Premio Ramiro Melendreras

Información del trabajo presentado

TÍTULO:	Basis expansion approaches for functional analysis of variance with repeated measures
AUTORES:	Christian Acal - Ana M. Aguilera
REVISTA:	Advances in Data Analysis and Classification
	Factor de impacto: 2.134 (2020) Statistics & Probability: 40/125 (1º Tercil)
DOI:	https://doi.org/10.1007/s11634-022-00500-y
AÑO:	2022

Resumen y estado del arte

El problema del análisis de la varianza funcional (FANOVA) puede ser visto como la extensión natural del problema del análisis de la varianza clásico cuando los datos disponibles son funciones que evolucionan a lo largo de algún argumento continuo como puede ser el tiempo o el espacio. Por tanto, el objetivo principal del FANOVA es comprobar si existen diferencias significativas entre las funciones medias de una variable de respuesta funcional en distintos grupos independientes. Esta metodología ha sido ampliamente abordada tanto para el caso univariante [1, 2, 3, 4] como multivariante, es decir, más de una variable funcional en el estudio [5, 6, 7].

Cuando la información se mide repetidamente sobre los mismos sujetos, el marco teórico subyacente es el de medidas repetidas. Centrándonos en el caso univariante, el objetivo del FANOVA con medidas repetidas (FANOVA-RM) es comprobar si se verifica la igualdad de las funciones medias de una variable funcional cuando ésta es observada en diferentes periodos de tiempo bajo distintas condiciones. A pesar del notable interés por sus múltiples aplicaciones con datos reales, el diseño de medidas repetidas para datos funcionales es raramente considerado en la literatura. De hecho, la gran mayoría de trabajos solo se centran en el caso de muestras apareadas (observaciones en dos condiciones o periodos de tiempo diferentes). En [8] se introdujo el primer estadístico para resolver este problema teniendo en cuenta solo la variabilidad entre grupo. Para aproximar la distribución nula del estadístico, se propusieron diferentes métodos paramétricos y no paramétricos a través de tests de permutación y bootstrap. En esta línea, [9] presentó otra aproximación tipo Box-type, la cual resultó ser más rápida computacionalmente. Asimismo, con el fin de obtener estadísticos más potentes que tengan en cuenta también la variabilidad dentro del grupo, en [10] se adaptaron dos nuevos estadísticos basados en el estadístico t-Student clásico para medidas repetidas. Las distribuciones de estos estadísticos fueron aproximadas también a través de métodos paramétricos basados en la derivación de distribuciones asintóticas y mediante técnicas no paramétricas de permutación y bootstrap.

En este trabajo, abordamos el problema del FANOVA de dos vías, en el cual los sujetos son clasificados en grupos independientes y la variable de respuesta es observada bajo diferentes condiciones para cada individuo. Por tanto, un factor representa el efecto de medidas repetidas (tratamientos) y el segundo denota la contribución de cada grupo. Desde un punto de vista teórico, la variable de respuesta funcional X es medida repetidamente en cada sujeto en m periodos de tiempo diferentes o tratamientos. Entonces $\{x_{ijk}(t) : i = 1, 2, \ldots, m; j = 1, 2, \ldots, g; k = 1, 2, \ldots, n_j; t \in T\}$ representa g muestras independientes de curvas (una por grupo) definidas en el intervalo continuo T, donde $x_{ijk}(t)$ denota la respuesta del k-ésimo sujeto en el j-ésimo grupo bajo el i-ésimo tratamiento. Se asume que el diseño es balanceado (cada tratamiento es aplicado sobre todos los sujetos) y que las curvas muestrales pertenecen al espacio de Hilbert $L^2[T]$ de funciones de cuadrado integrable con el producto interno usual $< f|g >= \int_T f(t)g(t)dt,\ \forall f, g \in L^2[T]$.

En estas condiciones, en el modelo FANOVA-RM de dos vías las curvas muestrales verifican el siguiente modelo lineal funcional

$$x_{ijk}(t) = \mu(t) + \alpha_i(t) + \beta_j(t) + \theta_{ij}(t) + \epsilon_{ijk}(t),\ \forall t \in T,$$

con $\mu(t)$ siendo la función media general; $\alpha_i(t)$ y $\beta_j(t)$ siendo las funciones i-ésima y j-ésima de los efectos-principales de los tratamientos y grupos, respectivamente; $\theta_{ij}(t)$ el (i, j)-ésimo efecto-interacción entre tratamientos y grupos; y $\epsilon_{ijk}(t)$ siendo errores i.i.d. con distribución $SP(0, \gamma(s, t))$ $\forall i = 1, 2, \ldots, m; j = 1, 2, \ldots, g; k = 1, 2, \ldots, n_j$.

Los principales contrastes de hipótesis asociados a este modelo son los siguientes:

- ¿Hay diferencias significativas entre los tratamientos?

$$H_0 : \alpha_1(t) = \alpha_2(t) = \ldots = \alpha_m(t) = 0,\ \forall t \in T;$$

- ¿Hay diferencias significativas entre los grupos?

$$H_0 : \beta_1(t) = \beta_2(t) = \ldots = \beta_g(t) = 0,\ \forall t \in T;$$

- ¿Existe interacción entre los grupos y tratamientos?

$$H_0 : \theta_{ij}(t) = 0,\ \forall i, j;\ \forall t \in T;$$

contra la alternativa, de en cada caso, que la igualdad no se mantiene.

Para resolver estos contrastes, la idea principal es asumir la expansión básica de las curvas propuesta por [2]. En este contexto, las cuervas muestrales pertenecen a un espacio finito-dimensional generado por una base $\{\phi_1(t), \ldots, \phi_p(t)\}$, de manera que las trayectorias pueden ser expresadas a través de la siguiente expresión

$$x_{ijk}(t) = \sum_{h=1}^{p} y_{ijkh}\phi_h(t),$$

con p suficientemente grande para garantizar una correcta representación de las curvas.

Asumiendo este resultado, en el trabajo se demuestra que el modelo FANOVA-RM es reducido a un modelo MANOVA con medidas repetidas, en el que la respuesta multivariante queda definida por los coeficientes básicos y_{ijkh} de las curvas muestrales. En este punto, tanto el modelo doble multivariante (*doubly multivariate model*) como el modelo mixto multivariante (*mixed multivariate model*) podrían ser aplicados para resolver los contrastes planteados [11, 12]. Hasta donde llega nuestro conocimiento, este marco teórico no ha sido aún abordado en la literatura; solo en [8] se comenta brevemente cómo algunos los estadísticos propuestos podrían ser generalizados para el modelo de una vía con más de dos muestras apareadas, pero no se hace referencia al caso de disponer también de grupos independientes.

La contribución metodológica de este trabajo viene motivada por estudios biomecánicos en los que los datos disponibles normalmente reflejan la aceleración lineal o posición de alguna articulación (ángulo formado con algún eje) en términos del tiempo o el porcentaje del ciclo de marcha. Uno de los principales objetivos del campo de la biomecánica es detectar posibles diferencias en los patrones de marcha cuando los sujetos realizan múltiples actividades (diseño de medidas repetidas). Históricamente, el análisis de las curvas de movimiento humano se ha realizado mediante métodos multivariantes a partir de las observaciones discretas de las curvas o incluso a través de medidas descriptivas de las mismas. En este sentido, la metodología propuesta en el presente trabajo no solo evita la pérdida de información que puede causar los enfoques clásicos al no considerar todo el comportamiento de la curva sino que por consiguiente, también proporciona resultados mucho más rigurosos y precisos. En particular, esta nueva metodología es aplicada sobre dos conjuntos de datos biomecánicos reales.

En conclusión, las principales aportaciones del trabajo son:

1. Se introduce un nuevo enfoque funcional basado en la expansión básica de las curvas para resolver el problema de comprobar la igualdad de las curvas medias de una variable funcional observada en varios grupos independientes bajo diferentes tratamientos o periodos de tiempo.

2. Dado que el problema del FANOVA-RM de dos vías es reducido a un MANOVA para la respuesta multivariante de coeficientes básicos, se consideran dos enfoques diferentes para resolver el MANOVA con medidas repetidas.

3. Se adapta un nuevo test de permutación para el caso en el que no se verifiquen las condiciones del modelo MANOVA con medidas repetidas.

4. Dada la naturaleza estocástica de los datos biomecánicos, la metodología propuesta resulta ser una herramienta más adecuada que los enfoques clásicos utilizados en el ámbito de la biomecánica para estudiar si existen diferencias significativas en el ciclo de marcha.

Referencias

[1] Cuevas A, FebreroM, Fraiman R (2004) An anova test for functional data. Comput Stat Data An 47(1):111–122

[2] Ramsay JO, Silverman BW (2005) Functional data analysis. Springer, Berlin

[3] Górecki T, Smaga L (2015) Comparison of tests for the one-way anova problem for functional data. Comput Stat 30(4):987–1010

[4] Aguilera AM, Acal C, Aguilera-Morillo MC, Jiménez-Molinos F, Roldán JB (2021) Homogeneity problem for basis expansion of functional data with applications to resistive memories. Math Comput Simulat 186:41–51

[5] Jacques J, Preda C (2014) Model-based clustering for multivariate functional data. Comput Stat Data An 71:92–106

[6] Górecki T, Smaga L (2017) Multivariate analysis of variance for functional data. J Appl Stat 44(12):2172–2189

[7] Acal C, Aguilera AM, Sarra A, Evangelista A, Di-Battista T, Palermi S (2022) Functional anova approaches for detecting changes in air pollution during the covid-19 pandemic. Stoch Env Res Risk A 36:1083–1101

[8] Martínez-Camblor P, Corral N (2011) Repeated measures analysis for functional data. Comput Stat Data An 55:3244–3256

[9] Smaga L (2019) Repeated measures analysis for functional data using box-type approximation with applications. REVSTAT-Stat J 17(4):523–549

[10] Smaga L (2020)Anote on repeated measures analysis for functional data. AStAAdv Stat Anal 104:117–139

[11] Boik RJ (1988) Themixed model for multivariate repeated measures: validity conditions and an approximate test. Psychometrika 53(4):469–486

[12] Boik RJ (1991) Scheffés mixed model for multivariate repeated measures: a relative efficiency evaluation. Commun Stat Theor M 20(4):1233–1255