FOLLOW-UP TESTING IN FUNCTIONAL ANALYSIS OF VARIANCE

by

Olga Vsevolozhskaya

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

MONTANA STATE UNIVERSITY
Bozeman, Montana

May, 2013

APPROVAL

of a dissertation submitted by

Olga Vsevolozhskaya

This dissertation has been read by each member of the dissertation committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to The Graduate School.

Dr. Mark C. Greenwood

Approved for the Department of Mathematical Sciences

Dr. Kenneth L. Bowers

Approved for The Graduate School

Dr. Ronald W. Larsen

STATEMENT OF PERMISSION TO USE

In presenting this dissertation in partial fulfillment of the requirements for a doctoral degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for extensive copying or reproduction of this dissertation should be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Olga Vsevolozhskaya

May, 2013

ACKNOWLEDGEMENTS

First, I am most thankful to my advisor, Mark Greenwood, for his continuous help and support during my research and studies. For the six years that I have known him, I have grown as a scholar largely due to him. I am very fortunate to have had him as my advisor.

Second, I am grateful to the rest of my committee, especially those who have also been my professors. More specifically, thanks to Jim Robison-Cox for teaching me basics of R and linear models, to John Borkowski for introducing me to LaTeX and for tirelessly trying to convert me into the SAS® user. I thank the rest of my committee for their valuable comments on this thesis.

I thank the members of the Mathematical Sciences department for their kindness and help throughout my studies. More specifically, thanks to Ben Jackson for convincing me to use Linux and Emacs, to Shari Samuels and Kacey Diemert for valuable advises on my writing skills, and to the rest of my office mates for their friendship that I could always rely on.

I thank Dave for his continuous love and support in all of my ventures.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF FIGURES – CONTINUED

Figure Page

ABSTRACT

Sampling responses at a high time resolution is gaining popularity in pharmaceutical, epidemiological, environmental and biomedical studies. For example, investigators might expose subjects continuously to a certain treatment and make measurements throughout the entire duration of each exposure. An important goal of statistical analysis for a resulting longitudinal sequence is to evaluate the effect of the covariates, which may or may not be time dependent, on the outcomes of interest. Traditional parametric models, such as generalized linear models, nonlinear models, and mixed effects models, are all subject to potential model misspecification and may lead to erroneous conclusions in practice. In semiparametric models, a time-varying exposure might be represented by an arbitrary smooth function (the nonparametric part) and the remainder of the covariates are assumed to be fixed (the parametric part). The potential drawbacks of the semiparametric approach are uncertainty in the smoothing function interpretation, and ambiguity in the parametric test (a particular regression coefficient being zero in the presence of the other terms in the model).

Functional linear models (FLM), or the so called structural nonparametric models, are used to model continuous responses per subject as a function of time-variant coefficients and a time-fixed covariate matrix. In recent years, extensive work has been done in the area of nonparametric estimation methods, however methods for hypothesis testing in the functional data setting are still undeveloped and greatly in demand. In this research we develop methods that address hypotheses testing problem in a special class of FLMs, namely the Functional Analysis of Variance (FANOVA). In the development of our methodology, we pay a special attention to the problem of multiplicity and correlation among tests. We discuss an application of the closure principle to the follow-up testing of the FANOVA hypotheses as well as computationally efficient shortcut arising from a combination of test statistics or p-values. We further develop our methods for pair-wise comparison of treatment levels with functional data and apply them to simulated as well as real data sets.

CHAPTER 1

INTRODUCTION.

The purpose of this research is to develop and study statistical methods for functional data analysis (FDA). Most of the motivation arises from the problem of Functional Analysis of Variance (FANOVA), however the applicability of certain approaches described here is broader.

Ramsay and Silverman (1997) define FDA as "*analysis of data whose observations are themselves functions*". In the functional data paradigm, each observed time series is seen as a realization of an underlying stochastic process or smooth curve that needs to be estimated. In practice, the infinite dimensional function $f(t)$ (conventionally, a function of time), is projected onto a finite K-dimensional set of basis functions:

$$f(t) = \sum_{k=1}^{K} \alpha_k \theta_k(t),$$

where α_k's are coefficients (weights) and $\theta_k(t)$ are basis functions. A common choice for basis functions is $1, t, t^2, \ldots, t^k$ which fits a low degree polynomial (regression spline) (de Boor (1978)) to represent $f(t)$. If $f(t)$ is known to have some periodic oscillations, Fourier functions – $1, sin(\omega t), cos(\omega t), sin(2\omega t), cos(2\omega t), \ldots, sin(k\omega t), cos(k\omega t)$ – can be used for the basis. Alternatively, a B-spline base system (Green and Silverman (1994)) can be employed to fit smoothing splines. With B-splines, *knots* are typically equally spaced over the range of t (the two exterior knots are placed at the end point of the functional domain). B-spline basis functions, $\theta_k(t)$, are polynomials of order m pieced together so that $\theta_k(t), \theta_k'(t), \theta_k''(t), \ldots, \theta_k^{(m-1)}(t)$ are continuous at each knot. The coefficients α_k's are fit using penalized least-squares wich includes a

constraint in the curve's smoothness, controlled by a single non-negative smoothing parameter λ.

There are a number of advantages to using a functional data approach over a conventional time series analysis. First, it can handle missing observations. In instances of varying time grids among units (subjects), smoothing techniques can be used to reconstruct the missing time points (Faraway (1997), Xu et al. (2011)). Second, functional data techniques are designed to handle temporally correlated non-linear responses (Ramsay and Silverman (2005)). Finally, it can potentially handle extremely short time series (Berk et al. (2011)).

In a designed experiment with k groups of curves, functional analysis of variance (FANOVA) methods are used to test for a treatment effect. The FANOVA model is written as

$$y_{ij}(t) = \mu_i(t) + \epsilon_{ij}(t),$$

where $i = 1, \ldots, k$, $j = 1, \ldots n_i$ is the number of observations per subject, $\mu_i(t)$ is assumed to be fixed, but unknown, population mean function, and $\epsilon_{ij}(t)$ is the residual error function. There are two distinct ways of modeling error in the FANOVA setting: the *discrete noise model* and the *functional noise model*. In the discrete noise model (Ramsay and Silverman (2005), Luo et al. (2012)), for each $i = 1, \ldots, k$, ϵ_{ij} is considered *independent for different measurement points* and identically distributed normal random variable with mean 0 and constant variance σ^2. In the functional noise model (Zhang et al. (2010), Berk et al. (2011), Xu et al. (2011)), ϵ_{ij} is a Gaussian stochastic process with mean zero and covariance function $\gamma(s, t)$. This choice of model implies that for a discretized error curve, the random errors are independent among subjects, normally distributed within each subject with mean zero and non-diagonal covariance matrix $\mathbf{\Sigma}$, which implies *dependency among different measurement points*. Little re-

search has been done on the impact of a particular noise model on the corresponding inferential method. Hitchcock et al. (2006) provided preliminary results on the effect of the noise model on functional cluster analysis, however more research is required in this direction.

None of the methods that we propose in the current work are affected by the choice of the noise model. Whenever we work with the discretized curves we take a resampling-based approach, which automatically incorporates the correlation structure at the nearby time points into the analysis. There is another advantage to the resampling-based approach which is discussed further in the outline of Chapter 4 in the context of the multiple testing problem.

The FANOVA null and alternative hypotheses are

$$\begin{aligned} H_0 : &\quad \mu_1(t) = \mu_2(t) = \ldots = \mu_k(t) \\ H_a : &\quad \mu_i(t) \neq \mu_{i'}(t), \text{ for at least one } t \text{ and } i \neq i'. \end{aligned}$$

The problem is to assess evidence for the existence or not of differences among population mean curves under k different conditions (treatment levels) somewhere in the entire functional domain. Different approaches have been taken to solve the FANOVA problem. Ramsay and Silverman (2005) as well as Cox and Lee (2008) take advantage of the fact that the measurements are usually made on a finite set of time points. Cuevas et al. (2004) and Shen and Faraway (2004) approach FANOVA testing based on the analysis of the squared norms. An overview of these methods is provided in the beginning of Chapters 2 and 3.

Our initial interest in functional data analysis came from the experiment conducted by Gabriel Bellante as a part of his master's thesis (Bellante (2011)). Bellante was studying methods by which a soil CO_2 leak from a geological carbon seques-

tration (GCS) site can be detected. Since vegetation is the predominant land cover over GCS sites, remote sensing, like periodic airborne imaging, can aid in identifying CO_2 leakage through the detection of plant stress caused by elevated soil CO_2 levels. More specifically, aerial images taken with a hyperspectral camera were proposed for the analysis. Hyperspectral imaging collects information across the electromagnetic spectrum withing continuous narrow reflectance bands. In practice, images collected in this study had 80 radiance measurements at each pixel and these measurements reflected smooth variation over electromagnetic spectrum (see Figure 4.5). The methods by Cuevas et al. (2004) and Shen and Faraway (2004) would have allowed us to decide on the existence or not of differences in mean spectral curves *somewhere* across the electromagnetic spectrum. The methods by Ramsay and Silverman (2005) and Cox and Lee (2008) would have identified points at which the mean curves deviate (additional drawbacks of these two methods are detailed in Chapter 4). However, a research question was to assess evidence for differences over *a priori* specified electromagnetic regions.

In Chapter 2, we develop a follow-up testing procedure for the FANOVA test that addresses the research question described above. The procedure begins by splitting the entire functional domain into mutually exclusive and exhaustive sub-intervals and performing a global test. The null hypothesis of the global test is that there is no difference among mean curves at any of the sub-intervals. The alternative hypothesis is that there is at least one sub-interval where at least one group of mean curves deviate. If the global null hypothesis involving all sub-intervals (i.e., the entire domain) is rejected, it is of interest to "follow-up" and localize one or more sub-intervals where there is evidence of a difference in the mean curves. The procedure that starts with the global test (over the entire functional domain) and proceeds with

subsets of hypotheses (over unions of sub-intervals) and individual hypotheses (over a single sub-interval), is called the "closure" principle of Marcus et al. (1976).

Since with the closure principle the global null hypothesis is expressed as an intersection of the individual null hypotheses (no difference on the entire domain is equivalent to no difference at any of the sub-intervals), it is reasonable to express the test statistic for the global null in terms of the test statistics at the sub-interval level. The procedures that combine evidence against the null hypothesis (either test statistics or p-values) are called "combination methods" (Pesarin (1992), Basso et al. (2009)). We propose a test statistic and perform a series of simulations to study performance of the proposed combination test along with the closure principle in the FANOVA setting. Application of our procedure addressed the research question in the data collected by Bellante (2011). Using our approach we were able to detect evidence for differences over the entire electromagnetic spectrum, as well as over *a priori* specified electromagnetic regions.

In Chapter 3, we extend this research and developed a method for multiple testing of pair-wise differences among treatment levels within regions of significant statistical difference. The motivation for this research came from data collected during a pharmacological experiment. The goal of the experiment was to detect differences in the process of mice red blood cells breakdown (hemolysis) under different dosages of treatment (more on this in Chapter 3). The specific research question was to identify pair-wise differences among mean hemolysis curves. We developed a two-stage follow-up method to the FANOVA problem that allows one to (i) identify regions of time with some difference among curves; (ii) perform comparisons of pairs of treatments within these regions. To the best of our knowledge, there are no existing competing procedures to the proposed methodology. Thus, our numerical results reported in this chapter do not include a comparison of the proposed method to other alternatives.

Nevertheless, the simulations reveal that our procedure has satisfactory power and does a good job of picking out the differences between population means for different combinations of true and false null hypotheses.

In Chapter 4, we focus on a challenging problem of point-wise testing with functional data, which is rather misleadingly termed "naive" in the literature (Abramovich et al. (2002), Cuesta-Albertos and Febrero-Bande (2010), Xu et al. (2011)). The idea is to take advantage of the fact that the measurements are typically made on a finite grid of points. The "naive" approach is to examine the point-wise t or F-statistics at each time point. This approach carries serious problems of multiple testing inflation of error rates along with highly correlated tests over time. Abramovich et al. (2002), Cuesta-Albertos and Febrero-Bande (2010), Xu et al. (2011) all suggested a Bonferroni-type procedure to correct for simultaneous tests, but then concluded that it would yield an extremely low-powered test. This is not a surprising result, since the Bonferroni procedure is designed to correct for *independent* simultaneous tests and becomes extremely conservative with a large number of correlated tests (Cribbie (2007), Smith and Cribbie (2013)).

We propose a powerful method that both provides a decision for the overall hypothesis and adequately adjusts the individual p-values to account for simultaneous tests. The method first uses two different p-value combining methods to summarize the associated evidence across time points; defines a new test statistic based on the smallest p-value from the two combination methods; and applies the closure principle of Marcus et al. (1976) to individually adjust the point-wise p-values. The problem of correlated tests is addressed by using permutation instead of a parametric distribution for finding p-values. More specifically, Cohen and Sackrowitz (2012) note that stepwise multiple testing procedures (including the closure principle) are not designed to account for a correlation structure among hypotheses being tested. That

is, test statistics for an intersection hypothesis will always be the same regardless of the correlation structure among tests considered. Thus, the shortcoming of the stepwise procedures is determining a correct critical value. The resampling-based approach alleviates this shortcoming by accounting for dependency in its calculation of the critical values.

The idea of using the minimum p-value as the test statistic for the overall test across different combination methods has been used in multiple genetics studies (Hoh et al. (2001), Chen et al. (2006), Yu et al. (2009)). A challenge for the proposed analysis was the individual adjustment performed using the closure principle. The closure principle generally requires $2^L - 1$ intersection hypotheses to consider. To overcome this obstacle, we describe a computational shortcut which allows individual adjustments using the closure method even for a large number of tests. We also provide an `R` script (R Core Team (2013)) for the implementation of our method, which makes our methodology available to mass users of `R`.

Most of our work is concentrated around the classical definition of functional data as "*analysis of data whose observations are themselves functions*". Chapter 5 attempts to look at application of our methods in association studies between a quantitative trait with genetic variants (both common and rare) in a genomic region. Similar work exists by Luo et al. (2012) where they considered the model

$$Y_i = \mu + \int_0^T X_i(t)\alpha(t)dt + \epsilon_i,$$

with quantitative discrete response (such as BMI – body mass index) and functional explanatory variable. We suggest to "flip the relationship" and use the FANOVA methods to address the same research question. Chapter 5 also outlines a direction

for future research. Specifically, certain ideas on the drawbacks of the proposed methods as well as ways to overcome them are discussed.

References

Abramovich, F., Antoniadis, A., Sapatinas, T., Vidakovic, B., 2002. Optimal testing in functional analysis of variance models. Tech. rep., Georgia Institute of Technology.

Basso, D., Pesarin, F., Solmaso, L., Solari, A., 2009. Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applications with R. Springer.

Bellante, G. J., 2011. Hyperspectral remote sensing as a monitoring tool for geologic carbon sequestration. Master's thesis, Montana State University.

Berk, M., Ebbels, T., Montana, G., 2011. A statistical framework for biomarker discovery in metabolomic time course data. Bioinformatics 27 (14), 1979–1985.

Chen, B., Sakoda, L., Hsing, A., Rosenberg, P., 2006. Resamplingbased multiple hypothesis testing procedures for genetic casecontrol association studies. Genetic Epidemiology 30, 495–507.

Cohen, A., Sackrowitz, H., 2012. The interval property in multiple testing of pairwise-differences. Statistical Science 27 (2), 294–307.

Cox, D. D., Lee, J. S., 2008. Pointwise testing with functional data using the westfall-young randomization method. Biometrika 95 (3), 621–634.

Cribbie, R. A., 2007. Multiplicity control in structural equation modeling. Structural Equation Modeling 14 (1), 98–112.

Cuesta-Albertos, J. A., Febrero-Bande, M., 2010. Multiway anova for functional data. TEST 19, 537–557.

Cuevas, A., Febrero, M., Fraiman, R., 2004. An anova test for functional data. Computational Statistics and Data Analysis 47, 111–122.

de Boor, C., 1978. A Practical Guide to Splines. Springer, New York.

Faraway, J., 1997. Regression analysis for a functional response. Technometrics 39, 254–261.

Green, P., Silverman, B., 1994. Nonparametric Regression and Generalized Linear Models. Chapman and Hall, London.

Hitchcock, D., Casella, G., Booth, J., 2006. Improved estimation of dissimilarities by presmoothing functional data. Journal of the American Statistical Association 101 (473), 211–222.

Hoh, J., Wille, A., Ott, J., 2001. Trimming, weighting, and grouping snps in human case-control association studies. Genome Research 11, 2115–2119.

Luo, L., Zhu, Y., M., X., 2012. Quantitative trait locus analysis for next-generation sequencing with the functional models. Journal of Medical Genetics 49, 513–524.

Marcus, R., Peritz, E., Gabriel, K. R., 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63 (3), 655–660.

Pesarin, F., 1992. A resampling procedure for nonparametric combination of several dependent tests. Statistical Methods & Applications 1 (1), 87–101.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org`

Ramsay, J., Silverman, B., 1997. Functional Data Analysis. Springer-Verlag, New York.

Ramsay, J. O., Silverman, B. W., 2005. Functional Data Analysis, Second Edition. Springer.

Shen, Q., Faraway, J., 2004. An f test for linear models with functional responses. Statistica Sinica 14, 1239–1257.

Smith, C., Cribbie, R., 2013. Multiplicity control in structural equation modeling: incorporating parameter dependencies. Structural Equation Modeling 20 (1), 79–85.

Xu, H., Shen, Q., Yang, X., Shptaw, S., 2011. A quasi f-test for functional linear models with functional covariates and its application to longitudinal data. Statistics in Medicine 30 (23), 2842–2853.

Yu, K., Li, Q., Bergen, A., Pfeiffer, R., Rosenberg, P., Caporasi, N., Kraft, P., Chatterjee, N., 2009. Pathway analysis by adaptive combination of p-values. Genetic Epidemiology 33, 700–709.

Zhang, C., Peng, H., Zhang, J., 2010. Two sample tests for functional data. Communications in Statistics – Theory and Methods 39 (4), 559–578.

CHAPTER 2

COMBINING FUNCTIONS AND THE CLOSURE PRINCIPLE FOR PERFORMING FOLLOW-UP TESTS IN FUNCTIONAL ANALYSIS OF VARIANCE.

Contribution of Authors and Co-Authors

Author: Olga A. Vsevolozhskaya

Contributions: Responsible for the majority of the writing.

Co-Author: Dr. Mark C. Greenwood

Contributions: Provided feedback on statistical analysis and drafts of the manuscript.

Co-Author: Gabriel J. Bellante

Contributions: Data collection.

Co-Author: Dr. Scott. L. Powell

Contributions: Provided application expertise and feedback on drafts of the manuscript.

Co-Author: Rick L. Lawrence

Contributions: Provided application expertise.

Co-Author: Kevin S. Repasky

Contributions: Provided funding.

Manuscript Information Page

Olga A. Vsevolozhskaya, Mark C. Greenwood, Gabriel J. Powell, Scott L. Powell, Scott L. Powell, Rick L. Lawrence

Journal of Computational Statistics and Data Analysis

Status of Manuscript:

_____ Prepared for submission to a peer-reviewed journal

_____ Officially submitted to a peer-review journal

__X__ Accepted by a peer-reviewed journal

_____ Published in a peer-reviewed journal

Published by Elsevier.

Submitted March, 2013

Abstract

Functional analysis of variance involves testing for differences in functional means across k groups in n functional responses. If a significant overall difference in the mean curves is detected, one may want to identify the location of these differences. Cox and Lee (2008) proposed performing a point-wise test and applying the Westfall-Young multiple comparison correction. We propose an alternative procedure for identifying regions of significant difference in the functional domain. Our procedure is based on a region-wise test and application of a combining function along with the closure multiplicity adjustment principle. We give an explicit formulation of how to implement our method and show that it performs well in a simulation study. The use of the new method is illustrated with an analysis of spectral responses related to vegetation changes from a CO_2 release experiment.

1. Introduction

Functional data analysis (FDA) concerns situations in which collected data are considered a realization of an underling stochastic process. Modern data recording methods often allow researchers to observe a random variable densely in time from t_{min} to t_{max}. Even though each data point is a measure at a discrete point in time, overall these values can reflect smooth variation. Therefore, instead of basing inference on a set of dense time series, it is often desirable to analyze these records as continuous functions.

Situations in which the responses are random functions and the predictor variable is the group membership can be analyzed using Functional Analysis of Variance (FANOVA). The FANOVA model can be written as

$$y_{ij}(t) = \mu_i(t) + \epsilon_{ij}(t), \tag{2.1}$$

where $\mu_i(t)$ is the mean function of group i at time t, $i = 1, \ldots, k$, j indexes a functional response within a group, $j = 1, \ldots, n_i$, and $\epsilon_{ij}(t)$ is the residual function.

In practice, one does not observe $y_{ij}(t)$ for all t but only on a dense grid of points between t_{min} and t_{max}. To construct a functional observation $y_{ij}(t)$ from the discretely observed data one can employ a standard smoothing technique such as smoothing cubic B-splines. An implementation of the smoothing techniques is readily available in R (R Core Team (2013)) in the `fda` package (Ramsay et al. (2012)).

The prime objective of FANOVA is the extension of the ideas of typical analysis of variance. Specifically, within the FANOVA framework, one wants to test for a difference in mean curves from k populations anywhere in t.

$$
\begin{aligned}
H_0 : &\quad \mu_1(t) = \mu_2(t) = \ldots = \mu_k(t) \\
H_a : &\quad \mu_i(t) \neq \mu_{i'}(t), \ \text{ for at least one } t \text{ and } i \neq i'.
\end{aligned}
$$

There are two distinct approaches to solve the FANOVA problem. One approach, considered by Ramsay and Silverman (2005), Ramsay et al. (2009), and Cox and Lee (2008), is *point-wise*. The idea is to evaluate the functional responses on a finite grid of points $\{t_1, \ldots, t_L\} \in [t_{min}, t_{max}]$ and perform a univariate F-test at each t_l, $l = 1, \ldots, L$. The other approach, taken by Shen and Faraway (2004), Cuevas et al. (2004), and Delicado (2007), is *region-wise*. It is based on the L_2 norms among continuous, versus point-wise, functional responses.

In the next section we provide a more detailed overview of these two approaches and distinct issues these approaches can address in the FANOVA setting.

2. Methods for Functional ANOVA

Suppose that functional responses have been evaluated on a finite grid of points $\{t_1, \ldots, t_L\} \in [t_{min}, t_{max}]$. Ramsay and Silverman (2005) suggested to consider the

F-statistic at each point

$$\begin{aligned} F(t_l) &= \frac{\left[\sum_{ij}(y_{ij}(t_l)-\hat{\mu}(t_l))^2 - \sum_{ij}(y_{ij}(t_l)-\hat{\mu}_i(t_l))^2\right]/(k-1)}{\sum_{ij}(y_{ij}(t_l)-\hat{\mu}_i(t_l))^2/(n-k)}, \\ &= MST(t_l)/MSE(t_l). \end{aligned} \tag{2.2}$$

Here, $\hat{\mu}(t)$ is an estimate of the overall mean function, $\hat{\mu}_i(t)$ is an estimate of group i's mean function, $j = 1, \ldots, n_i$, and n is the total number of functional responses. To perform inference across time t, Ramsay and Silverman (2005) suggested plotting the values of $F(t_l)$, $l = 1, \ldots, L$, as a line (which can be easily accomplished if the evaluation grid is dense) against the permutation α-level critical value at each t_l. If the obtained line is substantially above the permutation critical value over a certain time region, significance is declared at that location. This approach does not account for the multiplicity problem, generating as many tests as the number of evaluation points L.

To perform the overall test Ramsay et al. (2009) suggested using the maximum of the F-ratio in (2.2). The test is overall in a sense that it is designed to detect differences anywhere in t instead of performing inference across t as was described above (i.e., identifying specific regions of t with significant difference among functional means). The null distribution of the statistic for the overall test is obtained by permuting observations across groups and tracking $max\{F(t_l)\}$ across the permutations.

Cox and Lee (2008) suggested using a univariate F-test at each single evaluation point t_l, $l = 1, \ldots, L$, and correct for multiple testing using the Westfall-Young multiplicity correction method (Westfall and Young (1993)). This provides pointwise inferences for differences at L times but does not directly address the overall FANOVA hypotheses.

Alternative inferential approaches were considered by Shen and Faraway (2004), Cuevas et al. (2004), and Delicado (2007). Suppose a smoothing technique was applied to obtain a set of continuous response functions. They each proposed test statistics that accumulate differences across the entire time region $[t_{min}, t_{max}]$ and thus detect deviations from the null hypothesis anywhere within the domain of the functional response. In particular, Shen and Faraway (2004) proposed a functional F-ratio

$$\begin{aligned} \mathcal{F} &= \frac{\left[\sum_{ij}\int_{t_{min}}^{t_{max}}(y_{ij}(t)-\hat{\mu}(t))^2dt - \sum_{ij}\int_{t_{min}}^{t_{max}}(y_{ij}(t)-\hat{\mu}_i(t))^2dt\right]/(k-1)}{\sum_{ij}\int_{t_{min}}^{t_{max}}(y_{ij}(t)-\hat{\mu}_i(t))^2dt/(n-k)} \quad (2.3) \\ &= \frac{\sum_i n_i\int_{t_{min}}^{t_{max}}(\hat{\mu}_i(t)-\hat{\mu}(t))^2dt/(k-1)}{\sum_{ij}\int_{t_{min}}^{t_{max}}(y_{ij}(t)-\hat{\mu}_i(t))^2dt/(n-k)}, \end{aligned}$$

where n is the total number of functional responses and k is the number of groups. Shen and Faraway (2004) derived the distribution of the functional $\mathcal{F}$ statistic under the null hypothesis on the region $[t_{min}, t_{max}]$, but significance can also be assessed via permutations. Cuevas et al. (2004) noted that the numerator of $\mathcal{F}$ accounts for the "external" variability among functional responses. This led Cuevas et al. (2004) to base their test statistic on the numerator of $\mathcal{F}$ since the null hypothesis of FANOVA should be rejected based on a measure of differences among group means. They proposed a test statistic

$$V_n = \sum_{i<j}^{k} n_i||\hat{\mu}_i(t)-\hat{\mu}_j(t)||^2,$$

where $||f|| = \left(\int_a^b f^2(x)dx\right)^{1/2}$. To derive the null distribution of the test statistic, Cuevas et al. (2004) used the Central Limit Theorem as the number of functional responses, n, goes to infinity or, once again, significance can be assessed via permutation methods. Delicado (2007) noted that for a balanced design, V_n differs from the

numerator of $\mathcal{F}$ only by a multiplicative constant. Delicado (2007) also showed equivalence between (3.2) and the Analysis of Distance approach in Gower and Krzanowski (1999).

The region-wise approach, like in Shen and Faraway (2004) and Cuevas et al. (2004), performs an overall FANOVA test, i.e., detects a significant difference anywhere in $[t_{min}, t_{max}]$. However, once overall significance is established, one may want to perform a follow-up test across t to identify specific regions of time where the significant difference among functional means has occurred. The point-wise approaches of Ramsay and Silverman (2005) and Cox and Lee (2008) can be considered as follow-up tests but both techniques have their caveats. Ramsay and Silverman (2005) fail to account for the multiplicity issue while performing L tests across the evaluation points. Cox and Lee (2008) account for multiplicity but their method can not assess overall significance. Using either point-wise approach as a follow-up test could produce results that are inconsistent with the overall test inference.

The remainder of the paper is organized in the following way. Section 3 discusses the problem of multiplicity that has been briefly mentioned above. In Section 4 we propose a new method to perform a follow-up test in the FANOVA setting and contrast it to the existing method of Cox and Lee (2008). Sections 5 and 6 present simulation study results, Section 7 applies the methods to data from a study of CO_2 impact on spectral measurements of vegetation, and Section 8 concludes with a discussion.

3. Multiple Testing Procedures

In hypothesis testing problems involving a single null hypothesis, the statistical tests are chosen to control the Type I error rate of incorrectly rejecting H_0 at a

prespecified significance level α. If L hypotheses are tested simultaneously, the probability of at least one Type I error increases in L, and will be close to one for large L. That is, a researcher will commit a Type I error almost surely and thus wrongly conclude that results are significant. To avoid these situations with misleading findings, the p-values based on which the decisions are made should be adjusted for L simultaneous tests.

A common approach to the multiplicity problem calls for controlling the family-wise error rate (FWER), the probability of committing at least one Type I error. Statistical procedures that properly control for the FWER, and thus adjust the p-values based on which a decision is made, are called multiple comparison or multiple testing procedures. Generally, multiple comparison procedures can be classified as either single-step or stepwise. Single-step multiple testing procedures, e.g., Bonferroni, reject or fail to reject a null hypothesis without taking into account the decision for any other hypothesis. For stepwise procedures, e.g., Holm (1979), the rejection or non-rejection of a null hypothesis may depend on the decision of other hypotheses. Simple single-step and stepwise methods produce adjusted p-values of 1 whenever the number of tests, L, goes to ∞. Since, in the functional response setting, the possible number of tests is potentially infinite, one needs to employ more sophisticated multiplicity adjustment methods. Two possibilities are reviewed below.

The Westfall-Young method (Westfall and Young (1993)) is a step-down resampling method, i.e., the testing begins with the first ordered hypothesis (corresponding to the smallest unadjusted p-value) and stops at the first non-rejection. To implement this method first find unadjusted p-values and order them from *min* to *max*, $p_{(1)} \leq \ldots \leq p_{(L)}$. Generate a vector $(p^*_{(1),n}, \ldots, p^*_{(L),n})$, $n = 1, \ldots, N$, from the same, or at least, approximately the same, distribution as the original p-values under the global null. That is, randomly permute observations N times. For each permu-

tation compute the unadjusted p-values $(p^*_{1,n}, \ldots, p^*_{L,n})$, where n indexes a particular permutation. Put the $p^*_{l,n}$'s, $l = 1, \ldots, L$, in the same order as p-values for the original data. Next, compute successive minima $q^*_{(l),n} = min\{p^*_{(s),n} : s \geq l\}$, $l = 1, \ldots, L$ for all permutations $n = 1, \ldots N$. Finally, the adjusted p-value is the proportion of the $q^*_{(l),n}$ less than or equal to $p_{(l)}$, with an additional constraint of enforced monotonicity (successive ordered adjusted p-values should be greater or equal than one another). See Westfall and Young (1993) Algorithm 2.8 for a complete description of the method.

Another approach is the closure method, which is based on the union-intersection test. The union-intersection test was proposed by Roy (1953) as a method of constructing a test of any global hypothesis H_0 that can be expressed as an intersection of the collection of individual (or elementary) hypotheses. If the global null is rejected, one has to decide which individual hypothesis H_l is false. Marcus et al. (1976) introduced the closure principle as a construction method which leads to a step-wise test adjustment procedure, and allows one to draw conclusions about the individual hypotheses. The closure principle can be summarized as follows. Define a set $\mathcal{H} = \{H_1, \ldots, H_L\}$ of individual hypotheses and the closure set $\bar{\mathcal{H}} = \{H_J = \cap_{j \in J} H_j : J \subset \{1, \ldots, L\}, H_J \neq \emptyset\}$. For each intersection hypothesis $H_J \in \bar{\mathcal{H}}$, perform a test and reject individual H_j if all hypotheses $H_J \in \bar{\mathcal{H}}$ with $j \in J$ are rejected. For example, if $L = 5$ then the closure set is $\bar{\mathcal{H}} = \{H_1, H_2, \ldots, H_5, H_{12}, H_{13}, \ldots, H_{45}, H_{123}, H_{124}, \ldots, H_{345}, H_{1234}, H_{1235}, \ldots, H_{2345}, H_{12345}\}$. The entire closure set for $L = 5$ is shown in Figure 3.2. A rejection of H_1 requires rejection of all intersection hypotheses that include H_1, which are highlighted in Figure 3.2. See Hochberg and Tamhane (1987) for a discussion of closed testing procedures.

In the closure principle, the global null hypothesis is defined as an intersection of the individual null hypotheses and therefore one would like to base the global test statistic on a combination of the individual test statistics. The mapping of the

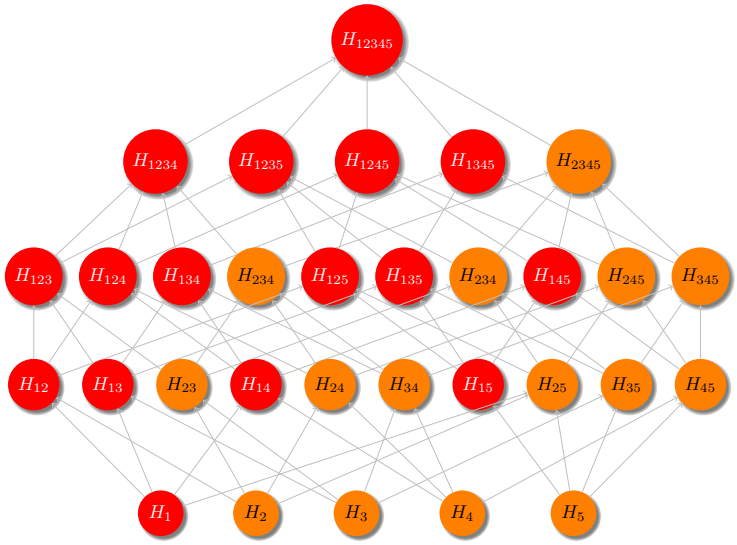

Figure 2.1: Closure set for five elementary hypotheses $H_1, \ldots, H_5$ and their intersections. A rejection of all intersection hypotheses highlighted in colors is required to reject H_0.

individual test statistics to a global one is obtained via a combining function. Pesarin (1992) and Basso et al. (2009) state that a suitable combining function should satisfy the following requirements: (i) it must be continuous in all its arguments, (ii) must be non-decreasing in its arguments, (iii) must reach its supremum when one of its arguments rejects the corresponding partial null hypothesis with probability one. Basso et al. (2009) suggest the following combining functions in the comparison of means of two groups:

1. The unweighted sum of T-statistics

$$T_{sum} = \sum_{h=1}^{m} T_h,$$

where T_h is the standard Student's t-test statistic.

2. A weighted sum of T-statistics

$$T_{wsum} = \sum_{h=1}^{m} w_h T_h,$$

where w_h are the weights with $\sum w_h = 1$.

3. A sum of signed T squared statistics

$$T_{ssT^2} = \sum_{h=1}^{m} sign(T_h) T_h^2.$$

Note that the $max\{F(t_l)\}$ in Ramsay et al. (2009) is an extreme case of the weighted sum combining function with all of the weights equal to zero except one for the largest observed test statistic. Also, the numerator of the $\mathcal{F}$ statistic, defined in (3.2), can be viewed in the context of an unweighted sum combining function. We employ this $\mathcal{F}$ numerator property in the development of our method.

In the next section we propose a new procedure to perform a follow-up test in the FANOVA setting based on the ideas of the closure principle and combining functions. The closure principle will allow us to make a decision for both the overall test, to detect a difference anywhere in time t, and adjust the p-values for the follow-up test, to test across t. By using a combining function we will be able to easily find the value of the test statistic for the overall null based on the values of the individual test statistics.

4. Follow-Up Testing in FANOVA

There are two ways in which one can perform follow-up testing to identify regions with significant differences. One possibility, as in Ramsay and Silverman (2005) and

Cox and Lee (2008), is to evaluate the functional responses on a finite, equally spaced grid of L points from t_{min} to t_{max} (see Figure 2.2a). Another possibility, proposed here, is to split the domain into L mutually exclusive and exhaustive subintervals, say $[a_l, b_l], l = 1, \ldots, L$ (see Figure 2.2b). Based on these two possibilities, we considered follow-up tests for the following four possibilities:

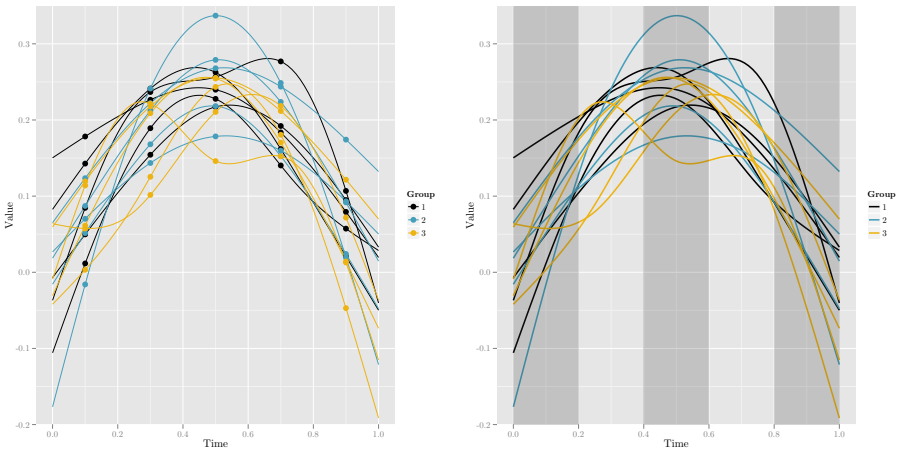

Figure 2.2: Two follow-up testing methods illustrated on simulated data with three groups, five curves per group, and five evaluation points or regions.

1. The procedure proposed by Cox and Lee (2008), which is to evaluate continuous functional responses on a finite grid of points, and at each evaluation point t_l, $l = 1, \ldots, L$, perform a parametric F-test. The individual p-values are adjusted using the Westfall-Young method. We do not consider the Ramsay and Silverman (2005) procedure because it fails to adjust for L simultaneous tests.

2. We propose performing a test based on subintervals of the functional response domain and use the closure principle to adjust for multiplicity. The method

is implemented as follows. Apply a smoothing technique to obtain continuous functional responses. Split the domain of functional responses into L mutually exclusive and exhaustive intervals such that $[t_{min}, t_{max}] = \cup_{l=1}^{L}[a_l, b_l]$. Let the elementary null hypothesis H_l be of no significant difference among functional means anywhere in t on the subinterval $[a_l, b_l]$. For each subinterval, find the individual test statistic T_l as a numerator of $\mathcal{F}$ in Equation (3.2)

$$T_l = \int_{a_l}^{b_l} n_i(\hat{\mu}_i(t) - \hat{\mu}(t))^2 dt/(k-1).$$

Because significance is assessed using permutations, only the numerator of $\mathcal{F}$ is required to perform the tests. The other reason for this preference is the fact that the numerator of $\mathcal{F}$ nicely fits with the idea of the unweighted sum combining function. That is

$$\begin{aligned}
\sum_{l=1}^{L} T_l &= \sum_{l=1}^{L} \int_{[a_l,b_l]} \sum_{i=1}^{k} n_i(\hat{\mu}_j(t) - \hat{\mu}(t))^2 dt/(k-1) \\
&= \int_{t_{min}}^{t_{max}} \sum_{i=1}^{k} n_i(\hat{\mu}_i(t) - \hat{\mu}(t))^2 dt/(k-1) \\
&= T.
\end{aligned}$$

Thus, to test the intersection of two elementary hypotheses, say H_l and $H_{l'}$, of no difference in groups over $[a_l, b_l] \cup [a_{l'}, b_{l'}]$, construct the test statistic $T_{(ll')}$ as a sum of $T_l + T_{l'}$ and find the p-value via permutations. The number of permutations, B, should be chosen such that $(B+1)\alpha$ is an integer to insure that the test is not liberal (Boos and Zhang (2000)). The p-values of the individual hypotheses H_l are adjusted according to the closure principle by taking the maximum p-value

of all hypotheses in the closure set involving H_l. Intermediate intersections of hypotheses are adjusted similarly.

3. We also considered performing the test based on the subregions of the functional domain with the Westfall-Young multiplicity adjustment. To implement the method, first find the unadjusted p-values for each subregion $[a_l, b_l]$, $l = 1, \ldots, L$, by computing $\mathcal{F}^*_{l_b}$ for $b = 1, \ldots, B$ permutations and then counting $(\#\text{ of } (\mathcal{F}^*_{l_b} \geq \mathcal{F}_{l_0}))/B$, where $\mathcal{F}_{l_0}$ is the value of $\mathcal{F}$ for a given sample on the interval $[a_l, b_l]$. Then correct the unadjusted p-values using the Westfall-Young method. Note that to obtain a vector $(p^*_{(1),n}, \ldots, p^*_{(L),n})$, $n = 1, \ldots, N$, the values $(\mathcal{F}^*_{(1),n}, \ldots, \mathcal{F}^*_{(L),n})$ can be computed based on a *single* permutation and then compared to the distribution of $\mathcal{F}^*_{l_b}$, $b = 1, \ldots, B$, and $l = 1, \ldots, L$, obtained previously. Thus, instead of simulating L *separate* permutation distributions of $\mathcal{F}^*_{(l),n}$'s for each $n = 1, \ldots, N$ in the Westfall-Young algorithm, one can use the *same* permutation distribution that was generated to calculate the unadjusted p-values. This dual use of one set of permutations dramatically reduces the computational burden of this method without impacting the adjustment procedure.

4. Finally, we considered a combination of the point-wise test with the closure method for multiplicity adjustment. The procedure is implemented as follows. First, evaluate functional responses on a grid of L equally spaced points and obtain individual test statistics at each of L evaluation points based on the regular univariate F-ratio. Then calculate the unadjusted p-values based on B permutations and use the unweighted sum combining function to obtain the global test statistic and all of the test statistics for the hypotheses in the closure set. In other words, to obtain a test statistic for the overall null hypothesis of

no difference anywhere in t simply calculate $\sum_{l=1}^{L} F_l$. Note that this combining method is equivalent to the sum of signed T-squared statistics, T_{ssT^2}, suggested by Basso et al. (2009). The adjusted p-values of the elementary hypothesis H_l are once again found by taking the maximum p-value of all hypotheses in the closure set involving H_l.

5. Simulation Study

Now, we present a small simulation study to examine properties of the point-wise follow-up test proposed by Cox and Lee (2008), the region-based method with the closure adjustment, the region-based method with the Westfall-Young adjustment, and the point-wise test with the closure adjustment. The properties of interest were the weak control of the FWER, the strong control of the FWER, and power. Hochberg and Tamhane (1987) define the error control as weak if the Type I error rate is controlled only under the global null hypothesis, $H = \cap_{k=1}^{m} H_k$, which assumes that all elementary null hypotheses are true. Hochberg and Tamhane (1987) define the error control as strong if the Type I error rate is controlled under any partial configurations of true and false null hypotheses. To study the weak control of the FWER, we followed the setup of Cuevas et al. (2004) and simulated 25 points from $y_{ij}(t) = t(1-t) + \epsilon_{ij}(t)$ for $i = 1, 2, 3$, $j = 1, \ldots, 5$, $t \in [0, 1]$, and $\epsilon_{ij} \sim N(0, 0.15^2)$. Once the points were generated, we fit these data with smoothing cubic B-splines, with 25 equally spaced knots at times $t_1 = 0, \ldots, t_{25} = 1$. A smoothing parameter, λ, was selected by generalized cross-validation. To study the strong control of the FWER, the observations for the third group were simulated as $y_{3j}(t) = t(1-t) + 0.05 beta_{(37,37)}(t) + \epsilon_{3j}(t)$, where $beta_{a,b}(t)$ is the density of the $Beta(a, b)$ distribution. In our simulation study, this setup implied a higher proportion of H_a's in the partial configuration of

true and false hypotheses as the number of tests increased. To investigate the power, we considered a shift alternative, where the observations for the third group were simulated as $y_{3j}(t) = t(1-t) + \delta + \epsilon_{3j}(t)$ and $\delta = 0.03, 0.06, 0.09$, and 0.12. We also wanted to check whether the two methods are somewhat independent of the number of evaluation points or evaluation intervals. To check this condition, we performed follow-up testing at either $m = 5$ or $m = 10$ intervals/evaluation points.

For this study, we needed two simulation loops. The outside loop was of size $O = 1000$ replications. For each iteration, the permutation-based p-values for the point-wise method with the Westfall-Young adjustment were calculated using the `mt.minP` function from the `multtest` R package (Pollard et al. (2011)). We would like to point out that, unlike the suggestion in Cox and Lee (2008) to use a parametric F distribution to find the unadjusted p-values, the `mt.minP` function finds the unadjusted p-values via permutations. For the region-based method with the closure adjustment, the unadjusted p-values were calculated using the `adonis` function from the `vegan` package (Oksanen et al. (2011)). We wrote an R script to adjust the p-values according to the closure principle. The calculation of the p-values based on the region method with the Westfall-Young adjustment required computation of m unadjusted p-values based on $B = 999$ permutations and a consecutive simulation of N vectors $(p^*_{(1),n}, \ldots, p^*_{(m),n})$, $n = 1, \ldots, N$. To reduce computation time during power investigation for the third scenario, we used a method of power extrapolation based on linear regression described by Boos and Zhang (2000). The method is implemented by first finding three $1 \times m$ vectors of the adjusted p-values based on the Westfall-Young algorithm for $(N_1, N_2, N_3) = (59, 39, 19)$ for each iteration of the outside loop.

Method	5 intervals/evaluations	10 intervals/evaluations
Region-based/Closure	$\mathbf{0.020 \pm 0.009}$	$\mathbf{0.008 \pm 0.006}$
Point-wise/Closure	$\mathbf{0.028 \pm 0.010}$	$\mathbf{0.008 \pm 0.006}$
Region-based/Westfall-Young	0.043 ± 0.013	0.034 ± 0.011
Point-wise/Westfall-Young	0.045 ± 0.013	0.045 ± 0.013

Table 2.1: Estimates of the Type I error ($\pm$ margin of error) control in the weak sense for $\alpha = 0.05$.

Then the estimated power is computed at each subregion as

$$\widehat{\text{pow}}_{k,N_r} = \frac{1}{O}\sum_{j=1}^{O} I(p_{k,N_r} \leq \alpha),$$

where $I()$ is an indicator function, $r = 1, 2, 3$, $k = 1, \ldots, m$, $O = 1000$, and p_k is the adjusted p-value for the k^{th} subregion based on the Westfall-Young algorithm. Finally, the adjusted power based on the linear extrapolation was calculated as

$$\widehat{\text{pow}}_{k,\text{lin}} = 1.01137(\widehat{\text{pow}}_{k,59}) + 0.61294(\widehat{\text{pow}}_{k,39}) - 0.62430(\widehat{\text{pow}}_{k,19}).$$

The p-values for the point-wise test with the closure adjustment were also found based on $B = 999$ inner permutations. For all scenarios an R script is available upon request.

6. Simulation Results

Tables 2.1 and 2.2 report estimates of the family-wise error rate in the weak and the strong sense respectively for the nominal significance level of 5%. The margin of errors from 95% confidence intervals have been calculated based on the normal approximation of the binomial distribution.

Method	5 intervals/evaluations	10 intervals/evaluations
Region-based/Closure	0.042 ± 0.012	0.035 ± 0.011
Point-wise/Closure	0.047 ± 0.013	0.049 ± 0.013
Region-based/Westfall-Young	0.050 ± 0.014	$\mathbf{0.111 \pm 0.019}$
Point-wise/Westfall-Young	0.039 ± 0.012	$\mathbf{0.071 \pm 0.016}$

Table 2.2: Estimates of the Type I error ($\pm$ margin of error) control in the strong sense for $alpha = 0.05$.

Table 2.1 indicates that both testing methods tend to be conservative whenever the closure multiplicity adjustment is applied with the simulations under the global null (highlighted in bold). From Table 2.2 it is evident that both testing methods with the Westfall-Young multiplicity adjustment become liberal as the proportion of H_a's increases in the configuration of the true and false null hypotheses (highlighted in bold). We offer the following explanation for this phenomenon. The test for the overall significance, i.e., whether or not a difference in mean functions exists anywhere in t, is not always rejected if the observations are coming from a mixture of the hypotheses. The closure principle rejects an individual hypothesis only if all hypotheses implied by it (including the overall null) are rejected. Thus, whenever the overall null is accepted, the individual p-values are adjusted accordingly – over the level of significance – and control of the FWER in the strong sense is maintained. With the Westfall-Young method the overall test is not performed. Only the individual p-values are penalized for multiplicity, but the penalty is not "large" enough which likely causes the method to be liberal.

The results of the power investigation for 5 intervals/evaluation points are illustrated in Figure 2.3 and for 10 intervals/evaluation points in Figure 2.4. Solid lines correspond to power of the region-based method with the closure adjustment, dashed lines to the region-based method with the Westfall-Young adjustment, solid circles

to the point-wise test with the Westfall-Young adjustment, and solid triangles to the point-wise method with the closure adjustment. The grouping of power results based on the shift amount, δ, is pretty apparent but a transparency effect is added to aid visualization. The most solid objects (lower graph) correspond to a shift of $\delta = 0.03$, and the most transparent objects (upper graph) to $\delta = 0.12$.

From Figure 2.3 it appears that a combination of the closure multiplicity correction with either testing method provides higher power across all testing points/intervals for moderate values of the shift deviation ($\delta = 0.06$ and $\delta = 0.09$) than the Wetsfall-Young method. There does not seem to be any striking visual difference in power of the four methods for the lowest and highest shift amount ($\delta = 0.03$ and $\delta = 0.12$). Although the powers were very close at the extreme values of δ, it appears that the closure multiplicity correction provides higher overall power across different values of δ while maintaining its conservative nature under the global null. Similar conclusions can be drawn based on Figure 2.4.

A contrast of Figure 2.4 to Figure 2.3 reveals that all methods tend to lose power as the number of evaluation points/intervals increases. This observation implies an intuitive result that a region-based method should be more powerful than a point-wise method. That is, in a real application of a point-wise method one would want to employ many more than $m = 10$ evaluation points. With the region-based application one may not have more than a few *a priori* specified subintervals of interest. Since the power of methods decreases with an increase in m, a region-wise method with a modest number of intervals provides a higher-powered alternative to the point-wise procedures, as they would be used. Additional simulation results for larger values of m provided in the supplementary material support this conclusion.

Both Figures 2.3 and 2.4 indicate that a point-wise test in a combination with the closure procedure provides the highest power. However, there is a caveat in a potential

application of this method. The cardinality of the closure set with m testing points is $2^m - 1$. Therefore, if one would like to perform point-wise tests on a dense grid of evaluation points, the closure principle might become impractical. For example, if one wants to perform a test at $m = 15$ points, $|\bar{\mathcal{H}}| = 32,767$, where $|\bar{\mathcal{H}}|$ denotes the cardinality of the closure set $\bar{\mathcal{H}}$. Zaykin et al. (2002) proposed a computationally feasible method for isolation of individual significance through the closure principle even for a large number of tests. However, since in our application the region-based follow-up test directly addresses research questions and the number of elementary hypotheses is typically small, we left an implementation of this computational shortcut for future study.

As mentioned above, the closure multiplicity correction provides an additional advantage over the Westfall-Young correction of being able to assess the overall significance. Cox and Lee (2008) suggest taking a leap of faith that when the Westfall-Young corrected p-values are below the chosen level of significance, then there is evidence of overall statistical significance. A use of any combining method along with the closure principle allows one to perform a global test as well as to obtain multiplicity adjusted individual p-values. The closure method also provides adjusted p-values for all combinations of elementary hypotheses and the union of some sub-intervals may be of direct interest to researchers.

7. Application

Data from an experiment related to the effect of leaked carbon dioxide (CO_2) on vegetation stress conducted at the Montana State University Zero Emissions Research and Technology (ZERT) site in Bozeman, MT are used to motivate these methods. Further details may be found in Bellante et al. (2013). One of the goals of the

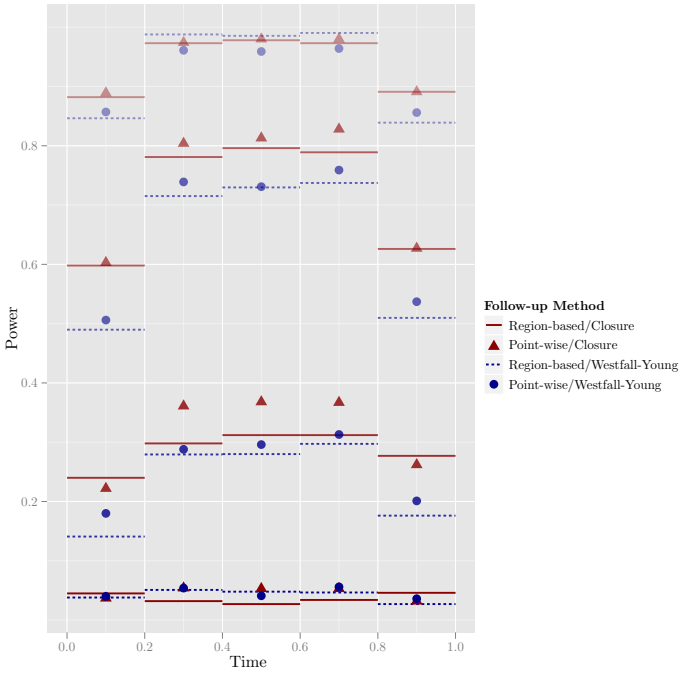

Figure 2.3: Power of the four methods at different values of the shift amount. The solid objects in the lower graph correspond to $\delta = 0.03$. The three groups of objects above that correspond to $\delta = 0.06, 0.09$, and 0.12 respectively.

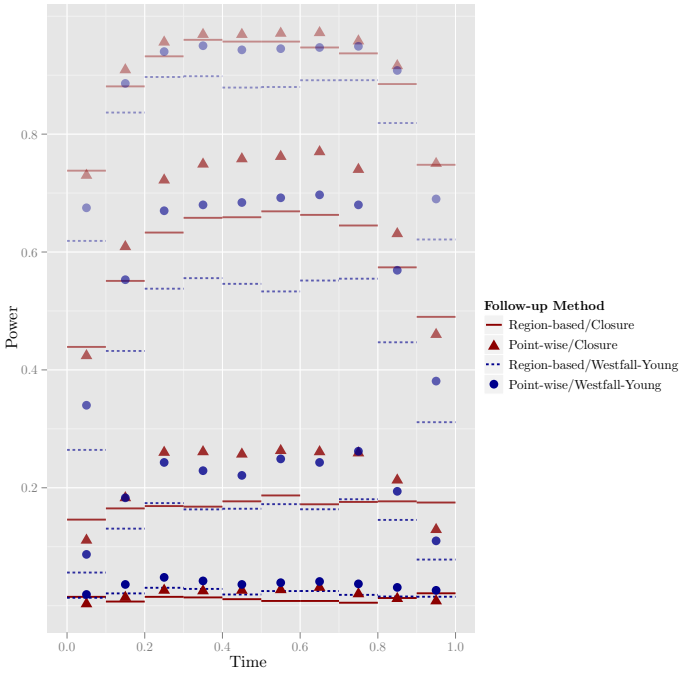

Figure 2.4: Power of the four methods with 10 intervals/evaluation points.

experiment was to investigate hyperspectral remote sensing for monitoring geologic sequestration of carbon dioxide. A safe geologic carbon sequestration technique must effectively store large amounts of CO_2 with minimal surface leaks. Where vegetation is the predominant land cover over geologic carbon sequestration sites, remote sensing is proposed to indirectly identify subsurface CO_2 leaks through detection of plant stress caused by elevated soil CO_2. During the course of the month long controlled CO_2 release experiment, an aerial imaging campaign was conducted with a hyperspectral imager mounted to a small aircraft. A time series of images was generated over the shallow CO_2 release site to quantify and characterize the spectral changes in overlying vegetation in response to elevated soil CO_2.

We analyzed measurements acquired on June 21, 2010 during the aerial imaging campaign over the ZERT site. The pixel-level measurements consisted of 80 spectral reflectance responses between 424.46 and 929.27 nm. For each pixel, we calculated the horizontal distance of the pixel to the CO_2 release pipe. We hypothesized that the effect of the CO_2 leak on plant stress would diminish as we moved further away from the pipe. To test this, we binned the continuous measurements of distance into five subcategories: (0,1], (1,2], (2,3], (3,4], and (4,5] meters to the CO_2 release pipe. Our null hypothesis was that the spectral responses obtained at different distances are indistinguishable. Thus, we could assume exchangeability and permute observations across distances under the null hypothesis. Since the entire image consisted of over 30,000 pixels, we randomly selected 500 pixels from each of the binned distance groups. The spectral responses in 80 discrete wavelengths were generally smooth, providing an easy translation to functional data. There were 2500 spectral response curves in total, with a balanced design of a sample of 500 curves per binned distance. Overall significance was detected (permutation p-value=0.0003), so we were interested in identifying the regions of the electromagnetic spectrum where the sig-

nificant differences occurred. In particular, we were interested in whether there were significant differences in the visible (about 400 nm to 700 nm), "red edge" (about 700 nm to 750 nm), and near infrared (about 750 nm to 900 nm) portions of the electromagnetic spectrum. Since our spectral response ranged to 929.27 nm, we also included the additional region of >900 nm. Because of our interest in specific regions of the electromagnetic spectrum, the regionalized analysis of variance based on the $\mathcal{F}$ test statistic was performed for each of the four spectral regions. The corresponding unadjusted p-values were found based on the permutation approximation. For each region we applied the two multiplicity correction methods, namely the closure and the Westfall-Young method. The results are shown in Figure 4.6.

The p-values adjusted by the two methods are quite similar to each other. Both methods returned the lowest p-value corresponding to the "red edge" spectral region. This is a somewhat expected result since the "red edge" spectral region is typically associated with plant stress. In addition, significant differences were detected in both the visible and near infrared regions. The observed difference between the two adjustments is probably due to the fact that the p-values adjusted with the closure method cannot be lower than the overall p-value, while the Westfall-Young method does not have this restriction. These results demonstrate the novelty and utility of our approach with regards to this application. A previous attempt at examining spectral responses as a function of distance to the CO_2 release pipe relied on a single spectral index as opposed to the full spectral function (Bellante et al. (2013)). Identification of significant differences among spectral regions could prove to be an important analysis technique for hyperspectral monitoring of geologic carbon sequestration. By using a method that provides strong Type I error control, we can reduce false detection of plant stress which could lead to unneeded and costly examination of CO_2 sequestration equipment in future applications of these methods.

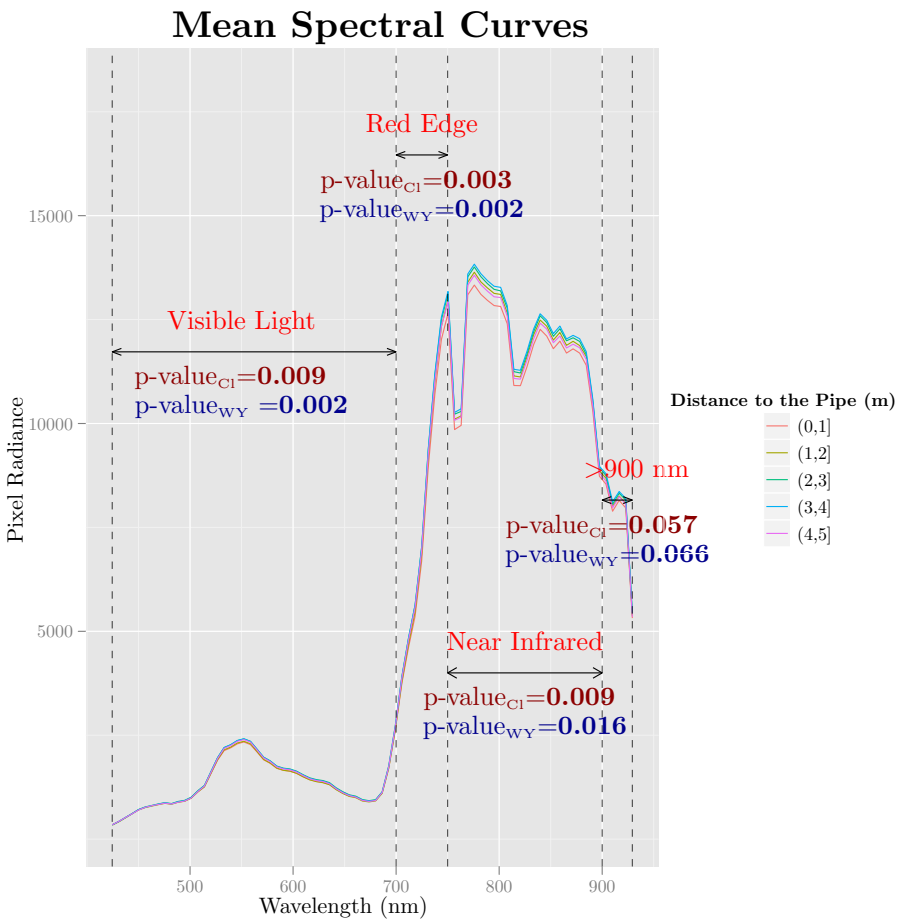

Figure 2.5: Plot of mean spectral curves at each of the five binned distances to the CO_2 release pipe. p-value$_{WY}$ represents a p-value obtained by a combination of the regionalized testing method with the Westfall-Young multiplicity correction. p-value$_{CL}$ represents a p-value obtained by the regionalized method with the closure multiplicity adjustment.

8. Discussion

We have suggested an alternative procedure to the method proposed by Cox and Lee (2008) to perform follow-up testing in the functional analysis of variance setting. Although there is no single approach that is superior in every situation, we have shown in our simulation study that the method for the individual p-value adjustment based

on combining functions via the closure principle provides higher power than that based on the Westfall-Young adjustment. We have shown that the multiplicity adjustment method based on the closure principle tends to be conservative assuming a common mean function, $\mu(t)$, for all t (i.e., on the entire functional domain). The Westfall-Young method was shown to be liberal assuming heterogeneous mean functions, $\mu_i(t)$, on some subregions of the functional domain.

The point-wise follow-up testing method provides slightly higher power than the region-based method. However, we would like to stress one more time that these two methods should not be considered as direct competitors. The choice of one follow-up testing method over the other should be application driven. In our application, we were interested in significant differences in regions of the electromagnetic spectrum and applied the region-based method. In this case it showed similar results with the two multiplicity adjustment corrections despite their differences in performance in simulations.

References

Basso, D., Pesarin, F., Solmaso, L., Solari, A., 2009. Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applications with R. Springer.

Bellante, J., Powell, S., Lawrence, R., Repasky, K., Dougher, T., 2013. Aerial detection of a simulated co2 leak from a geologic sequestration site using hyperspectral imagery. International Journal of Greenhouse Gas Control 13, 124–137.

Boos, D. D., Zhang, J., 2000. Monte carlo evaluation of resampling-based hypothesis tests. Journal of the American Statistical Association 95, 486–492.

Cox, D. D., Lee, J. S., 2008. Pointwise testing with functional data using the westfall-young randomization method. Biometrika 95 (3), 621–634.

Cuevas, A., Febrero, M., Fraiman, R., 2004. An anova test for functional data. Computational Statistics and Data Analysis 47, 111–122.

Delicado, P., 2007. Functional k-sample problem when data are density functions. Computational Statistics 22 (3), 391–410.

Gower, J. C., Krzanowski, W. J., 1999. Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. Journal of the Royal Statistical Society 48 (4), 505–519.

Hochberg, Y., Tamhane, A. C., 1987. Multiple Comparison Procedures. Wiley.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

Marcus, R., Peritz, E., Gabriel, K. R., 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63 (3), 655–660.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Wagner, H., 2011. vegan: Community Ecology Package. R package version 2.0-1.
URL `http://CRAN.R-project.org/package=vegan`

Pesarin, F., 1992. A resampling procedure for nonparametric combination of several dependent tests. Statistical Methods & Applications 1 (1), 87–101.

Pollard, K. S., Gilbert, H. N., Ge, Y., Taylor, S., Dudoit, S., 2011. multtest: Resampling-based multiple hypothesis testing. R package version 2.10.0.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org

Ramsay, J. O., Hooker, G., Graves, S., 2009. Functional Data Analysis with R and MATLAB. Springer.

Ramsay, J. O., Silverman, B. W., 2005. Functional Data Analysis, Second Edition. Springer.

Ramsay, J. O., Wickham, H., Graves, S., Hooker, G., 2012. fda: Functional Data Analysis. R package version 2.3.2.
URL http://CRAN.R-project.org/package=fda

Roy, S. N., 1953. On a heuristic method of test construction and its use in multivariate analysis. The Annals of Mathematical Statistics 23 (220-238).

Shen, Q., Faraway, J., 2004. An F test for linear models with functional responses. Statistica Sinica 14, 1239–1257.

Westfall, P. H., Young, S. S., 1993. Resampling-based Multiple Testing: Examples and Methods for P-Value Adjustment. Wiley.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., Weir, B. S., 2002. Truncated product method for combining p-values. Genetic Epidemiology 22 (2), 170–185.

CHAPTER 3

PAIRSWISE COMPARISON OF TREATMENT LEVELS IN FUNCTIONAL ANALYSIS OF VARIANCE WITH APPLICATION TO ERYTHROCYTE HEMOLYSIS.

Contribution of Authors and Co-Authors

Author: Olga A. Vsevolozhskaya

Contributions: Wrote the majority of the manuscript.

Co-Author: Dr. Mark C. Greenwood

Contributions: Provided feedback on statistical analysis and drafts of the manuscript.

Co-Author: Dmitri Holodov

Contributions: Collected the data. Provided field expertise.

Manuscript Information Page

Olga A. Vsevolozhskaya, Mark C. Greenwood, Dmitri Holodov

Journal of Annals of Applied Statistics.

Status of Manuscript:

_____ Prepared for submission to a peer-reviewed journal

__X__ Officially submitted to a peer-review journal

_____ Accepted by a peer-reviewed journal

_____ Published in a peer-reviewed journal

Published by Institute of Mathematical Statistics.

Submitted April, 2013

Abstract

Motivated by a practical need for the comparison of hemolysis curves at various treatment levels, we propose a novel method for pairwise comparison of mean functional responses. The hemolysis curves – the percent hemolysis as a function of time – of mice erythrocytes (red blood cells) by hydrochloric acid have been measured among different treatment levels. This data set fits well within the functional data analysis paradigm, in which a time series is considered as a realization of the underlying stochastic process or a smooth curve. Previous research has only provided methods for identifying some differences in mean curves at different times. We propose a two-level follow-up testing framework to allow comparisons of pairs of treatments within regions of time where some difference among curves is identified. The closure multiplicity adjustment method is used to control the family-wise error rate of the proposed procedure.

1. Introduction

The use of non-steroidal anti-inflammatory drugs (NSAIDs) is wide-spread in the treatment of various rheumatic conditions (Nasonov and Karateev (2006)). Gastrointestinal symptoms are the most common adverse events associated with the NSAID therapy (Garcia-Rodriguez et al. (2001)). Holodov and Nikolaevski (2012) suggested oral administration of procaine (novocaine) solution in low concentration (0.25 to 1%) to reduce the risk of upper gastrointestinal ulcer bleeding associated with NSAIDs. To validate the effectiveness of the proposed therapy, an experiment was conducted to study the effect of novocaine on the resistance of the red blood cells (erythrocytes) to hemolysis by hydrochloric acid. Hydrochloric acid is a major component of gastric juice and a lower rate of erythrocyte hemolysis should indicate a protective effect of novocaine.

Hemolytic stability of erythrocytes for the control and three different dosages of novocaine (4.9×10^{-6} mol/L, 1.0×10^{-5}mol/L, and 2.01×10^{-5}mol/L) was measured as a percentage of hemolysed cells. The data for the analysis were curves of hemolysis

(erythrograms) that were measured as functions of time. Figure 3.1 illustrates a sample of percent hemolysis curves. The goal of the statistical analysis was to summarize the associated evidence across time of the novocaine effect including performing pairwise comparisons of novocaine dosages.

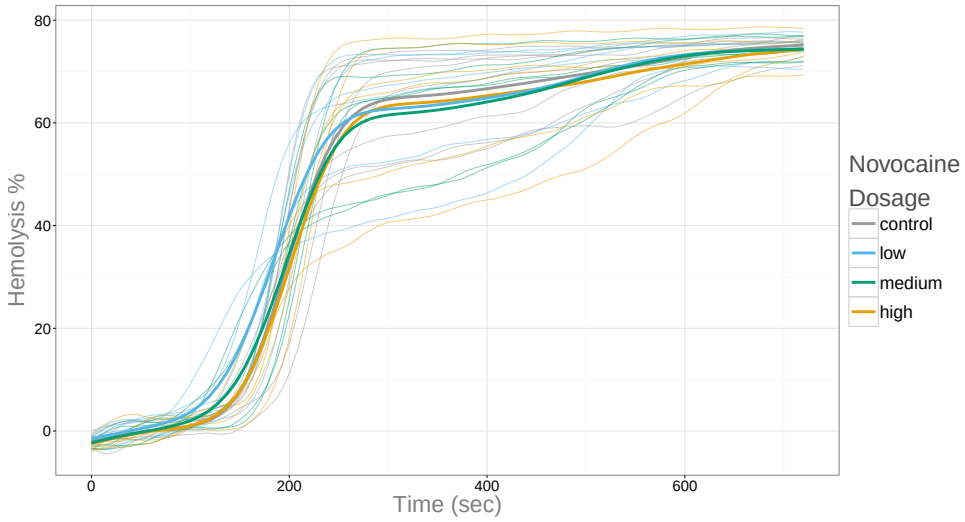

Figure 3.1: Hemolysis curves of mice erythrocytes by hydrochloric acid with superimposed estimated mean functions.

Most current approaches essentially evaluate differences among groups of curves point-wise (typically, with many one-way ANOVA tests). For such approaches, when testing is performed at a large number of points simultaneously, the type I error rate is going to be inflated. Cox and Lee (2008) proposed a method that utilizes a point-wise approach, while properly controlling Type I error, and can be used for investigating specific subregions of the functional domain (time) for a significant difference. Alternatively, the functional analysis of variance (FANOVA) can be employed to perform testing among k groups of curves. The overall functional testing methods, such as the functional $\mathcal{F}$ of Shen and Faraway (2004) or the functional V_n of Cuevas et al. (2004),

can be utilized to test for associated evidence across the entire functional domain (across all time). However, none of these methods allow for pairwise comparisons of functional means. Thus, the challenge for the current analysis was to determine differences among novocaine dosages withing specific intervals of time, where significant difference among hemolysis curves is present.

In this paper, we introduce a new two-step procedure: first, to detect regions in time of "significant" differences among mean curves, and second, to perform a pairwise comparison of treatment levels within those regions. Our approach utilizes two ideas: (i) combining methods to map the test statistics of the individual hypotheses, $H_1, \ldots, H_m$, to the global one, $\cap_{i=1}^{m} H_i$, and (ii) the closure principle of Marcus et al. (1976) to control the family-wise error rate (FWER), the probability of at least one false rejection. The rest of the article is organized in the following manner. We give an overview of the FANOVA problem and the existing methods for investigating the functional domain for significant differences. We discuss the proposed procedure for investigating regions of time for significant differences and detail a computational shortcut that allows isolation of individual significance even for a large number of tests. We extend the proposed procedure to perform pairwise comparisons of the treatment levels within identified functional regions of statistical significance. The protective effect of novocaine is demonstrated based on the different patterns between groups detected in certain regions of time.

2. Methods

Functional analysis of variance involves testing for some difference among k functional means. In functional data analysis, t is used to denote a real-valued variable (usually of time), and $y(t)$ denotes a continuous outcome, which is a function of t.

Then, the FANOVA model is written as:

$$y_{ij}(t) = \mu_i(t) + \epsilon_{ij}(t), \tag{3.1}$$

where $\mu_i(t)$ is the mean function of group i at time t, $i = 1, \ldots, k$, j indexes a functional response within a group, $j = 1, \ldots, n_i$, and $\epsilon_{ij}(t)$ is the residual function. Each $\epsilon_{ij}(t)$ is assumed to be a mean zero and independent Gaussian stochastic process. The FANOVA hypotheses are written as:

$$\begin{aligned} H_0 : &\quad \mu_1(t) = \mu_2(t) = \ldots = \mu_k(t) \\ H_a : &\quad \mu_i(t) \neq \mu_{i'}(t), \text{ for at least one } t \text{ and } i \neq i'. \end{aligned}$$

The alternative hypothesis considers any difference anywhere in t among k population means of $y_{ij}(t)$.

In recent years two different general approaches have emerged to perform the FANOVA test. In Shen and Faraway (2004), as well as many other papers (see Cuevas et al. (2004), Ramsay et al. (2009) and Cuesta-Albertos and Febrero-Bande (2010)), a global test statistic has been developed to perform the FANOVA test. The statistic is "global" because it is used to detect differences anywhere in the entire functional domain (anywhere in t). An alternative approach (Ramsay and Silverman (2005) and Cox and Lee (2008)) is to use a point-wise (or individual) test statistic to perform inference across t, i.e., identify specific regions of t with significant difference among functional means.

2.1. "Global" Approach

Suppose the domain $[a, b]$ of functional responses can be split into m pre-specified mutually exclusive and exhaustive intervals such that $[a, b] = \cup_{i=1}^{m} [a_i, b_i]$. For instance, in the novocaine experiment the researchers were interested in the effect of novocaine during specific time intervals associated with hemolysis of different erythrocyte populations: hemolysis of the least stable population ($[a_2, b_2]$ = 61-165 sec.), general population ($[a_3, b_3]$ = 166-240 sec.), and most stable ($[a_4, b_4]$ = over 240 sec.). For each interval $[a_i, b_i]$, $i = 1, \ldots, m$, an individual functional statistic of Shen and Faraway (2004), $\mathcal{F}_i, i = 1, \ldots m$, can be calculated as

$$\mathcal{F}_i = \frac{\int_{[a_i,b_i]} \sum_{j=1}^{k} n_j(\hat{\mu}_j(t) - \hat{\mu}(t))^2 dt/(k-1)}{\int_{[a_i,b_i]} \sum_{j=1}^{k} \sum_{s=1}^{n} (y_{js}(t) - \hat{\mu}_j(t))^2 dt/(n-k)}, \tag{3.2}$$

where n is the total number of functional responses and k is the number of groups. The numerator of the $\mathcal{F}$ statistic accounts for "external" variability among functional responses and the denominator for the "internal" variability. Cuevas et al. (2004) argues that the null hypothesis should be rejected based on the measure of the differences among groups, i.e., the "external" variability. Hence, Cuevas et al. (2004) proposed a statistic V_n based on the numerator of $\mathcal{F}$:

$$V_n = \sum_{i<j}^{k} n_i ||\hat{\mu}_i(t) - \hat{\mu}_j(t)||^2, \tag{3.3}$$

where $|| \cdot ||$ is the L_2 norm. Gower and Krzanowski (1999) also argue that in a permutation setting a test can be based just on the numerator of the test statistic. That is, if only the numerator of the functional $\mathcal{F}$ is used, the changes to the test statistic are monotonic across all permutations and thus probabilities obtained are

identical to the ones obtained from the original $\mathcal{F}$. Additionally, Delicado (2007) points out that for a balanced design, the numerator of the functional $\mathcal{F}$ and V_n differ by only a multiplicative constant.

2.2. Point-wise Approach

Suppose that a set of smooth functional responses is evaluated on a dense grid of points, $t_1, \ldots, t_m$. For instance, the percentage of hemolysed cells can be evaluated every second. Cox and Lee (2008) propose a test for differences in the mean curves from several populations, i.e., perform functional analysis of variance, based on these discretized functional responses. First, at each of the m evaluation points, the regular one-way analysis of variance test statistic, $F_i, i = 1, \ldots, m$, is computed. For each test the p-value is calculated based on the parametric F-distribution and then the Westfall-Young randomization method (Westfall and Young (1993)) is applied to correct the p-values for multiplicity. The implementation of the method can be found in the `multtest` (Pollard et al. (2011)) R package (R Core Team (2013)).

Certain criticisms may be raised for both the *"global"* and the *point-wise* approaches. First, the *point-wise* approach can determine regions of the functional domain with a difference in the means, but can not determine which pairs of populations are different. Second, for the Cox and Lee (2008) procedure, the p-value for the global test can not be obtained, which is an undesirable property since the method might be incoherent between the global and point-wise inference. We suggest a procedure that overcomes both of these issues. By using a combining function along with the closure principle of Marcus et al. (1976) we are able to obtain the p-value for the overall test as well as adjust the individual p-values for multiplicity. This method also allows us to perform a pairwise comparison of the group's functional means and therefore determine which populations are different in each region.

2.3. Proposed Methodology

Once again, suppose the domain $[a, b]$ is split into m pre-specified mutually exclusive and exhaustive intervals. We propose to use the numerator of the functional $\mathcal{F}$ as the test statistic T_i, $i = 1, \ldots, m$, for each $[a_i, b_i]$, and then utilize a combining function to obtain the test statistic for the entire $[a, b]$. Typical combining functions have the same general form: the global statistic is defined as a weighted sum, $T = \sum w_i T_i$, of the individual statistics with some w_i weights (see Pesarin (1992) and Basso et al. (2009)). A p-value for the overall null hypothesis (that all individual null hypotheses are true) is based either on the distribution of the resulting global statistic T or on a permutation approximation. If the unweighted sum combining function is applied to the proposed T_i, then

$$\begin{aligned} T &= \int_{[a,b]} \sum_{j=1}^{k} n_j (\hat{\mu}_j(t) - \hat{\mu}(t))^2 dt/(k-1) \\ &= \sum_{i=1}^{m} \int_{[a_i,b_i]} \sum_{j=1}^{k} n_j (\hat{\mu}_j(t) - \hat{\mu}(t))^2 dt/(k-1) \\ &= \sum_{i=1}^{m} T_i. \end{aligned}$$

The closure procedure is then applied to perform the overall test based on this combining functions as well as adjust the individual p-values for multiplicity. The closure method is based on testing all nonempty intersections of the set of m individual hypotheses, which together form a closure set. The procedure rejects a given hypothesis if all intersections of hypotheses that contain it as a component are rejected. Hochberg and Tamhane (1987) show that the closure procedure controls the family-wise error

rate (FWER) at a strong level, meaning that the type I error is controlled under any partial configuration of true and false null hypotheses.

When the number of individual tests m is relatively large, the use of the closure method becomes computationally challenging. For example, setting $m = 15$ results in $2^{15}-1 = 32,767$ intersections of hypotheses. Hochberg and Tamhane (1987) described a shortcut for the $T = max\{T_i\}$ combining function, where T_i stands for the i^{th} test statistic for i in the set of H_i pertinent to a particular intersection hypothesis. For this combining function they showed that the significance for any given hypothesis in the closure set can be determined using only m individual tests. Zaykin et al. (2002) described a shortcut for the closure principle in the application of their truncated p-value method (TPM) that uses an unweighted sum combining function. In the next section we exploit the shortcut described by Zaykin et al. (2002) and show that for the $T = \sum T_i$ combining function the required number of evaluations is $m(m+1)/2$.

2.3.1. The Shortcut Version of the Closure Procedure The shortcut version of the closure method for the unweighted sum combining function should be implemented as follows. First, order the individual test statistics from minimum to maximum as $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(m)}$, where

$$T_i = \int_{[a_i,b_i]} \sum_{j=1}^{k} n_j(\hat{\mu}_j(t) - \hat{\mu}(t))^2 dt/(k-1). \tag{3.4}$$

Let $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ be the corresponding ordered individual hypotheses of no significant difference among functional means on the interval $\left[a_{(i)}, b_{(i)}\right]$, $i = 1, \ldots, m$. Now, among intersection hypotheses of size two:

$$T_{(1)} + T_{(2)} \leq T_{(1)} + T_{(3)} \leq \ldots \leq T_{(1)} + T_{(m)},$$

$$T_{(2)} + T_{(3)} \leq T_{(2)} + T_{(4)} \leq \ldots \leq T_{(2)} + T_{(m)},$$

$$\ldots$$

Here, the statistic $T_{(i)} + T_{(j)}$ corresponds to intersection hypotheses $H_{(ij)}$ of no significant difference on both intervals $\left[a_{(i)}, b_{(i)}\right] \cup \left[a_{(j)}, b_{(j)}\right]$. Among intersections of size three:

$$T_{(1)} + T_{(2)} + T_{(3)} \leq T_{(1)} + T_{(2)} + T_{(4)} \leq \ldots \leq T_{(1)} + T_{(2)} + T_{(m)},$$

$$T_{(2)} + T_{(3)} + T_{(4)} \leq T_{(2)} + T_{(3)} + T_{(5)} \leq \ldots \leq T_{(2)} + T_{(3)} + T_{(m)},$$

$$\ldots$$

Thus, significance for the hypothesis $H_{(m)}$ can be determined by looking for the largest p-value among m tests

$$T_{(m)}, T_{(m)} + T_{(1)}, \ldots, \sum_{i=1}^{m} T_{(i)}.$$

For the hypothesis $H_{(m-1)}$, the significance can be determined by investigating the p-values corresponding to $(m-1)$ tests

$$T_{(m-1)}, T_{(m-1)} + T_{(1)}, \ldots, \sum_{i=1}^{m-1} T_{(i)}$$

along with the p-value for the test $\sum_{i=1}^{m} T_{(i)}$ which is already found. Finally, for the first ordered hypothesis $H_{(1)}$, the significance can be determined by evaluating a single test $T_{(1)}$ and then looking for the largest p-value among it and the p-values of the hypotheses $H_{(12)}$, $H_{(123)}, \ldots, H_{(12\ldots m)}$, which are already evaluated. Thus, significance of any individual hypothesis $H_{(i)}$ is determined using m p-values, but the number of unique evaluations to consider is $m + (m-1) + \cdots + 1 = m(m+1)/2$.

The described shortcut assumes that all distributions corresponding to the test statistics are the same and the magnitude of the test statistic has a monotonic relationship with its p-value. If the p-values for the individual tests are determined from permutational distributions (as in our situation), a bias will be introduced. The bias is caused by a mismatch between the minimum value of the test statistics and the maximum p-value. That is, the minimum statistic is not guaranteed to correspond to the maximum p-value. The procedure becomes liberal since the individual p-values are not always adjusted adequately. To reduce and possibly eliminate the bias, we made the following adjustment to the shortcut. First, we adjusted the individual p-values according to the shortcut protocol described above and obtained a set of adjusted individual p-values, $p_1, p_2, \ldots, p_m$. Then, we ordered the individual test statistics based on the ordering of the unadjusted individual p-values. That is, we order the unadjusted p-values from maximum to minimum and get a corresponding ordering of the test statistics $T^*_{(1)}, T^*_{(2)} \ldots, T^*_{(m)}$. Now the inequality $T^*_{(1)} \leq T^*_{(2)} \leq \ldots \leq T^*_{(m)}$ will not necessarily hold. We applied the shortcut based on this new ordering and obtained another set of adjusted individual p-values, $p^*_1, p^*_2, \ldots, p^*_m$. Finally, the adjusted individual p-values were computed as $max\{p_i, p^*_i\}, i = 1, \ldots, m$. This correction to the shortcut increases the number of the required computations by a factor of two, however it is still of the order m^2 instead of 2^m.

A small simulation study was used to check whether this version of the correction provides results comparable to adjustments generated by the entire set of intersection hypotheses. For the four multiplicity adjustment schemes: (i) correction based on the ordered test statistics shortcut, (ii) correction based on the ordered unadjusted p-values shortcut, (iii) correction based on $max\{p_i, p^*_i\}$ (combination of both corrections (i) and (ii)), and (iv) the full closure method, we obtained p-values under the global null based on 1000 permutations, $m = 5$, and conducted 1000 simulations, providing

5000 corrected p-values. First, we were interested in how many times the p-values adjusted by various shortcuts were "underestimated" (not corrected enough) relative to the full closure method. The p-values adjusted by a shortcut based on the ordered test statistics, $p_1, p_2, \ldots, p_m$, were underestimated 554 out of 5000 times. The p-values adjusted by a shortcut based on the ordered unadjusted p-values, $p_1^*, p_2^*, \ldots, p_m^*$, were underestimated 60 out of 5000 times. The p-values adjusted using both corrections, $max\{p_i, p_i^*\}$, $i = 1, \ldots, m$, were underestimated 38 out of 5000 times. Second, we compared Type I error rates under the $max\{p_i, p_i^*\}$ shortcut and the full closure method and found that they were *exactly* the same. The above results allowed us to conclude that the multiplicity adjustment based on $max\{p_i, p_i^*\}$ shortcut is adequate.

2.3.2. Pairwise Comparison of Functional Means Above, we provided details on how to implement the proposed methodology to isolate regions of the functional domain with statistically significant differences and showed that with a computational shortcut the closed testing scheme is computable even for a large number of individual tests m. Now, we show how to further use the proposed methodology to find pairs of functional means that are different within the regions where statistical significance was identified. The procedure is implemented as follows:

i. Within an interval $[a_i, b_i]$ with a statistically significant difference among functional means, set the p-value for the "global" null of no difference among functional means to the adjusted individual p-value corresponding to that interval.

ii. Compute the pairwise statistic as well as statistics for the intersection hypotheses as in (3.4).

iii. Find the p-values based on the permutation algorithm and adjust them using the closure principle.

Figure 3.2 illustrates the closure set for pairwise comparison of four populations. The p-value of the top node hypothesis, H_{ABCD}, of no significant difference among the four population means would be set equal to the adjusted p-value of the interval level individual hypothesis of interest H_i, $i = 1, \ldots, m$. The bottom node individual hypotheses, $H_{AB}, \ldots, H_{CD}$, are of no significant pairwise difference between groups AB, $AC, \ldots, CD$ in this interval. Note that now the indexing of the hypotheses corresponds to *population means* instead of *intervals in the functional domain*. The closure principle is used to adjust the individual p-values.

Certain issues may arise with a test of pairwise comparison conducted by global randomization. Petrondas and Gabriel (1983) noted that for the overall equality hypothesis all permutations are assumed to be equally probable, that is, the exchangeability among all treatment groups is assumed. However, for the hypothesis of equality of a particular subset of treatments, the global permutation distribution can not be used because differences in variability among the treatment groups can cause bias in the statistical tests. The results of the simulation study, presented in the next section, did not reveal any noticeable bias in the permutation test. In the case of the pairwise comparison, our method maintained good control of the Type I error rate as well as had enough power to correctly identify groups of unequal treatments. The minimal bias observed might be due to a relatively small (three) number of treatments that we chose to consider in our simulation study. Petrondas and Gabriel (1983) and Troendle and Westfall (2011) provided ways to perform permutation tests correctly in the case of the pairwise comparison. We leave implementation of these solutions for future research.

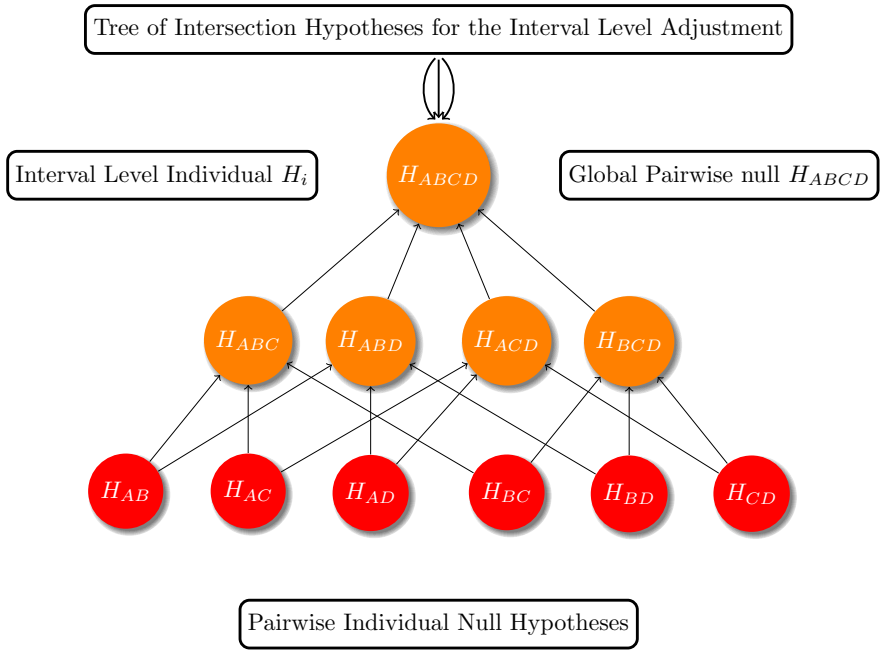

Figure 3.2: Example of the closure set for the pairwise comparison of four groups. The darker nodes represent individual hypotheses for pairwise comparison.

3. Simulations

A simulation study has been carried out in order to evaluate the performance of our approach. The set up of simulations was inspired by a Monte Carlo study in Cuesta-Albertos and Febrero-Bande (2010). We considered

(M1) $f_i(t) = 30(1-t)t - 3\beta|sin(16\pi t)|I_{\{0.325<t<0.3575\}} + \epsilon_i(t)$,

(M2) $f_i(t) = 30(1-t)t - \beta|sin(\pi t/4)| + \epsilon_i(t)$,

where $t \in [0,1]$, $\beta \in \{0.000, 0.045, 0.091, 0.136, 0.182, 0.227, 0.273, 0.318, 0.364, 0.409, 0.455, 0.500\}$, and random errors $\epsilon_i(t)$ are independent between curves, but normally distributed within a curve with mean zero and variance 0.3. Case M1 corresponds to a situation where a small set of observations was generated under H_A

to create a spike. In M2, a large number of observations were generated under H_A but the differences are less apparent (a deviation along the entire range of t that gradually increases from $min(t)$ to $max(t)$). The parameter β controls the strength of the deviation from the global null. The reason for considering these two cases was to check the performance of our method for different ranges of false null hypotheses.

In each case (M1 and M2), we generated three samples of functional data with 5 observations from each group. The first two samples had the same mean ($\beta = 0$) and the third sample's mean was deviating ($\beta \neq 0$). Once the functional data were generated for different values of $\beta \neq 0$, we split the functional domain into different numbers of equal-length intervals (m=5 and m=10) and evaluated the power of rejecting the null hypotheses $H_0 : \mu_1(t) = \mu_2(t) = \mu_3(t)$ at the 5% level. We used 1000 simulations to obtain a set of power values for each combination of β and m values.

Figure 3.3 presents results of power evaluation for model M1 and five intervals (m=5). Under this model, a set of observations generated under H_A fell into the second interval. That is, the functional mean of the third sample had a spike deviation from the functional mean of the first two samples over the second interval. The magnitude of the spike increased monotonically as a function of β. The plot shows that the proportion of rejections reveals a peak over the region of the true deviation, while being conservative over the locations with no deviations. Thus, we conclude that the proposed methodology provides satisfactory power over the region with true differences, while being conservative over the regions where the null hypothesis is true.

Once we identified the region of the functional domain with differences in means (i.e., the second interval), we used the extension of the proposed methodology to perform a pairwise comparison and determine which populations are different. Figure

3.4 provides the results of power evaluation of the pairwise comparisons at the 5% significance level. In the case of H_{AB} (where the null $\mu_1 = \mu_2$ is true) the simulation output tells us that the procedure is a bit conservative, maintaining the Type I error rate right below the 5% level for the higher values of β. In case of H_{AC} and H_{BC} (where the null is false) it can be seen that the power of the pairwise comparison is satisfactory.

The results for the M2 case, where the number of true effects is large and the magnitude of the effect gradually increases from $min(t)$ to $max(t)$, are provided in Tables 3.1-3.5 and Figure 3.5. The plot shows that for a fixed value β, the proportion of rejections of the hypothesis $H_0 : \mu_1(t) = \mu_2(t) = \mu_3(t)$ gradually increases with the magnitude of the effect. Across different values of β, power values are also increasing, attaining the value of 1 for the fifth interval and $\beta = 0.5$. The results of the pairwise comparisons are provided in Tables 3.1-3.5. Power is the highest for the highest value of β (0.5) but overall the method does a good job of picking out the differences between μ_1 and μ_3, and μ_2 and μ_3, while maintaining control of spurious rejections for μ_1 and μ_2.

Results based on $m = 10$ intervals are similar to those based $m = 5$ intervals and can be found in the supporting information.

4. Analysis of Hemolysis Curves

In this section we illustrate the proposed methodology by applying it to a study of the effect of novocaine conducted by Holodov and Nikolaevski (2012). The motivation behind the study was to investigate pharmaceutical means of preventing the formation of stomach erosive and ulcerative lesions caused by a long-term use of nonsteroidal

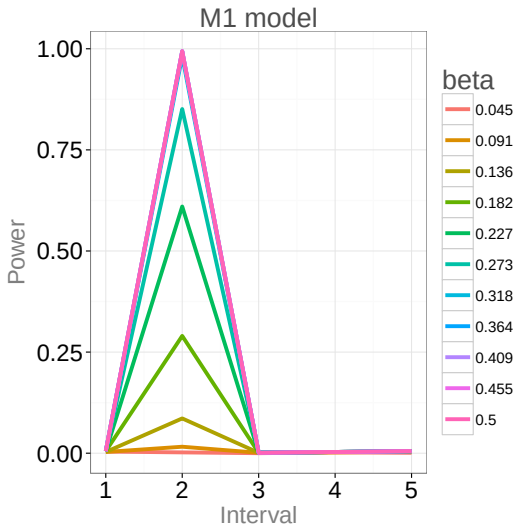

Figure 3.3: The probability of rejecting the null hypothesis $H_0 : \mu_1(t) = \mu_2(t) = \mu_3(t)$ for $m = 5$ intervals.

anti-inflammatory drugs (NSAIDs). Internal use of a novocaine solution was proposed as a preventative treatment for NSAID-dependent complications.

During the course of the experiment, blood was drawn from male rats to obtain an erythrocyte suspension. Then, four different treatments were applied: control, low (4.9×10^{-6} mol/L), medium (1.0×10^{-5} mol/L), and high (2.01×10^{-5} mol/L) dosages of procaine. After treatment application, the erythrocyte suspension was incubated for 0, 15, 30, 60, 120, or 240 minutes. At the end of each incubation period, hemolysis was initiated by adding 0.1 M of hydrochloric acid to the erythrocyte suspension. The percent of hemolysis or the percent of red blood cells that had broken down was measured every 15 seconds for 12 minutes. The experiment was repeated 5 times for each dosage/incubation combination using different rats. Therefore, the dataset consists of 120 separate runs with 49 discretized observations per run and involves four experimental conditions with six incubation times, replicated 5 times for each

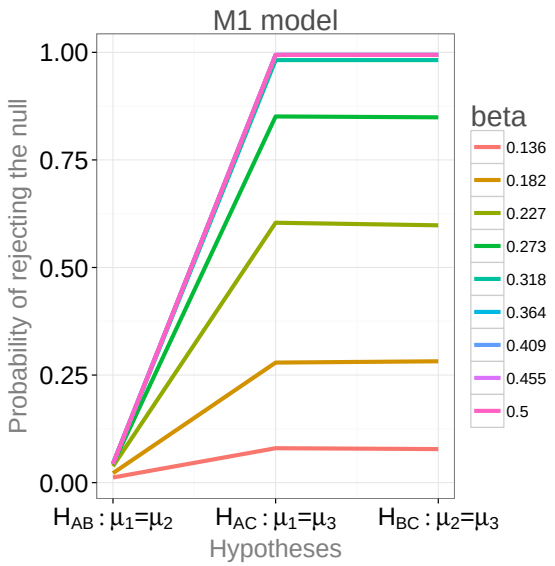

Figure 3.4: The probability of rejecting individual pairwise hypotheses $H_{AB} : \mu_1(t) = \mu_2(t)$, $H_{AC} : \mu_1(t) = \mu_3(t)$, and $H_{BC} : \mu_2(t) = \mu_3(t)$.

treatment/incubation combination. For more details see Holodov and Nikolaevski (2012).

We fit the data with smoothing cubic B-splines with 49 equally spaced knots at times $t_1 = 0, \ldots, t_{49} = 720$ seconds to generate the functional data. A smoothing parameter was selected by generalized cross validation (GCV) for each functional observation with an increased penalty for each effective degree of freedom in the GCV, as recommended in Wood (2011).

To keep the analysis as simple as possible, each incubation data set was analyzed for treatment effects separately. Our initial test was to check for a significant difference in mean erythrograms (mean hemolysis curves) anywhere in time among novocaine dosages. A Bonferroni correction was applied to these initial p-values to adjust for multiplicity at this level. The results indicated strong evidence of differences for the 15 and 30 minute incubation times (p-value$_{Bonf} = 0.006$ and p-value$_{Bonf} = 0.018$ respectively). Figure 4.2 illustrates the results for these incubation times. For the

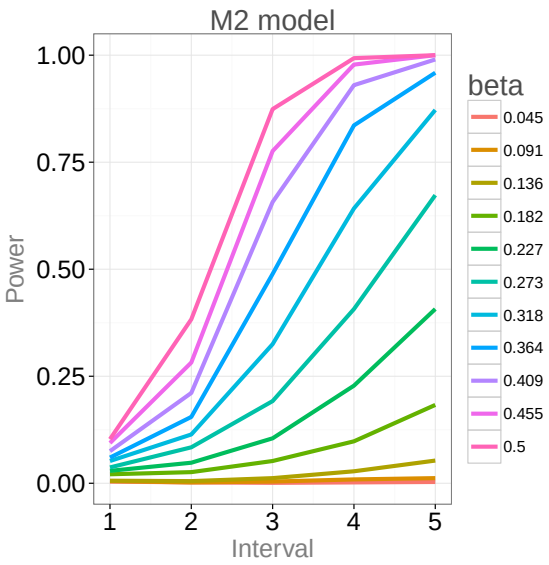

Figure 3.5: The probability of rejecting the null hypothesis $H_0 : \mu_1(t) = \mu_2(t) = \mu_3(t)$ in case of M2 model and 5 intervals.

rest of the incubation times, we found no evidence against the null hypothesis that the four erythrogram means coincided so no further analysis was conducted.

Next, we examined the 15 and 30 minute incubation results in more detail to asses the nature of the differences. For both incubation times, four time intervals of interest were pre-specified: (i) the latent period (0-60 sec), (ii) hemolysis of the population of the least stable red blood cells (61-165 sec), (iii) hemolysis of the general red blood cell population (166-240 sec), and (iv) the plateau (over 240 sec). The latent period is associated with erythrocytes spherulation and occurs between addition of the hemolytic agent and initiation of hemolysis. The names of the next two periods are self-explanatory. The plateau period is associated with deterioration of the population of the most stable erythrocytes.

We applied our method to determine if statistical significance is present in each of the four time intervals. In the application of our method, we set the p-values for the global hypotheses H_{1234} of no significant difference on all four intervals to the Bonfer-

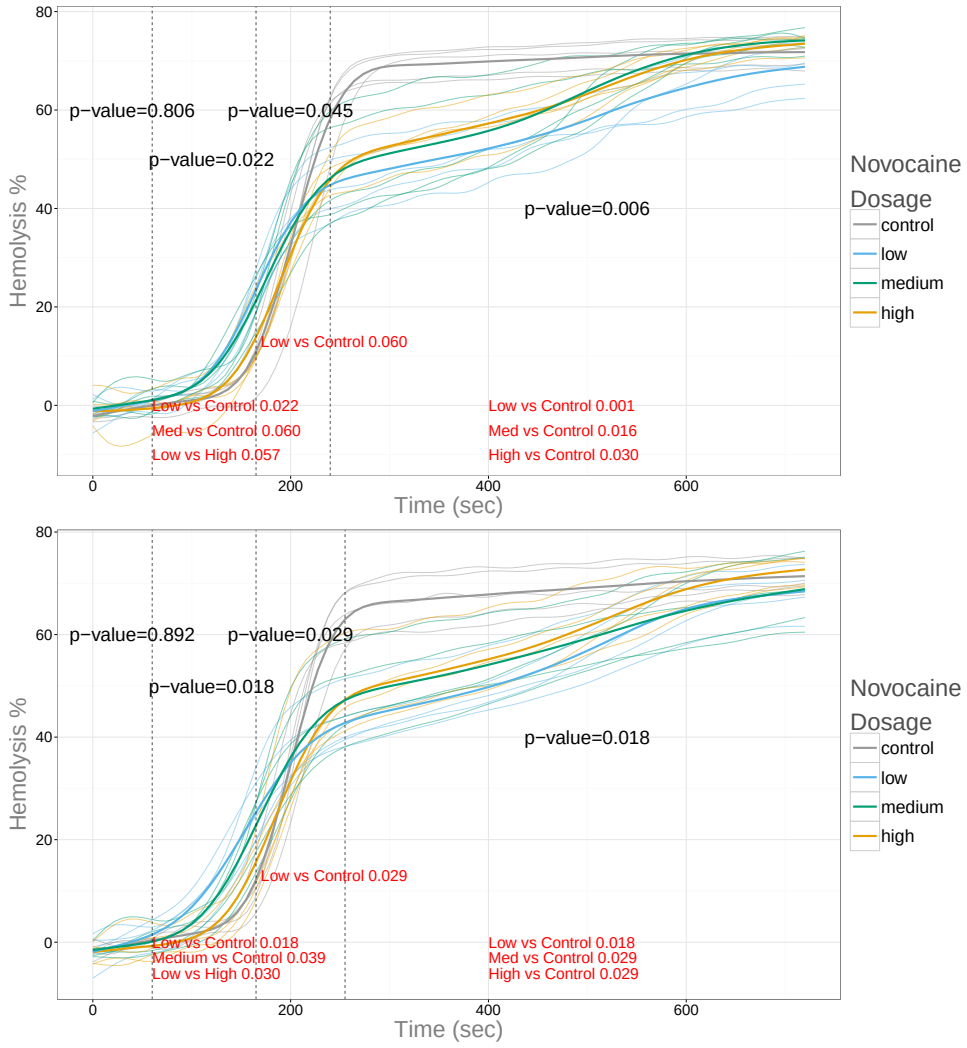

Figure 3.6: Erythrogram means for the control group and the treatment groups for 15 (top graph) and 30 (bottom graph) minute incubation times

roni adjusted p-values obtained on the previous step. For the 15 minute incubation time, no statistical significance was found during the latent period (p-value=0.806), statistically significant results were found during hemolysis of the least stable red blood cell population (p-value=0.022), general red blood cell population (marginal significance with the p-value=0.060), and plateau (p-value=0.006). The same results were obtained from the 30 minute incubation, i.e., no statistical significance during

the latent period (p-value=0.892) and statistical significance for the rest of the time intervals with p-values of 0.018, 0.029, and 0.018 for the periods of hemolysis of the least stable population, general population, and plateau respectively.

Finally, we were interested in pairwise comparison of treatment levels within the time intervals of statistical significance. Once again, similar results were found for both incubation times although the p-values were often larger for the 15 minute incubation time. During the hemolysis of the least stable red blood cell population, at least some evidence was found of a difference between low dosage and control (p-value$_{15}$=0.020, p-value$_{30}$=0.018), medium dosage and control (p-value$_{15}$=0.060, p-value$_{30}$=0.039), and low dosage and high dosage (p-value$_{15}$=0.057, p-value$_{30}$=0.030). During the hemolysis of the general population, at least some evidence of a significant difference was found between the low dose and control (p-value$_{15}$=0.060, p-value$_{30}$=0.029). During the plateau interval, there was a significant difference between low dose and control (p-value$_{15}$=0.001, p-value$_{30}$=0.018), medium dose and control (p-value$_{15}$=0.016, p-value$_{30}$=0.029), and high dose and control (p-value$_{15}$=0.030, p-value$_{30}$=0.029).

The results of the analysis can be summarized as follows. The rate of hemolysis increases with the dosage of novocaine. That is, the structural and functional modifications in the erythrocyte's membrane induced by novocaine are dosage dependent. The results also indicate the distribution of erythrocytes into sub-populations with low, medium and high resistance to hemolysis. These populations modified by novocaine react differently with the hemolytic agent. After 15 and 30 minutes of incubation, the "old" erythrocytes (least stable) modified by low (4.9×10^{-6} mol/L) and medium (1.0×10^{-5} mol/L) doses of procaine react faster to the hemolytic agent than those under the control or the high (2.01×10^{-5} mol/L) dose. However, reaction of the general and "young" (most stable) erythrocyte population modified by the

same (low and medium) dosages is characterized by higher stability of the membrane and thus have higher resistance to the hemolytic agent. Thus, novocaine in low and medium doses has a protective effect on the general and "young" erythrocyte populations. However, an increase in procaine dosage does not lead to increase of erythrocyte resistance to the hemolytic agent. The effect of the high dose of novocaine (2.01×10^{-5} mol/L) does not differ significantly from the control and thus is destructive rather than protective.

Conclusions of our statistical analysis confirm certain findings reported in a patent by Holodov and Nikolaevski (2012). Specifically, our analysis confirm that novocaine in low dosages tends to have a protective effect. However, Holodov and Nikolaevski (2012) reported a significant difference among erythrograms for all incubation times but zero minutes. This inconsistency is due to a failure to properly control the tests for the multiplicity in the original analysis. The findings reported in the current paper provide a higher assurance that a replication experiment will be able to detect the same differences reported here.

5. Discussion

We have suggested a procedure which allows researchers to find regions of significant difference in the domain of functional responses as well as to determine which populations are different over these regions. To the best of our knowledge, there are no existing competing procedures to the proposed methodology. Thus, our numerical results reported in Section 3 do not include a comparison of the proposed method to other alternatives. Nevertheless, the simulations revealed that our procedure has satisfactory power and does a good job of picking out the differences between population means. Also, in our simulation study, a relatively small number of regions

($m = 5$ and $m = 10$) was considered. A higher number of individual tests (intervals) can be easily implemented with the described shortcut to the closure principle.

Note that the regions of interest in the functional domain should be pre-specified prior to the analysis. However, in our experience researchers have never had a problem with *a priori* region identification. From previous research, expected results as well as specific regions of interest are typically known. We also mentioned that in the application of our method the intervals should be mutually exclusive and exhaustive. If researchers are interested in a test over overlapping intervals, the solution is to split the functional domain into smaller mutually exclusive intervals for individual tests (terminal nodes of the hypotheses tree). The decision for the overlapping region would be provided by a test of an intersection hypothesis ("higher" node in the hypotheses tree). We also expect the intervals to be exhaustive since it would be unexpected for researchers to collect data over time periods that they have no interest in. Finally, if for some reason distinct regions can not be prespecified, a large number of equal sized intervals can easily be employed.

We could not find a method directly comparable to the proposed procedure, but the present work has two open issues that suggest a direction for future research. First, the method is conservative and so a more powerful approach may be possible. Second, the permutation strategy for the pairwise comparison test may lead to biased inference. Solutions to the latter problem were suggested both by Petrondas and Gabriel (1983) and Troendle and Westfall (2011). We leave implementation of these solutions for future research as this seems to be a minor issue with a small number of treatment groups as are most often encountered in FANOVA applications.

β	$H_{AB} : \mu_1 = \mu_2$	$H_{AC} : \mu_1 = \mu_3$	$H_{BC} : \mu_2 = \mu_3$
0.318	0.027	0.021	0.026
0.364	0.029	0.024	0.028
0.409	0.031	0.034	0.038
0.455	0.036	0.041	0.047
0.500	0.036	0.049	0.054

Table 3.1: Power of the pairwise comparison assuming common means μ_1 and μ_2 over the 1st interval, (M2) model

β	$H_{AB} : \mu_1 = \mu_2$	$H_{AC} : \mu_1 = \mu_3$	$H_{BC} : \mu_2 = \mu_3$
0.273	0.018	0.049	0.057
0.318	0.025	0.074	0.086
0.364	0.031	0.104	0.116
0.409	0.037	0.145	0.164
0.455	0.041	0.214	0.224
0.500	0.045	0.298	0.323

Table 3.2: Power of the pairwise comparison assuming common means μ_1 and μ_2 over the 2nd interval, (M2) model

β	$H_{AB} : \mu_1 = \mu_2$	$H_{AC} : \mu_1 = \mu_3$	$H_{BC} : \mu_2 = \mu_3$
0.182	0.015	0.038	0.040
0.227	0.021	0.077	0.084
0.273	0.027	0.160	0.155
0.318	0.037	0.289	0.275
0.364	0.041	0.437	0.434
0.409	0.048	0.610	0.600
0.455	0.048	0.731	0.735
0.500	0.049	0.839	0.835

Table 3.3: Power of the pairwise comparison assuming common means μ_1 and μ_2 over the 3rd interval, (M2) model

β	$H_{AB}: \mu_1 = \mu_2$	$H_{AC}: \mu_1 = \mu_3$	$H_{BC}: \mu_2 = \mu_3$
0.182	0.017	0.082	0.080
0.227	0.023	0.207	0.196
0.273	0.030	0.375	0.365
0.318	0.036	0.618	0.611
0.364	0.039	0.817	0.807
0.409	0.041	0.920	0.915
0.455	0.041	0.971	0.971
0.500	0.041	0.993	0.993

Table 3.4: Power of the pairwise comparison assuming common means μ_1 and μ_2 over the 4th interval, (M2) model

β	$H_{AB}: \mu_1 = \mu_2$	$H_{AC}: \mu_1 = \mu_3$	$H_{BC}: \mu_2 = \mu_3$
0.136	0.012	0.044	0.042
0.182	0.020	0.164	0.160
0.227	0.030	0.380	0.383
0.273	0.038	0.640	0.645
0.318	0.041	0.858	0.859
0.364	0.042	0.955	0.957
0.409	0.042	0.986	0.988
0.455	0.042	0.997	1.000
0.500	0.042	1.000	1.000

Table 3.5: Power of the pairwise comparison assuming common means μ_1 and μ_2 over the 5th interval, (M2) model

References

Basso, D., Pesarin, F., Solmaso, L., Solari, A., 2009. Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applications with R. Springer.

Cox, D. D., Lee, J. S., 2008. Pointwise testing with functional data using the westfall-young randomization method. Biometrika 95 (3), 621–634.

Cuesta-Albertos, J. A., Febrero-Bande, M., 2010. Multiway anova for functional data. TEST 19, 537–557.

Cuevas, A., Febrero, M., Fraiman, R., 2004. An anova test for functional data. Computational Statistics and Data Analysis 47, 111–122.

Delicado, P., 2007. Functional k-sample problem when data are density functions. Computational Statistics 22 (3), 391–410.

Garcia-Rodriguez, L. A., Hernandez-Diaz, S., de Abajo, F. J., 2001. Association between aspirin and upper gastrointestinal complications: Systematic review of epidemiologic studies. British Journal of Clinical Pharmacology 52, 563–571.

Gower, J. C., Krzanowski, W. J., 1999. Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. Journal of the Royal Statistical Society 48 (4), 505–519.

Hochberg, Y., Tamhane, A. C., 1987. Multiple Comparison Procedures. Wiley.

Holodov, D. B., Nikolaevski, V. A., 2012. A method for preventing damages to the stomach mucous membrane when taking non-steroidal anti-inflammatory drugs. Patent RU 2449784.
URL `http://www.findpatent.ru/patent/244/2449784.html`

Marcus, R., Peritz, E., Gabriel, K. R., 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63 (3), 655–660.

Nasonov, E. L., Karateev, A. E., 2006. The use of non-steroidal anti-inflammatory drugs: clinical recommendations. Russian Medical Journal 14 (25), 1769–1777.

Pesarin, F., 1992. A resampling procedure for nonparametric combination of several dependent tests. Statistical Methods & Applications 1 (1), 87–101.

Petrondas, D. A., Gabriel, K. R., 1983. Multiple comparisons by rerandomization tests. Journal of the American Statistical Association 78 (384), 949–957.

Pollard, K. S., Gilbert, H. N., Ge, Y., Taylor, S., Dudoit, S., 2011. multtest: Resampling-based multiple hypothesis testing. R package version 2.10.0.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org

Ramsay, J. O., Hooker, G., Graves, S., 2009. Functional Data Analysis with R and MATLAB. Springer.

Ramsay, J. O., Silverman, B. W., 2005. Functional Data Analysis, Second Edition. Springer.

Shen, Q., Faraway, J., 2004. An F test for linear models with functional responses. Statistica Sinica 14, 1239–1257.

Troendle, J. F., Westfall, P. H., 2011. Permutational multiple testing adjustments with multivariate multiple group data. Journal of Statistical Planning and Inference 141, 2021–2029.

Westfall, P. H., Young, S. S., 1993. Resampling-based Multiple Testing: Examples and Methods for p-Values Adjustment. Wiley.

Wood, S. N., 2011. mgcv: generalized additive model method. R package version 1.7-19.
URL http://CRAN.R-project.org/package=mgcv

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., Weir, B. S., 2002. Truncated product method for combining p-values. Genetic Epidemiology 22 (2), 170–185.

CHAPTER 4

RESAMPLING-BASED MULTIPLE COMPARISON PROCEDURE WITH APPLICATION TO POINT-WISE TESTING WITH FUNCTIONAL DATA.

Contribution of Authors and Co-Authors

Author: Olga A. Vsevolozhskaya

Contributions: Wrote the majority of the manuscript.

Co-Author: Dr. Mark C. Greenwood

Contributions: Provided feedback on statistical analysis and drafts of the manuscript.

Co-Author: Dr. Scott. L. Powell

Contributions: Provided application expertise and feedback on drafts of the manuscript.

Co-Author: Dr. Dmitri V. Zaykin

Contributions: Provided feedback on statistical analysis and drafts of the manuscript.

Manuscript Information Page

Olga A. Vsevolozhskaya, Mark C. Greenwood, Scott L. Powell, Dmitri V. Zaykin

Environmental and Ecological Statistics

Status of Manuscript:

__X__ Prepared for submission to a peer-reviewed journal

_____ Officially submitted to a peer-review journal

_____ Accepted by a peer-reviewed journal

_____ Published in a peer-reviewed journal

Published Springer.

Submitted April, 2013

Abstract

In this paper we describe a coherent multiple testing procedure for correlated test statistics such as are encountered in functional linear models. The procedure makes use of two different p-value combination methods: the Fisher combination method and the Šidák correction-based method. The distribution of Fisher's and Šidák's test statistics are estimated through resampling to cope with the correlated tests. Building upon these two existing combination methods, we propose the smallest p-value as a new test statistic for each hypothesis. The closure principle is incorporated along with the new test statistic to obtain the overall p-value and appropriately adjust the individual p-values. Furthermore, a shortcut version for the proposed procedure is detailed, so that individual adjustments can be obtained even for a large number of tests. The motivation for developing the procedure comes from a problem of point-wise inference with smooth functional data where tests at neighboring points are related. A simulation study verifies that the methodology performs well in this setting. We illustrate the proposed method with data from a study on the aerial detection of the spectral effect of below ground carbon dioxide leakage on vegetation stress via spectral responses.

1. Introduction

High-dimensional data analysis is a current emphasis in statistical methodology development. High-dimensional data consisting of observations measured "continuously" in time are typically called functional data. Examples include longitudinal data with subjects exposed continually to a certain treatment (Coull et al. (2000)) and more recently data obtained through Next Generation Sequencing (NGS) (Luo et al. (2012)) with the position of a genetic variant in a genomic region playing the role of time. Because in practice the continuous measurements are approximated by a vector – a continuous function evaluated on a grid of L points t_i, $i = 1, \ldots, L$ – point-wise inference provides an intuitive and easy way to analyze functional data. For example, Godwin et al. (2010) were interested in variability observed in human motion patterns. By discretizing kinematic and kinetic lifting curves on a grid of

$L = 100$ points (and performing inference point-wise), they were able to demonstrate additional areas of difference in motion patterns beyond those identified by traditional analysis based solely on peak values. However, conclusions based on a set of L point-wise p-values may lead to far too many falsely significant tests (see Rice (1988) for some numerical examples). In particular, although Godwin et al. (2010) say that "additional areas outside of the peaks were significantly different," they concluded significance for *all* $L = 100$ points and *all* types of lifting curves. These conclusions made the interpretation of findings troublesome. An adequate method for simultaneous point-wise testing needs to account for potentially inflated false positive results.

The commonly used Bonferroni correction for false positive decisions is not ideal for point-wise inference with functional data. The Bonferroni procedure is designed to correct for L *independent* simultaneous tests. If functional inference is performed on a point-wise grid, the corresponding p-values at nearby time points are correlated and the Bonferroni correction becomes overly conservative (Cribbie (2007)). Some methods suggest replacing the number of tests L in the Bonferroni method by an estimate of the effective number of independent tests (Cheverud (2001), Nyholt (2004), Li and Ji (2005)). The idea is to estimate the effective number of tests based on the eigenvalue variance of the correlation matrix. However, the suggestion that a single parameter, i.e., the number of independent tests, can fully capture the correlation structure is rather simplistic.

Geologic carbon sequestration (GCS) is a carbon capture and storage technique that could play a major role in climate mitigation strategy. Our work is motivated by a problem of CO_2 surface leakage detection from a GCS site. Since vegetation is a predominant land cover over a GCS cite, Bellante et al. (2013) analyzed areal hyperspectral images of the simulated CO_2 leak site in attempt to identify differ-

ences in mean spectral signatures of healthy vegetation and vegetation under stress. Specifically, Bellante et al. (2013) proposed the Red Edge Index (REI) – a single test statistic that summarizes differences between the spectral signatures of healthy and stressed vegetation. We used the data collected by Bellante et al. (2013) in an attempt to identify specific wavelength regions where the mean spectral signatures (mean spectral responses) of healthy vegetation and vegetation under stress differ (see Figure 4.5). Our approach was to perform the analyses on a discretized grid of 80 points because the original spectral data were collected in 80 bands throughout the visible and near infrared wavelengths (see Bellante et al. (2013) for a detailed data collection description).

Although interest in point-wise inference is obvious, few viable approaches exist in this direction that account for inflated false positive results and correlation structure among tests. Ramsay and Silverman (2005) proposed a method for performing L point-wise tests simultaneously, however fail to adjust the results to control the family-wise error rate (FWER), the probability of at least one false rejection of all the tests. A more promising approach was introduced by Cox and Lee (2008) who used the multiplicity correction procedure proposed by Westfall and Young (1993) to control the FWER. Additionally, neither of the proposed methods provide a decision regarding the overall null hypothesis that all single L hypotheses are true. This is an undesirable property since a multiple comparison procedure may be non-coherent (Gabriel (1969)), i.e., the rejection of at least one individual hypothesis may not imply the rejection of the global null, which might lead to interpretation problems.

In this paper, we propose a point-wise procedure that both provides a decision for the overall hypothesis and adequately adjusts the individual p-values to account for L simultaneous tests. The method first uses two different p-value combining methods to summarize the associated evidence across L points, defines a new test

statistic, W, based on the smallest p-value from the two combination methods, and applies the closure principle of Marcus et al. (1976) to individually adjust the L point-wise p-values. The idea of using the minimum p-value as the test statistic for the overall test across different combination methods has been used in multiple genetics studies (Hoh et al. (2001), Chen et al. (2006), Yu et al. (2009)). A challenge for the proposed analysis was the individual adjustment performed using the closure principle. The closure principle generally requires $2^L - 1$ tests. To overcome this obstacle, we describe a computational shortcut which allows individual adjustments using the closure method even for large L. Accordingly, the paper is organized as follows. We give an overview of the closure principle and detail the computational shortcut to it. We give an explicit strategy for the proposed approach and compare its performance to other possibilities in a simulation study. We apply the proposed methodology in order to identify regions of the electromagnetic spectrum that differ based on distances to a simulated underground CO_2 leak.

2. Multiple Tests and Closure Principle

2.1. The General Testing Principle

It is well known that by construction all inferential methods have a nonzero probability of Type I error. Therefore, when L multiple tests are conducted simultaneously, the probability of finding at least one spurious result is greater than the threshold α. A multiple test adjustment procedure, which controls a family-wise error rate for the family of individual hypotheses, $H_1, \ldots, H_L$, at a pre-specified level α, can be obtained through the closure principle of Marcus et al. (1976). The closure principle considers all possible combination hypotheses obtained via the intersection of the set of L individual hypotheses $H_I = \cap\{H_i : i \in I\}$, $I \subseteq \{1, \ldots, L\}$. The coherence

of the procedure is enforced by rejecting an individual hypothesis H_i, $i = 1, \ldots, L$, only if all intersection hypotheses that contain it as a component are rejected. Most researchers prefer the results of a multiple test procedure to be presented in terms of L individually adjusted p-values. The individually adjusted p-value for the hypothesis H_i is set to the maximum p-value of all intersection hypotheses implied by H_i.

A valuable feature of the closure principle is its generality, i.e., any suitable α-level test can be used to test the intersection hypotheses. However, the implementation of the method becomes computationally challenging for a large number of tests. The total number of intersection hypotheses is $2^L - 1$ which grows quickly with L and limits the applicability of the method. Grechanovsky and Hochberg (1999) exhaustively discussed the conditions under which the closure procedure admits a shortcut. However, the discussion in Grechanovsky and Hochberg (1999) is motivated by the case of joint multivariate normal test statistics and the question remains of how to reduce the computational burden of the closure method in the case of non-normal correlated tests.

2.2. Closure in a Permutation Context

Permutation-based methods are becoming more popular for multiple testing corrections with high-dimensional data. They do not require normality assumptions and utilize the data-based correlation structure. That is, the resulting procedure for false positive decision corrections based on permutation test is exact despite unknown covariance structure (unlike the Bonferroni procedure that tends to be over conservative for correlated tests). The closure method easily admits permutation-based tests, all that is required is an α-level permutation test for each intersection hypothesis. Westfall and Troendle (2008) described a computational shortcut for the closure principle with a permutation test that reduces the number of required computations from the

order of 2^L to L. The drastic reduction in computational burden is achieved by (i) testing each intersection hypothesis H_I with either a min-p test statistic ($\min_{i\in I} p_i$, where p_i is the individual p-value) or a max-t statistic ($\max_{i\in I} t_I$, where t_i is the individual test statistic), and (ii) the assumption of subset pivotality. However, there are other more powerful test statistics one can use to test an intersection hypothesis H_I. Here, we show how to implement a computational shortcut for the Šidák (Šidák (1967)) and Fisher (Fisher (1932)) permutation-based tests to reduce the number of computations from the order of 2^L to the order of L^2.

Suppose K tests are conducted and the resulting p-values are $p_1, \ldots, p_K$. Denote the ordered p-values by $p_{(1)} \leq \cdots \leq p_{(K)}$. The test based on the Šidák correction for the intersection of K hypothesis, $\cap_{i=1}^K H_i$, is

$$S_K = 1 - (1 - p_{(1)})^K. \tag{4.1}$$

The Fisher test statistic for the same intersection hypothesis is

$$F_K = -2 \sum_{i=1}^{K} \ln p_i. \tag{4.2}$$

The permutation p-values based on the Šidák correction are equivalent to the p-values based on the min-p test statistic and the rank truncated product statistic (RTP), $W(K) = \prod_{i=1}^K p_{(i)}$, of Dudbridge and Koeleman (2003) with truncation at $K = 1$. The equivalence is due to the fact that $1 - (1 - p_{(1)})^K$ is a monotonic transformation of $p_{(1)}$. Similarly, $-2\sum_{i=1}^K \ln p_i$ is a monotonic transformation of $\prod_{i=1}^K p_{(i)}$, and the permutation p-values based on these two test statistics are equivalent.

The idea behind the shortcut is to consider only the "worst" (the smallest) test statistic in the subsets of the same cardinality. Note that, for both the Šidák

correction-based test and the Fisher test, the values of the test statistics are monotonically decreasing among intersection hypotheses of the same size. Thus, for the ordered p-values, $p_{(1)} \leq \cdots \leq p_{(L)}$, the hypotheses that will be used for individual adjustments are:

$$\begin{aligned}
&\text{for } H_1, && \{\mathbf{H_1}, \mathbf{H_{1L}}, \mathbf{H_{1L(L-1)}}, \ldots, \mathbf{H_{1L\ldots2}}\};\\
&\text{for } H_2, && \{\mathbf{H_2}, \mathbf{H_{2L}}, \mathbf{H_{2L(L-1)}}, \ldots, H_{2L\ldots1}\};\\
&\text{for } H_3, && \{\mathbf{H_3}, \mathbf{H_{3L}}, \mathbf{H_{3L(L-1)}}, \ldots, H_{3L\ldots1}\};\\
&\quad\vdots && \quad\vdots\\
&\text{for } H_L, && \{\mathbf{H_L}, H_{L(L-1)}, H_{L(L-1)(L-2)}, \ldots, H_{L(L-1)\ldots1}\}.
\end{aligned}$$

Here, the hypothesis H_1 has p-value $p_{(1)}$, H_2 has p-value $p_{(2)}$, etc. The unique intersection hypotheses to consider are highlighted in bold, and the number of unique tests to consider is $1 + 2 + \cdots L = \frac{L(L+1)}{2}$. One need not necessarily use resampling to apply the above shortcut, however it also works well if permutations are used to find p-values for each intersection hypothesis. Figure 4.1 illustrates correspondence between p-values calculated based on the full closure "resampling-based" procedure and the "resampling-based" shortcut. The graphs show an excellent agreement between the two adjusted p-values for both Šidák and Fisher test statistics. To obtain the plots, $L = 10$ p-values were simulated $B = 20$ times from the $Unif(0,1)$ distribution and corrected using the full closure and the computational shortcut. Considering the adjustments for more than $L = 10$ p-values was impractical due to the computational burden of the full closure procedure.

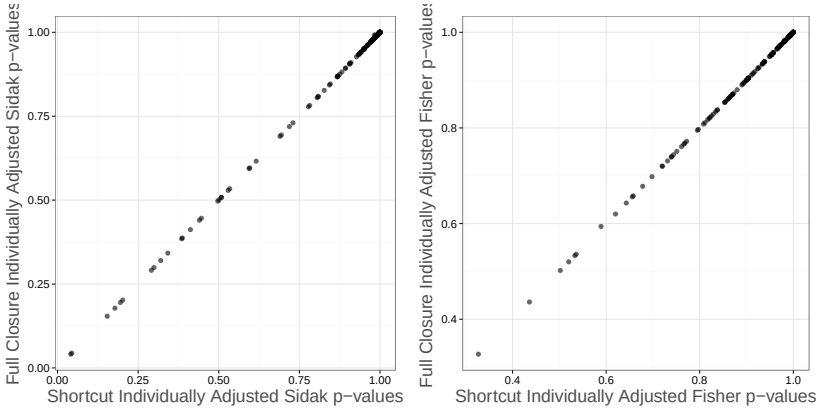

Figure 4.1: Correspondence between individually adjusted p-values using the full closure algorithm and the computational shortcut ($L = 10$). The Šidák p-values are illustrated in the left panel, and the Fisher p-values in the right panel.

3. Proposed Methodology

We now describe a permutation algorithm that ensures coherency when estimating adjusted individual p-values. Suppose L correlated tests are conducted simultaneously. Apply the following steps to obtain the corresponding individually adjusted p-values.

1. Construct $\frac{L(L+1)}{2} \times (B+1)$ matrices of Šidák and Fisher permuted test statistics, S_{ij}, F_{ij}, $i = 1, \ldots, \frac{L(L+1)}{2}$ (indexes intersection hypotheses), $j = 1, \ldots, B+1$ (indexes permutations). The first column of the matrices contains the observed test statistics.

2. Construct $\frac{L(L+1)}{2} \times (B+1)$ matrices of the permuted p-values based on the algorithm from Ge et al. (2003) – discussed below. The Šidák p-value for the i-th intersection hypothesis and j-th permutation is

$$P_{ij}^{s} = \frac{1}{B} \sum_{k=1, k \neq j}^{B+1} I(S_{ik} \leq S_{ij}).$$

The Fisher p-value for the i-th intersection hypothesis and j-th permutation is

$$P_{ij}^{f} = \frac{1}{B} \sum_{k=1, k \neq j}^{B+1} I(F_{ik} \geq F_{ij}).$$

3. Define the statistic $W_{ij} = \min(P_{ij}^{S}, P_{ij}^{f})$ and obtain its p-value as

$$P_{i}^{W} = \frac{1}{B} \sum_{k=2}^{B+1} I(W_{ik} \leq W_{i1}), \quad i = 1, \ldots, \frac{L(L+1)}{2}.$$

4. Make an overall decision and obtain L individually adjusted p-values by applying the closure principle to the set of P_{i}^{W}'s.

To avoid nested permutations in Step 2, we used the algorithm by Ge et al. (2003) to compute permutational p-values for each permutation $j = 2, \ldots, B+1$. More specifically, the algorithm allows one to obtain permutational p-values in the closure based on just B permutations instead of B^2. Also, in Step 3, testing the i^{th} intersection hypothesis with W_{ij} at a threshold α would lead to inflated Type I error rate, because choosing the smallest of the two p-values P_{ij}^{S} and P_{ij}^{f} leads to yet another multiple testing problem. To overcome this issue, one can use either the Bonferroni correction and define the test statistic as $2\min(P^{s}, P^{f})$ or, as suggested, determine the significance of W on the basis of permutations. Finally, setting $W = \min(P^{S}, P^{f})$ is the same as $\min(\mathrm{RTP}(1), \mathrm{RTP}(L))$, where $\mathrm{RTP}(\cdot)$ is the rank truncated product statistic of Dudbridge and Koeleman (2003) but also considered in Zaykin (2000) and Zaykin et al. (2007). Thus, W incorporates two extremes: the combination of all p-values and a min-p adjustment procedure. Simulation studies are used to show it retains desired properties of both type of statistics.

4. Simulations

4.1. Simulation Study Setup

We were motivated by a problem of identifying differences in mean spectral signatures of healthy vegetation and vegetation under stress across electromagnetic spectra. We approach the problem by evaluating functional responses on a grid of 80 points across wavelengths and performing tests point-wise. More generally, we were interested in evaluating k groups of functional responses on a grid of L points, $t_1, \ldots, t_L$, and performing point-wise inference in the functional data setting. The goal of the simulation study was to investigate the power of the proposed procedure to detect departures from (1) the global null hypothesis $\cap_{i=1}^{L} H_i$ of no difference anywhere in t and (2) the point-wise null hypotheses $H_0 : \mu_1(t_i) = \mu_2(t_i)$ for all t_i, $i = 1, \ldots, L$. We followed the setup of Cox and Lee (2008) and for all simulations generated two samples of functional data with $n_1 = n_2 = 250$ observations in each group ($N = 500$). The mean function of the first sample was constant and set to zero, $\mu_1(t) \equiv 0$, $t \in [0, 1]$. The mean of the second sample was either set to $\mu_2(t) = \gamma Beta(1000, 1000)(t)$ or $\mu_3(t) = \gamma Beta(5, 5)(t)$, where $Beta$ represents probability density function of the Beta distribution. Figure 4.2 illustrates $\mu_2(t)$ and $\mu_3(t)$ for the range of different γ values explored.

First, we simulated the case where all L point-wise hypotheses were true ($\mu_1(t_i) \equiv \mu_2(t_i)$ $\forall t_i$). To obtain functional data, we evaluated the mean functions on a grid of 140 equally spaced points ranging from -0.2 to 1.2 and added random noise, $\epsilon_{ij} \sim N(0, 0.01^2)$. Then, we fitted a smoothing spline using the `smooth.spline` R function (R Core Team (2013)) with the 0.95 smoothing parameter for each functional observation as suggested in Cox and Lee (2008). The output of the `smooth.spline` function is the fitted values of functional responses evaluated on the original grid of

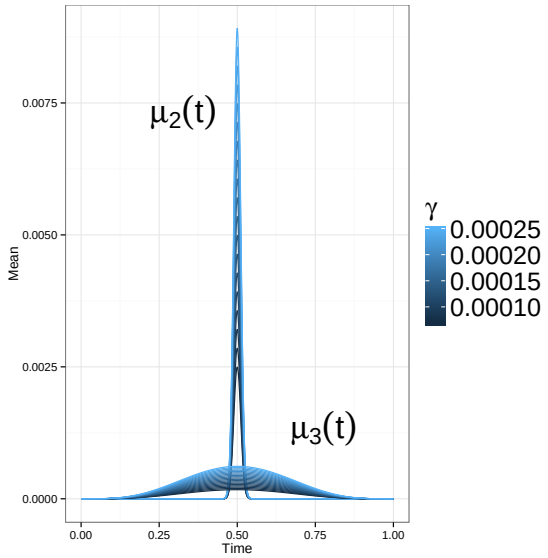

Figure 4.2: Two choices for the mean of the second sample.

points. We disposed of 20 points from each end to remove excessive boundary variability from the estimated splines and for each curve sub-sampled 50 equally spaced values on the grid between 0 and 1. At the 0.05 level, we evaluated the empirical Type I error rate for the global null and the control of the FWER in a weak sense (meaning that all observations come from the null hypothesis) for the proposed procedure and five alternative statistical methods: the Šidák correction based test, the Fisher test, the Cox and Lee method (Cox and Lee (2008)), the functional $\mathcal{F}$ statistic (Shen and Faraway (2004)), and the functional V_n (Cuevas et al. (2004)).

The functional test statistics of Shen and Faraway (2004) and Cuevas et al. (2004) are designed to perform the overall functional analysis of variance (FANOVA) test. The FANOVA null and alternative hypotheses are

$$H_0: \quad \mu_1(t) = \mu_2(t) = \ldots = \mu_k(t)$$

$$H_a: \quad \mu_i(t) \neq \mu_{i'}(t), \quad \text{for at least one } t \text{ and } i \neq i',$$

where $\mu_i(t)$ is assumed to be fixed, but unknown, population mean function of group i, $i = 1, \ldots, k$. Parametric distributions are available for both $\mathcal{F}$ and V_n from the original papers. The FANOVA test assesses evidence for the existence of differences among population mean curves in the entire functional domain. The test across the entire t is a global test. Thus, we considered these two methods as competitive to the proposed methodology.

Second, we investigated two properties of our method: (1) power to detect deviations from the combined null hypothesis $\cap_{i=1}^{L} H_i$ and (2) power to detect deviations from point-wise hypotheses $H_1, H_2, \ldots, H_L$. To calculate power for the combined null hypotheses, we simulated $B = 1000$ sets of functional observations for the specified range of γ values, performed the overall test and calculated the empirical probability of rejecting $\cap_{i=1}^{L} H_i$. At the point-wise level, the concept of power is not as clear cut. For example, one may calculate conjunctive power – the probability of rejecting all false null hypotheses – or disjunctive – the probability of rejecting at least one false hypothesis. For a detailed discussion of these different choices see Bretz et al. (2010). Here, we adopted the approach of Cox and Lee (2008) to be able to directly compare to their results. We considered a single simulated set of functional observations for a specific choice of γ; calculated the unadjusted point-wise p-values; performed the multiplicity adjustment using W, as well as by Fisher's, Šidák's, and Westfall-Young method; then we compared the adjusted p-values by plotting them on a single graph.

4.2. Results

Table 4.1 summarizes control of the Type I error rate for the overall null hypothesis, $\cap_{i=1}^{L} H_i$, and the family-wise error rate (FWER) in the weak sense for the point-wise tests (i.e., $\cap_{i=1}^{L} H_i$ is true). All methods tend to be liberal in terms of the Type I error rate control ("combined null" line). The family-wise error rate is inflated

for Šidák's test, too conservative for Fisher's test, and right on the 0.05 margin for the Westfall-Young adjustment.

	Šidák	Fisher	W	Cox and Lee	$\mathcal{F}$	V_n
combined null	0.059	0.065	0.060	NA	0.059	0.057
FWER	0.059	0.000	0.036	0.049	NA	NA

Table 4.1: The Type I error for the global null ($\cap_{i=1}^{L} H_i$) and the FWER for $L = 50$ tests, 1000 simulations, and $\alpha = 0.05$.

Figure 4.3 illustrates power for the global null hypothesis ($\cap_{i=1}^{L} H_i$). We see that in both graphs Fisher's method outperforms all of the other methods, however W has similar power for this realization. The performance of the functional $\mathcal{F}$ of Shen and Faraway (2004) is very similar to the functional V_n of Cuevas et al. (2004). The Šidák test is the clear laggard.

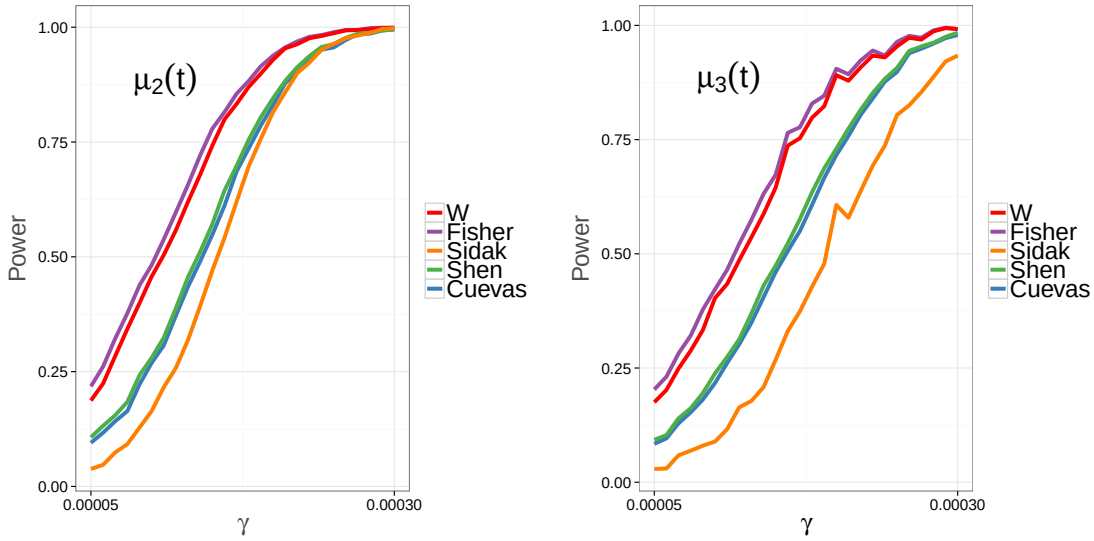

Figure 4.3: Plots of empirical power for the combined null hypothesis with $\alpha = 0.05$.

Figure 4.4 shows the unadjusted and the adjusted p-values for a single set of functional observations. To compute the unadjusted p-values, we simulated 250 curves with mean $\mu_1(t) = 0$ and $\mu_2(t) = 0.0003Beta(1000, 1000)(t)$ (left graph) or $\mu_1(t) = 0$ and $\mu_3 = 0.0003Beta(5, 5)(t)$ (right graph) and performed a t-test on a grid of 50 equally spaced points $t_1 = 0, \ldots, t_{50} = 1$. From both graphs, it is evident that Fisher's method has the lowest power. The performance of W is very similar to Šidák's test. The Westfall–Young method has the highest power.

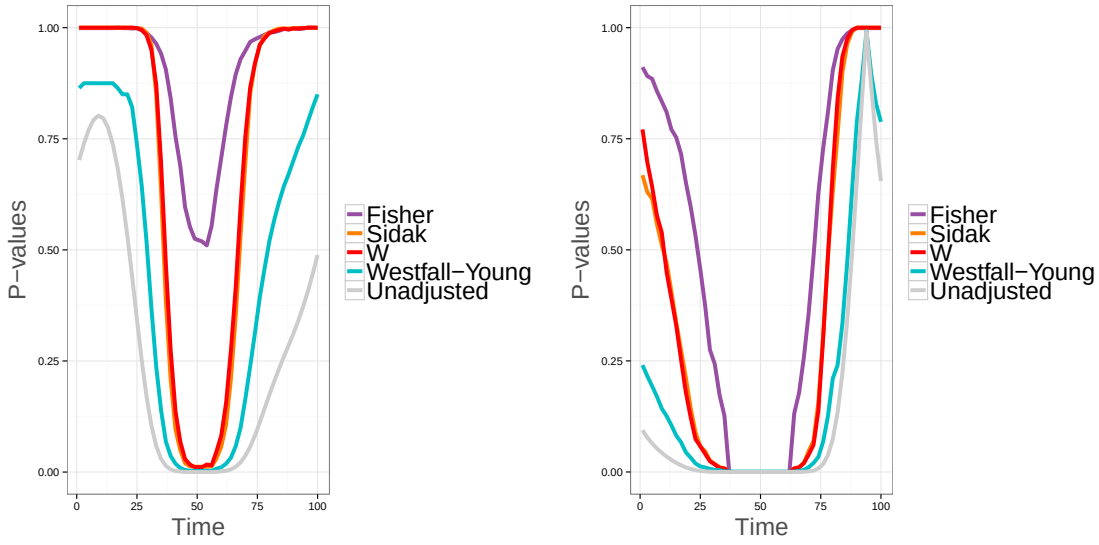

Figure 4.4: Plots of point-wise adjusted p-values for $\gamma = 0.0003$. Left graph: $H_i : \mu_1(t_i) = \mu_2(t_i)$, $i = 1, \ldots, L$. Right graph: $H_i : \mu_1(t_i) = \mu_3(t_i)$, $i = 1, \ldots, L$.

5. Application to Carbon Dioxide Data

Bellante et al. (2013) conducted an experiment to study the effect of carbon dioxide (CO_2) surface leak on vegetation stress at the Montana State University Zero Emission Research and Technology (ZERT) site in Bozeman, MT. To study the spec-

tral changes in overlying vegetation in response to elevated soil CO_2 levels, a time series of aerial images were acquired over a buried carbon dioxide release pipe. A single image acquired on June 21, 2010 was the focus of the current analysis. The pixel-level measurements (with nearly 32,000 pixels) of the image consist of 80 spectral responses ranging from 424 to 929 nm. For each pixel, a horizontal distance to the CO_2 release pipe was calculated and 500 spectral responses were randomly chosen from five distance subcategories: (0,1], (1,2], (2,3], (3,4], and (4,5] meters (see Figure 4.5). To obtain a functional response for each pixel, we used the penalized cubic B-spline smoother with a smoothing parameter determined by generalized cross-validation (Ramsay et al. (2012)). The functional responses were evaluated on the original grid of $L = 80$ points and subsequently the analysis of variance test was performed point-wise to obtain the unadjusted p-values.

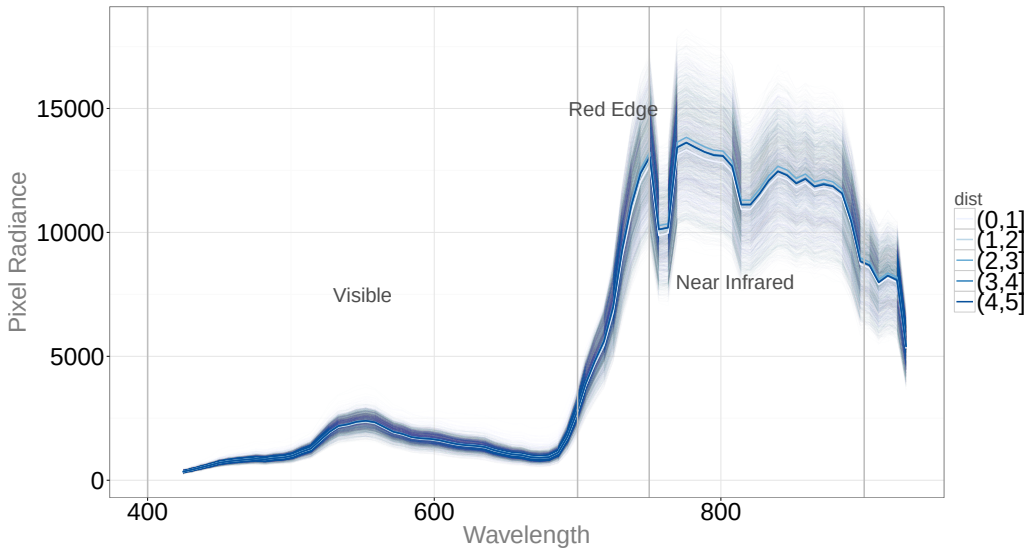

Figure 4.5: Spectral responses from 2,500 pixels corresponding to five different binned distances with superimposed fitted mean curves.

First, we tested the global null hypothesis of no difference in the entire range of spectral responses based on the distance from the CO_2 release pipe and obtained the corresponding overall p-value of 0.001 (from 1000 permutations) using W. We then obtained the corrected point-wise p-values, which are illustrated in Figure 4.6. The adjusted p-values from 700 to 750 nm were below $\alpha = 0.05$ and correspond to the "red edge" spectral region, which indicates that the spectral responses among binned distances differ significantly within this region. This is an encouraging result since previous research has indicated that the "red edge" spectral region is typically associated with plant stress (Carter and Knapp (2001)).

The method proposed by Cox and Lee (2008), which employs the Westfall-Young correction for multiplicity, identifies a much larger region of wavelengths than the other methods. On the one hand, these additional discoveries may contribute to the higher power of the method. On the other hand, these result may be due to inflated FWER control in the strong sense. That is, we suspect that in the application the p-values come from a mixture of the null and the alternative hypotheses. Our simulations provided the FWER control only for a situation when *all* observations came from the null hypothesis. More research is required in this direction to make more conclusive statements.

6. Discussion

Modern data recording techniques allow one to sample responses at a high time resolution. In many applications it is of interest to utilize all of the recorded information and perform a test at each point, while accounting for the correlation of the test statistics at nearby times, properly controlling the probability of false positive findings, and providing information on the overall difference. Here, we suggested a

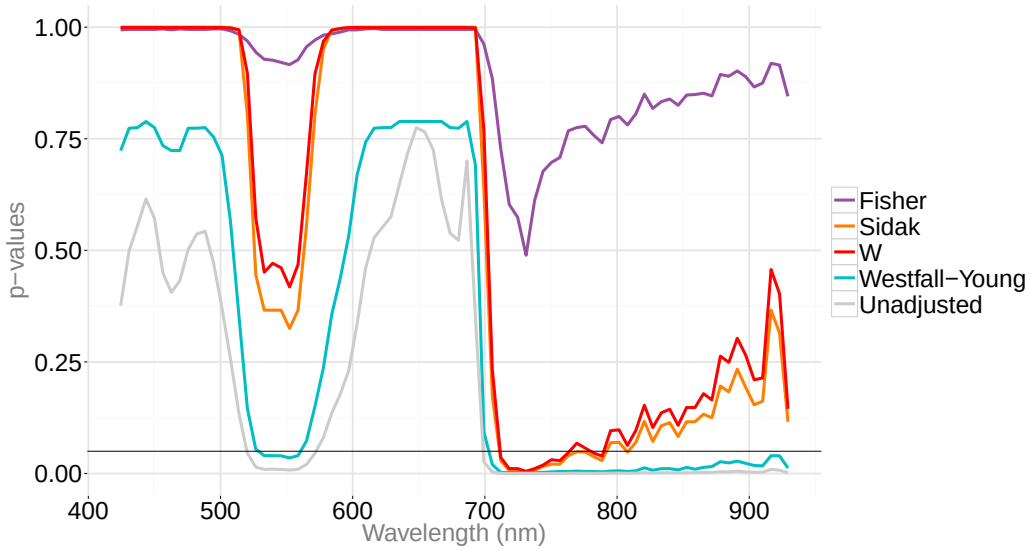

Figure 4.6: Plots of unadjusted and adjusted p-values. A horizontal line at 0.05 is added for a reference.

coherent method for point-wise testing with the desired properties. Our approach was inspired by methods used in genetic association studies, but extends these ideas and allows for obtaining individually adjusted p-values in the case of multiple correlated tests.

Our method capitalizes on the evidence based on the minimum p-value (the Šidák method) and the product (or the sum on the logarithmic scale) of all p-values (the Fisher method). This results in a procedure that has high power for the combined null hypothesis, $\cap_{i=1}^{L} H_i$, and for the individual tests $H_1, H_2, \ldots, H_L$. These characteristics of our procedure can be better understood by examining rejection regions of Fisher's and Šidák's tests. In general, rejection regions for L tests are hypervolumes in L-dimensional space, however some conclusions can be drawn from considering just two p-values. The two-dimensional rejection regions for Fisher's and Šidák's tests are provided in Loughin (2004). Based on the rejection regions, a clear difference is evident between the Fisher method and the Šidák method. In particular, the Fisher

method will reject the combined null hypothesis, $\cap_{i=1}^{L} H_i$, if at least some p-values are "small enough", but not necessarily significant. The Šidák method will reject the combined null hypothesis only if min-p is significant. Thus, Fisher's method is higher-powered than Šidák's method for the overall null hypothesis. On the other hand, Fisher's test along with the closure principle is lower-powered than Šidák's method for the individual adjustments. Envision a situation where the smallest p-value, $p_{(1)}$, is just above α. The adjusted value of $p_{(1)}$ by the closure principle is the maximum p-value of all hypotheses implied by $H_{(1)}$. To test an intersection hypothesis of size K, Fisher's test considers the combination of $p_{(1)}, p_{(L)}, \ldots, p_{(L-K+1)}$. All $p_{(L)}, \ldots, p_{(L-K+1)}$ are greater than $p_{(1)}$ and Fisher's test will not be able to reject $\cap_{i=1}^{K} H_i$ and thus $H_{(1)}$. Conversely, the decision for $\cap_{i=1}^{K} H_i$ based on Šidák's test is made regardless of the magnitudes of $p_{(L)}, \ldots, p_{(L-K+1)}$ but solely on the magnitude of $p_{(1)}$. Thus, the Šidák method along with closure principle has higher power than the Fisher method for the individual tests $H_1, H_2, \ldots, H_L$. Since our approach combines the Fisher and the Šidák method, it possesses desirable properties of both tests and has high power for all $\cap_{i=1}^{L} H_i$ and $H_1, H_2, \ldots, H_L$.

Our method is permutation-based. Generally, a drawback of the permutations methods is their computational intensity. However, there is a big advantage to using permutation-based methods. Cohen and Sackrowitz (2012) note that stepwise multiple testing procedures (including the closure principle) are not designed to account for a correlation structure among hypotheses being tested. That is, test statistics for an intersection hypothesis will always be the same regardless of the correlation structure among tests considered. Thus, the shortcoming of the stepwise procedures is determining a correct critical value. The permutation-based approach alleviates this shortcoming and allows for dependency to be incorporated into the calculation of the critical values.

Another advantageous property of our method is that it does not require access to the original data but only to the L unadjusted point-wise p-values. The matrices of the test statistics in Step 1 can be found based on the Monte Carlo algorithm described in Zaykin et al. (2002). The test statistics are found by first obtaining $L \times 1$ vectors, $\mathbf{R}^*$, of independent random values from the $Unif(0,1)$ distribution and then transforming them to $\mathbf{R}$ – vectors with components that have the same correlation structure as the observed p-values. Since functional observations are evaluated on a dense grid of points, the correlation structure among observed p-values can be estimated with reasonable precision. Thus, our method efficiently employs information contained just in the p-values and is more flexible than methods that require access to the original observations.

In summary, we proposed a coherent p-value combination method that allows researchers to obtain individually adjusted p-values for multiple simultaneous correlated tests. We hope that our work will promote new research in this direction. In particular, in our approach we treated all p-values as equally important. It might be possible to incorporate some weights that would optimize desirable properties of the procedure based on a particular application. Alternatively, adaptive selection of the test statistic is possible. That is, instead of considering just min-p (RTP(1)) and the combination of all p-values (RTP(L)), one might optimize power and size of the proposed method by considering RTP(K) across all possible values of $K = 1, \ldots, L$.

Software

A sample script to adjust the point-wise p-values with the proposed method is available at `http://www.math.montana.edu/~vsevoloz/fanova/minSF/`. The

script requires users to provide a vector of unadjusted point-wise p-values. The authors welcome questions regarding script usage.

References

Bellante, J., Powell, S., Lawrence, R., Repasky, K., Dougher, T., 2013. Aerial detection of a simulated co2 leak from a geologic sequestration site using hyperspectral imagery. International Journal of Greenhouse Gas Control 13, 124–137.

Bretz, F., Hothorn, T., Westfall, P., 2010. Multiple Comparisons Using R. Chapman & Hall/CRC.

Carter, G., Knapp, A., 2001. Leaf optical properties in higher plants: linking spectral characteristics to stress and chlorophyll concentration. American Journal of Botany 88, 677–684.

Chen, B., Sakoda, L., Hsing, A., Rosenberg, P., 2006. Resamplingbased multiple hypothesis testing procedures for genetic casecontrol association studies. Genetic Epidemiology 30, 495–507.

Cheverud, J., 2001. A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87, 52–58.

Cohen, A., Sackrowitz, H., 2012. The interval property in multiple testing of pairwise-differences. Statistical Science 27 (2), 294–307.

Coull, B., Catalano, P., Godleski, J., 2000. Semiparametric analysis of cross-over data with repeated measures. Journal of Agricultural, Biological, and Environmental Statistics 5 (4), 417–429.

Cox, D. D., Lee, J. S., 2008. Pointwise testing with functional data using the westfall-young randomization method. Biometrika 95 (3), 621–634.

Cribbie, R. A., 2007. Multiplicity control in structural equation modeling. Structural Equation Modeling 14 (1), 98–112.

Cuevas, A., Febrero, M., Fraiman, R., 2004. An anova test for functional data. Computational Statistics and Data Analysis 47, 111–122.

Dudbridge, F., Koeleman, B., 2003. Rank truncated product of p-values, with application to genomewide association scans. Genetic Epidemiology 25, 360–366.

Fisher, R., 1932. Statistical Methods for Research Workers. Oliver and Boyd, London.

Gabriel, K. R., 1969. Simultaneous test procedures – some theory of multiple comparison. Annals of Mathematical Statistics 40, 224–250.

Ge, Y., Dudoit, S., Speed, T., 2003. Resampling-based multiple testing for microarray data analysis. Test 12, 1–44.

Godwin, A., Takaharab, G., Agnewc, M., Stevensond, J., 2010. Functional data analysis as a means of evaluating kinematic and kinetic waveforms. Theoretical Issues in Ergonomics Science 11 (6), 489–503.

Grechanovsky, E., Hochberg, Y., 1999. Closed procedures are better and often admit a shortcut. Journal of Statistical Planning and Inference 76, 79–91.

Hoh, J., Wille, A., Ott, J., 2001. Trimming, weighting, and grouping snps in human case-control association studies. Genome Research 11, 2115–2119.

Li, J., Ji, L., 2005. Adjusting multiple testing in multilocus analysis using the eigenvalues of a correlation matrix. Heredity 95, 221–227.

Loughin, T., 2004. A systematic comparison of methods for combining p-values from independent tests. Computational Statistics & Data Analysis 47, 467–485.

Luo, L., Zhu, Y., M., X., 2012. Quantitative trait locus analysis for next-generation sequencing with the functional models. Journal of Medical Genetics 49, 513–524.

Marcus, R., Peritz, E., Gabriel, K. R., 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63 (3), 655–660.

Nyholt, D., 2004. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. American Journal of Human Genetics 74, 765–769.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org`

Ramsay, J. O., Silverman, B. W., 2005. Functional Data Analysis, Second Edition. Springer.

Ramsay, J. O., Wickham, H., Graves, S., Hooker, G., 2012. fda: Functional Data Analysis. R package version 2.3.2.
URL `http://CRAN.R-project.org/package=fda`

Rice, W., 1988. Analyzing tables of statistical tests. Evolution 43 (1), 223–225.

Shen, Q., Faraway, J., 2004. An F test for linear models with functional responses. Statistica Sinica 14, 1239–1257.

Šidák, Z., 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 78, 626–633.

Westfall, P., Troendle, J., 2008. Multiple testing with minimal assumptions. Biometrical Journal 50, 745–755.

Westfall, P. H., Young, S. S., 1993. Resampling-based Multiple Testing: Examples and Methods for p-Values Adjustment. Wiley.

Yu, K., Li, Q., Bergen, A., Pfeiffer, R., Rosenberg, P., Caporasi, N., Kraft, P., Chatterjee, N., 2009. Pathway analysis by adaptive combination of p-values. Genetic Epidemiology 33, 700–709.

Zaykin, D. V., 2000. Statistical analysis of genetic associations. Ph.D. thesis. North Carolina State University.

Zaykin, D. V., Zhivotovsky, L. A., Czika, W., Shao, S., Wolfinger, R. D., 2007. Combining p-values in large-scale genomics experiments. Pharmaceutical Statistics 6 (3), 217–226.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., Weir, B. S., 2002. Truncated product method for combining p-values. Genetic Epidemiology 22 (2), 170–185.

CHAPTER 5

GENERAL DISCUSSION.

In this work, we presented procedures that allow for the extension of the overall functional analysis of variance hypothesis testing. Our procedures capitalize on the closure method along with the different combination methods of the test statistics or the p-values. The closure multiplicity testing proved to be highly flexible since different α-level tests can be employed to test different intersection hypotheses. Next, we discuss limitations of the proposed methodology and outline direction for future research.

In Chapter 2, we introduced the idea of a combining function as a weighted sum of the observed test statistics. We also mentioned an extreme case of weights (all but one are zeros) used by Ramsay et al. (2009). In our procedure we employ an equal-weight combining function, however we think that incorporation of different weights is really promising for improving the procedure in terms of size and power. Also, in Chapter 2 we gave little weight to the discussion of the computational intensity of the simulation studies with functional data. Despite escalating computing power, it is still hard to handle extensive simulations and investigate all the desired properties of a procedure. We briefly mentioned the Boos and Zhang (2000) method of power extrapolation based on a set of 59/39/19 permutations. We also talked about a need for an efficient computational shortcut to the closure procedure. Based on referee feedback, we removed a discussion of the permutation strategy with the functional responses. To speed up permutations, we employed a distance-based permutational multivariate analysis of variance (perMANOVA) method of Anderson (2001) implemented in the `vegan R` package(Oksanen et al. (2012)). The package allows one to

compute a pseudo-F statistic (Anderson (2001)) – equivalent to the functional $\mathcal{F}$ (Shen and Faraway (2004)) – and provides an efficient distance-based permutation strategy. For future research, it might be of interest to investigate further connections between FANOVA and perMANOVA and make some generalizations.

In Chapter 3, we discussed a method that controls for multiple testing of all pair-wise differences. It would be of interest to see if the proposed methodology satisfies the "interval property". If a procedure lacks the interval property, it could reject the null hypothesis in one instance, and fail to reject it in a situation when intuitively we have "stronger" evidence against the null hypothesis. The lack of the interval property in a one-way ANOVA for testing all pair-wise differences is shown in Cohen et al. (2010). In Cohen and Sackrowitz (2012), a residual based multiple testing procedure for pairwise differences is introduced that does have the interval property. It would be of interest to explore the interval property in the FANOVA setting and develop a procedure that satisfies it. Another possibility of further extension of the proposed methodology involves testing of treatment versus control problems and change point problems.

In Chapter 4, we introduced a procedure that allows one to adjust L correlated p-values for multiplicity as well as to combine information across K multiple tests, i.e., test an intersection hypothesis $\cap_{i=1}^{K} H_i$. The motivation for the study came from a problem of point-wise testing in the FANOVA setting. For the two-group comparison the point-wise null hypothesis is $H_0 : \mu_1(t) = \mu_2(t)$ for each t versus a two-sided alternative $H_a : \mu_1 \neq \mu_2$. However, in certain situations combining two-sided p-values is undesirable because they disregard the effect direction. For example, imagine a replication study in which the effect direction is flipped but both two-sided p-values are small. The resulting combined p-value is going to be small and will promote false conclusions. However, the combined result of the corresponding one-sided p-

values will properly reflect the change in the effect size direction. A simple way of taking the effect size into consideration is presented in Zaykin (2011). This method allows one to convert a two-sided p-value into a one-sided p-value and vice versa. It also allows incorporation of different weights like the square root of the sample size. Integration of this method into the procedure presented in Chapter 4 might broaden its applicability.

Chapter 4 also discussed the issue of individual adjustment of many correlated tests. Our solution to the test dependency problem was to use a permutation-based method to find the p-values for each intersection hypothesis. The solution to the plethora of tests was presented in a form of a shortcut to the closure principle of Marcus et al. (1976). However, the temporal correlation might be directly incorporated into the calculation of a test statistic. The idea is to combine p-values that are separated up to the points in time at which the correlation extinguishes. If L non-independent tests are performed and p_i, $i = 1, \ldots, L$, are the corresponding p-values, the test statistic will be

$$W = \prod_{i=1}^{L} p_i^{w_i},$$

where w_i are some weights based on the distance between points. For example, assume that we have a temporal correlation up to lag 2 and we are interested in testing five individual hypotheses over time points $t_1, \ldots, t_5$. We would incorporate distance between time points into the construction of the intersection test statistics (Figure 5.1) and assign zero weights to the p-values that are "far apart" (Figure 5.2). The idea is inspired by the work in Zaykin (2000), Zaykin et al. (2002), and Dudbridge and Koeleman (2003). In Zaykin (2000) and Zaykin et al. (2002) the weights were

assigned based on a truncation point τ, i.e., $w_i = I(p_i \leq \tau)$. In Zaykin (2000) and Dudbridge and Koeleman (2003) the p-values up to rank k were combined.

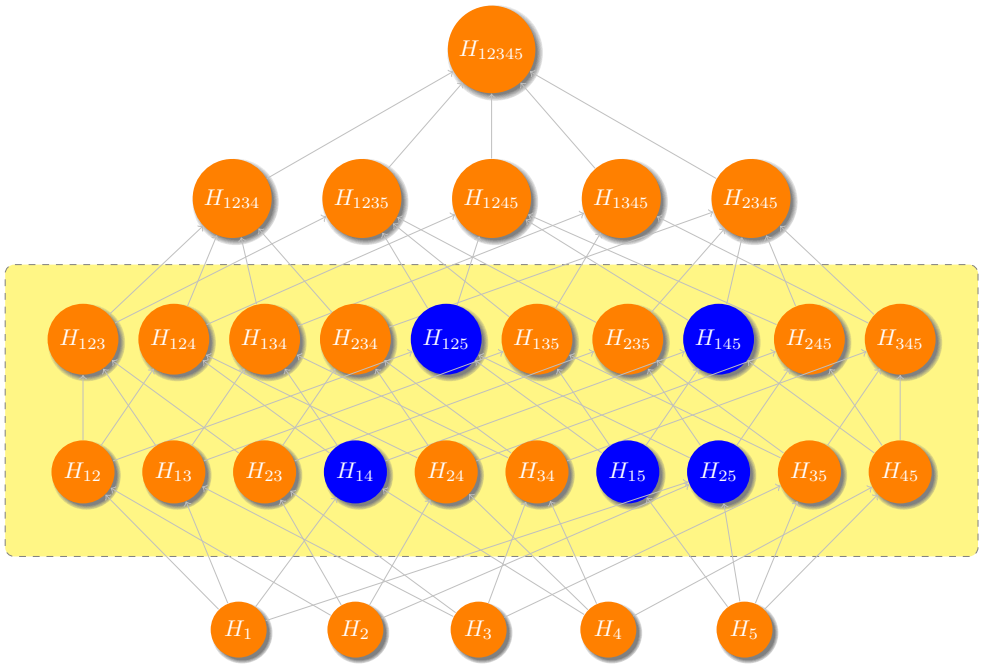

Figure 5.1: The closure set formed by five individual hypotheses. The intersection hypotheses that correspond to time points "far apart" are highlighted in blue.

Finally, applicability of the FANOVA can be extended to data sets that traditionally are not considered "functional." Specifically, Luo et al. (2012) studied the association between a quantitative trait and genetic variants in a genomic region. Quantitative trait was treated as a scalar response. Genotype profile was considered to be a function of a genomic position. Luo et al. (2012) tested for the additive effect of a marker at the genomic position t across the entire genomic region. That is, if t

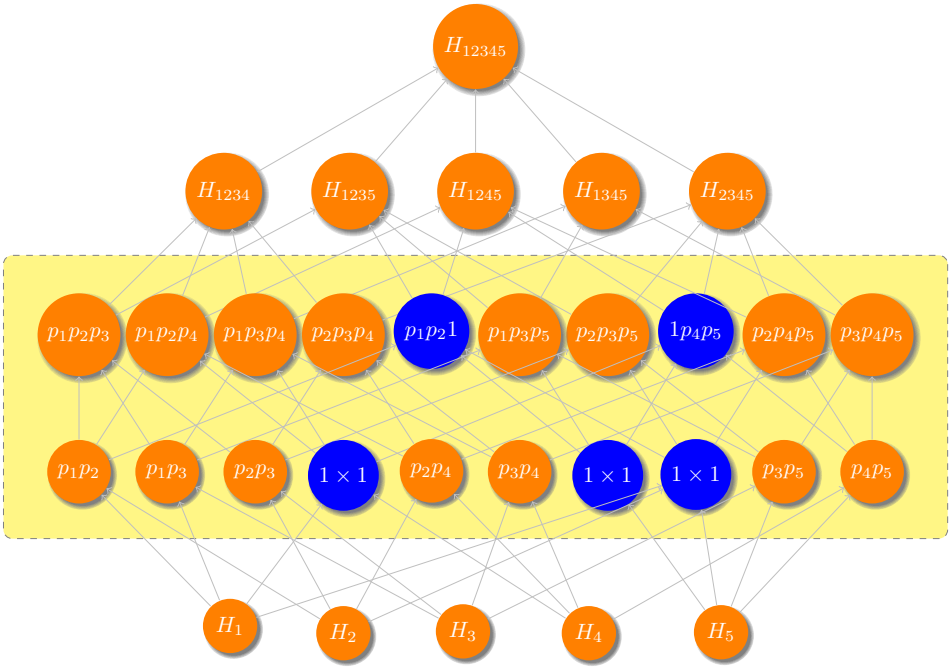

Figure 5.2: The p-values corresponding to time points "far apart" are assigned zero weights.

is a genomic position, a genomic profile $X_i(t)$ of the i-th individual is:

$$X_i(t) = \begin{cases} 1, & MM \\ 0, & Mm \\ -1, & mm \end{cases}.$$

Y_i is a quantitative phenotype value of the i-th individual and a functional linear model for a quantitative trait can be written as:

$$Y_i = \mu + \int_0^T X_i(t)\alpha(t)dt + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$, T is the length of the genome region, and $\alpha(t)$ is a function of the genetic additive effect of the marker at the genomic position t. What can we do differently? We can "flip" the relationship, i.e., try to determine if there is a significant difference among SNP's with minor and major allele frequency based on a certain categorical phenotype (like the presence or absence of a disease). We are planning to investigate the performance of the FANOVA methodology in this setting in future research.

References

Anderson, M. J., 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26, 32–46.

Boos, D. D., Zhang, J., 2000. Monte carlo evaluation of resampling-based hypothesis tests. Journal of the American Statistical Association 95, 486–492.

Cohen, A., Sackrowitz, H., 2012. The interval property in multiple testing of pairwise-differences. Statistical Science 27 (2), 294–307.

Cohen, A., Sackrowitz, H. B., Chen, C., 2010. Multiple testing of pairwise comparisons. Borrowing Strength: Theory Powering Applications – A Festschrift for Lawwrence D. Brown. IMS Collections 6 (144-157).

Dudbridge, F., Koeleman, B., 2003. Rank truncated product of p-values, with application to genomewide association scans. Genetic Epidemiology 25, 360–366.

Luo, L., Zhu, Y., M., X., 2012. Quantitative trait locus analysis for next-generation sequencing with the functional models. Journal of Medical Genetics 49, 513–524.

Marcus, R., Peritz, E., Gabriel, K. R., 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63 (3), 655–660.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Wagner, H., 2012. vegan: Community Ecology Package. R package version 2.0-5.
URL `http://CRAN.R-project.org/package=vegan`

Ramsay, J. O., Hooker, G., Graves, S., 2009. Functional Data Analysis with R and MATLAB. Springer.

Shen, Q., Faraway, J., 2004. An f test for linear models with functional responses. Statistica Sinica 14, 1239–1257.

Zaykin, D. V., 2000. Statistical analysis of genetic associations. Ph.D. thesis. North Carolina State University.

Zaykin, D. V., 2011. Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis. Journal of Evolutionary Biology 24 (8), 1836–1841.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., Weir, B. S., 2002. Truncated product method for combining p-values. Genetic Epidemiology 22 (2), 170–185.

REFERENCES CITED

Abramovich, F., Antoniadis, A., Sapatinas, T., Vidakovic, B., 2002. Optimal testing in functional analysis of variance models. Tech. rep., Georgia Institute of Technology.

Anderson, M. J., 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26, 32–46.

Basso, D., Pesarin, F., Solmaso, L., Solari, A., 2009. Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applications with R. Springer.

Bellante, G. J., 2011. Hyperspectral remote sensing as a monitoring tool for geologic carbon sequestration. Master's thesis, Montana State University.

Bellante, J., Powell, S., Lawrence, R., Repasky, K., Dougher, T., 2013. Aerial detection of a simulated co2 leak from a geologic sequestration site using hyperspectral imagery. International Journal of Greenhouse Gas Control 13, 124–137.

Berk, M., Ebbels, T., Montana, G., 2011. A statistical framework for biomarker discovery in metabolomic time course data. Bioinformatics 27 (14), 1979–1985.

Boos, D. D., Zhang, J., 2000. Monte carlo evaluation of resampling-based hypothesis tests. Journal of the American Statistical Association 95, 486–492.

Bretz, F., Hothorn, T., Westfall, P., 2010. Multiple Comparisons Using R. Chapman & Hall/CRC.

Carter, G., Knapp, A., 2001. Leaf optical properties in higher plants: linking spectral characteristics to stress and chlorophyll concentration. American Journal of Botany 88, 677–684.

Chen, B., Sakoda, L., Hsing, A., Rosenberg, P., 2006. Resamplingbased multiple hypothesis testing procedures for genetic casecontrol association studies. Genetic Epidemiology 30, 495–507.

Cheverud, J., 2001. A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87, 52–58.

Cohen, A., Sackrowitz, H., 2012. The interval property in multiple testing of pairwise-differences. Statistical Science 27 (2), 294–307.

Cohen, A., Sackrowitz, H. B., Chen, C., 2010. Multiple testing of pairwise comparisons. Borrowing Strength: Theory Powering Applications – A Festschrift for Lawwrence D. Brown. IMS Collections 6 (144-157).

Coull, B., Catalano, P., Godleski, J., 2000. Semiparametric analysis of cross-over data with repeated measures. Journal of Agricultural, Biological, and Environmental Statistics 5 (4), 417–429.

Cox, D. D., Lee, J. S., 2008. Pointwise testing with functional data using the westfall-young randomization method. Biometrika 95 (3), 621–634.

Cribbie, R. A., 2007. Multiplicity control in structural equation modeling. Structural Equation Modeling 14 (1), 98–112.

Cuesta-Albertos, J. A., Febrero-Bande, M., 2010. Multiway anova for functional data. TEST 19, 537–557.

Cuevas, A., Febrero, M., Fraiman, R., 2004. An anova test for functional data. Computational Statistics and Data Analysis 47, 111–122.

de Boor, C., 1978. A Practical Guide to Splines. Springer, New York.

Delicado, P., 2007. Functional k-sample problem when data are density functions. Computational Statistics 22 (3), 391–410.

Dudbridge, F., Koeleman, B., 2003. Rank truncated product of p-values, with application to genomewide association scans. Genetic Epidemiology 25, 360–366.

Faraway, J., 1997. Regression analysis for a functional response. Technometrics 39, 254–261.

Fisher, R., 1932. Statistical Methods for Research Workers. Oliver and Boyd, London.

Gabriel, K. R., 1969. Simultaneous test procedures – some theory of multiple comparison. Annals of Mathematical Statistics 40, 224–250.

Garcia-Rodriguez, L. A., Hernandez-Diaz, S., de Abajo, F. J., 2001. Association between aspirin and upper gastrointestinal complications: Systematic review of epidemiologic studies. British Journal of Clinical Pharmacology 52, 563–571.

Ge, Y., Dudoit, S., Speed, T., 2003. Resampling-based multiple testing for microarray data analysis. Test 12, 1–44.

Godwin, A., Takaharab, G., Agnewc, M., Stevensond, J., 2010. Functional data analysis as a means of evaluating kinematic and kinetic waveforms. Theoretical Issues in Ergonomics Science 11 (6), 489–503.

Gower, J. C., Krzanowski, W. J., 1999. Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. Journal of the Royal Statistical Society 48 (4), 505–519.

Grechanovsky, E., Hochberg, Y., 1999. Closed procedures are better and often admit a shortcut. Journal of Statistical Planning and Inference 76, 79–91.

Green, P., Silverman, B., 1994. Nonparametric Regression and Generalized Linear Models. Chapman and Hall, London.

Hitchcock, D., Casella, G., Booth, J., 2006. Improved estimation of dissimilarities by presmoothing functional data. Journal of the American Statistical Association 101 (473), 211–222.

Hochberg, Y., Tamhane, A. C., 1987. Multiple Comparison Procedures. Wiley.

Hoh, J., Wille, A., Ott, J., 2001. Trimming, weighting, and grouping snps in human case-control association studies. Genome Research 11, 2115–2119.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

Holodov, D. B., Nikolaevski, V. A., 2012. A method for preventing damages to the stomach mucous membrane when taking non-steroidal anti-inflammatory drugs. Patent RU 2449784.
URL http://www.findpatent.ru/patent/244/2449784.html

Li, J., Ji, L., 2005. Adjusting multiple testing in multilocus analysis using the eigenvalues of a correlation matrix. Heredity 95, 221–227.

Loughin, T., 2004. A systematic comparison of methods for combining p-values from independent tests. Computational Statistics & Data Analysis 47, 467–485.

Luo, L., Zhu, Y., M., X., 2012. Quantitative trait locus analysis for next-generation sequencing with the functional models. Journal of Medical Genetics 49, 513–524.

Marcus, R., Peritz, E., Gabriel, K. R., 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63 (3), 655–660.

Nasonov, E. L., Karateev, A. E., 2006. The use of non-steroidal anti-inflammatory drugs: clinical recommendations. Russian Medical Journal 14 (25), 1769–1777.

Nyholt, D., 2004. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. American Journal of Human Genetics 74, 765–769.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Wagner, H., 2011. vegan: Community Ecology Package. R package version 2.0-1.
URL http://CRAN.R-project.org/package=vegan

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Wagner, H., 2012. vegan: Community Ecology Package. R package version 2.0-5.
URL http://CRAN.R-project.org/package=vegan

Pesarin, F., 1992. A resampling procedure for nonparametric combination of several dependent tests. Statistical Methods & Applications 1 (1), 87–101.

Petrondas, D. A., Gabriel, K. R., 1983. Multiple comparisons by rerandomization tests. Journal of the American Statistical Association 78 (384), 949–957.

Pollard, K. S., Gilbert, H. N., Ge, Y., Taylor, S., Dudoit, S., 2011. multtest: Resampling-based multiple hypothesis testing. R package version 2.10.0.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org`

Ramsay, J., Silverman, B., 1997. Functional Data Analysis. Springer-Verlag, New York.

Ramsay, J. O., Hooker, G., Graves, S., 2009. Functional Data Analysis with R and MATLAB. Springer.

Ramsay, J. O., Silverman, B. W., 2005. Functional Data Analysis, Second Edition. Springer.

Ramsay, J. O., Wickham, H., Graves, S., Hooker, G., 2012. fda: Functional Data Analysis. R package version 2.3.2.
URL `http://CRAN.R-project.org/package=fda`

Rice, W., 1988. Analyzing tables of statistical tests. Evolution 43 (1), 223–225.

Roy, S. N., 1953. On a heuristic method of test construction and its use in multivariate analysis. The Annals of Mathematical Statistics 23 (220-238).

Shen, Q., Faraway, J., 2004. An F test for linear models with functional responses. Statistica Sinica 14, 1239–1257.

Smith, C., Cribbie, R., 2013. Multiplicity control in structural equation modeling: incorporating parameter dependencies. Structural Equation Modeling 20 (1), 79–85.

Troendle, J. F., Westfall, P. H., 2011. Permutational multiple testing adjustments with multivariate multiple group data. Journal of Statistical Planning and Inference 141, 2021–2029.

Šidák, Z., 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 78, 626–633.

Westfall, P., Troendle, J., 2008. Multiple testing with minimal assumptions. Biometrical Journal 50, 745–755.

Westfall, P. H., Young, S. S., 1993. Resampling-based Multiple Testing: Examples and Methods for P-Value Adjustment. Wiley.

Wood, S. N., 2011. mgcv: generalized additive model method. R package version 1.7-19.
URL `http://CRAN.R-project.org/package=mgcv`

Xu, H., Shen, Q., Yang, X., Shptaw, S., 2011. A quasi f-test for functional linear models with functional covariates and its application to longitudinal data. Statistics in Medicine 30 (23), 2842–2853.

Yu, K., Li, Q., Bergen, A., Pfeiffer, R., Rosenberg, P., Caporasi, N., Kraft, P., Chatterjee, N., 2009. Pathway analysis by adaptive combination of p-values. Genetic Epidemiology 33, 700–709.

Zaykin, D. V., 2000. Statistical analysis of genetic associations. Ph.D. thesis. North Carolina State University.

Zaykin, D. V., 2011. Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis. Journal of Evolutionary Biology 24 (8), 1836–1841.

Zaykin, D. V., Zhivotovsky, L. A., Czika, W., Shao, S., Wolfinger, R. D., 2007. Combining p-values in large-scale genomics experiments. Pharmaceutical Statistics 6 (3), 217–226.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., Weir, B. S., 2002. Truncated product method for combining p-values. Genetic Epidemiology 22 (2), 170–185.

Zhang, C., Peng, H., Zhang, J., 2010. Two sample tests for functional data. Communications in Statistics – Theory and Methods 39 (4), 559–578.