Computational Statistics and Data Analysis 67 (2013) 175–184

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Combining functions and the closure principle for performing follow-up tests in functional analysis of variance

O.A. Vsevolozhskaya [a,*], M.C. Greenwood [a], G.J. Bellante [b], S.L. Powell [b], R.L. Lawrence [b], K.S. Repasky [c]

[a] Mathematical Sciences, Montana State University, Bozeman, MT 59717, United States

[b] Land Resources and Environmental Sciences, Montana State University, Bozeman, MT 59717, United States

[c] Electrical and Computer Engineering, Montana State University, Bozeman, MT 59717, United States

ARTICLE INFO

Article history:
Received 16 May 2012
Received in revised form 8 May 2013
Accepted 8 May 2013
Available online 17 May 2013

Keywords:
Functional data analysis
Multiple comparison procedure
Permutation method
Distance-based method

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 curves. 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

$$y_{ij}(t) = \mu_i(t) + \epsilon_{ij}(t), \tag{1}$$

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 Development Core Team, 2012) in the `fda` package (Ramsay et al., 2011).

* Corresponding author. Tel.: +1 406 994 1962.
E-mail addresses: vsevolozhskaya@gmail.com, vsevoloz@math.montana.edu (O.A. Vsevolozhskaya), greenwood@math.montana.edu (M.C. Greenwood), gabrielbellante@fs.fed.us (G.J. Bellante), spowell@montana.edu (S.L. Powell), rickl@montana.edu (R.L. Lawrence), repasky@ece.montana.edu (K.S. Repasky).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.05.005

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.

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

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

$$F(t_l) = \frac{\left[\sum_{ij}(y_{ij}(t_l) - \hat{\mu}(t_l))^2 - \sum_{ij}(y_{ij}(t_l) - \hat{\mu}_i(t_l))^2\right] \Big/ (k-1)}{\sum_{ij}(y_{ij}(t_l) - \hat{\mu}_i(t_l))^2/(n-k)},$$
$$= MS_T(t_l)/MS_E(t_l). \tag{2}$$

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). 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 correcting for multiple testing using the Westfall–Young multiplicity correction method (Westfall and Young, 1993). This provides point-wise 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 significance anywhere within the domain of the functional response. In particular, Shen and Faraway (2004) proposed a functional F-ratio

$$\mathcal{F} = \frac{\left[\sum_{ij}\int_{t_{\min}}^{t_{\max}}(y_{ij}(t) - \hat{\mu}(t))^2 dt - \sum_{ij}\int_{t_{\min}}^{t_{\max}}(y_{ij}(t) - \hat{\mu}_i(t))^2 dt\right] \Big/ (k-1)}{\sum_{ij}\int_{t_{\min}}^{t_{\max}}(y_{ij}(t) - \hat{\mu}_i(t))^2 dt/(n-k)}$$
$$= \frac{\sum_i n_i \int_{t_{\min}}^{t_{\max}}(\hat{\mu}_i(t) - \hat{\mu}(t))^2 dt/(k-1)}{\sum_{ij}\int_{t_{\min}}^{t_{\max}}(y_{ij}(t) - \hat{\mu}_i(t))^2 dt/(n-k)}, \tag{3}$$

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 difference 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) 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 cannot 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 for significant results. 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 re-sampling 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, $p_{(1)} \leq \cdots \leq p_{(L)}$. 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 permutation 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 Fig. 1. A rejection of H_1 requires rejection of all intersection hypotheses that include H_1, which are highlighted in Fig. 1. 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 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

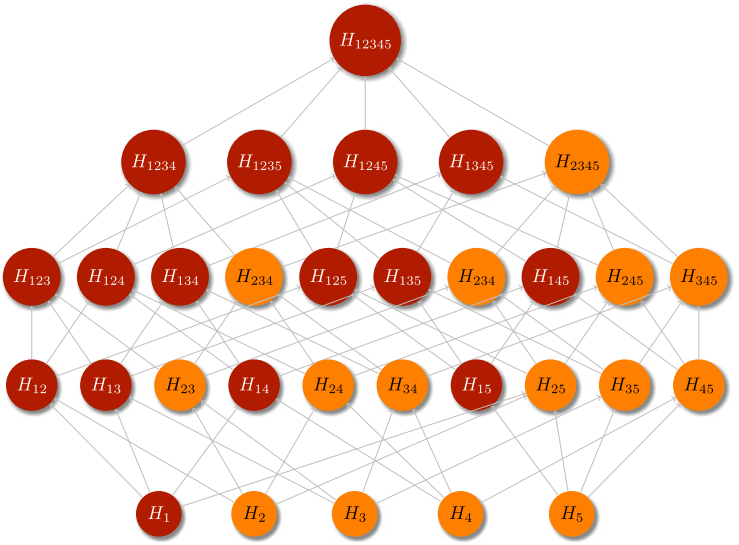

Fig. 1. Closure set for five elementary hypotheses $H_1, \ldots, H_5$ and their intersections.

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), 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 of significant difference. 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 Fig. 2(a)). 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 Fig. 2(b)). Based on these two possibilities and two correction methods, we considered follow-up tests for the following four combinations:

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.

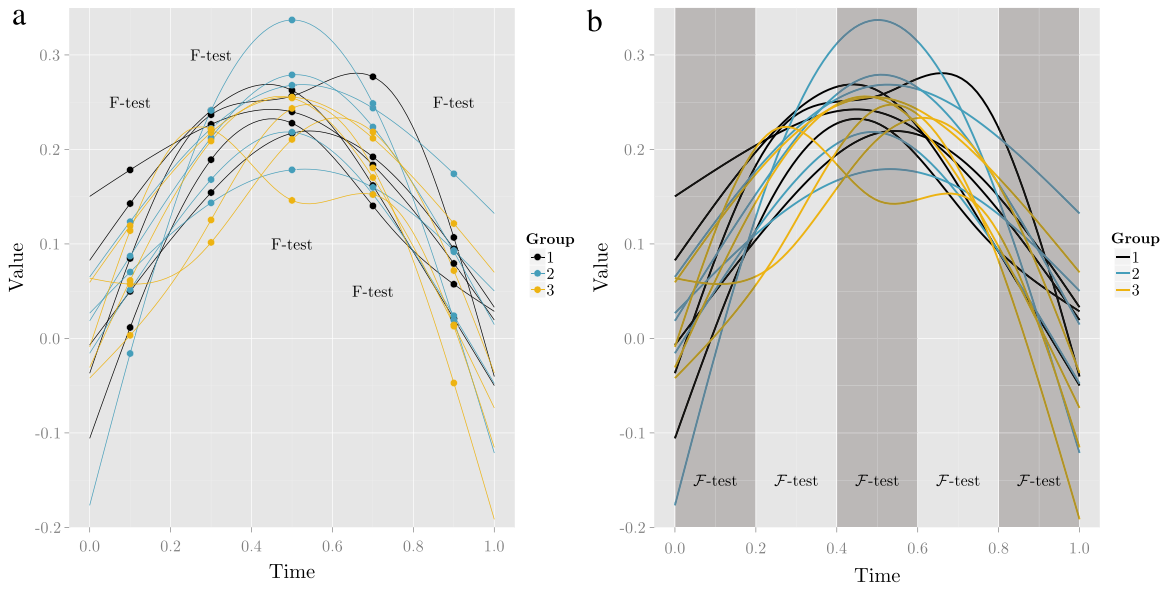

Fig. 2. Two follow-up testing methods illustrated on simulated data with three groups, five curves per group, and five evaluation points or regions.

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 Eq. (3)

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

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

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

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 the 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

Table 1
Estimates of the Type I error (±margin of error) control in the weak sense for $\alpha = 0.05$.

Method	5 intervals/evaluations	10 intervals/evaluations
Region-based/closure	**0.020 ± 0.009**	**0.008 ± 0.006**
Point-wise/closure	**0.028 ± 0.010**	**0.008 ± 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

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 $y_{3j}(t) = t(1-t) + p + \epsilon_{3j}(t)$ and $p = 0.03, 0.06, 0.09$, 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. Then the estimated power is computed at each subregion as

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

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 kth 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 1 and 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 95% confidence intervals of the estimates have been calculated based on the normal approximation of the binomial distribution.

Table 2
Estimates of the Type I error ($\pm ME$) control in the strong sense for $\alpha = 0.05$.

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	$\mathbf{0.071 \pm 0.016}$

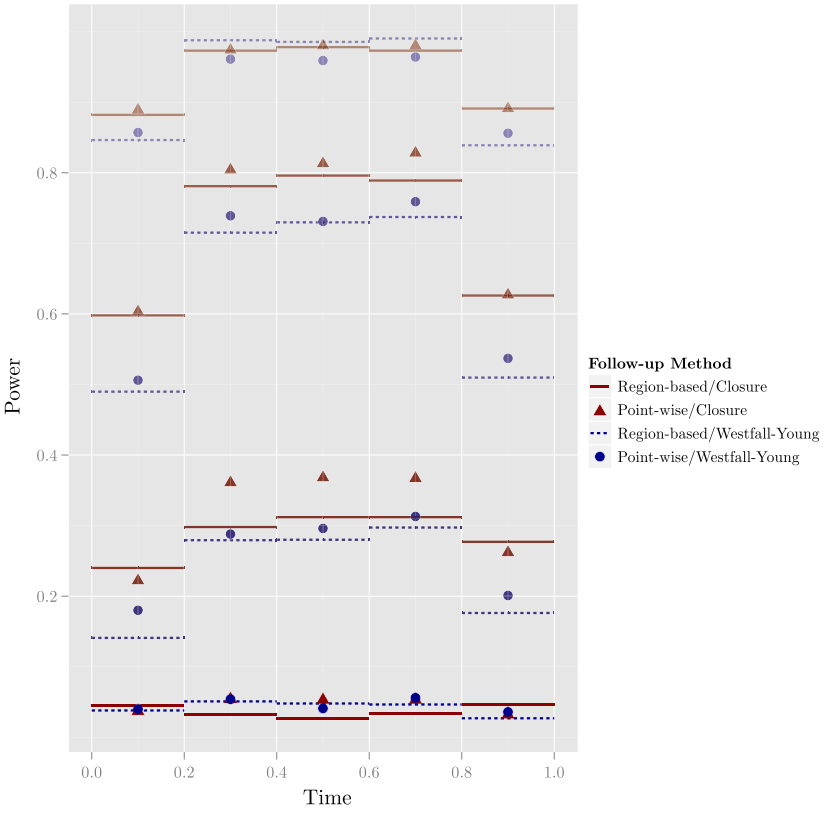

Fig. 3. Power of the four methods at different values of the shift amount. The solid objects in the lower graph correspond to $p = 0.03$. The three groups of objects above that correspond to $p = 0.06$, 0.09, and 0.12 respectively.

Table 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 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 Fig. 3 and for 10 intervals/evaluation points in Fig. 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, p, is pretty apparent but a transparency effect is added to aid visualization. The most solid objects (lower graph) correspond to a shift of $p = 0.03$, and the most transparent objects (upper graph) to $p = 0.12$.

From Fig. 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 ($p = 0.06$ and $p = 0.09$) than the Westfall–Young method. There does not seem to be any striking visual difference in the power of the four methods for the

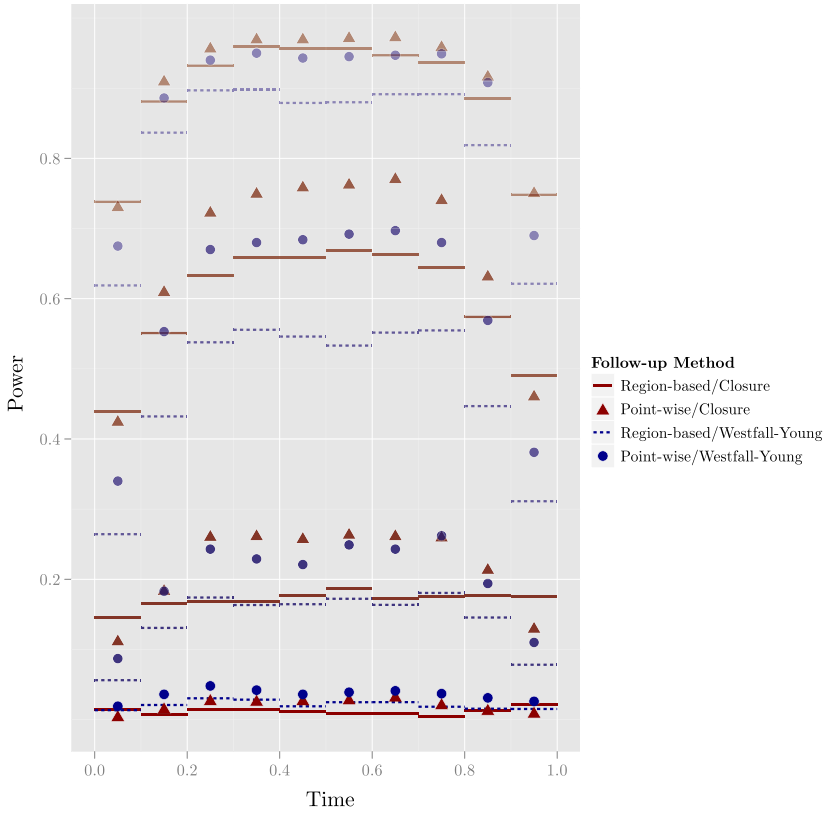

Fig. 4. Power of the four methods with 10 intervals/evaluation points.

lowest and highest shift amount ($p = 0.03$ and $p = 0.12$). Although the powers were very close at the extreme values of p, it appears that the closure multiplicity correction provides higher overall power across different values of p while maintaining its conservative nature under the global null. Similar conclusions can be drawn based on Fig. 4.

A contrast of Fig. 4 to Fig. 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 Figs. 3 and 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

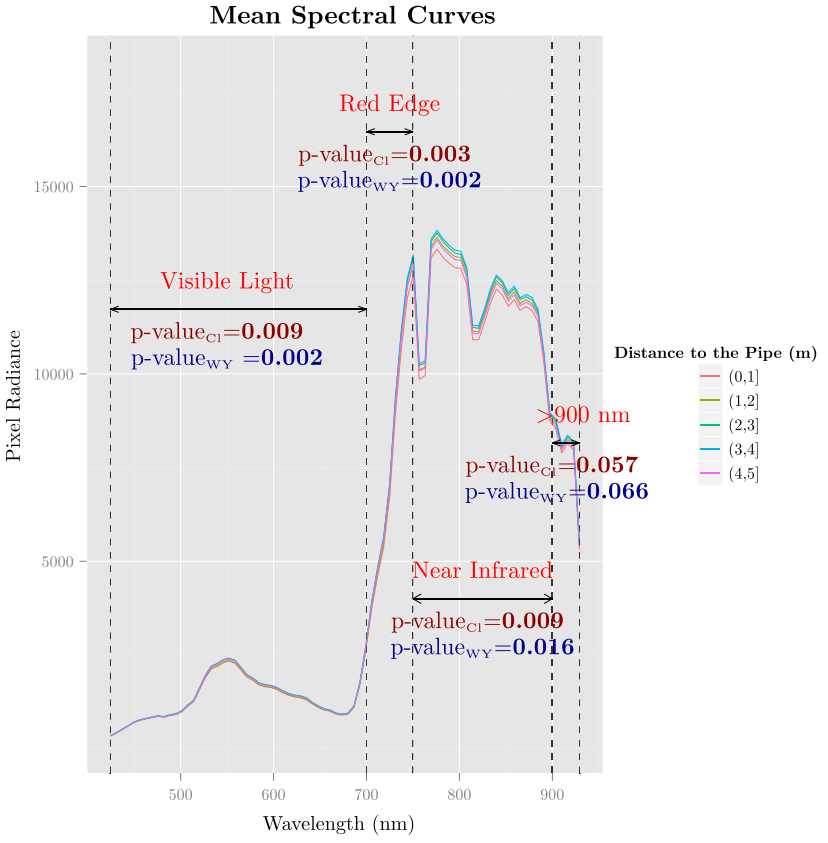

Fig. 5. Plot of mean spectral curves at each of the five binned distances to the CO_2 release pipe. p-value$_{WY}$ represents a p-value obtained by a combination of the regionalized testing method with the Westfall–Young multiplicity correction. p-value$_{Cl}$ represents a p-value obtained by the regionalized method with the closure multiplicity adjustment.

these methods. Further details may be found in Bellante et al. (2013). One of the goals of the experiment was to investigatehyperspectral 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] m 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 significant differences occurred. In particular, we were interested in whether there were significant differences in the visible (about 400–700 nm), "red edge" (about 700–750 nm), and near infrared (about 750–900 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 Fig. 5.

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.

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 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.

Acknowledgments

The authors would like to thank the two anonymous referees for their valuable feedback on the manuscript. Additional feedback from Megan Higgs and Jim Robison-Cox was very helpful. This work was carried out within the ZERT II project, with the support of the U.S. Department of Energy and the National Energy Technology Laboratory, under Award No. DE-FE0000397. However, any opinions, findings, conclusions, or recommendations expressed herein are those of the author(s) and do not necessarily reflect the views of the DOE.

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.G., Powell, S.L., Lawrence, R.L., Repasky, K., Dougher, T., 2013. Aerial detection of a simulated CO_2 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 Development Core Team,, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0, 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 ed. Springer.
Ramsay, J.O., Wickham, H., Graves, S., Hooker, G., 2011. fda: Functional Data Analysis. R package version 2.2.7. 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.