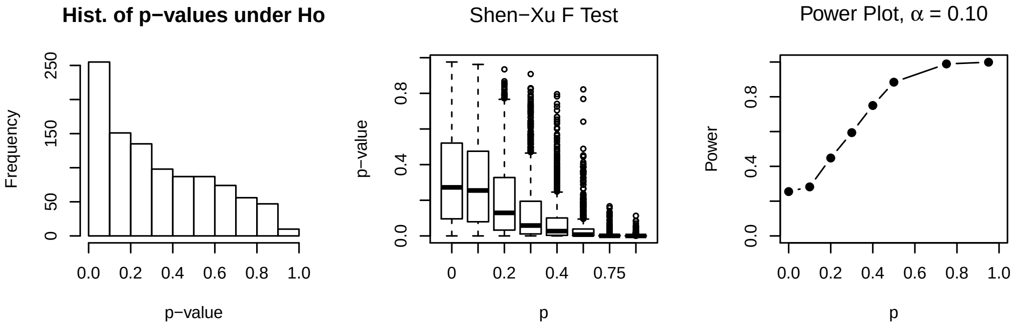

Figure B.7: Histogram, boxplots, and power plot for parametric Shen-Xu F test on monotone curves, plotting over p.

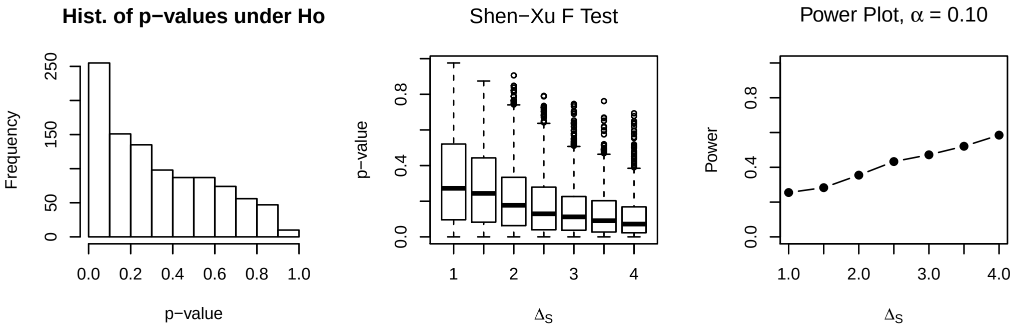

Figure B.8: Histogram, boxplots, and power plot for parametric Shen-Xu F test on monotone curves, plotting over Δ_S.

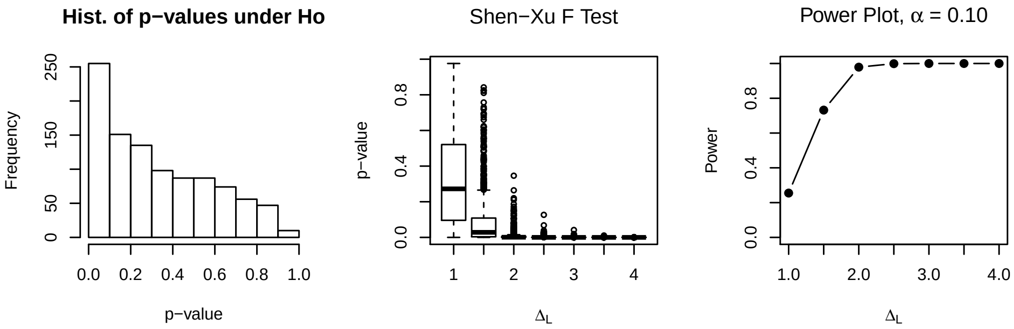

Figure B.9: Histogram, boxplots, and power plot for parametric Shen-Xu F test on monotone curves, plotting over Δ_L.

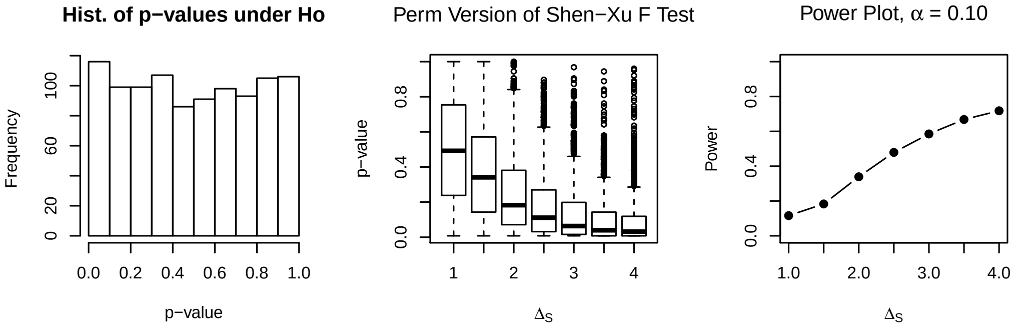

Figure B.10: Histogram, boxplots, and power plot for permutation version of Shen-Xu F test on unimodal curves, plotting over Δ_S.

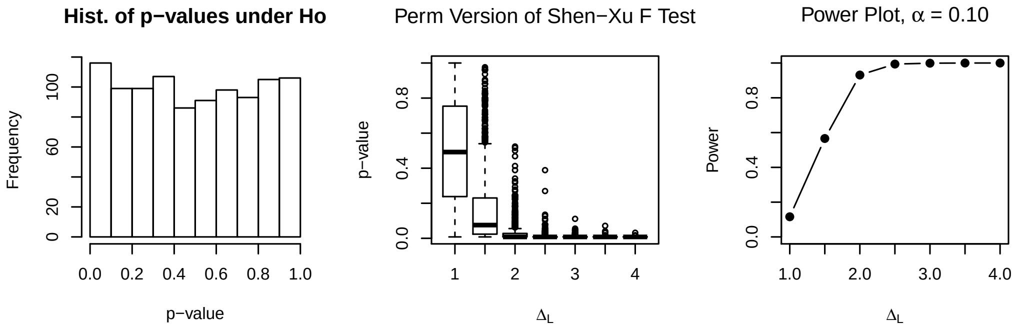

Figure B.11: Histogram, boxplots, and power plot for permutation version of Shen-Xu F test on unimodal curves, plotting over Δ_L.

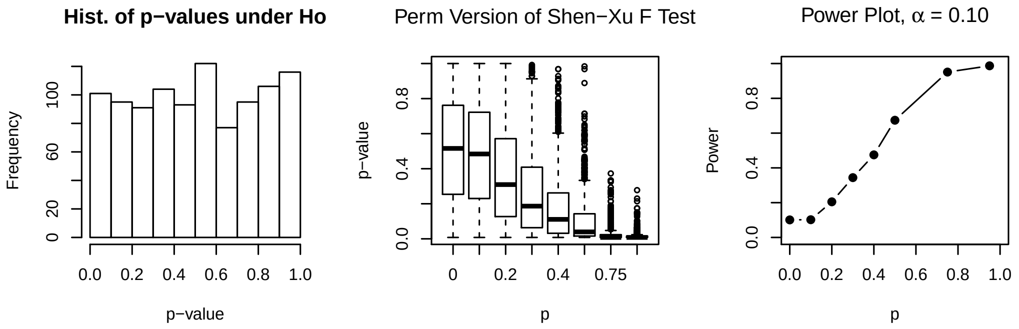

Figure B.12: Histogram, boxplots, and power plot for permutation version of Shen-Xu F test on monotone curves, plotting over p.

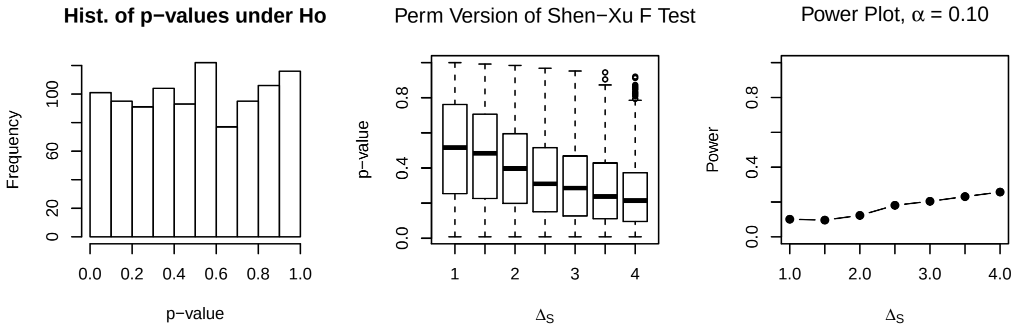

Figure B.13: Histogram, boxplots, and power plot for permutation version of Shen-Xu F test on monotone curves, plotting over Δ_S.

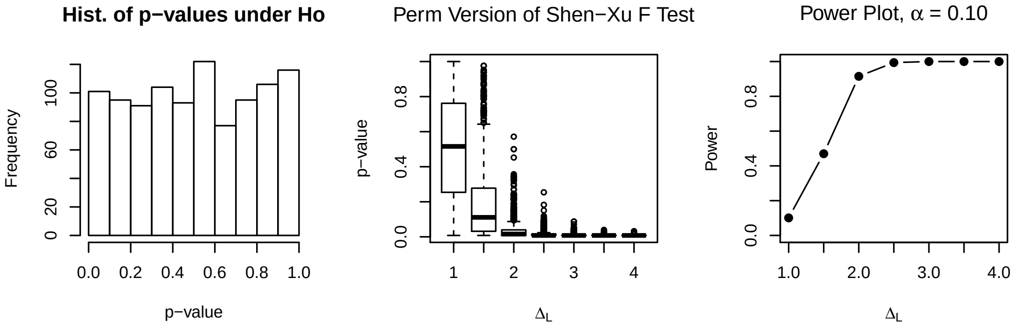

Figure B.14: Histogram, boxplots, and power plot for permutation version of Shen-Xu F test on monotone curves, plotting over Δ_L.

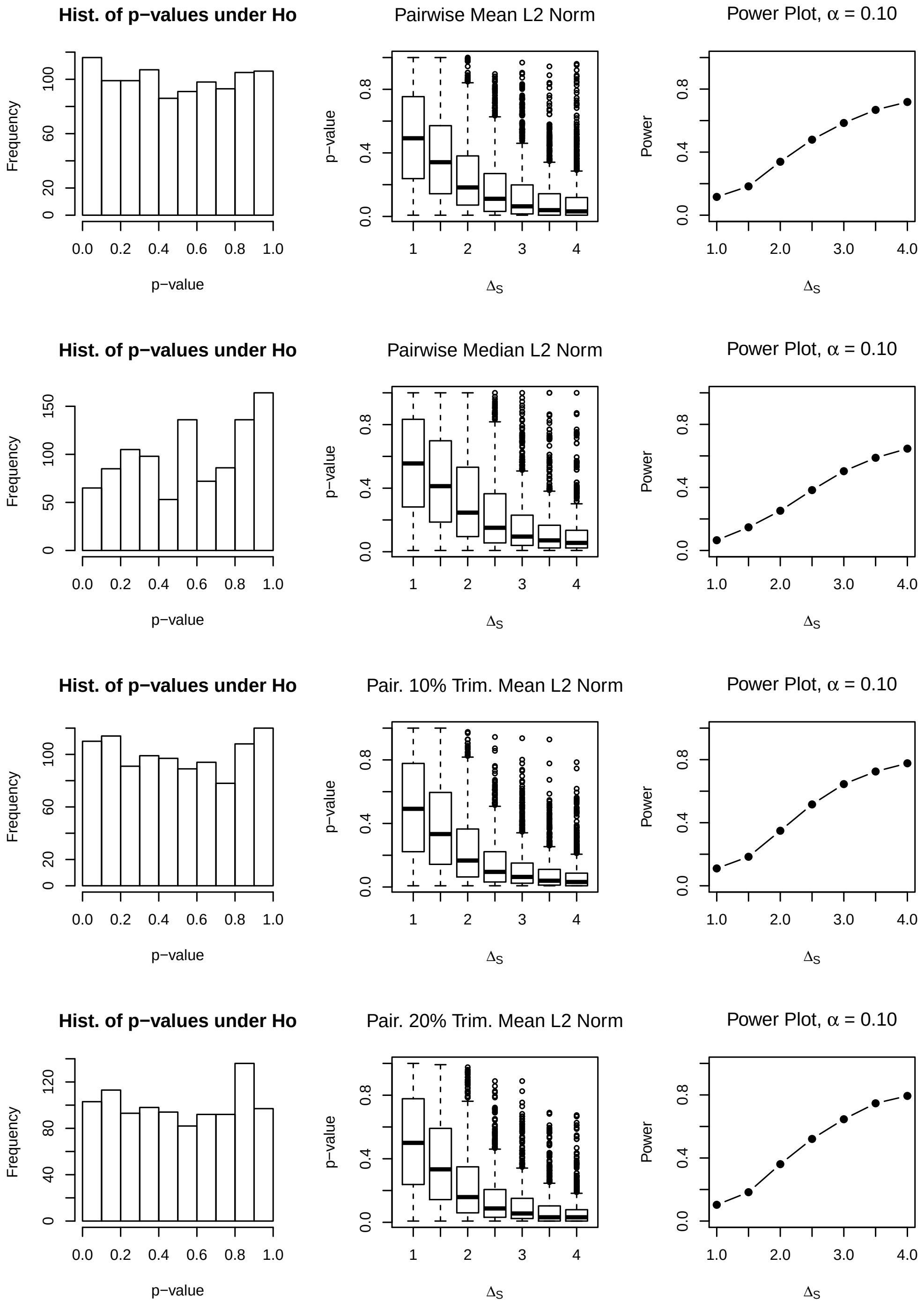

Figure B.15: Histogram, boxplots, and power plots for L_2 norms with varying summary statistics on unimodal curves, plotting over Δ_S.

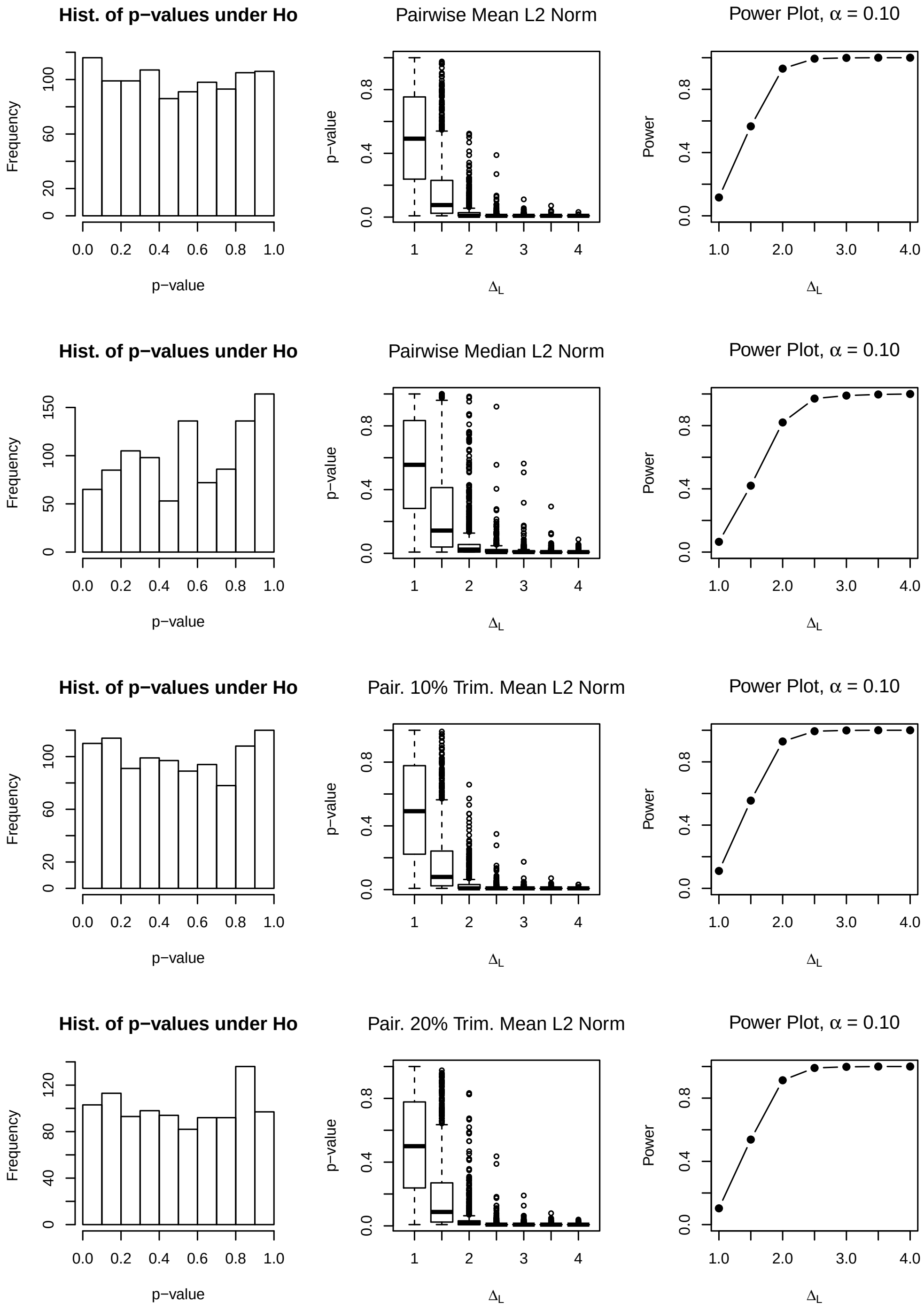

Figure B.16: Histogram, boxplots, and power plots for L_2 norms with varying summary statistics on unimodal curves, plotting over Δ_L.

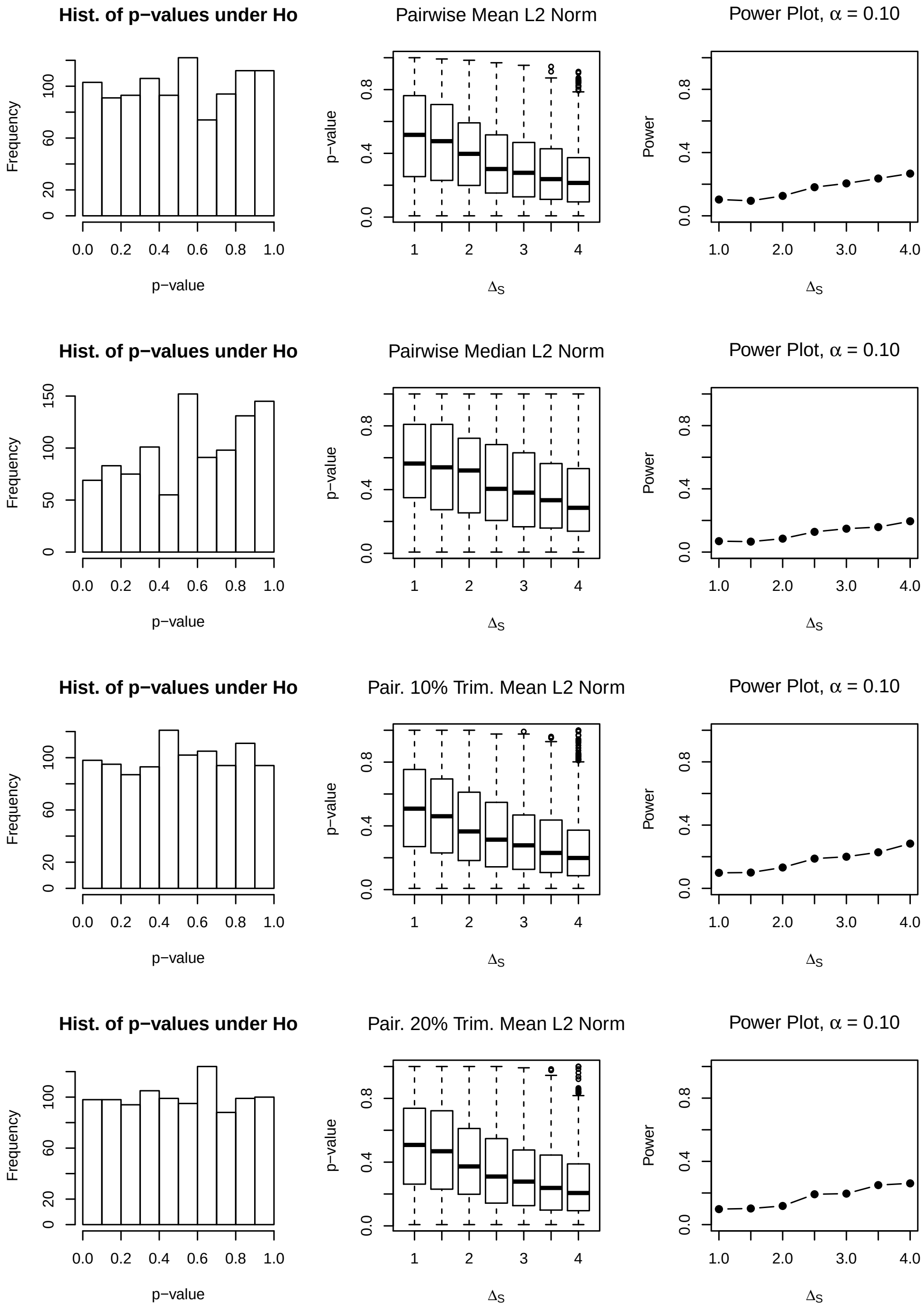

Figure B.17: Histogram, boxplots, and power plots for L_2 norms with varying summary statistics on monotone curves, plotting over Δ_S.

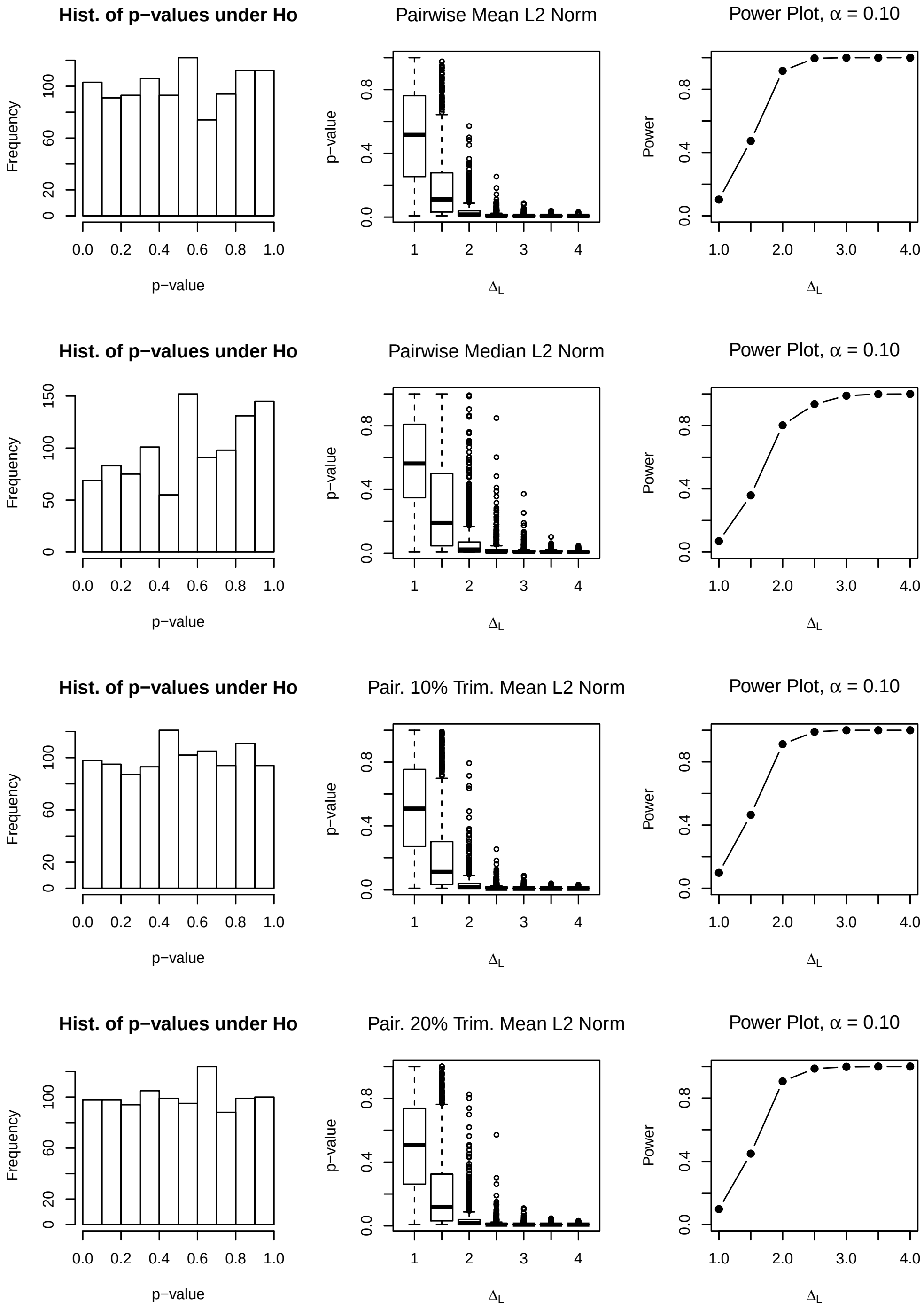

Figure B.18: Histogram, boxplots, and power plots for L_2 norms with varying summary statistics on monotone curves, plotting over Δ_L.

BIBLIOGRAPHY

Felix Abramovich, Anestis Antoniadis, Theofanis Sapatinas, and Brani Vidakovic. Optimal testing in a fixed-effects functional analysis of variance model. *International Journal of Wavelets, Multiresolution and Information Processing*, 2:323–349, 2004. (Cited on pages 11 and 16.)

Katrine Almind and C. Ronald Kahn. Genetic determinants of energy expenditure and insulin resistance in diet-induced obesity in mice. *Diabetes*, 53:3274–3285, 2004. (Cited on page 34.)

Anestis Antoniadis and Theofanis Sapatinas. Estimation and inference in functional mixed-effects models. *Computational Statistics & Data Analysis*, 51:4793–4813, 2007. (Cited on page 18.)

Stephan Arndt, Ted Cizadlo, Nancy C. Andreasen, Dan Heckel, Sherri Gold, and Daniel S. O'Leary. Tests for comparing images based on randomization and permutation methods. *Journal of Cerebral Blood Flow and Metabolism*, 16:1271–1279, 1996. (Cited on page 34.)

W. E. Bardsley, M. A. Jorgensen, P. Alpert, and T. Ben-Gai. A significance test for empty corners in scatter diagrams. *Journal of Hydrology*, 219:1–6, 1999. (Cited on page 33.)

William T. Barry, Andrew B. Nobel, and Fred A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21:1943–1949, 2005. (Cited on page 34.)

William T. Barry, Andrew B. Nobel, and Fred A. Wright. A statistical framework for testing functional categories in microarray data. *The Annals of Applied Statistics*, 2:286–315, 2008. (Cited on page 34.)

Sam Behseta and Robert E. Kass. Testing equality of two functions using BARS. *Statistics in Medicine*, 24:3523–3534, 2005. (Cited on page 17.)

Sam Behseta, Robert E. Kass, and Garrick L. Wallstrom. Hierarchical models for assessing variability among functions. *Biometrika*, 92:419–434, 2005. (Cited on page 17.)

Sam Behseta, Robert E. Kass, David E. Moorman, and Carl R. Olson. Testing equality of several functions: Analysis of single-unit firing-rate curves across multiple experimental conditions. *Statistics in Medicine*, 26:3958–3975, 2007. (Cited on pages 11, 17 and 18.)

Matthew Belmonte and Deborah Yurgelun-Todd. Permutation testing made practical for functional magnetic resonance image analysis. *IEEE Transactions of Medical Imaging*, 20:243–248, 2001. (Cited on page 34.)

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300, 1995. (Cited on page 34.)

Michal Benko, Wolfgang Härdle, and Alois Kneip. Common functional principal components. *The Annals of Statistics*, 37:1–34, 2009. (Cited on page 16.)

Phillippe C. Besse and Hervé Cardot. Approximation spline de la prévsion d'un processus fonctionnel autorégressif d'ordre 1. *The Canadian Journal of Statistics*, 24:467–487, 1996. (Cited on page 10.)

R. Clifford Blair, James J. Higgins, Walt Karniski, and Jeffrey D. Kromrey. A study of multivariate permutation tests which may replace Hotelling's t^2 test in prescribed circumstances. *Multivariate Behavioral Research*, 2:141–163, 1994. (Cited on page 35.)

Federico Bugni, Peter Hall, Joel L. Horowitz, and George R. Neumann. Goodness-of-fit tests for functional data. *Econometrics Journal*, 12:S1–S18, 2009. (Cited on page 35.)

Ed Bullmore, Chris Long, John Suckling, Jalal Fadili, Gemma Calvert, Fernando Zelaya, T. Adrian Carpenter, and Mick Brammer. Colored noise and computational inference in neurophysiological (fMRI) time series analysis: Resampling methods in time and wavelet domains. *Human Brain Mapping*, 12:61–78, 2001. (Cited on page 34.)

Edward T. Bullmore, John Suckling, Stephan Overmeyer, Sophia Rabe-Hesketh, Eric Taylor, and Michael J. Brammer. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Transactions on Medical Imaging*, 18:32–42, 1999. (Cited on page 34.)

William B. Capra and Hans-Georg Müller. An accelerated-time model for response curves. *Journal of the American Statistical Association*, 92:72–83, 1997. (Cited on page 27.)

Hervé Cardot, Frédéric Ferraty, André Mas, and Pascal Sarda. Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics*, 30:241–255, 2003. (Cited on page 9.)

Hervé Cardot, Aldo Goia, and Pascal Sarda. Testing for no effect in functional linear regression models, some computational approaches. *Communications in Statistics - Simulation and Computation*, 33:179–199, 2004. (Cited on pages 9 and 36.)

Hervé Cardot, Christophe Crambes, Alois Kneip, and Pascal Sarda. Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics & Data Analysis*, 51:4832–4848, 2007a. (Cited on page 9.)

Hervé Cardot, Luboš Prchal, and Pascal Sarda. No effect and lack-of-fit permutation tests for functional regression. *Computational Statistics*, 22:371–390, 2007b. (Cited on page 36.)

E. T. Castano, E. S. Mercado, F. G. Leon, C. H. Gorrostieta, J. J. Chamorro, E. B. Vazquez, and V. T. Aguirre. Statistical functional modeling of quality changes of garlic under different storage regimes. *Journal of Data Science*, 4:233–246, 2006. (Cited on pages 11, 35 and 36.)

Stephane Champley, Bruno Guinand, Jean Thioulouse, and Annabelle Clermidy. Functional data analysis of curve asymmetry with applications to the color pattern of *Hydropsyche contubernalis* head capsule. *Biometrics*, 53:294–305, 1997. (Cited on page 33.)

Wilkin Chau, Anthony R. McIntosh, Stephen E. Robinson, Matthias Schulz, , and Christo Pantev. Improving permutation test power for group analysis of spatially filtered MEG data. *NeuroImage*, 23:983–996, 2004. (Cited on page 34.)

Probal Chaudhuri and J. S. Marron. SiZer for exploration of structure in curves. *Journal of the American Statistical Association*, 94:807–823, 1999. (Cited on page 29.)

Probal Chaudhuri and J. S. Marron. Curvature vs. slope inference for features in nonparametric curve estimates. Unpublished manuscript, 2002. (Cited on page 29.)

Jeng-Min Chiou and Hans-Georg Müller. Diagnostics for functional regression via residual processes. *Computational Statistics & Data Analysis*, 51:4849–4863, 2007. (Cited on page 36.)

Jeng-Min Chiou, Hans-Georg Müller, and Jane-Ling Wang. Functional response models. *Statistica Sinica*, 14:675–693, 2004. (Cited on page 14.)

J. H. Chung and D. A. S. Fraser. Randomization tests for a multivariate two-sample problem. *Journal of the American Statistical Association*, 53:729–735, 1958. (Cited on page 35.)

G. A. Churchill and R. W. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138:963–971, 1994. (Cited on page 34.)

Dennis D. Cox and Jong Soo Lee. Pointwise testing with functional data using the WestfallYoung randomization method. *Biometrika*, 95:621–634, 2008. (Cited on page 34.)

Levent Şendur, Voichiţa Maxim, Brandon Whitcher, and Ed Bullmore. Multiple hypothesis mapping of functional MRI data in orthogonal and complex wavelet domains. *IEEE Transactions on Signal Processing*, 53:3413–3426, 2005. (Cited on page 34.)

J. A. Cuesta-Albertos, E. del Barrio, R. Fraiman, and C. Matrán. The random projectional method in goodness of fit for functional data. *Computational Statistics & Data Analysis*, 51:4814–4831, 2007. (Cited on page 18.)

Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. An ANOVA test for functional data. *Computational Statistics & Data Analysis*, 47:111–122, 2004. (Cited on pages 14, 35 and 36.)

Kathryn S. Dawson, Chris Gennings, and Walter H. Carter. Two graphical techniques useful in detecting correlation structure in repeated measures data. *The American Statistician*, 51:275–283, 1997. (Cited on page 29.)

Miguel A. Delgado. Testing the equality of nonparametric regression curves. *Statistics & Probability Letters*, 17:199–204, 1993. (Cited on page 17.)

Pedro Delicado. Functional k-sample problem when data are density functions. *Computational Statistics*, 22:391–410, 2007. (Cited on page 36.)

Holger Dette and Stephan Derbort. Analysis of variance in nonparametric regression models. *Journal of Multivariate Analysis*, 76:110–137, 2001. (Cited on page 17.)

Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of Longitudinal Data, Second Edition*. Oxford University Press, 2002. (Cited on pages 12, 20 and 24.)

Yu Ding, Li Zeng, and Shiyu Zhou. Phase I analysis for monitoring nonlinear profiles in manufacturing processes. *Journal of Quality Technology*, 38:199–216, 2006. (Cited on pages 22 and 23.)

Angela Distaso, Luca Abatangelo, Rosalia Maglietta, Teresa Maria Creanza, Ada Piepoli, Massimo Carella, Annarita D'Addabbo, and Nicola Ancona. Biological and functional analysis of statistically significant pathways deregulated in colon cancer by using gene expression profiles. *International Journal of Biological Sciences*, 4:368–378, 2008. (Cited on page 34.)

R. W. Doerge and G. A. Churchill. Permutation tests for multiple loci affecting a quantitive character. *Genetics*, 142:285–294, 1996. (Cited on page 34.)

Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, and Terence P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–139, 2002. (Cited on page 34.)

Eugene S. Edgington. Randomization tests. *The Journal of Psychology*, 57:445–449, 1964. (Cited on page 37.)

Eugene S. Edgington. Approximate randomization tests. *The Journal of Psychology*, 72:143–149, 1969. (Cited on page 39.)

Eugene S. Edgington. *Randomization Tests*. Marcel Dekker, Inc., 1980. (Cited on page 33.)

Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1:107–129, 2007. (Cited on page 34.)

Jianqing Fan. Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association*, 91:674–688, 1996a. (Cited on pages 5, 15, 42, 51 and 52.)

Jianqing Fan. Adaptive Neyman test code. `http://www.orfe.princeton.edu/%7Ejqfan/papers/pub/aneyman.s`, 1996b. (Cited on page 77.)

Jianqing Fan and Sheng-Kuei Lin. Test of significance when data are curves. *Journal of the American Statistical Association*, 93:1007–1021, 1998. (Cited on pages 5, 15, 42, 51, 52, 54 and 59.)

Jianqing Fan and Jin-Ting Zhang. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62:303–322, 2000. (Cited on pages 25 and 29.)

Yixin Fang and Yuanjia Wang. Testing for familial aggregation of functional traits. *Statistics in Medicine*, 28:3611–3625, 2009. (Cited on page 33.)

Julian J. Faraway. Regression analysis for a functional response. *Technometrics*, 39:254–261, 1997. (Cited on pages 2, 7, 11, 13 and 14.)

Michael P. Fay and Joanna H. Shih. Permutation tests using estimated distribution functions. *Journal of the American Statistical Association*, 93:387–396, 1998. (Cited on page 35.)

Frédéric Ferraty and Yves Romain. *The Oxford Handbook of Functional Data Analysis*. Oxford University Press, 2011. (Cited on page 24.)

Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis*. Springer, 2006. (Cited on page 13.)

Frédéric Ferraty, hilippe Vieu, and Sylvie Viguier-Pla. Factor-based comparison of groups of curves. *Computational Statistics & Data Analysis*, 51:4903–4910, 2007. (Cited on page 16.)

B. Flury. *Common Principal Components and Related Models*. Wiley, New York, 1988. (Cited on page 16.)

Ricardo Fraiman and Graciela Muniz. Trimmed means for functional data. *Test*, 10:419–440, 2001. (Cited on page 19.)

Martha M. Gardner, Jye-Chyi Lu, Ronald S. Gyurcsik, Jimmie J. Wortman, Brian E. Hornung, Holger H. Heinisch, Eric A. Rying, Suraj Rao, Joseph C. Davis, and Purnendu K. Mozumder. Equipment fault detection using spatial signatures. *IEEE Transactions of Components, Packaging, and Manufacturing Technology - Part C*, 20:295–304, 1997. (Cited on page 22.)

Theo Gasser and Alois Kneip. Searching for structure in curve samples. *Journal of the American Statistical Association*, 90:1179–1188, 1995. (Cited on page 27.)

Theo Gasser, Hans-Georg Müller, Walter Köhler, Andrea Prader, Remo Largo, and Luciano Molinari. An analysis of the mid-growth and adolescent spurts of height based on acceleration. *Annals of Human Biology*, 12:129–148, 1985. (Cited on pages 11 and 27.)

José E. González-García, José L. Sanz-González, and Francisco Álvarez Vaquero. Nonparametric permutation tests versus parametric tests in radar detection under k-distributed clutter. In *Radar Conference, 2005 IEEE International*, pages 250 – 255, May 2005. (Cited on page 33.)

Phillip I. Good. *Resampling Methods A Practical Guide to Data Analysis*. Birkhäuser, third edition, 2006. (Cited on pages 33, 37, 39 and 130.)

Mark C. Greenwood, Richard S. Sojda, and Todd M. Preston. Functional linear models to test for differences in prairie wetland hydraulic gradients. *International Environmental Modelling and Software Society (iEMSs) 2010 International Congress on Environmental Modelling and Software Modelling for Environments Sake*, 2010. (Cited on page 36.)

Cortland K. Griswold, Richard Gomulkiewicz, and Nancy Heckman. Hypothesis testing in comparative and experimental studies of function-valued traits. *Evolution*, 62:1229–1242, 2008. (Cited on page 33.)

Wensheng Guo. Inference in smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B*, 64:887–898, 2002. (Cited on page 18.)

Daniel B. Hall and Thomas A. Severini. Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association*, 93:1365–1375, 1998. (Cited on page 11.)

Peter Hall and Nader Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89:359–374, 2002. (Cited on page 35.)

Peter Hall and Ingrid Van Keilegom. Two-sample tests in functional data analysis starting from discrete data. Technical report, IAP Statistics Network, Interuniversity Attraction Pole, 2006. (Cited on page 26.)

Peter Hall and Ingrid Van Keilegom. Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica*, 17:1511–1531, 2007. (Cited on page 26.)

Peter Hall, James Reimann, and John Rice. Nonparametric estimation of a periodic function. *Biometrika*, 87:545–557, 2000. (Cited on page 11.)

Peter Hall, D. S. Poskitt, and Brett Presnell. A functional data-analytic approach to signal discrimination. *Technometrics*, 43:1–9, 2001a. (Cited on pages 11 and 25.)

Sharon M. Hall, Kevin L. Delucchi, Wayne F. Velicer, Christopher W. Kahler, James Ranger-Moore, Donald Hedeker, Janice Y. Tsoh, and Ray Niaura. Statistical analysis of randomized trials in tobacco treatment: Longitudinal designs with dichotomous outcome. *Nicotine & Tobacco Research*, 3:193–202, 2001b. (Cited on pages 11 and 19.)

David Hand and Martin Crowder. *Practical Longitudinal Data Analysis*. Chapman & Hall, 1996. (Cited on page 20.)

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55:757–796, 1993. (Cited on page 18.)

Dan Heckel, Stephan Arndt, Ted Cizadlo, and Nancy C. Andreasen. An efficient procedure for permutation tests in imaging research. *Computers and Biomedical Research*, 31:164–171, 1998. (Cited on page 39.)

Nancy E. Heckman and Ruben H. Zamar. Comparing the shapes of regression functions. *Biometrika*, 87:135–144, 2000. (Cited on page 18.)

Daniel Hlubinka and Luboš Prchal. Changes in atmospheric radiation from the statistical point of view. *Computational Statistics & Data Analysis*, 51:4926–4941, 2007. (Cited on page 11.)

A. P. Holmes, R. C. Blair, J. D. G. Watson, and I. Ford. Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow and Metabolism*, 16:7–22, 1996. (Cited on page 33.)

Lajos Horváth, Piotr Kokoszka, and Matthew Reimherr. Two sample inference in functional linear models. *The Canadian Journal of Statistics*, 37:571–591, 2009. (Cited on page 16.)

Rob J. Hyndman and Md. Shahid Ullah. Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51:4942–4956, 2007. (Cited on page 11.)

Gareth M. James and Ashish Sood. Performing hypothesis tests on the shape of functional data. *Computational Statistics & Data Analysis*, 2005. (Cited on page 36.)

Gareth M. James and Ashish Sood. Performing hypothesis tests on the shape of functional data. *CompStatDatAn*, 2006. (Cited on page 19.)

Gareth M. James, Trevor J. Hastie, and Catherine A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87:587–602, 2000. (Cited on pages 11 and 26.)

Willis A. Jensen, Jeffrey B. Birch, and William H. Woodall. High breakdown estimation methods for Phase I multivariate control charts. *Quality and Reliability Engineering International*, 23:615–629, 2007. (Cited on page 23.)

Willis A. Jensen, Jeffrey B. Birch, and William H. Woodall. Monitoring correlation within linear profiles using mixed models. *Journal of Quality Technology*, 40: 167–183, 2008. (Cited on page 23.)

Jionghua Jin and Jianjun Shi. Automatic feature extraction of waveform signals for in-process diagnostic performance improvement. *Journal of Intelligent Manufacturing*, 12:257–268, 2001. (Cited on pages 10 and 22.)

M. C. Jones and John A. Rice. Displaying the important features of large collections of similar curves. *The American Statistician*, 46:140–145, 1992. (Cited on page 26.)

Lan Kang and Susan L. Albin. On-line monitoring when the process yields a linear profile. *Journal of Quality Technology*, 32:418–426, 2000. (Cited on pages 10 and 23.)

Frédérique Kazi-Aoual, Simon Hitier, Robert Sabatier, and Jean-Dominique Lebreton. Refined approximations to permutation tests for multivariate inference. *Computational Statistics & Data Analysis*, 20:643–656, 1995. (Cited on page 39.)

A. Kelemen, G. Szkely, and G. Gerig. Three-dimensional model-based segmentation. Technical Report 178, Image Science Lab, ETH Zürich, 1997. (Cited on page 11.)

Sallie Keller-McNulty and James J. Higgins. Effect of tail weight and outliers on power and type-I error of robust permutation tests for location. *Communications in Statistics - Simulation and Computation*, 16:17–35, 1987. (Cited on page 39.)

Keunpyo Kim, Mahmoud A. Mahmoud, and William H. Woodall. On the monitoring of linear profiles. *Journal of Quality Technology*, 35:317–327, 2003. (Cited on page 23.)

A. Kneip, X. Li, K. B. MacGibbon, and J. O. Ramsay. Curve registration by local regression. *The Canadian Journal of Statistics*, 28:19–29, 2000. (Cited on page 27.)

Alois Kneip and Theo Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, 20:1266–1305, 1992. (Cited on page 27.)

Jim Kuelbs and Anand N. Vidyashankar. Asymptotic inference for high-dimensional data. *The Annals of Statistics*, 38:836–869, 2010. (Cited on page 18.)

Jobst Landgrebe, Wolfgang Wurst, and Gerhard Welzl. Permutation-validated principal components analysis of microarray data. *Genome Biology*, 3:1–11, 2002. (Cited on page 34.)

J. F. Lawless, R. J. Mackay, and J. A. Robinson. Analysis of variation transmission in manufacturing process - part I. *Journal of Quality Technology*, 31:131–142, 1999. (Cited on page 10.)

Ker-Chau Li, Ming Yan, and Shinsheng Yuan. A simple statistical model for depicting the CDC15-synchronized yeast cell-cycle regulated gene expression data. *Statistica Sinica*, 12:141–158, 2002. (Cited on pages 11 and 26.)

Xueli Liu and Hans-Georg Müller. Modes and clustering for time-warped gene expression profile data. *Bioinformatics*, 19:1937–1944, 2003. (Cited on page 27.)

Joseph J. Locascio, Peggy J. Jennings, Christopher I. Moore, and Suzanne Corkin. Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. *Human Brain Mapping*, 5:168–193, 1997. (Cited on page 34.)

N. Loncatore, J. S. Marron, D. G. Simpson, N. Tripoli, J. T. Zhang, and K. L. Cohen. Robust principal component analysis for functional data. *Test*, 8:1–73, 1999. (Cited on page 11.)

Jesú A. López, Emilio Benfenati, and Werner Dubitzky. *Knowledge Exploration in Life Science Informatics*. Springer, 2004. (Cited on page 11.)

Sara López-Pintado and Juan Romo. Depth-based inference for functional data. *Computational Statistics & Data Analysis*, 51:4957–4968, 2007. (Cited on page 36.)

Mahmoud A. Mahmoud and William H. Woodall. Phase I analysis of linear profiles with calibration applications. *Technometrics*, 46:380–391, 2004. (Cited on page 23.)

James S. Marron, Hans-Georg Müller, Lohn Rice, Jane-Ling Wang, Naisyin Wang, and Yuedong Wang. Discussion of nonparametric and semiparametric regression. *Statistica Sinica*, 14:615–624, 2004. (Cited on pages 12 and 30.)

A. R. McIntosh, W. K. Chau, and A. B. Protzner. Spatiotemporal analysis of event-related fMRI data using partial least squares. *NeuroImage*, 23:764–775, 2004. (Cited on page 34.)

Oto Mestek, Jiří Pavlík, and Miroslav Suchánek. Multivariate control charts: Control charts for calibration curves. *Fresenius' Journal of Analytical Chemistry*, 350:344–351, 1994. (Cited on page 11.)

Jeffrey S. Morris and Raymond J. Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68: 179–199, 2006. (Cited on page 18.)

Yolanda Muñoz Maldonado, Joan G. Staniswalis, Louis N. Irwin, and Donna Byers. A similarity analysis of curves. *The Canadian Journal of Statistics*, 30:373–381, 2002. (Cited on pages 11, 19, 33 and 35.)

Hans-Georg Müller. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32:223–240, 2005. (Cited on pages 10, 11, 12, 28 and 31.)

Axel Munk and Holger Dette. Nonparametric comparison of several regression functions: Exact and asymptotic theory. *The Annals of Statistics*, 26:2339–2368, 1998. (Cited on page 17.)

D. Nain, M. Styner, M. Niethammer, J. J. Levitt, M. E. Shenton, G. Gerig, A. Bobick, and A. Tannenbaum. Statistical shape analysis of brain structures using spherical wavelets. *IEEE*, 2007. (Cited on page 34.)

Vijayan N. Nair, Winson Taam, and Kenny Q. Ye. Analysis of functional responses from robust design studies. *Journal of Quality Technology*, 34:355–370, 2002. (Cited on pages 10 and 84.)

NASA. `http://www.nasa.gov/mission_pages/constellation/main/index2.html`, 2012. (Cited on page 62.)

David Nerini and Badih Ghattas. Classifying densities using functional regression trees: Applications in oceanology. *Computational Statistics & Data Analysis*, 51:4984–4993, 2007. (Cited on page 11.)

Dan Nettleton, Justin Recknor, and James M. Reecy. Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*, 24:192–201, 2008. (Cited on page 34.)

Natalie Neumeyer and Holger Dette. Nonparametric comparison of regression curves: An empirical approach. *The Annals of Statistics*, 31:880–920, 2003. (Cited on page 17.)

Thomas E. Nichols and Andrew P. Holmes. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15:1–25, 2001. (Cited on pages 34 and 35.)

Emma O'Connor, Nick Fieller, Andrew Holmes, John C. Waterton, and Edward Ainscow. Functional principal component analyses of biomedical images as outcome measures. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 59:57–76, 2010. (Cited on page 34.)

Fortunato Pesarin and Luigi Salmaso. *Permutation Tests for Complex Data*. Wiley, 2010. (Cited on page 33.)

E. J. G. Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4:199–130, 1937a. (Cited on pages 5, 33 and 39.)

E. J. G. Pitman. Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*, 4:225–232, 1937b. (Cited on pages 5 and 33.)

E. J. G. Pitman. Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, 29:322–335, 1938. (Cited on page 33.)

Stan Pounds, Xueyuan Cao, Cheng Cheng, Jun Yang, Dario Campana, William E. Evans, Ching-Hon Pui, , and Mary V. Relling. Integrated analysis of pharmacokinetic, clinical, and SNP microarray data using projection onto the most interesting statistical evidence with adaptive permutation testing. *IEEE International Conference on Bioinformatics and Biomedicine*, 52:203–209, 2009. (Cited on page 34.)

Alkes L. Price, Gregory V. Kryukov, Paul I.W. de Bakker, Shaun M. Purcell, Jeff

Staples, Lee-Jen Wei, and Shamil R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86:1–7, 2010. (Cited on page 33.)

Andrey A. Ptitsyn, Sanjin Zvonic, and Jeffrey M. Gimble. Permutation test for periodicity in short time series data. *BMC Bioinformatics*, 7(Suppl 2):S10, 2006. (Cited on page 34.)

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org/`. ISBN 3-900051-07-0. (Cited on pages 62, 70, 77 and 94.)

J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B*, 53:539–572, 1991. (Cited on pages 10 and 24.)

J. O. Ramsay and Xiaochun Li. Curve registration. *Journal of the Royal Statistical Society. Series B*, 60:351–363, 1998. (Cited on page 27.)

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer-Verlag, 1997. (Cited on pages 7, 13 and 24.)

J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis, Methods and Case Studies*. Springer-Verlag, 2002. (Cited on page 13.)

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis, Second Edition*. Springer-Verlag, 2005. (Cited on pages 12, 13, 14, 24, 26, 27 and 84.)

J. O. Ramsay, Hadley Wickham, Spencer Graves, and Giles Hooker. *fda: Functional Data Analysis*, 2011. URL `http://CRAN.R-project.org/package=fda`. R package version 2.2.6. (Cited on pages 48 and 79.)

J.O. Ramsay, Giles Hooker, and Spencer Graves. *Functional Data Analysis with R and Matlab*. Springer, 2009. (Cited on pages 6, 15, 42, 48, 59 and 124.)

C. Radhakrishna Rao. Some statistical methods for comparison of growth curves. *Biometrics*, 14:1–17, 1958. (Cited on page 10.)

Sarah J. Ratcliffe, Gillian Z. Heller, and Leo R. Leader. Functional data analysis with application to periodically stimulated foetal heart rate data II: Functional logistic regression. *Statistics in Medicine*, 21:1115–1127, 2002a. (Cited on pages 11 and 28.)

Sarah J. Ratcliffe, Leo R. Leader, and Gillian Z. Heller. Functional data analysis with application to periodically stimulated foetal heart rate data I: Functional regression. *Statistics in Medicine*, 21:1103–1114, 2002b. (Cited on page 11.)

Jonathan Raz. Testing for no effect when estimating a smooth function by nonparametric regression: A randomization approach. *Journal of the American Statistical Association*, 85:132–138, 1990. (Cited on page 39.)

John A. Rice. Functional and longitudinal data analysis: Perspectives on smoothing. *Statistica Sinica*, 14:631–647, 2004. (Cited on pages 12, 25 and 31.)

John A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society Series B*, 53:233–243, 1991. (Cited on pages 10 and 26.)

John A. Rice and Colin O. Wu. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57:253–259, 2001. (Cited on page 25.)

Narinder Singh Sahni, Are Halvor Aasveit, and Tormod Naes. In-line process and product control using spectroscopy and multivariate calibration. *Journal of Quality Technology*, 37:1–20, 2005. (Cited on page 10.)

José L. Sanz-González, Francisco Álvarez Vaquero, and José E. González-García. Permutation test algorithms for nonparametric radar detection. *IET Conference Publications*, 2007(CP530):66–66, 2007. (Cited on page 33.)

Richard L. Schmoyer. Permutation tests for correlation in regression errors. *Journal of the American Statistical Association*, 89:1507–1516, 1994. (Cited on page 36.)

Qing Shen. *Linear models for a functional response.* PhD thesis, University of Michigan, 1999. (Cited on pages 7 and 131.)

Qing Shen and Julian Faraway. An F test for linear models with functional responses. *Statistica Sinica*, 14:1239–1257, 2004. (Cited on pages 7, 11, 14, 20, 36, 42, 55, 56, 57 and 91.)

Qing Shen and Hongquan Xu. Diagnostics for linear models with functional responses. *Technometrics*, 49:26–33, 2007. (Cited on pages 4, 42, 55, 56, 57, 75 and 84.)

Jian Qing Shi and Taeryon Choi. *Gaussian Process Regression Analysis for Functional Data.* CRC Press, Taylor & Francis Group, 2011. (Cited on page 28.)

Minggao Shi, Robert E. Weiss, and Jeremy M. G. Taylor. An analysis of paediatric cd4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics*, 45:151–163, 1996. (Cited on page 25.)

Joanna H. Shih and Michael P. Fay. A class of permutation tests for stratified survival data. *Biometrics*, 55:1156–1161, 1999. (Cited on page 35.)

Bill Shipley. Inferential permutation tests for maximum entropy models in ecology. *Ecology*, 91:2794–2805, 2010. (Cited on page 33.)

Bernard W. Silverman. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24:1–24, 1996. (Cited on page 26.)

Terry Speed. Discussion of "That BLUP is a good thing: the estimation of random effects" by G. K. Robinson. *Statistical Science*, 6:42–44, 1999. (Cited on page 18.)

Christina Staudhammer, Thomas C. Maness, and Robert A. Kozak. Profile charts for monitoring lumber manufacturing using laser range sensor data. *Journal of Quality Technology*, 39:224–240, 2007. (Cited on page 10.)

Christina L. Staudhammer, Valerie M. Lemay, Robert A. Kozak, and Thomas C. Maness. Mixed-model development for real-time statistical process control data in wood products manufacturing. *Forest Biometry, Modelling, and Information Sciences*, 1:19–35, 2005. (Cited on pages 10 and 23.)

Frederick S. Stover and Robert V. Brill. Statistical quality control applied to ion chromatography calibrations. *Journal of Chromatography A*, 804:37–43, 1998. (Cited on pages 10 and 23.)

Joseph Sturino, Ivan Zorych, Bani Mallick, Karina Pokusaeva, Ying-Ying Chang, Raymond J. Carroll, and Nikolay Bliznuyk. Statistical methods for comparative phenomics using high-throughput phenotype microarrays. *The International Journal of Biostatistics*, 6(1):Article 29, 2010. (Cited on pages 5, 6, 7, 16, 19, 37, 41, 42, 46, 47, 49, 50, 58, 59, 81 and 112.)

John Suckling and Ed Bullmore. Permutation tests for factorially designed neuroimaging experiments. *Human Brain Mapping*, 22:193–205, 2004. (Cited on page 34.)

Inga K. Teismann, Olaf Steinstraeter, Kati Stoeckigt, Sonja Suntrup, Andreas Wollbrink, Christo Pantev, and Rainer Dziewas. Functional oropharyngeal sensory disruption interferes with the cortical control of swallowing. *BMC Neuroscience*, 8:62, 2007. (Cited on page 34.)

Timothy B. Terriberry, Sarang C. Joshi, and Guido Gerig. Hypothesis testing with nonlinear shape models. In *Information Processing in Medical Imaging (IPMI'05)*, pages 15–26, 2005. (Cited on page 33.)

Kongming Wang and Theo Gasser. Synchronizing sample curves nonparametrically. *The Annals of Statistics*, 27:439–460, 1999. (Cited on page 27.)

William H. Woodall. Current research on profile monitoring. *Revista Produção*, 17: 420–425, 2007. (Cited on page 22.)

William H. Woodall, Dan J. Spitzner, Douglas C. Montgomery, and Shilpa Gupta. Using control charts to monitor process and product quality profiles. *Journal of Quality Technology*, 36:309–320, 2004. (Cited on pages 12, 22 and 24.)

Rongling Wu and Min Lin. Functional mapping - how to map and study the genetic architecture of dynamic complex traits. *Nature Reviews Genetics*, 7:229–237, 2006. (Cited on page 34.)

Xiaowei Yang, Qing Shen, Hongquan Xu, and Steven Shoptaw. Functional regression analysis using an F test for longitudinal data with large numbers of repeated measures. Technical report, Department of Statistics Papers, University of California, Los Angeles, 2005. (Cited on pages 31 and 32.)

Xiaowei Yang, Qing Shen, Hongquan Xu, and Steven Shoptaw. Functional regression analysis using an F test for longitudinal data with large numbers of repeated measures. *Statistics in Medicine*, 26:1552–1566, 2007. (Cited on pages 11, 21 and 28.)

Fang Yao, Hans-Georg Müller, Andrew J. Clifford, Steven R. Dueker, Jennifer Follet, Yumei Lin, Bruce A. Buchholz, and John S. Vogel. Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59:676–685, 2003. (Cited on pages 11 and 26.)

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100: 577–590, 2005a. (Cited on page 27.)

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33:2873–2903, 2005b. (Cited on page 28.)

John Stephen Yap, Jianqing Fan, and Rongling Wu. Nonparametric modeling of longitudinal covariance structure in functional mapping of quantitative trait loci. *Biometrics*, 65:1068–1077, 2009. (Cited on page 34.)

Paul Yushkevich, Stephen M. Pizer, Sarang Joshi, and J. S. Marron. Intuitive, localized analysis of shape variability. *Information Processing in Medical Imaging*, pages 402–408, 2001. (Cited on page 11.)

Gary O. Zerbe and Strother H. Walker. A randomization test for comparison of groups of growth curves with different polynomial design matrices. *Biometrics*, 33:653–657, 1977. (Cited on page 36.)

Xin Zhao, J. S. Marron, and Martin T. Wells. The functional data analysis view of longitudinal data. *Statistica Sinica*, 14:789–808, 2004. (Cited on pages 29 and 31.)

Chunxiao Zhou and Yongmei Michelle Wang. New blockwise permutation tests preserving exchangeability in functional neuroimaging. *31st Annual International Conference of the IEEE EMBS*, pages 6977–6980, 2009. (Cited on page 39.)

Chunxiao Zhou, Huixia Judy Wang, and Yongmei Michelle Wang. Efficient moments-based permutation tests. *Advances in Neural Information Processing Systems*, 22:2277–2285, 2009. (Cited on page 39.)