Signal identification by local functional ANOVA

Rémi Mahmoud [1] & Ching-Fan Sheu [2] & David Causeur [3]

[1] *Institut Agro Rennes-Angers, Département Statistique et Informatique, UMR 6625 IRMAR CNRS remi.mahmoud@agrocampus-ouest.fr*
[2] *Institute of Education, National Cheng Kung University, csheu@ncku.edu.tw*
[3] *Institut Agro Rennes-Angers, Département Statistique et Informatique, UMR 6625 IRMAR CNRS david.causeur@agrocampus-ouest.fr*

Résumé. L'analyse des données fonctionnelles permet d'étudier des phénomènes à variations continues à partir de mesures discrètes. En particulier, l'analyse de la variance fonctionnelle peut être vue comme une transposition au cadre des données fonctionnelles des tests usuels de Fisher pour la comparaison de modèles linéaires. La problématique d'analyse de variance fonctionnelle la plus connue est probablement celle à un facteur du test de comparaison de courbes moyennes par groupes. Le test d'un effet sur l'ensemble de la plage d'observation des courbes est l'objet de nombreuses discussions, notamment sur la manière optimale d'agréger les statistiques de tests ponctuelles pour calculer une statistique de test global optimale et sur la gestion de la dépendance entre celles-ci. La détection par ce type de test d'un signal d'association entre des courbes moyennes et une ou plusieurs variables explicatives ouvre la voie à la question de l'identification de ce signal d'association, en d'autres termes à la recherche du support sur lequel le signal d'association est non-nul.

Dans de nombreux domaines d'application, cette question est abordée par les méthodes de tests multiples. Or, d'une part, ces méthodes ne prennent en général pas en compte la nature continue du signal d'association, ce qui peut conduire à identifier des temps de mesures isolés, dont l'interprétation est impossible. D'autre part, les méthodes de contrôle du nombre de faux positifs sont en général trop conservatrices, ce qui conduit souvent à n'identifier aucun intervalle support du signal d'association alors que le test global a conclu à l'existence de celui-ci à l'échelle de la plage d'observation. Ce problème se pose notamment en agronomie, où il est crucial d'identifier les périodes temporelles au cours desquelles des variations biologiques notables surviennent en fonction de covariables environnementales.

L'approche hybride que nous proposons pour l'identification des intervalles supports du signal d'association repose sur le F-test généralisé proposé par Causeur et al., 2020 pour l'analyse de la variance fonctionnelle de courbes d'électroencéphalogrammes dans des dispositifs expérimentaux impliquant plusieurs variables explicatives. À partir de l'application locale de ce test sur des intervalles de même amplitudes dont le centre parcourt la plage d'observation des courbes, la méthode vise à identifier la plus grande portion du signal dans laquelle le signal d'association n'est pas significatif. L'amplitude de l'intervalle glissant fait le compromis entre une procédure de tests multiples (intervalles d'un seul point) et un test global (intervalle couvrant toute la plage, fANOVA).

Après avoir détaillé l'algorithme et discuté des choix méthodologiques, nous illustrons son application sur des données simulées ainsi que sur des données réelles portant sur l'étude des

liens entre les émissions de mycotoxines par un champignon phytopathogène et l'excédent hydrique mesuré régulièrement au cours du temps pendant la croissance de la plante. L'application aux données réelles met en évidence la capacité de la méthode à identifier des périodes clés propices à l'émission de mycotoxines.

Mots-clés. ANOVA fonctionnelle, Test global, Identification d'un signal.

Abstract. Functional Data Analysis (FDA) focuses on studying time-continuous variations using discretized observations. Analogous to the standard F-test used to compare linear models when the response is scalar, Functional Analysis of Variance (fANOVA) provides a global test for assessing the effect of covariates on functional data. A well-known application of fANOVA is the comparison of mean curves in a one-way design. A central challenge in this context is the construction of a global test across the entire time domain, particularly regarding the aggregation of individual test statistics into an optimal global test statistic and the handling of temporal dependence across these statistics. Detecting a significant association between mean curves and one or more explanatory variables opens up the complementary objective of identifying time intervals where this association occurs.

In many cases, this identification problem is addressed using multiple testing procedures. However, these methods are not designed to account for the intrinsic functional nature of the association signal, leading to the detection of spurious, isolated significant time points. Additionally, the control of false positives tends to be overly conservative, frequently resulting in no significant intervals being identified, even when the global test indicates the presence of a non-zero association. This issue is particularly relevant in agricultural studies, where it is crucial to pinpoint time intervals during which biological variations are linked to environmental covariates.

To address this limitation, we propose a hybrid approach for identifying the temporal support of the association signal. Our method builds upon the generalized functional F-test introduced by Causeur et al., 2020 for fANOVA in electroencephalogram (EEG) studies involving multiple explanatory variables. The approach aims to identify the largest portion of the time domain where the association signal is not significant. This is achieved by applying the fANOVA test locally over intervals of fixed amplitude, with centers regularly distributed across the entire time frame. The amplitude of the sliding interval strikes a balance between a multiple testing procedure (intervals of a single point) and a global test (interval covering the entire range, fANOVA).

Following a detailed presentation of the algorithm, with a focus on key methodological choices, we assess its performance on both simulated and real data. Specifically, we analyse the emission of mycotoxins by a phytopathogenic fungus in relation to water excess, measured at regular intervals throughout plant growth. This case study demonstrates the method's ability to identify key time intervals that favour mycotoxin emissions, showcasing its potential for applications in agricultural and environmental research.

Keywords. Functional ANOVA, Global test, Signal identification

1 Functional Analysis of Variance

Functional data consist of observations of continuous curves measured on a discretized grid of points. The increasing availability of instruments and techniques — such as sensors, and spectroscopy — has led to a surge in high-throughput curve data, posing new challenges for statistical methodologies in Functional Data Analysis (e.g., Ramsay and Silverman, 2002, Ferraty, 2006). A notable challenge arises in neuroscience, where Event-Related Potentials (ERP) designs study the effects of experimental covariates on brain activity measured by Electroencephalography time-locked to an event of interest. This motivated Causeur et al., 2020 to propose a Functional Analysis of Variance (fANOVA) test, extending the standard F-test used for comparing linear models to accommodate functional responses.

Let $(Y(t))_{t\in\mathcal{T}}$ denote a functional response variable defined on a time domain $\mathcal{T}$. Let $x = (x_1, \ldots, x_p)'$ stand for a p-vector of time-independent explanatory variables. For all $t \in \mathcal{T}$, the following linear model is assumed:

$$Y(t) \quad = \quad \beta_0(t) + \beta_1(t)x_1 + \ldots + \beta_p(t)x_p + \varepsilon(t), \tag{1}$$

where $\beta_0(t)$ is the pointwise intercept parameter at time t and $\beta(t) = (\beta_1(t), \ldots, \beta_p(t))'$ is the p-vector of pointwise regression parameters at time t.

Analogously to the standard model comparison issue in linear model, the objective of fANOVA is to test general null hypotheses of the form H_0: $R\beta(t) = 0$, for all $t \in \mathcal{T}$, where R is a $k \times p$ matrix defining k linear contrasts and $\beta(t) = (\beta_1(t), \ldots, \beta_p(t))'$. For instance, in one-way designs where the only explanatory variable is a two-group covariate, the primary contrast is typically the mean difference curve. The fANOVA test proposed in Causeur et al., 2020 is designed to handle arbitrarily complex null hypotheses that may involve many explanatory variables. Rejecting H_0 implies that the association signal $t \mapsto \beta(t)$ is nonzero at some point, which is why this issue is often referred to as signal detection or global testing.

Model (1) is typically supplemented with specific assumptions regarding time-continuous variables. A common assumption is that the functions $t \mapsto \beta_j(t),\ j = 0, 1, \ldots, p$, exhibit a certain degree of regularity over time. This smoothness can be captured using linear decompositions on basis functions, such as B-splines. In this case, if $\phi(t) = (\phi_1(t), \ldots, \phi_q(t))$ represents the values of q basis functions $\phi_j(t)$, $j = 1, \ldots, q$ at time t, then $\beta(t) = \phi'(t)B$, where B is a $q \times p$ vector of basis functions coefficients. In the following, the signal identification method will be presented without this smoothing option. Finally, we assume $\varepsilon(t) \sim \mathcal{N}(0, \sigma^2(t))$ with the autocorrelation function $\rho(t, t') = \text{cor}(\varepsilon(t), \varepsilon(t'))$. The functional responses $Y(t)$ are observed on a discrete time grid $t_1, \ldots, t_m$. For each $t \in \mathcal{T}$, model (1) corresponds to a standard linear model for a scalar response variable. Thus, pointwise F-test statistics $(F_{t_1}, \ldots, F_{t_m})$ can be used to test the null hypotheses H_{0,t_j}: $R\beta(t_j) = 0,\ j = 1, \ldots m$. Most fANOVA test statistics rely on simple aggregations – such as sums, weighted sums or maximum – of the pointwise F-statistics $F_{t_j},\ j = 1, \ldots, m$. Surprisingly, the time dependence between these pointwise F-statistics is often overlooked in the aggregation process (Górecki and Smaga, 2019). To ensure proper control of the type I error rate under arbitrary time dependence, the null distribution of the global test statistic is however estimated using random permutation techniques.

To account for the strong temporal dependence typically observed in functional data, the fANOVA test proposed by Causeur et al., 2020 addresses this dependence by constructing global test statistics as sums of fully or partially decorrelated pointwise test statistics. Partial decorrelation is achieved through linear operators derived from a k-factor approximation of the inverse correlation matrix of the pointwise test statistics. The number of factors, k, serves as a tuning parameter, ranging from 0 (no whitening) to the rank of the residual covariance matrix of model (1) (full whitening).

While Causeur et al., 2020 introduces a strategy for selecting the optimal number of factors, we propose an enhanced version of the test. In this new approach, a global omnibus test statistic is derived by aggregating the fANOVA test statistics computed for different choices of k, thereby incorporating information across multiple levels of decorrelation.

2 Signal identification by local fANOVA

When a global functional F-statistic detects a significant association signal over the entire time frame, it is natural to further investigate which specific segments contribute most to this finding. However, multiple testing procedures applied to the pointwise F-tests $F_{t_j},\ j = 1, \ldots, m$ often struggle to identify meaningful significant intervals. This limitation arises from two key factors: the strong time dependence between pointwise test statistics and the regularity of the association signal, which multiple testing fails to account for. As a result, these methods tend to either miss significant regions or detect spurious isolated time points.

Instead of attempting to identify significant time points, we propose a local functional ANOVA testing procedure that takes the opposite approach, *i.e* identifying the largest set of time intervals where no significant effect is detected using the functional ANOVA test introduced in Section 1. This procedure ensures consistency with the global fANOVA test. If the global test is not significant, then the largest set of time points where no significant effect is found spans the entire time frame, meaning no significant time points are identified. Conversely, if the global test is significant, then the largest non-significant set is necessarily smaller, ensuring coherence between local and global analyses.

To determine the maximal set of non-significant time points, we propose the following approach (Algorithm 1). First, we define a grid of regularly spaced time points across the entire time domain and test for the presence of an association signal within their neighbourhoods using the functional F-test introduced earlier. The neighborhood amplitude acts as a hyper-parameter, whose choice will be discussed during the presentation. If no significant intervals are found, the algorithm terminates, concluding no significant findings. Otherwise, a second step, backward elimination, sequentially removes significant intervals from the list until the fANOVA test outside these intervals becomes significant.

Algorithm 1 Signal identification

Input: l (window size), I (number of intervals), α (critical p-value threshold)

Compute interval centers $\mathcal{C} := (C_i)_{1 \leq i \leq I}$ with $C_i := 1 + (i-1) \cdot \frac{T-1}{I-1}$
for $i = 1$ to I **do**
    Define interval $J_i := \{t_j : |t_j - C_i| \text{ among the } l \text{ smallest}\}$
    Compute $p_i \leftarrow$ fANOVA p-value for J_i
end for
if $\min_i(p_i) \geq \alpha$ **then**
    return No interval identified
else
    Rank p_i values as $\tilde{p} := \{p_{(1)}, \ldots, p_{(S)}\}$ where $p_{(s)} < \alpha$ for $s \leq S$
    Initialize $\tilde{I}_1 \leftarrow I_1$, $s \leftarrow 1$, and $q_s \leftarrow 1$
    while $s \leq S \ \wedge \ q_s \geq \alpha$ **do**
        $s \leftarrow s + 1$
        Perform fANOVA on $T \setminus \tilde{I}_s$
        $q_s \leftarrow$ associated p-value
        if $q_s < \alpha$ **then**
            return $\bigcup_{s=1}^{S} I_s$ (significant intervals)
        else
            $\tilde{I}_s \leftarrow \tilde{I}_s \cup I_s$
        end if
    end while
end if

2.1 Application of the algorithm

2.1.1 Simulation

We test our method on simulated data, exploring parameters such as the number of individuals ($n \in \{100, 1000, 5000\}$), the temporal resolution ($T \in \{50, 500, 2000\}$), the presence of a temporal dependance (yes / no). For each of the combination we generate 100 datasets and compare the results to other methods ((Fused) LASSO, Tibshirani et al., 2005).

2.1.2 Application of the method on real-world data

Context: Fusarium Head Blight is a cereal disease that causes significant damage. The fungi responsible emit mycotoxins, such as Nivalenol (NIV), which pose risks to human and animal health. When contamination exceeds legal thresholds, wheat is devalued. Using a dataset from Arvalis (technical research institute) with data from 320 farms, we analyzed mycotoxin threshold exceedances alongside climatic time series (rainfall, evapotranspiration, global radiation, etc.). We applied our method to detect signals in water excess time series (P - ETP, mm) around flowering, linked to NIV threshold exceedances. The multiple testing procedure (pointwise F-tests) found no differences between curves ($p = 0.44$, Benjamini-Hochberg correction), while the fANOVA test introduced in Section 1 was significant ($p = 0$) but did not pinpoint specific segments responsible for this difference.

The two intervals identified (Fig. 1, $[-675°Cd, -350°Cd]$ et $[175°Cd, 525°Cd]$) correspond to a pre-flowering period where the moisture conditions are favorable for the growth of the

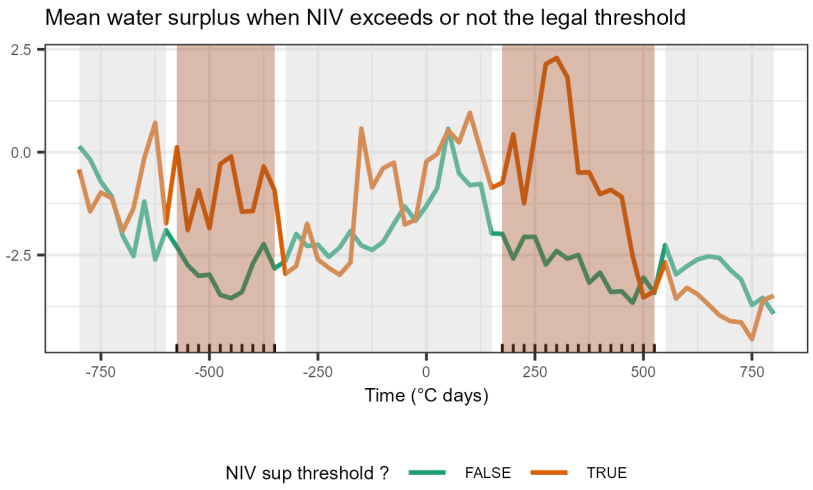

Figure 1: Application of the method to real-world data: Mean water surplus curves over time for cases exceeding (orange) or not (blue) the NIV legal threshold. The grey zone indicates time intervals with no significant effect (functional ANOVA), while the orange zone represents the complementary intervals likely causing the difference.

fungus before infection, and a post-flowering period where moisture around the spikes could promote the development of Fusarium as well as mycotoxin emission (Cowger et al., 2009).

3 Discussion and perspectives

The algorithm presented in this communication will be made available as an R package. A suitable criterion for selecting hyperparameters (l, I) still needs to be developed.

Références

Causeur, D., Sheu, C.-F., Perthame, E., & Rufini, F. (2020). A functional generalized f-test for signal detection with applications to event-related potentials significance analysis. *Biometrics*, *76*(1), 246–256. https://doi.org/10.1111/biom.13118

Cowger, C., Patton-Özkurt, J., Brown-Guedira, G., & Perugini, L. (2009). Post-anthesis moisture increased fusarium head blight and deoxynivalenol levels in north carolina winter wheat [PMID: 19271972]. *Phytopathology®*, *99*(4), 320–327. https://doi.org/10.1094/PHYTO-99-4-0320

Ferraty, F. (2006). *Nonparametric functional data analysis*. Springer. https://doi.org/https://doi.org/10.1007/0-387-36620-2

Górecki, T., & Smaga, Ł. (2019). Fdanova: An r software package for analysis of variance for univariate and multivariate functional data. *Computational Statistics*, *34*(2), 571–597. https://doi.org/10.1007/s00180-018-0842-7

Ramsay, J. O., & Silverman, B. W. (2002). *Applied functional data analysis: Methods and case studies*. Springer. https://doi.org/https://doi.org/10.1007/b98888

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(1), 91–108.