Letter to the Editor on: "Comparing Groups of Time Dependent Data Using Locally Weighted Scatterplot Smoothing Alpha-Adjusted Serial T-tests" by Niiler (2020)

Dominik Liebl*

University Bonn

1. Introduction

The main contribution in Niiler (2020) is a new α level correction method for multiple testing in a functional data context. A secondary contribution is the proposal to use the nonparametric LOESS smoothing method of Cleveland & Devlin (1988) to overcome problems with irregularly sampled functional data. Unfortunately, both contributions generally lead to invalid statistical inferences and I will explain these issues in the following.

In Section 2, I briefly explain the α level correction method of Niiler (2020) and introduce necessary notation. In Section 3, I explain the statistical shortcomings in Niiler (2020) and demonstrate that the α level correction of Niiler (2020) leads to invalid inferences. A short conclusion is given in Section 4.

2. The α level correction method of Niiler (2020)

Niiler (2020) considers the statistical problem of testing for differences between mean functions of two groups (a and b) using two independent samples of biomechanical functional curve data. Exemplary data are shown, for instance, in Fig 1 in Niiler (2020). The testing is done using a series of M many two-sample test statistics

$$t(g_j) = \frac{\hat{\mu}_a(g_j) - \hat{\mu}_b(g_j)}{\sqrt{\frac{\hat{\sigma}_a^2(g_j)}{N_a} + \frac{\hat{\sigma}_b^2(g_j)}{N_b}}}, \quad j = 1, \dots, M, \tag{1}$$

where each test statistic $t(g_j)$ conducts a statistical hypothesis test specific to a grid point g_j with, for instance, $0\% \leq g_1 < \cdots < g_j < \cdots < g_M \leq 100\%$ when using a standardized time (gait cycle) domain $[0\%, 100\%]$. The estimates $\hat{\mu}_a(g_j)$, $\hat{\mu}_b(g_j)$, $\hat{\sigma}_a^2(g_j)$, and $\hat{\sigma}_b^2(g_j)$ denote the mean and variance estimates of groups a and b at grid point g_j, and N_a and N_b denote the samples sizes. If one uses the classic sample mean and variance estimates, the test statistic in (1) becomes the classic Welch's t-test. Niiler (2020),

*Corresponding author: Dominik Liebl, University Bonn, Institute of Finance and Statistics, Adenauerallee 24-26, 53113 Bonn (Germany), dliebl@uni-bonn.de. This research was supported by the Hausdorff Center of Mathematics (HCM), a Cluster of Excellence at the University financed by the German Research Foundation (DFG).

however, suggests using the mean and variance estimates as computed by the R function `loess` for local polynomial regression (Cleveland & Devlin, 1988; R Core Team, 2021).

The jth test statistic, $t(g_j)$, tests the null hypothesis, H_0, of equal means against the two-sided alternative, H_1:

$$\begin{aligned} \mathrm{H}_0(g_j)&: \quad \mu_a(g_j) = \mu_b(g_j) \\ \mathrm{H}_1(g_j)&: \quad \mu_a(g_j) \neq \mu_b(g_j) \end{aligned}$$

However, one is generally not interested in the test decision at a single grid point, g_j, but one uses the whole family of M many test statistics $\{t(g_1), \ldots, t(g_M)\}$ to find regions in $[0\%, 100\%]$ over which the mean functions are statistically different from each other. This leads to a severe multiple testing problem since the number of tests, M, can be arbitrarily large.

Statistical multiple testing procedures must control the *family-wise* type I error rate. I.e., the probability of observing type I errors in at least one of the tests $\{t(g_1), \ldots, t(g_M)\}$ must be bounded from above by the pre-chosen significance level α,

$$P_{\mathrm{H}_0}\big(\text{reject } \mathrm{H}_0(g_j) \text{ for at least one } j \in \{1, \ldots, M\}\big) \leq \alpha \tag{2}$$

with, for instance, $\alpha = 0.05$.

To control the family-wise type I error rate, one could use, for instance, the classic Bonferroni correction, where each point-wise null hypothesis $\mathrm{H}_0(g_j)$ is tested at the reduced significance level of $\alpha' = \alpha/M$. However, for biomechanical curve data, Bonferroni corrections can result in unnecessarily conservative (low power) testing procedures since biomechanical curve data are typically relatively smooth and, therefore, the point-wise test statistics $\{t(g_1), \ldots, t(g_M)\}$ are typically strongly correlated with each other. These correlations are ignored by the Bonferroni correction. Similar issues arise with other standard α level corrections such as, for instance, the Holm-Bonferroni or the Hochberg correction.

The main contribution in Niiler (2020) is the proposal of a new, less conservative α level correction which tries to take into account the correlations between the test statistics $\{t(g_1), \ldots, t(g_M)\}$. The proposed correction is given in equation (2) in Niiler (2020), but also presented here for convenience:

$$\alpha' = \frac{\alpha}{M(1 - \hat{\rho}^2) + \hat{\rho}^2}, \tag{3}$$

where Niiler (2020) sets $\alpha = 0.05$ and where $\hat{\rho}$ denotes the sample auto-correlation coefficient between adjacent test statistics $t(g_j)$ and $t(g_{j+1})$

$$\hat{\rho} = \frac{(M-1)^{-1} \sum_{j=1}^{M-1} (t(g_j) - \bar{t}\,)(t(g_{j+1}) - \bar{t}\,)}{M^{-1} \sum_{j=1}^{M} (t(g_j) - \bar{t}\,)^2}$$

with $\bar{t} = M^{-1} \sum_{j=1}^{M} t(g_j)$. Niiler (2020) motivates his proposal as follows: While the case of perfect auto-correlation $\hat{\rho} = 1$ leads to no α level correction ($\alpha' = \alpha$), the case of no auto-correlation $\hat{\rho} = 0$ leads to the Bonferroni correction ($\alpha' = \alpha/M$).

3. Main critique: Invalid α level correction

The auto-correlation coefficient $\hat{\rho}$ is not a meaningful statistic in case of non-stationary time series.[1] But even if $t(g_1), \ldots, t(g_M)$ were a stationary series of test statistics, $\hat{\rho}$ would only measure the correlation between adjacent test statistics $t(g_j)$ and $t(g_{j+1})$. All the other pair-wise correlations are not considered. Therefore, one cannot expect that this method is able to control the family-wise type I error rate – except for trivial and practically irrelevant special cases. Indeed, Niiler (2020) does not provide any theoretical justification for his α correction method in (3), and, as far as I know, there is no similar correction in the statistical literature.

Niiler (2020) uses a Monte Carlo simulation study to demonstrate that his α level correction is able to control the family-wise type I error rate; see Appendix C of the supplementary material of Niiler (2020). Simulations, however, cannot replace theoretical considerations since the considered simulation scenarios may only reflect non-generalizing special cases – which is exactly what happened in Niiler (2020).

The only reason, why Niiler (2020) was able to demonstrate that his α level correction is able to control the family-wise type I error rate, is the very specific choice of his simulation study. Niiler (2020) considers the following overly simple type of random functions

$$\sin(x) + Z, \quad \text{with} \quad x \in [0, 2\pi] \quad \text{and} \quad Z \sim \mathcal{N}(0,1) \tag{4}$$

which are just random vertical shifts of a deterministic sinus function; see Figure 1 (A) and (B).

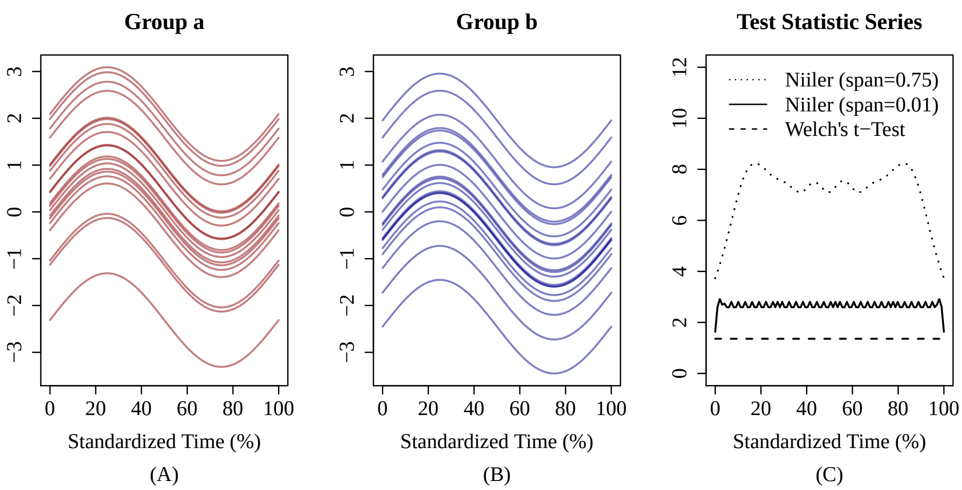

Figure 1: Plots (A) and (C): Exemplary functions from the sinus-shift random process in (4) as used in the Monte Carlo simulation study in Niiler (2020) for checking the familiy-wise type I error rate. Plot (C): test statistic series $t(g_1), \ldots, t(g_M)$ computed using the LOESS smoothing method suggested by Niiler (2020) with smoothing parameters spar $= 0.01$ and spar $= 0.75$. Additionally, the classic Welch's t-test statistics series is shown.

For this special case, the Welch's t-test statistics $t(g_1), \ldots, t(g_M)$ are all exactly equal to each other (see Fig. 1 (C)) which demonstrates that

[1] Biomechanical curve data are usually not stationary.

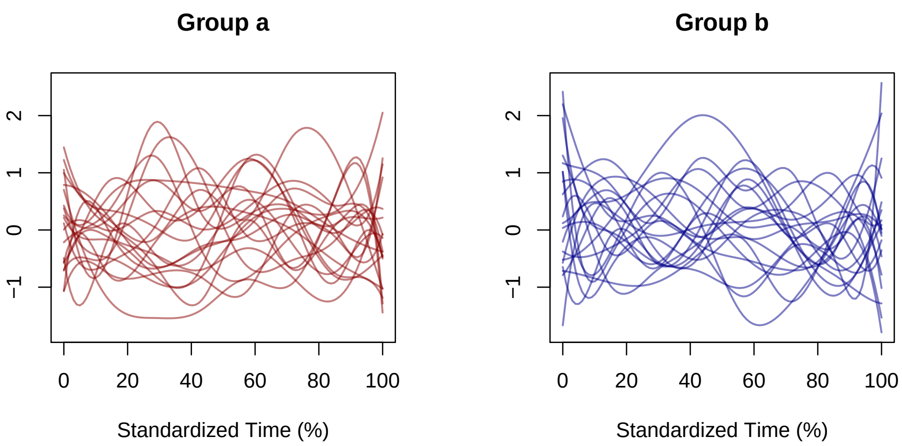

Figure 2: Exemplary functions from the B-splines random process in (5) as used in my Monte Carlo simulation for checking the familiy-wise type I error rate.

there is no multiple testing problem in this special case. All t-tests are either *simultaneously* significant or not, and, therefore, the family-wise type I error rate coincides with the type I error rate of a single t-test which makes the stochastic process in (4) unsuitable for checking α level correction methods.

In the following, I consider a practically more relevant situation by drawing random functions from

$$f(t) = \sum_{k=1}^{10} Z_k B_{k,10}(t), \quad \text{with} \quad t \in [0\%, 100\%], \tag{5}$$

where $Z_k \sim \mathcal{N}(0,1)$ and where $B_{k,10}(t)$ denotes the kth cubic B-spline function based on equidistant knots in $[0, 100]$; see Figure 2. B-spline functions have compact supports which guarantees that the random functions in (5) consist of independent as well as dependent stochastic components making this case suitable for checking α level correction methods.

To check the family-wise type I error rate of the α level correction in Niiler (2020), I simulte N_a and N_b many functions from (5). Both groups a and b have the same population mean (zero) such that all M null hypotheses $\mathrm{H}_0(g_j)$: $\mu_a(g_j) = \mu_b(g_j)$, $j = 1, \dots, M$, are fulfilled. Under this scenario, the family-wise type I error rate must be smaller or equal to the pre-selected significance level $\alpha = 0.05$; otherwise, the α correction method is invalid and cannot be used in practice.

Table 1 shows my simulation results based on 10,000 Monte Carlo replications. To check the effect of different choices for the number of sampling grid points M, I consider the values $M \in \{50, 75, 100\}$. To check the effect of different sample sizes, I consider the values $N_a = N_b \in \{10, 20, 50\}$. The LOESS smoother used by Niiler (2020) involves setting a smoothing parameter, where I use a small smoothing parameter span $= 0.01$ and a relatively large smoothing parameter span $= 0.75$. I compare the inference method of Niiler (2020) with a Bonferroni adjusted series of Welch's t-tests and with the random field theory based method SPM1d (Pataky, 2016). For the latter method, I use the R package `ffscb` which contains

SPM1d based bands, but also more general simultaneous confidence bands as proposed by Liebl & Reimherr (2020).

Table 1: False positive (type I error) rates under the null-hypothesis for a pre-selected significance level $\alpha = 0.05$

		Niiler (2020)		t-test	SPM1d
$N_a = N_b$	M	(span= 0.01)	(span= 0.75)	(Bonferroni)	
10	50	0.10	0.82	0.00	0.05
10	75	0.60	0.92	0.00	0.05
10	100	0.81	0.97	0.00	0.06
20	50	0.19	0.88	0.01	0.05
20	75	0.73	0.95	0.00	0.05
20	100	0.88	0.98	0.00	0.05
50	50	0.25	0.90	0.01	0.05
50	75	0.77	0.96	0.01	0.05
50	100	0.91	0.98	0.00	0.05

The only two methods that are able to control the type I error rate in this simulation study are the point-wise t-tests with Bonferroni correction and the random fields theory based method SPM1d. While the Bonferroni correction is overly conservative, SPM1d is able to exploit the significance level $\alpha = 0.05$ very well. The inference procedure proposed by Niiler (2020) fails to control the type I error rate severely and, therefore, leads to invalid inferences.

Further issues. Niiler (2020) misses several further issues. For instance, the standard errors computed by the `loess` function in R are invalid for smoothing functional data when M becomes large (see Liebl, 2019). Moreover, nonparametric smoothing methods like LOESS have biased estimates and can suffer from boundary problems both leads to distorted estimation results as indicated in Figure 2 (C). To get valid inference in finite samples, one would need bias and boundary corrections – both are not considered in Niiler (2020).

4. Conclusion

The development of simultaneous inference methods for functional data is an active research field in statistics (see Degras (2011), Cao et al. (2012), Wang et al. (2020), Pini & Vantini (2017), Choi & Reimherr (2018), Liebl & Reimherr (2020), and many others). New steps forward in this literature are often published in the most prestigious statistical science journals. In my humble opinion, the work of Niiler (2020) fails to make a contribution to this literature.

Every statistical testing procedure can lead to invalid inferences when misapplied. However, this is different here. The proposed α level testing procedure of Niiler (2020) will lead to false inferences even when "correctly" applied. This is a serious issue since the method is already applied in the literature (see Shoja et al., 2020).

References

Cao, G., Yang, L., & Todem, D. (2012). Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics*, *24*, 359–377.

Choi, H., & Reimherr, M. (2018). A geometric approach to confidence regions and bands for functional parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*, 239–260.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, *83*, 596–610.

Degras, D. A. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*, *21*, 1735–1765.

Liebl, D. (2019). Inference for sparse and dense functional data with covariate adjustments. *Journal of Multivariate Analysis*, *170*, 315–335.

Liebl, D., & Reimherr, M. (2020). Fast and fair simultaneous confidence bands for functional parameters. *Preprint arXiv:1910.00131*, (pp. 1–37).

Niiler, T. (2020). Comparing groups of time dependent data using locally weighted scatterplot smoothing alpha-adjusted serial t-tests. *Gait & Posture*, *76*, 58–63.

Pataky, T. C. (2016). Rft1d: Smooth one-dimensional random field upcrossing probabilities in python. *Journal of Statistical Software*, *71*, 1–22.

Pini, A., & Vantini, S. (2017). Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, *29*, 407–424.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: `https://www.R-project.org/`.

Shoja, O., Farsi, A., Towhidkhah, F., Feldman, A. G., Abdoli, B., & Bahramian, A. (2020). Visual deprivation is met with active changes in ground reaction forces to minimize worsening balance and stability during walking. *Experimental Brain Research*, *238*, 369–379.

Wang, Y., Wang, G., Wang, L., & Ogden, R. T. (2020). Simultaneous confidence corridors for mean functions in functional data analysis of imaging data. *Biometrics*, *76*, 427–437.