

Cumulant-Based Robustness of ANOVA-like Tests in Multiple Additive Models

CRISTINA DIAS^{1,3,*}, CARLA SANTOS^{2,3,*}, JOÃO TIAGO MEXIA³

¹Polytechnic University of Portalegre,
PORTUGAL

²Polytechnic University of Beja,
PORTUGAL

³NOVA Math - Center of Mathematics and Applications,
FCT- NOVA University of Lisbon,
PORTUGAL

**Corresponding Author*

Abstract: - Multiple additive models provide a robust framework for analysing families of additive models associated with the treatments of a base design. In this context, ANOVA-like procedures can be used to assess the effects of factors through corresponding components of principal estimable functions. However, the validity of the resulting F-tests is typically justified under approximate homoscedasticity and distributional assumptions that may be violated in practice. This paper studies the robustness of ANOVA-like inference in multiple additive models under non-normal error distributions. Using cumulants of order two, three and four, we characterise departures from normality in terms of variance, skewness and excess kurtosis. A simulation study evaluates the empirical type I error and power of the proposed ANOVA-like tests under normal, heavy-tailed and skewed error distributions, including Gumbel-type components. The results identify conditions under which the F approximation remains reliable and cases where non-normality leads to distorted inference. The proposed cumulant-based perspective provides both a diagnostic and methodological framework for robust inference in multiple additive models.

Key-Words: - ANOVA-like procedures; Cumulants; F-tests; Kurtosis; Multiple additive models; Skewness.

Received: May 9, 2026. Revised: June 11, 2026. Accepted: July 12, 2026. Published: October 7, 2026.

1 Introduction

Analysis of variance and its associated F-tests constitute one of the most widely used inferential frameworks for assessing the effects of factors and interactions in experimental and observational designs. In their classical formulation, these methods rely on the assumptions of independence and normality of errors and homogeneity of variances. In real-world applications, where these assumptions are only approximately met, the sensitivity of F-tests to deviations from normality and homoscedasticity should be carefully assessed, particularly in the presence of asymmetric error distributions, heavy tails, or unequal variances.

The robustness of classical ANOVA and F-tests to deviations from normality has been widely studied, with results that vary according to the specific design and assumptions considered. Classical F-tests are often reasonably robust to non-

normality under many conditions, particularly in balanced designs. Foundational contributions suggest that F-tests may maintain reasonable performance under moderate deviations from normality, particularly when sample sizes are balanced and variances are homogeneous [1], [2], [3], [4]. More recent studies have confirmed the F-test's robustness throughout a range of conditions. For example, a study using Monte Carlo simulation [5] found that the F-test showed good control of Type I error across a range of manipulated conditions, such as equal and unequal group sample sizes, total sample size, distribution shape, and others. In non-normal settings observed that sample size influences the performance of the F-test, with larger sample sizes tending to enhance the power of the test and to support better control of Type I error rates [6]. Other studies concluded that extreme violations of assumptions, such as high variance

heterogeneity or severe non-normality, degrade F-test performance. For example, [7] reported that the F-test may become less robust when the variance ratio exceeds 1.5, especially in the presence of unequal sample sizes. [8], [9] noted that the F-test may have more difficulty maintaining its nominal Type I error rate under heteroscedasticity. [10], [11] showed that when non-normality is combined with unequal variances, unbalanced designs, skewness, or heavy-tailed distributions, the effective Type I error rate can deviate from the nominal significance level.

Thus, the behaviour of F-tests is influenced by sample size, the balance of the design, the degree of heterogeneity of variances, and the shape of the underlying distribution of errors. This issue becomes particularly relevant in the context of multiple additive models, which extend the classical additive model framework by associating several individual additive models with the treatments in a base experimental design.

In the context of multiple additive models, see for example [12], ANOVA-like inference is performed not on scalar means, as in traditional ANOVA, but on the corresponding components of estimable principal functions. This is enabled by conditions that aim to ensure the estimators of the parameter vectors associated with the individual models share approximately the same covariance matrix, a property closely linked to assumptions about fixed effects, the orthogonality of design matrices and covariance structures, and commutativity conditions.

The eigenvectors of the common covariance matrix can then be used to define the principal estimable functions used in the subsequent ANOVA-like analysis. Thus, although the inferential strategy is inspired by classical ANOVA, it operates in a more general context, involving vector estimators, covariance structures and model parameter functions.

Despite the extensive literature on the robustness of classical ANOVA, the robustness of F-tests and ANOVA-like tests in multiple additive models have been less thoroughly explored. In particular, it is not yet clear whether the conditions underpinning robustness in classical ANOVA, such as balanced designs, homogeneous variances and moderate deviations from normality, are sufficient to ensure reliable inference when the tests are applied to corresponding components of estimable principal functions. This gap is particularly important when the underlying errors exhibit asymmetry, heavy tails or high kurtosis, as these distributional characteristics may affect the

behaviour of the cumulants and, consequently, the approximation underlying the proposed F-tests.

The contribution of this paper is distinct from the algebraic construction of multiple additive models in [12] and from earlier work on the derivation of complex linear models. [12] establish a family of treatment-indexed individual additive models and the conditions under which their parameter estimators can be analysed through principal estimable functions. The present study instead examines the finite-sample validity of the final ANOVA-like inferential stage when the component estimators are affected by non-Gaussian errors. More specifically, we (i) express departures from Gaussianity through standardised third- and fourth-order cumulants; (ii) relate these quantities to the accuracy of the approximate F reference distribution; and (iii) evaluate their interaction with sample-size imbalance and variance-size pairing. The contribution is therefore a robustness analysis and a reproducible diagnostic framework, rather than a new construction of the underlying multiple additive model or an exact finite-sample correction to the F test.

2 Multiple Additive Models

Mixed models are a natural extension of classical linear models, accommodating both fixed and random effects within a unified framework. This extension enables the modelling of complex, hierarchical, and unbalanced data, and underlies advanced methodologies in both linear and generalised linear contexts [13]. In its most common linear form, a mixed model can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \quad (1)$$

where $\mathbf{Y}$ represents the vector of observations, $\mathbf{X}\boldsymbol{\beta}$ corresponds to the fixed-effects component, $\mathbf{Z}\mathbf{b}$ represents the random-effects component, and $\boldsymbol{\varepsilon}$ is the vector of random errors. It is usually assumed that $\mathbf{b}$ has null mean and a covariance matrix $\mathbf{G}$, while $\boldsymbol{\varepsilon}$ has null mean and a covariance matrix $\mathbf{R}$, these two components being independent. Such models are particularly relevant when the data involve clustering, repeated measures, blocks, treatments, individuals or other sources of dependence, for which the assumption of independent observations may be inappropriate [14], [15].

In turn, additive models express the response as a sum of explanatory components. In general terms, an additive model can be written as

$$Y_i = \alpha + f_1(X_{1i}) + f_2(X_{2i}) + \dots + f_p(X_{pi}) + \varepsilon_i, \quad (2)$$

where Y_i is the observed response for unit i , α is a constant, $X_{1i}, \dots, X_{pi}$ are explanatory variables, $f_1, \dots, f_p$ are functions associated with each explanatory variable, and ε_i represents the random error. When the functions f_j are linear, this formulation approximates the classical linear model; when they are smooth functions, the framework of generalised additive models is obtained, which allows for the representation of non-linear relationships between the predictors and the response, see e.g. [16], [17].

The combination of these two ideas leads to generalised mixed additive models, in which the additive structure is supplemented by random effects. A simplified form of this model can be represented by

$$Y_i = f_1(X_{1i}) + f_2(X_{2i}) + \dots + f_p(X_{pi}) + \mathbf{Z}_i \mathbf{b} + \varepsilon_i, \quad (3)$$

where $\mathbf{Z}_i \mathbf{b}$ represents the contribution of random effects to observation i . This formulation allows for the simultaneous consideration of potentially non-linear relationships between predictors and the response, as well as dependence, clustering or heterogeneity among observational units. This is precisely the rationale presented in the literature on GAMMs, according to which these models combine the flexibility of generalised additive models with the ability of mixed models to incorporate random effects and capture variability between groups or clusters.

The term “multiple additive models” used below should not be confused with a generalised additive model composed of smooth predictive functions. Here, “multiple” refers to a family of individual models indexed by the treatment, linked by a common basic experimental design.

In this paper, we consider a distinct, though conceptually related, structure: multiple additive models. In this approach, the models that make up the multiple model correspond to the different treatments in a basic design. Thus, for each treatment $h = 1, \dots, d$, we consider an individual additive model, which can be generically represented by

$$Y_h = \mathbf{X}_h \boldsymbol{\beta}_h + \varepsilon_h, \quad h = 1, \dots, d \quad (4)$$

where Y_h is the vector of observations associated with treatment h , $\mathbf{X}_h$ is the individual model matrix, $\boldsymbol{\beta}_h$ is the corresponding parameter vector, and ε_h represents the vector of random errors. It is assumed that these errors have independent and

identically distributed components, with mean zero and finite cumulants up to at least the fourth order. This assumption will be formalised in the following section, where it is shown that the second-, third- and fourth-order cumulants allow the variance, skewness and kurtosis of the errors to be characterised, respectively.

Thus, the multiple additive models considered in this work are based on a family of individual additive models associated with the treatments of a base design. The subsequent inference focuses on the action of the factors of this base design on the estimated parameter vectors. In particular, an ANOVA-like analysis is performed on corresponding components of principal estimable functions; this is possible because the estimators of the parameter vectors of the individual models have, approximately, the same covariance matrix; the eigenvectors of this matrix then allow the definition of the principal estimable functions used in the inferential analysis.

Thus, in the next chapter, we consider a family of individual additive models indexed by the treatments of a basic design. It is upon this structure that the ANOVA-like inference developed in this work is built, together with the analysis of the impact of skewness and kurtosis on the behaviour of the approximate F-tests.

3 ANOVA-Like Inference

Consider a basic design consisting of d treatments. Each treatment $l = 1, \dots, d$ is associated with an individual additive model, such that the multiple model is formed by the family of these d models. For treatment l , consider the model

$$Y(l) = \mathbf{X}_0(l) \boldsymbol{\beta}_0(l) + \sum_{i=1}^w \mathbf{X}_i \mathbf{Z}_i(l), \quad l = 1, \dots, d, \quad (5)$$

where $\boldsymbol{\beta}_0(l)$ is fixed and the $\mathbf{Z}_i(l)$ are independent and have c_i i.i.d components $(\mathbf{z}_{i,1}(l) \dots \mathbf{z}_{i,c_i}(l))$ with null mean value and r -th cumulants $\kappa_{i,r}(l)$, $i = 1, \dots, w$, $r = 2, 3, 4$, $l = 1, \dots, d$. This formulation allows us to consider, in addition to variance, higher-order distributional characteristics, such as skewness and kurtosis, which play a significant role in assessing the robustness of ANOVA-like inference.

After considering estimators for the parameters $(\boldsymbol{\beta}_0(l), \mathbf{x}(l))$, $l = 1, \dots, d$ of the individual models, we study the action of the factors in the base design on the estimated vector

$$\boldsymbol{\psi}(l) = \mathbf{G} \boldsymbol{\beta}_0(l), \quad l = 1, \dots, d. \quad (6)$$

where $\mathbf{G}$ is a matrix that defines the estimable functions of interest. Namely, with

$$\boldsymbol{\psi}(l) = (\psi_1(l), \dots, \psi_h(l)), l = 1, \dots, d, \quad (7)$$

we will consider ANOVA-like inference on the values of

$$\eta(l) = \psi(l), l = 1, \dots, d \quad (8)$$

of a single component.

The analysis therefore focuses on the corresponding components of the main estimable functions, comparing the behaviour of that component across the different treatments in the base design.

We will single out the case in which u factors in the base design cross, having $J_1, \dots, J_u$ levels so that there will be

$$d = \prod_{h=1}^u J_h \quad (9)$$

treatments. The algebraic structure of these models has been studied using commutative Jordan algebras [13]. Furthermore, the classic treatment in [4] of the robustness of ANOVA and related techniques provides a justification for the use of approximate F-tests in this context. This justification should, however, be understood as approximate, since the inference is applied here to estimable function components derived from individual additive models, and not directly to scalar means as in classical ANOVA.

The construction of approximate F-tests is based on the fact that the estimators of the vectors $\boldsymbol{\psi}(l)$, for $l = 1, \dots, d$ have approximately the same covariance matrix. This property allows us to consider the vector

$$\boldsymbol{\eta} = (\eta_1, \dots, \eta_d)^T, \quad (10)$$

as approximately homoscedastic, making it possible to assess the statistical significance of hypotheses regarding the effects of factors and their interactions. Thus, the proposed ANOVA-like inference extends the logic of classical ANOVA to a more general context, in which the comparison between treatments is carried out via corresponding components of estimable principal functions associated with the individual additive models. Since the approximate validity of these procedures depends on the distributional structure of the errors, it is necessary to explicitly characterise the deviations from normality considered in this work. To this end, we use cumulants, which allow us to describe the variance, skewness and kurtosis of the error distributions in an integrated manner.

4 Cumulants and Non-Normality

An analysis of the robustness of ANOVA-like F-tests requires an explicit characterisation of the error distribution. In the normal case, the distribution is completely determined by the first two moments, that is, by the mean and the variance. However, when the errors exhibit asymmetry, heavy tails or high kurtosis, it becomes necessary to consider higher-order distributional characteristics. Cumulants provide a natural way of quantifying these characteristics.

Let U be a random variable with moment-generating function

$$M_U(t) = E(e^{tU}), \quad (11)$$

The generating function for the cumulants of U is defined by

$$K_U(t) = \log M_U(t). \quad (12)$$

The cumulant of order r is given by

$$k_r = K_U^{(r)}(0), \quad (13)$$

that is, by the r th-order derivative of the cumulant-generating function evaluated at $t = 0$. The first four cumulants have a straightforward statistical interpretation. The first cumulant corresponds to the mean,

$$k_1 = E(U), \quad (14)$$

the second cumulant is the variance,

$$k_2 = \text{Var}(U), \quad (15)$$

the third cumulant is linked to the asymmetry of the distribution,

$$k_3 = E[(U - E(U))^3], \quad (16)$$

and the fourth-order cumulant quantifies the deviation of the kurtosis from a Gaussian distribution,

$$k_4 = E[(U - E(U))^4] - 3 \text{Var}(U)^2. \quad (17)$$

Thus, the normal distribution has cumulants of order higher than two equal to zero. For this reason, non-null values of κ_3 and κ_4 indicate deviations from normality. In particular, $\kappa_3 \neq 0$ indicates asymmetry, while $\kappa_4 > 0$ is associated with leptokurtic distributions, which are often characterised by heavier tails than those of the normal distribution.

To compare distributions with different scales, it is common practice to use standardised measures.

Standardised skewness can be written as

$$\gamma_1 = \frac{\kappa_3}{\kappa_2^{3/2}} \quad (18)$$

whereas excess kurtosis is given by

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2}. \quad (19)$$

Thus, $\gamma_1 = 0$ characterises a symmetric distribution, whilst positive or negative values of γ_1 indicate right- or left-skewness. In turn, $\gamma_2 = 0$ is the excess kurtosis value of the normal distribution; positive values of γ_2 indicate high kurtosis and a greater weighting in the tails.

In the context of multiple additive models, these cumulants are particularly relevant because the errors of the individual models are not necessarily assumed to be Gaussian. By assuming that the random components have cumulants of order r , it becomes possible to study how variance, skewness and kurtosis influence the approximate distribution of the estimators of the main estimable functions. Consequently, third- and fourth-order cumulants provide a suitable tool for assessing the extent to which ANOVA-like inference based on approximate F-tests remains robust in the face of deviations from normality. The cumulant-based description is used as a diagnostic framework rather than as a closed-form derivation of the null distribution.

5 Robustness of ANOVA-like F-tests

ANOVA-like inference in multiple additive models aims to assess whether the factors in the base design exert a statistically significant effect on corresponding components of the main estimable functions. It should be noted that, for each treatment $l = 1, \dots, d$, an estimable component is considered

$$\eta(l) = \psi_j(l), \quad (20)$$

associated with the principal estimable functions vector

$$\boldsymbol{\psi}(l) = \mathbf{G}\boldsymbol{\beta}_0(l). \quad (21)$$

The comparison of treatments, main effects and interactions is then carried out on the vector

$$\tilde{\boldsymbol{\eta}}(l) = (\tilde{\eta}(1), \dots, \tilde{\eta}(d))^T, \quad (22)$$

where $\tilde{\eta}(l)$ represents the estimator of the component $\eta(l)$ in treatment l . This construction is consistent with the earlier formulation of the paper, in which ANOVA-like inference is applied to corresponding components of estimable principal functions associated with the different treatments in the base design.

The validity of approximate F-tests essentially rests on two conditions. The first is that the estimators $\tilde{\eta}(l)$, $l = 1, \dots, d$, have approximately the same variance. The second is that their covariances are either zero or sufficiently small for the vector $\tilde{\boldsymbol{\eta}}$ to be treated as approximately

homoscedastic and, in the ideal case, approximately normal. Formally, this is written as where

$$\begin{cases} E(\tilde{\boldsymbol{\eta}}) = \boldsymbol{\eta} \\ \text{Cov}(\tilde{\boldsymbol{\eta}}) \approx \sigma_{\tilde{\eta}}^2 \mathbf{I}_d, \\ \boldsymbol{\eta} = (\eta(1), \dots, \eta(d))^T \end{cases} \quad (23)$$

is the vector of true components, and $\sigma_{\tilde{\eta}}^2$ represents the approximate common variance of the component estimators. This assumption of approximate homoscedasticity is fundamental, as it allows the logic of the F-tests from classical ANOVA to be applied to this context.

Let us now consider an effect from the base design, for example a main effect or an interaction, represented by an orthogonal projection matrix $\mathbf{P}_A$. This matrix projects the vector $\tilde{\boldsymbol{\eta}}$ onto the subspace associated with the effect under study. The corresponding null hypothesis can be written as

$$H_0^{(A)}: \mathbf{P}_A \boldsymbol{\eta} = \mathbf{0}, \quad (24)$$

that is, the hypothesis that effect A has no influence on the corresponding component under consideration. The sum of squares associated with effect A is then given by

$$SS_A = \tilde{\boldsymbol{\eta}}^T \mathbf{P}_A \tilde{\boldsymbol{\eta}}, \quad (25)$$

with degrees of freedom

$$\nu_A = \text{rank}(\mathbf{P}_A). \quad (26)$$

Similarly, let $\mathbf{P}_E$ be the projection matrix associated with the residual subspace or the experimental error available in the base design. We define

$$SS_E = \tilde{\boldsymbol{\eta}}^T \mathbf{P}_E \tilde{\boldsymbol{\eta}}, \quad (27)$$

with

$$\nu_E = \text{rank}(\mathbf{P}_E). \quad (28)$$

The ANOVA-like statistic for testing effect A is then

$$F_A = \frac{SS_A/\nu_A}{SS_E/\nu_E}. \quad (29)$$

Given the null hypothesis $H_0^{(A)}$, and assuming approximate normality, approximate homoscedasticity and independence among the quadratic components under consideration, we have

$$F_A \sim F_{\nu_A, \nu_E}. \quad (30)$$

This approach forms the basis of ANOVA-like F-tests. However, unlike classical ANOVA, the F_A statistic is not constructed directly from independent scalar observations, but rather from estimators of the components of estimable principal functions. For this reason, its distribution depends

not only on the structure of the underlying design, but also on the distributional properties of the estimators $\tilde{\eta}(l)$.

When the errors of the individual models are Gaussian, the cumulants of order higher than two are null. In this case, the normal approximation of the estimators of the main estimable components is more readily justifiable, and the F_A statistic tends to behave in a manner close to the F_{ν_A, ν_E} distribution.

However, in non-Gaussian contexts, the third- and fourth-order cumulants may not be zero. The third cumulant, associated with skewness, and the fourth cumulant, associated with excess kurtosis, may affect the shape of the distribution of the estimators and, consequently, the quality of the F-approximation.

Thus, the distribution of F_A under $H_0^{(A)}$ should be understood as an approximation that depends on the cumulant structure of the errors. Conceptually, a formal Edgeworth expansion expresses the difference between the finite-sample distribution of F_A and its limiting F distribution by means of terms involving standardised cumulants of the component estimators,

$$\mathcal{L}(F_A) = \mathcal{L}_0(F_A) + \Delta_{\kappa_3} + \Delta_{\kappa_4} + o(\cdot), \quad (31)$$

where $\mathcal{L}_0(F_A)$ represents the behaviour of the statistic under normality, Δ_{κ_3} represents the effect of skewness, Δ_{κ_4} represents the effect of excess kurtosis, and $o(\cdot)$ encompasses higher-order terms or those that are asymptotically negligible.

Under certain regularity conditions, third-order cumulants generally contribute to order $n^{-1/2}$, while fourth-order cumulants and the squares of the third-order terms contribute to order n^{-1} . The exact coefficients depend on the estimator and projection matrices. These coefficients are not calculated here; the simulation study evaluates their combined effect on finite samples. See [18] for the classical case of one-way ANOVA.

The robustness of the ANOVA-like F-test can therefore be understood as the ability of the F_A statistic to maintain a Type I error rate close to the nominal level α , even when the errors in the individual models are not Gaussian. Formally, it is intended that

$$P_{H_0}(F_A > F_{\nu_A, \nu_E; 1-\alpha}) \approx \alpha, \quad (32)$$

where $F_{\nu_A, \nu_E; 1-\alpha}$ denotes the quantile of order $1 - \alpha$ of the F distribution with ν_A and ν_E degrees of freedom. If this probability is substantially greater than α , the test is liberal, rejecting the null hypothesis too frequently. If it is substantially less

than α , the test is conservative, having a lower ability to detect true effects.

In this context, skewness and kurtosis play distinct roles. Skewness can shift the distribution of the estimators and affect the symmetry of the quadratic terms that go into the F_A statistic. Excess kurtosis, in turn, tends to increase the likelihood of extreme observations, which may inflate the variability of the estimators and alter the behaviour of the sums of squares. Thus, leptokurtic or heavy-tailed distributions can lead to greater instability in the F_A statistic, particularly in small samples or in unbalanced designs.

The literature on classical ANOVA shows that F-tests can be relatively robust to moderate deviations from normality, particularly in balanced designs and under the assumption of homogeneity of variances [1], [2], [3], [4], [19]. However, this robustness diminishes when non-normality is combined with heteroscedasticity, sample imbalance or heavy tails [10], [11], [20], [21]. In the context of multiple additive models, these issues become even more delicate, since the inference focuses on estimators derived from the individual models and not directly on the original observations.

Based on this discussion, the following conceptual framework is proposed.

Heuristic robustness principle: Consider independent treatment-indexed additive models and a vector of approximately unbiased estimators of corresponding principal estimable-function components. If their covariance matrix satisfies

$$\text{Cov}(\tilde{\eta}) \approx \sigma_{\eta}^2 \mathbf{I}_d, \quad (33)$$

then, under the null hypothesis that effect A is absent, the ANOVA-like statistic

$$F_A = \frac{\tilde{\eta}' \mathbf{P}_A \tilde{\eta} / \nu_A}{\tilde{\eta}' \mathbf{P}_E \tilde{\eta} / \nu_E}, \quad (34)$$

follows approximately an F_{ν_A, ν_E} distribution. The approximation is most credible when the standardised higher-order cumulants of the component estimators are small, and the design is not strongly unbalanced.

The principle stems from the classical argument of the quadratic Gaussian form when the component estimation vector approximately follows a multivariate normal distribution. In the presence of non-Gaussian errors, the third and fourth-order cumulants describe the main deviations from this approximation, when they exist. The t_3 case is considered a heavy-tailed

limiting scenario, even though its fourth-order moment is infinite, and is therefore interpreted directly by simulation, rather than by expanding the fourth-order cumulant.

Three consequences guide the empirical study. First, balanced homoscedastic designs offer the most favourable scenario for the F approximation. Second, strong skewness or heavy distributional tails can impair the calibration of finite samples, even with equal variances. Third, in the case of unequally sized groups, the direction of variance-size matching can dominate the effect of the distribution shape: assigning the highest variance to the smallest group tends to induce liberal behaviour, while assigning it to the largest group tends to induce conservative behaviour.

These statements are evaluated for the specific designs and distributions described below. They are not extrapolated to all complete parameter configurations of a multiple additive model. The simulation therefore targets the final inferential step and is accompanied by a formal reduction demonstration in Section 6.1. This is precisely the motivation for the simulation study presented in the following section, in which type I error and the power of the ANOVA-like F-test are evaluated under different error distributions, including normal, thick-tailed symmetric, asymmetric and Gumbel distributions.

6 Simulation Study

This simulation study evaluates the calibration of finite samples and the power of the ANOVA-like F-test under controlled conditions of non-normality and homoscedasticity. The scenarios studied are representative, but not exhaustive. The main quantities analysed are the type I empirical error, the empirical power, and the Monte Carlo standard error.

6.1 Simulation Model

For each treatment $l = 1, \dots, d$, we consider the operational version of the individual additive model,

$$Y_{li} = \eta(l) + \varepsilon_{li}, i = 1, \dots, n_l \quad (35)$$

Here, Y_{li} represents the observation for treatment l , $\eta(l)$ represents the homogeneous component of the main estimable function associated with treatment l , and ε_{li} is the random error.

The reduction from the full individual model to (35) is justified when the estimator of the selected

principal estimable component admits a linear representation

$$\tilde{\eta}_l - \eta_l = \sum_{i=1}^{\eta_l} a_{li} \varepsilon_{li}, \quad (36)$$

With $\sum_{i=1}^{\eta_l} a_{li}^2$, approximately equal across treatments. In this representation, the final test, similar to an ANOVA-like test, depends on the joint distribution of the component estimators, and not on the other coordinates of the initial parameter vectors. The reduced simulation, therefore, reproduces the final comparison phase, but does not evaluate the errors resulting from the estimation of the complete covariance structure, the choice of eigenvectors, or the violation of the common covariance condition.

Under the null hypothesis,

$$H_0: \eta(1) = \eta(2) = \dots = \eta(d). \quad (37)$$

For the type I error simulations, the common component is set to zero without loss of location generality.

$$\eta(1) = \eta(2) = \dots = \eta(d) = 0$$

Under the alternative hypothesis, it is assumed that at least one of the components differs from the others. For example,

$\eta(1) = 0, \quad \eta(2) = \delta, \eta(3) = \dots = \eta(d) = 0$,
where $\delta > 0$ represents the magnitude of the treatment effect.

6.2 Distributions of Errors

To isolate the role of higher-order moments, all errors are centred and rescaled so that they have a null mean and variance one. Thus, differences in the test's behaviour can be attributed primarily to asymmetry and Excess kurtosis. The t_3 distribution has finite variance but an infinite fourth moment; therefore, it is a heavy-tailed limiting case, rather than a distribution with a finite fourth cumulant.

The following scenarios are considered:

$$\varepsilon_{li} \sim N(0,1),$$

as a Gaussian reference case, with

$$\gamma_1 = 0 \text{ and } \gamma_2 = 0.$$

We also consider a Student t-distribution with three degrees of freedom,

$$T \sim t_3,$$

centred and rescaled. This distribution is symmetric, so

$$\gamma_1 \approx 0,$$

but it has heavy tails and a high kurtosis. This scenario allows us to study the effect of kurtosis without asymmetry. Except for the t_3 scenario, the

error distributions are assumed to have finite cumulants up to order four. The t_3 distribution is included separately as a boundary heavy-tailed case with infinite fourth moment.

To study asymmetric errors, we consider a centred and rescaled chi-squared distribution:

$$\varepsilon = \frac{Q - E(Q)}{\sqrt{\text{Var}(Q)}}, Q \sim \chi_3^2. \quad (38)$$

In this case,

$$\gamma_1 > 0, \gamma_2 > 0.$$

A centred and rescaled log-normal distribution is also considered,

$$\varepsilon = \frac{L - E(L)}{\sqrt{\text{Var}(L)}}, L \sim \text{Lognormal}(0,1), \quad (39)$$

which exhibits strong positive skewness and high kurtosis.

Finally, we consider a centred and rescaled Gumbel distribution,

$$\varepsilon = \frac{G - E(G)}{\sqrt{\text{Var}(G)}}, G \sim \text{Gumbel}(0,1). \quad (40)$$

This scenario is particularly relevant because the Gumbel distribution arises naturally in contexts involving extreme values and exhibits non-null higher-order cumulants. It thus serves as a natural bridge between the cumulant-based formulation and the empirical assessment of the robustness of ANOVA-like F-tests.

6.3 Experimental Design of the Simulation

The study considers different combinations of the number of treatments, sample size and design balance.

$$d \in \{3, 5, 10\}.$$

For balanced designs, the following is taken into account

$$n_l \in \{5, 10, 30\}, l = 1, \dots, d.$$

For unbalanced designs, patterns such as

$$(n_1, n_2, n_3) = (5, 10, 30),$$

when $d = 3$, and similar patterns for $d = 5$ and $d = 10$, for example

$$(5, 10, 10, 20, 30)$$

and

$$(5, 5, 10, 10, 15, 15, 20, 20, 30, 30).$$

For each combination of factors, $B = 50\,000$ Monte Carlo replications are performed.

The nominal significance level is set at

$$\alpha = 0.05.$$

The magnitude of the effect under the alternative hypothesis is defined by

$$\delta \in \{0.25, 0.50, 0.75\}.$$

The case where $\delta = 0$ corresponds to the null hypothesis and is used to estimate the Type I error.

6.4 Test Statistics and Metrics Evaluated

For each Monte Carlo simulation, a one-way ANOVA is fitted to the simulated model. The statistic used is

$$F = \frac{SS_A/(d-1)}{SS_E/(N-d)}, \quad (41)$$

where SS_A is the sum of squares between treatments, SS_E is the sum of residual squares and

$$N = \sum_{l=1}^d n_l. \quad (42)$$

The null hypothesis is rejected when the simulated F statistic exceeds the 0.95 quantile of $F_{d-1, N-d}$

$$F > F_{d-1, N-d; 1-\alpha}. \quad (43)$$

The empirical rejection rate is calculated as

$$\hat{p} = \frac{1}{B} \sum_{b=1}^B I(F_b > F_{d-1, N-d; 1-\alpha}), \quad (44)$$

where F_b is the statistic obtained in replication b .

When $\delta = 0$, it estimates the Type I error. When $\delta > 0$, it estimates the empirical power.

To assess Monte Carlo precision, the standard error of the estimated rate is also calculated:

$$SE_{MC} = \sqrt{\frac{\hat{p}(1-\hat{p})}{B}}. \quad (45)$$

This value is important because it allows us to distinguish a statistically significant difference from a random fluctuation in the simulation. For example, if $B = 50\,000$ and $\hat{p} = 0.05$, then

$$SE_{MC} \approx 0.000975.$$

A simulation-based 95% interval is $\hat{p} \pm 1.96 SE_{MC}$ is $[0.04809, 0.05191]$.

In addition to the rejection rate, the skewness and kurtosis of the residuals in each replicate are estimated

$$\hat{\gamma}_1 = \frac{\hat{\kappa}_3}{\hat{\kappa}_2^2}, \quad (46)$$

and

$$\hat{\gamma}_2 = \frac{\hat{\kappa}_4}{\hat{\kappa}_2^2} \quad (47)$$

The simulation results can therefore be interpreted directly in the light of the cumulants approach.

6.5 Results

Table 1 summarises the simulation design. Table 2, Table 3 and Table 4 in Appendix, report empirical rejection rates. The classification scheme is based on [22], liberal robustness interval $[0.5\alpha, 1.5\alpha]$ which corresponds to $[0.025, 0.075]$ for $\alpha = 0.05$, together with a narrower descriptive central band. The classifications are defined as follows: strongly conservative, $\hat{\alpha} < 0.025$; slightly conservative, $0.025 \leq \hat{\alpha} < 0.045$; reliable, $0.045 \leq \hat{\alpha} < 0.055$; slightly liberal $0.055 < \hat{\alpha} \leq 0.075$; and strongly liberal $\hat{\alpha} > 0.075$. These descriptive labels are applied uniformly across all simulation scenarios. Monte Carlo uncertainty is reported using (45).

Table 1. Experimental design of the simulation study

Simulation Factor	Levels considered
Number of treatments	$d = 3$
Balanced sample size	(10, 10, 10) and (30, 30, 30)
Unbalance d sample size	(5, 10, 30)
Error distribution	$N(0,1), t_3, t_5, \chi_3^2, \text{lognormal}(0,1), \text{Gumbel}$
Nominal level	$\alpha = 0.05$
Monte Carlo replication	$B = 50\,000$
Effect sizes	$\delta = 0.25, 0.50, 0.75$
Evaluated metrics	Type I error ($\hat{\alpha}$), power, $\hat{\gamma}_1, \hat{\gamma}_2, SE_{MC}$

Source: created by the authors.

Under homogeneous variances, the empirical type I error is close to the nominal level under most of the conditions studied. In the small-sized balanced design, t_3 is slightly conservative (0.0415), showing that heavy distributional tails can affect the calibration of finite samples, even in the absence of skewness. The largest deviation for

small samples is observed for the lognormal distribution (0.0324). These effects diminish in the medium-sized balanced design. Under homogeneous but unbalanced sample sizes, all reported rates remain within the robustness interval [22].

The empirical rejection rates increase with delta, as expected. The lognormal distribution exhibits the highest crude rejection rates under the alternative assumptions, but this should not be interpreted as consistently superior test efficiency, as the calibration of zero differs across distributions.

In particular, the zero scenario corresponding to a small balanced lognormal distribution is conservative. Rigorous sensitivity comparisons would therefore require power adjusted for the specific size or calibrated critical values of the distribution. Table 3 (Appendix) should be interpreted only as a description of the crude rejection probabilities under the stated alternative assumptions, and not as evidence that skewness or kurtosis causally improves power.

The following analysis isolates the combined effect of unequal variances between groups and sample size imbalance. Table 4 (Appendix) presents the reproducible results obtained from the simulation design described above.

Three variance-sample-size configurations are compared: homogeneous variances (1,1,1), negative matching (4,2,1), and positive pairings (1,2,4), with respective sample sizes of (5,10,30).

Table 4 (Appendix) shows that variance matching is the predominant factor in calibrating the ANOVA-type F-test in this unbalanced design for finite samples. With homogeneous variances (1, 1, 1), all six distributions are considered reliable, with a Type I empirical error between 0.0484 and 0.0548. With negative pairing, a variance vector of (4, 2, 1), and sample sizes of (5, 10, 30), all distributions are strongly liberal; the empirical values range from 0.1546 for the lognormal distribution to 0.1840 for the normal distribution. With positive pairing, a variance vector of (1, 2, 4), the test is strongly conservative for the normal, t_3 , t_5 , χ_3^2 and Gumbel distributions. The rate for the lognormal distribution is 0.0326 and is considered slightly conservative. Although this value falls within the broad robustness range [22], it remains significantly below the nominal threshold of 0.05 and cannot be considered an exact calibration.

The comparison confirms that non-normality and heteroscedasticity are distinct characteristics whose effects interact. The simulations do not establish a causal cancellation mechanism for the

lognormal distribution. They only show that, for the experimental design and variance vector considered, its conservative skewness is less pronounced than that of other distributions under positive pairing.

5 Conclusion

The simulations confirm a limited and experimentally specific conclusion. Under the assumption of homogeneous variances, the F-test, similar to ANOVA, adequately controls for Type I error in the balanced and unbalanced configurations studied. In the reconstructed and unbalanced Table 4 (Appendix), all six distributions are reliable under the variance vector (1,1,1), with empirical rates ranging from 0.0484 to 0.0548. In the small balanced design of Table 2 (Appendix), the t_3 distribution is slightly conservative, and the lognormal distribution exhibits the most pronounced conservative bias; these biases are less pronounced in the medium-sized balanced sample.

The main flaw arises when unequal variances are combined with unequal group sizes. With sample sizes of (5, 10, 30) and negative pairing (4, 2, 1), the test is strongly liberal for every distribution, with an empirical type I error risk between 0.1546 and 0.1840. With positive pairing (1, 2, 4), it is strongly conservative for five distributions and slightly conservative for the lognormal distribution, with a rate of 0.0326. Thus, the direction of the variance-size pairing has a greater impact on calibration than the choice between the standardised error distributions studied.

The cumulant-based approach remains useful for describing non-Gaussianity and for justifying diagnoses on finite samples. For linear component estimators, their cumulants are inherited from the error cumulants via the estimator weights. However, the heuristic robustness principle does not constitute a theorem providing an exact F-distribution, and the t_3 case lies outside a finite fourth-order cumulant expansion. The conclusions are limited to the component comparison step, the sample sizes, the variance vectors, and the error distributions considered.

Therefore, ANOVA inference in multiple additive models must be accompanied by checks for the approximate equality of covariances, the balance of the experimental design, the adequacy of variance sizes, and the shape of the errors. Future work should extend the simulation to the complete parameter estimation procedure, including covariance estimation and eigenvector selection,

and compare the classical F-statistic with heteroscedasticity-robust alternatives, such as Welch-type and Brown-Forsythe procedures [23], [24].

Declaration of Generative AI and AI-assisted Technologies in the Writing Process The authors wrote, reviewed and edited the content as needed and they have not utilised artificial intelligence (AI) tools. The authors take full responsibility for the content of the publication.

References:

- [1] Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 3(1), 22–38. <https://doi.org/10.2307/3001535>
- [2] Box, G. E. P. (1954a). Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, 25(2), 290–302. <https://doi.org/10.1214/aoms/1177728786>
- [3] Box, G. E. P. (1954b). Some theorems on quadratic forms applied in the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two-way classification. *The Annals of Mathematical Statistics*, 25(3), 484–498. <https://doi.org/10.1214/aoms/1177728717>
- [4] Scheffé, H. (1959). *The analysis of variance*. Wiley. ISBN: 978-0-471-75834-1.
- [5] Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, 29(4), 552–557. <https://doi.org/10.7334/psicothema2016.383>
- [6] Thongteeraparp, A. (2019). The comparison of nonparametric statistical tests for interaction effects in factorial design. *Decision Science Letters*, 8, 309–316. <https://doi.org/10.5267/j.dsl.2018.11.003>
- [7] Biswas, P. Das & N. K. Mandal (2015) Designs Robust against Violation of Normality Assumption in the Standard F-test, *Communications in Statistics - Theory and Methods*, 44:14, 2898-2910. <https://doi.org/10.1080/03610926.2013.799697>
- [8] Nguyen, D., Kim, E., Wang, Y., Pham, T. V.&Chen, Y.H.(2019). Empirical comparison of tests for one-factor ANOVA under heterogeneity and non-normality: A

- Monte Carlo study. *Journal of Modern Applied Statistical Methods*, 18 (2), eP2906. <https://doi.org/10.22237/jmasm/1604190000>.
- [9] Zimmerman, D. W. (2006). Two separate effects of variance heterogeneity on the validity and power of significance tests of location. *Statistical Methodology*, 3(4), 351–374. <https://doi.org/10.1016/j.stamet.2005.10.002>.
- [10] Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42(3), 237–288. <https://doi.org/10.3102/00346543042003237>
- [11] Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research*, 66(4), 579–619. <https://doi.org/10.3102/00346543066004579>
- [12] Antunes, P., Ferreira, S. S., Ferreira, D., & Mexia, J. T. (2021). Multiple additive models. *Communications in Statistics Theory and Methods*, 50(19), 4649–4655. <https://doi.org/10.1080/03610926.2020.1723636>
- [13] Santos, C., Dias, C., Nunes, C., & Mexia, J. T. (2020). On the derivation of complex linear models from simpler ones. In *Proceedings of the 5th NA International Conference on Industrial Engineering and Operations Management*, Detroit, MI, USA (pp. 603–613). IEOM Society International. <http://www.ieomsociety.org/detroit2020/papers/144.pdf> (Accessed September 21, 2026)
- [14] Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. John Wiley & Sons. Print ISBN: 978-0-471-62162-1. Online ISBN: 978-0-470-31685-6. <https://doi.org/10.1002/9780470316856>
- [15] Khuri, A. I., Mathew, T., & Sinha, B. K. (1998). *Statistical tests for mixed linear models*. John Wiley & Sons. ISBN: 978-0-471-15653-6.
- [16] Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. Chapman & Hall. ISBN: 978-0-412-34390-2.
- [17] Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press. ISBN: 978-1-4987-2833-1. <https://doi.org/10.1201/9781315370279>
- [18] Harrar, S. W., & Gupta, A. K. (2007). Asymptotic expansion for the null distribution of the F-statistic in one-way ANOVA under non-normality. *Annals of the Institute of Statistical Mathematics*, 59(3), 531–556. <https://doi.org/10.1007/s10463-006-0055-7>.
- [19] Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In P. R. Krishnaiah (Ed.), *Handbook of Statistics: Analysis of Variance* (Vol. 1, pp. 199–236). North-Holland. ISBN 0-444-85335-9. [https://doi.org/10.1016/S0169-7161\(80\)01009-7](https://doi.org/10.1016/S0169-7161(80)01009-7)
- [20] Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods*, 50, 937–962. <https://doi.org/10.3758/s13428-017-0918-2>
- [21] Box, G. E. P., & Andersen, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(1), 1–26. <https://doi.org/10.1111/j.2517-6161.1955.tb00176.x>
- [22] Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144–152. <https://doi.org/10.1111/j.2044-8317.1978.tb00581.x>
- [23] Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346), 364–367. <https://doi.org/10.1080/01621459.1974.10482955>
- [24] Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking parametric assumptions seriously: Arguments for the use of Welch's F-test instead of the classical F-test in one-way ANOVA. *International Review of Social Psychology*, 32(1), Article 13. <https://doi.org/10.5334/irsp.198>

APPENDIX

Table 2. Type I error rate under approximate homoscedasticity

Error distribution	Design	(n_1, n_2, n_3)	$\hat{\alpha}$	SE_{MC}	$\hat{\gamma}_1$	$\hat{\gamma}_2$	Classification
Normal	Small balanced	(10, 10, 10)	0,0494	0.0010	0.000	0.000	Reliable
t_3	Small balanced	(10, 10, 10)	0.0415	0.0009	0.000	∞	Slightly Conservative
t_5	Small balanced	(10, 10, 10)	0.0474	0.0010	0.000	6.000	Reliable
χ_3^2	Small balanced	(10, 10, 10)	0.0450	0.0009	1.633	4.000	Reliable
Lognormal (0, 1)	Small balanced	(10, 10, 10)	0.0324	0.0008	6.185	110.936	Slightly Conservative
Gumbel	Small balanced	(10, 10, 10)	0.0479	0.0010	1.140	2.400	Reliable
Normal	Medium balanced	(30, 30, 30)	0.0495	0.0010	0.000	0.000	Reliable
t_3	Medium balanced	(30, 30, 30)	0.0455	0.0009	0.000	∞	Reliable
t_5	Medium balanced	(30, 30, 30)	0.0484	0.0010	0.000	6.000	Reliable
χ_3^2	Medium balanced	(30, 30, 30)	0.0481	0.0010	1.633	4.000	Reliable
Lognormal (0, 1)	Medium balanced	(30, 30, 30)	0.0412	0.0009	6.185	110.936	Slightly Conservative
Gumbel	Medium balanced	(30, 30, 30)	0.0493	0.0010	1.140	2.400	Reliable
Normal	Unbalanced	(5, 10, 30)	0.0482	0.0010	0.000	0.000	Reliable
t_3	Unbalanced	(5, 10, 30)	0.0542	0.0010	0.000	∞	Reliable
t_5	Unbalanced	(5, 10, 30)	0.0531	0.0010	0.000	6.000	Reliable
χ_3^2	Unbalanced	(5, 10, 30)	0.0483	0.0010	1.633	4.000	Reliable
Lognormal (0, 1)	Unbalanced	(5, 10, 30)	0.0548	0.0010	6.185	110.936	Reliable
Gumbel	Unbalanced	(5, 10, 30)	0.0498	0.0010	1.140	2.400	Reliable

Note: ∞ indicates that the fourth moment is not finite for this distribution

Source: created by the authors.

Table 3. Empirical power in a small balanced design with $d = 3, (10, 10, 10)$

Error distribution	$\delta = 0.25$	$\delta = 0.50$	$\delta = 0.75$	Comment
Normal	0.0766	0.1764	0.3526	Reference / Baseline
t_3	0.0879	0.2370	0.4623	Higher sensitivity to extreme values
t_5	0.0777	0.1917	0.3836	Slightly higher than the normal baseline
χ_3^2	0.0801	0.192	0.3886	Moderately higher power
Lognormal (0, 1)	0.0986	0.3020	0.5461	Highest power, but with a conservative Type I error rate
Gumbel	0.0769	0.1871	0.3714	Approximating the normal baseline

Source: created by the authors.

Table 4. Empirical Type I error under unbalanced sample sizes and explicit variance-sample-size pairings ($B = 50000$; $\alpha = 0.05$).

Error distribution	Variance size Pairing	(n_1, n_2, n_3)	Variance vector	$\hat{\alpha}$	SE_{MC}	95% MC interval	Classification
Normal	Homogeneous	(5, 10, 30)	(1, 1, 1)	0.0509	0.0010	[0.0490, 0.0528]	Reliable
t_3	Homogeneous	(5, 10, 30)	(1, 1, 1)	0.0531	0.0010	[0.0512, 0.0551]	Reliable
t_5	Homogeneous	(5, 10, 30)	(1, 1, 1)	0.0511	0.0010	[0.0492, 0.0530]	Reliable
χ_3^2	Homogeneous	(5, 10, 30)	(1, 1, 1)	0.0484	0.0010	[0.0466, 0.0503]	Reliable
Lognormal (0, 1)	Homogeneous	(5, 10, 30)	(1, 1, 1)	0.0548	0.0010	[0.0528, 0.0568]	Reliable
Gumbel	Homogeneous	(5, 10, 30)	((1, 1, 1)	0.0491	0.0010	[0.0472, 0.0510]	Reliable
Normal	Negative pairing	(5, 10, 30)	(4, 2, 1)	0.1840	0.0017	[0.1806, 0.1874]	Strongly liberal
t_3	Negative pairing	(5, 10, 30)	(4, 2, 1)	0.1656	0.0017	[0.1624, 0.1689]	Strongly liberal
t_5	Negative pairing	(5, 10, 30)	(4, 2, 1)	0.1732	0.0017	[0.1699, 0.1765]	Strongly liberal
χ_3^2	Negative pairing	(5, 10, 30)	(4, 2, 1)	0.1775	0.0017	[0.1742, 0.1808]	Strongly liberal
Lognormal (0, 1)	Negative pairing	(5, 10, 30)	(4, 2, 1)	0.1546	0.0016	[0.1514, 0.1578]	Strongly liberal
Gumbel	Negative pairing	(5, 10, 30)	(4, 2, 1)	0.1798	0.0017	[0.1764, 0.1831]	Strongly liberal
Normal	Positive pairing	(5, 10, 30)	(1, 2, 4)	0.0097	0.0004	[0.0088, 0.0105]	Strongly conservative
t_3	Positive pairing	(5, 10, 30)	(1, 2, 4)	0.0136	0.0005	[0.0126, 0.0147]	Strongly conservative
t_5	Positive pairing	(5, 10, 30)	(1, 2, 4)	0.0108	0.0005	[0.0099, 0.0117]	Strongly conservative
χ_3^2	Positive pairing	(5, 10, 30)	(1, 2, 4)	0.0159	0.0006	[0.0148, 0.0170]	Strongly conservative
Lognormal (0, 1)	Positive pairing	(5, 10, 30)	(1, 2, 4)	0.0326	0.0008	[0.0311, 0.0342]	Slightly conservative
Gumbel	Positive pairing	(5, 10, 30)	(1, 2, 4)	0.0121	0.0005	[0.0111, 0.0131]	Strongly conservative

Source: created by the authors.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Cristina Dias contributed to the conceptualisation of the study, literature review, simulation design, interpretation of results, and manuscript preparation. Carla Santos contributed to the methodological development, formal analysis, validation of the simulation framework, interpretation of results, and critical revision of the manuscript. João Tiago Mexia contributed to the theoretical formulation of the multiple additive model framework, mathematical validation, supervision, and critical revision of the manuscript. All authors have read and approved the final version of the manuscript.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This work was partially supported by the Portuguese Foundation for Science and Technology (FCT –Fundação para a Ciência e a Tecnologia, I.P.) under UID/00297/2025 (<https://doi.org/10.54499/UID/00297/2025>) and UID/PRR/00297/2025 (<https://doi.org/10.54499/UID/PRR/00297/2025>) (Center for Mathematics and Applications - NOVA Math, SST NOVA University of Lisbon),

Conflict of Interest

The authors declare no conflict of interest.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0 https://creativecommons.org/licenses/by/4.0/deed.en_US