REVSTAT – Statistical Journal
Supplementary file

Supplementary — Review of Methods for Functional One-Way Analysis of Variance

Authors: Łukasz Smaga ✉
– Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland
ls@amu.edu.pl

In this supplement, we present the simulation studies based on real data examples, which show a small comparison of the tests considered and a way to do this for a particular data set.

1. SIMULATION STUDIES BASED ON REAL DATA EXAMPLES

In the sections of the main paper, we have commented about the properties of the tests considered, in particular finite-sample properties established in simulation studies. Such studies are important as they give some practical recommendations about the tests. They describe the general properties of the tests and particular cases under which the tests have for example good power. However, it is not always easy to find such scenarios or verify them for real data sets. Therefore, in this section, we present the simulation studies based on real data sets, which we analyze. Such studies give information about the finite-sample properties of the tests but are also closely related to the data set by mimicking its properties, e.g., the expected value and variance of the variable considered. For the ANOVA for functional repeated measurements of Section 3.3 of the main paper, the simulation study based on the DTI data set was conducted in Kuryło and Smaga (2024). In the remainder of this section, we present such simulation studies for the other types of FANOVA.

We study the type I error control and power of the tests. We generate 1000 simulation samples to evaluate the performance of the tests and employ 1000 bootstrap or permutation samples for the bootstrap and permutation tests respectively. The remaining test hyperparameters were selected in the same way as in the previous sections. We set a significance level $\alpha = 0.05$. The experiments were conducted in the R programming language (R Core Team, 2024).

✉ Corresponding author

1.1. Two sample problem

In Section 2 of the main paper, the two-sample test by Horváth et al. (2014) was illustrated using the Canadian weather data. We compared the following pairs of regions separately: Eastern vs. Western, Eastern vs. Northern, and Western vs. Northern. To mimic these data, we generated the simulation data as follows:

- we had two samples with $(n_1, n_2) = \delta \cdot (15, 15), \delta \cdot (15, 5), \delta \cdot (15, 5)$ observations for comparisons Eastern vs. Western, Eastern vs. Northern, and Western vs. Northern respectively, where $\delta = 1, 2, 3, 4$;
- we used samples of functional observations at 365 design time points;
- for each functional observation, the covariance function $\gamma(s, t)$ was equal to the sample covariance function for the pooled sample;
- for checking the type I error control, in each group, the mean function was equal to the sample mean function for the pooled sample;
- for the power study, the mean function in the first and second group was equal to the sample mean function for the first and second sample respectively from the data set.

For given sample sizes n_1 and n_2, the 365-dimensional data were generated from the normal distribution having expected value and covariance matrix equal to the above sample mean functions and sample covariance function respectively.

The simulation results are presented in Table 1. First of all, we observe that for the sample sizes from the data set ($\delta = 1$), the test is too liberal. Fortunately, when sample sizes increase, the empirical sizes decrease and for $\delta = 3$, the test controls the type I error level in all cases. On the other hand, the power increases with the increase in the number of observations. However, for Eastern-Western comparison, where there was no significant difference, the power is quite small. In the other two comparisons, the power is close the 100%, which confirms the rejections of null hypotheses in these cases. These observations are in line with the properties noted in Section 2 of the main paper, i.e., the test controls the type I error level for moderate and large sample sizes. Returning to the original sample sizes ($\delta = 1$), though the test gives sensible decisions in this case, it is too liberal, which could result in inadequate conclusions. Then, the sample size should be increased or we could consider another test, for example, some of the FANOVA tests of the next section.

1.2. ANOVA for independent functional samples

In Section 3.1 of the main paper, we considered many tests for functional ANOVA for independent functional samples. Here, we compare their properties for the Canadian weather data. Namely, we consider the following test procedures:

Hypothesis	δ	Eastern-Western	Eastern-Northern	Western-Northern
H_0	1	8.7	10.0	10.4
	2	5.6	7.0	8.4
	3	5.0	6.4	5.2
	4	5.1	5.8	6.1
H_1	1	18.9	98.5	96.7
	2	20.7	100.0	100.0
	3	25.0	100.0	100.0
	4	27.5	100.0	100.0

Table 1: Empirical sizes and powers (as percentages) of the tests obtained in the simulation study in Section 1.1.

- FP test (Section 3.1.1 of the main paper),
- tests based on random projections with the Gaussian white noise with $k = 10, 20, 30$ projections and three tests applied to random projections: standard ANOVA, ANOVA-type test, WTPS (Section 3.1.2 of the main paper, denote these tests as G10A, G10At, G10W respectively for $k = 10$),
- tests based on random projections with the Brownian motion with $k = 10, 20, 30$ projections and three tests applied to random projections: standard ANOVA, ANOVA-type test, WTPS (Section 3.1.2 of the main paper, denote these tests as B10A, B10At, B10W respectively for $k = 10$),
- L^2-norm-based tests L^2N, L^2B, L^2b, F-type tests FN, FB, Fb, GPF test, and $F_{\max}$ tests (Section 3.1.3 of the main paper),
- graphical functional ANOVA test GFA (Section 3.1.4 of the main paper).

The simulation setup is similar to that in Section 1.1 with the following differences:

- we had three samples with sizes $n_1 = 15, n_2 = 15, n_3 = 5$;
- for the power study, the mean function in the i-th group was equal to the sample mean function for the i-th sample, $i = 1, 2, 3$.

Table 2 presents the simulation results. Most of the tests control the type I error level very well. The exemptions are the tests G10At, G20At, G30At, B10At, L^2N, L^2B, and L^2b, which are at least slighty too liberal. Most of the tests have also similar power, which is always above 92%. The projection-based tests with WTPS have smaller power than the other tests, but their power is still very large. To sum up, most of the tests have good finite sample properties, which confirms the statistical decisions they made.

Hypothesis	Tests and results								
H_0	FP								
	5.9								
	G10A	G20A	G30A	G10At	G20At	G30At	G10W	G20W	G30W
	3.4	3.0	3.0	7.1	7.6	7.0	3.9	4.0	3.6
	B10A	B20A	B30A	B10At	B20At	B30At	B10W	B20W	B30W
	4.2	3.8	3.8	6.7	6.4	5.7	4.2	3.3	3.7
	L^2N	L^2B	L^2b	FN	FB	Fb	GPF	$F_{\max}$	
	7.7	8.4	8.3	6.0	6.1	5.0	6.3	5.3	
	GFA								
	5.3								
H_1	FP								
	97.8								
	G10A	G20A	G30A	G10At	G20At	G30At	G10W	G20W	G30W
	97.2	98.1	97.6	95.6	95.8	97.1	93.1	92.5	92.9
	B10A	B20A	B30A	B10At	B20At	B30At	B10W	B20W	B30W
	98.2	98.2	98.1	96.4	96.7	96.6	92.9	93.1	92.9
	L^2N	L^2B	L^2b	FN	FB	Fb	GPF	$F_{\max}$	
	98.7	99.2	98.3	97.7	97.8	95.8	99.1	98.9	
	GFA								
	97.6								

Table 2: Empirical sizes and powers (as percentages) of the tests obtained in the simulation study in Section 1.2.

1.3. ANOVA for partially observed functional data

In Section 3.2 of the main paper, we considered the following three tests for ANOVA for partially observed functional data: the T_{L^2}, T_Q, and T_Q^b tests. The last one is the bootstrap version of the T_Q test. Let us now check their finite sample properties for the generated partially observed Canadian weather data. The simulation data were first generated in the same way as in Section 1.2, and later we randomly removed a part of these data as we did in the illustrative example in Section 3.2 of the main paper. Additionally, we were increasing the sample sizes, i.e., $(n_1, n_2, n_3) = \delta \cdot (15, 15, 5)$, where $\delta = 1, \ldots, 7$.

Table 3 contains the simulation results. The T_{L^2} test controls the type I error very well and has large power, even for the original data case ($\delta = 1$). On the other hand, the T_Q and T_Q^b tests need a larger sample size to appropriately control the type I error. Unfortunately, the empirical sizes of these tests are acceptable for about $\delta = 6, 7$ times the original sample sizes. Nevertheless, their powers are also very large.

Hypothesis	δ	T_{L^2}	T_Q	T_Q^b
H_0	1	5.3	17.6	12.2
	2	5.5	13.2	11.1
	3	5.6	11.1	10.3
	4	4.4	8.2	7.7
	5	5.8	8.9	8.0
	6	4.7	6.8	6.8
	7	5.2	6.4	6.2
H_1	1	83.8	92.4	86.8
	2	99.9	99.5	99.1
	3	100.0	100.0	100.0
	4	100.0	100.0	100.0
	5	100.0	100.0	100.0
	6	100.0	100.0	100.0
	7	100.0	100.0	100.0

Table 3: Empirical sizes and powers (as percentages) of the tests obtained in the simulation study in Section 1.3.

1.4. Multivariate analysis of variance

Finally, we consider the following tests for functional multivariate analysis of variance of Section 4 of the main paper:

- permutation tests based on a basis function representation and test statistics by Wilks, Lawley-Hotelling, Pillai, and Roy - denoted by W, LH, P, and R,
- tests based on random projections with the Gaussian white noise with $k = 10, 20, 30$ projections and four tests applied to random projections, i.e., these mentioned above - denoted by G10W, G10LH, G10P, and G10R respectively for $k = 10$,
- tests based on random projections with the Brownian motion with $k = 10, 20, 30$ projections and four tests applied to random projections, i.e., these mentioned above - denoted by B10W, B10LH, B10P, and B10R respectively for $k = 10$.

The simulation data are generated similarly to Section 1.2, but of course, we take into account both weather variables, i.e., temperature and precipitation.

The simulation results are presented in Table 4. We can observe that almost all tests control the type I error. The exemptions are the tests based on random projection with Roy's test, which are very liberal. The other projection-based tests can have a conservative character. Nevertheless, their power is large and comparable with the power of the tests based on a basis function representation. Therefore, all tests, except G10R, G20R, G30R, B10R, B20R, and B30R, are appropriate for testing the differences in the considered data set.

Hypothesis	Tests and results					
H_0	W	LH	P	R		
	4.7	4.8	4.7	5.2		
	G10W	G20W	G30W	G10LH	G20LH	G30LH
	3.4	3.7	3.3	4.8	4.4	4.7
	G10P	G20P	G30P	G10R	G20R	G30R
	2.5	3.1	2.1	17.1	14.8	14.5
	B10W	B20W	B30W	B10LH	B20LH	B30LH
	3.1	2.9	2.3	3.7	3.8	3.0
	B10P	B20P	B30P	B10R	B20R	B30R
	2.5	2.1	2.1	11.9	10.7	10.8
H_1	W	LH	P	R		
	98.3	98.3	98.6	97.0		
	G10W	G20W	G30W	G10LH	G20LH	G30LH
	99.1	99.2	99.0	99.0	99.3	99.2
	G10P	G20P	G30P	G10R	G20R	G30R
	98.8	98.9	98.8	99.7	99.9	100.0
	B10W	B20W	B30W	B10LH	B20LH	B30LH
	97.5	97.8	98.0	98.0	97.9	98.1
	B10P	B20P	B30P	B10R	B20R	B30R
	97.5	97.3	97.6	99.2	99.5	99.2

Table 4: Empirical sizes and powers (as percentages) of the tests obtained in the simulation study in Section 1.4.

REFERENCES

Horváth, L., Kokoszka, P., and Rice, G. (2014). Testing stationarity of functional time series. *Journal of Econometrics*, 179(1):66–82.

Kuryło, K. and Smaga, Ł. (2024). Functional repeated measures analysis of variance and its application. *Statistics in Transition new series*, 25:185–204.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.