Received: 24 November 2020 | Revised: 18 May 2021 | Accepted: 19 May 2021

DOI: 10.1002/sim.9092

RESEARCH ARTICLE

Statistics in Medicine WILEY

Analysis of covariance under variance heteroscedasticity in general factorial designs

Frank Konietschke[1,2] | Cong Cao[3] | Asanka Gunawardana[1,2] | Georg Zimmermann[4,5,6]

[1] Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute of Biometry and Clinical Epidemiology, Berlin, Germany

[2] Berlin Institute of Health (BIH), Berlin, Germany

[3] PPD Development, Hamilton, New Jersey, USA

[4] Team Biostatistics and Big Medical Data, IDA Lab Salzburg, Paracelsus Medical University, Salzburg, Austria

[5] Department of Mathematics, Paris-Lodron-University of Salzburg, Salzburg, Austria

[6] Department of Neurology, Christian Doppler University Hospital, Paracelsus Medical University, Salzburg, Austria

Correspondence
Frank Konietschke,
Charité-Universitätsmedizin Berlin,
Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute of Biometry and Clinical Epidemiology,
Charitéplatz 1, 10117 Berlin, Germany.
Email: frank.konietschke@charite.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: KO 4680/3-2; WISS 2025, Grant/Award Number: IDA-Lab Salzburg

Adjusting for baseline values and covariates is a recurrent statistical problem in medical science. In particular, variance heteroscedasticity is non-negligible in experimental designs and ignoring it might result in false conclusions. Approximate inference methods are developed to test null hypotheses formulated in terms of adjusted treatment effects and regression parameters in general analysis of covariance designs with arbitrary numbers of factors. Variance homoscedasticity is not assumed. The distributions of the test statistics are approximated using Box-type approximation methods. Extensive simulation studies show that the procedures are particularly suitable when sample sizes are rather small. A real data set illustrates the application of the methods.

KEYWORDS

ANCOVA, ANOVA-type statistic, Box-type approximation, experimental designs

1 | INTRODUCTION

General linear models play a major role in experimental sciences and especially in both preclinical (including translational) and clinical research. Hereby, not only the factor levels and their combinations might impact the response variable, but also other variables called *covariates*. For instance, baseline values, age, gender, and/or body weight might obscure the factor effects of the independent groups. *Analysis of covariance* (ANCOVA) is a general linear model that blends *analysis of variance* (ANOVA) and *regression* methods, and hence makes estimation of (adjusted) treatment effects as well as testing hypotheses formulated in terms of the (adjusted) treatment effects in such designs possible.[1] However, the performance of the methods (F-tests) with respect to maintaining the nominal type-1 error rate heavily depends on whether the data fulfill model assumptions, such as multivariate normality and homogeneous variances of the error term. Indeed, if the data are not in line with them, the methods tend to be liberal or conservative, depending on the amount of variance heteroscedasticity, sample size allocations, and general distributional shapes (see Section 5 for details). In statistical practice, however, verifying such assumptions is quasi impossible, especially when being confronted with small sample sizes.[2,3] We therefore prefer to use less stringent statistical methods for making inferences. The present article aims to introduce statistical inference methods for general ANCOVA designs with arbitrary (but fixed) numbers of factors and covariates without assuming multivariate normality and homoscedastic variances.

In recent years, several authors have proposed different methods to tackle the problem of variance heteroscedasticity in general ANOVA designs (without covariates). Pauly et al[4] provide a detailed overview of the different solutions and highlight the so-called *Wald-type statistic* (along with a permutation version) as well as the *ANOVA-type statistic* developed by Brunner et al.[5] In comparison, the Wald-type statistic is applicable for large samples only ($n_i \approx 50$), while the ANOVA-type statistic controls the type-1 error rate accurately even when sample sizes are rather small ($n_i \approx 10$). However, both of the statistics found their way into statistical practice and are implemented in prominent software packages like `SAS PROC MIXED` and in the `R`-package `GFD`.[6] For the analysis of general ANCOVA designs, Zimmermann et al[7] propose a Wald-type test along with a resampling version (wild-bootstrap approach) using *Heteroscedasticity Consistent Standard Error* (HCSE) estimators.[8-11] The methods even allow for complete variance heteroscedasticity, that is, every unit might have a different variance. Even though resampling methods share the advantages of being suitable methods particularly for small sample sizes, they might be numerically cumbersome and show their limitations in statistical planning purposes, in general. Statistical models that allow for complete heteroscedasticity might be very flexible and less stringent, however, in medicine and related sciences, reporting groupwise variance estimates is necessary, especially in translational research. We therefore focus on groupwise variance heteroscedasticity and discuss methods for complete heteroscedasticity as an aside. `SAS PROC MIXED` realizes an ANOVA-type statistic in such general ANCOVA designs by replacing the empirical means and variance estimators as used by Brunner et al[5] in ANOVA designs with estimators obtained from *Minimum Variance Quadratic Unbiased Estimation* (MIVQUE0)[12] algorithms in the general ANCOVA framework (see https://support.sas.com/documentation/onlinedoc/stat/141/mixed.pdf). The supporting site says *"This generalizes results in the appendix of Brunner, Dette, and Munk (1997) to a broader class of models,"* see https://support.sas.com/kb/24/516.html. The results, however, are only available in `SAS PROC MIXED`. We investigate whether the methods maintain the nominal type-1 error rate (and power) in extensive simulation studies and find that the methods make accurate inferences in fixed effects, while the tests for the covariate effects tend to be very liberal. In this article, we will therefore generalize the results and propose novel test statistics. We develop unbiased estimators of the variance components using methods of moments and approximate the distributions of the test statistics using Box-type approximation methods.[13,14] Numerical investigations and extensive simulation studies show that (1) the tests for the fixed effects are very similar and of similar quality as their siblings in `SAS` and control the type-1 error rate accurately, while (2) the new approximation for testing the effects of the covariates outperforms its competitor in `SAS`. Especially for skewed data, it turns out the ANOVA-type statistics seem to control the type-1 error rate better than the wild-bootstrap test.[7] We note that HCSE estimators are also available in `SAS PROC MIXED` using the *EMPIRICAL* option.

The remainder of the article is organized as follows. In Section 2, a factorial toxicological and carcinogenic study is discussed. The statistical model, point estimators, and hypotheses of interest are introduced in Section 3. In the following Section 4, procedures for testing the aforementioned hypotheses are explained. Their behavior in small sample size situations is investigated in extensive simulation studies in Section 5. The article closes with the evaluation of the data in Section 6 and a discussion about the results in Section 7. Technical derivations are provided in the Appendix.

Throughout the article we use the following notation. $\mathbf{1}_n$ denotes the $n \times 1$ vector of 1's and $\boldsymbol{I}_a$ the $a \times a$ unit matrix. The $a \times a$ centering matrix is $\boldsymbol{P}_a = \boldsymbol{I}_a - \frac{1}{a}\mathbf{1}_a\mathbf{1}_a'$. Furthermore, $\boldsymbol{A} \bigoplus \boldsymbol{B}$ and $\boldsymbol{A} \otimes \boldsymbol{B}$ denote the direct sum (block operator) and the Kronecker product of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively.

2 | A MOTIVATING EXAMPLE

As a motivating example, we consider a part of the toxicological study[15] on pyridine (number C55301B) obtained from the US National Toxicology Program (see https://manticore.niehs.nih.gov/cebssearch/test_article/110-86-1; assessed November 2020). This animal testing aims to investigate the impact of pyridine (a basic heterocyclic organic compound) on different clinical chemistry parameters. Here, we chose blood urea nitrogen (BUN) measured in mg/dL as the response variable. Initially, $N = 120$ rats (60 male and 60 female) were randomized to six different dose levels of pyridine (0, 50, 100, 250, 500, or 1000 ppm) ($n_{ij} = 10$ animals per dose) and BUN was measured at baseline and after 90 days. In addition, we also chose the change in body weight (in g) as a covariate. The data can thus be described in means of a two-way factorial design involving the factors *gender* with two levels and the factor *dose* with six levels with two covariates. We aim to estimate treatment effects of pyridine adjusted for baseline and change in body weight and to test whether the two factors and their interaction impact the results. We note that one male and two females in dose level 1000 died after the treatment, which makes the design imbalanced. Boxplots of BUN stratified by gender and pyridine concentration levels are displayed in Figure 1.

It can be seen from Figure 1 that the factor gender seems to have an impact on BUN for each pyridine concentration. It also appears that the effect of the pyridine concentration is homogeneous across gender (ie, no interaction effect). Since the sample sizes of the study are rather small, making any distributional assumptions (eg, normality) would be doubtful. The boxplots in Figure 1 indicate that the BUN data are not necessarily symmetrically distributed. Moreover, the empirical variances of each factor-level combinations of the two factors are rather different and thus, assuming variance homoscedasticity is unlikely. Therefore, we will analyze this factorial study without either assuming any specific distribution or homogeneous variances of the data using the methodologies presented in this article. First, a general ANCOVA model, point estimators, and hypotheses of interest will be introduced in the next section.

3 | STATISTICAL MODEL, HYPOTHESES, AND POINT ESTIMATORS

We consider a general ANCOVA model

$$Y_{ij} = b_i + \sum_{\ell=1}^{M} p_\ell M_{ij}^{(\ell)} + \epsilon_{ij}, \quad i = 1, \ldots, a; \quad j = 1, \ldots, n_i,$$

where Y_{ij} and $M_{ij}^{(\ell)}$ denote the response and ℓth covariate value from subject j under condition (treatment) i, b_i represents the fixed treatment effect of condition i, p_ℓ the ℓth regression parameter, and ϵ_{ij} the error term with $E(\epsilon_{i1}) = 0$

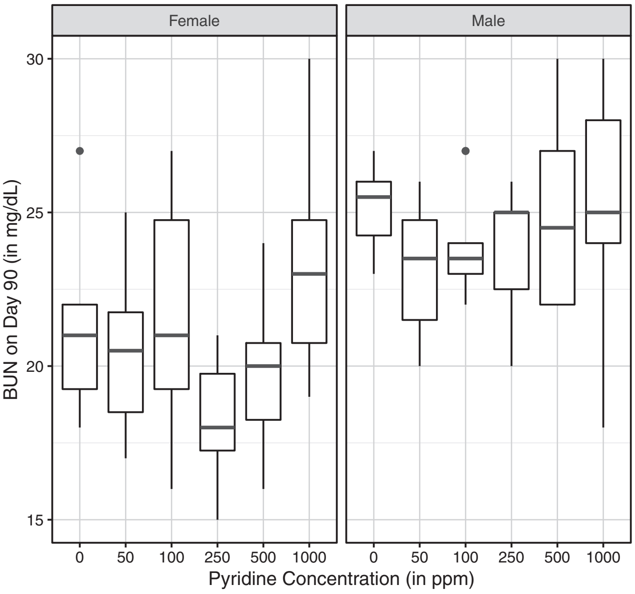

FIGURE 1 Boxplots of blood urea nitrogen (BUN) on day 90 stratified by the two factors: gender and pyridine concentration

and $Var(\epsilon_{i1}) = \sigma_i^2$, $i = 1, \dots, a$. In total, $N = \sum_{i=1}^{a} n_i$ subjects are enrolled in the trial, where n_i denotes the number of subjects in group $i \in \{1, \dots, a\}$. We note that two- and higher-way layouts can be modeled by subindexing the index i. In matrix notation, the model can be written as

$$\boldsymbol{Y} = \boldsymbol{Xb} + \boldsymbol{Mp} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{Y}$ denotes the $N \times 1$ response vector, $\boldsymbol{X} = \bigoplus_{i=1}^{a} \boldsymbol{1}_{n_i}$ the design matrix for the vector of fixed treatment effects $\boldsymbol{b} = (b_1, \dots, b_a)'$, $\boldsymbol{M} = (M_{ij}^{(\ell)})$ a $N \times M$ matrix of covariates, $\boldsymbol{p} = (p_1, \dots, p_M)'$ the $M \times 1$ vector of regression parameters, and $\boldsymbol{\epsilon}$ the vector of independent errors with

$$E(\boldsymbol{\epsilon}) = \boldsymbol{0} \text{ and } Var(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma} = \bigoplus_{i=1}^{a} \sigma_i^2 \boldsymbol{I}_{n_i}, \tag{2}$$

respectively. We furthermore assume the existence of fourth moments, that is, $E(\epsilon_{ij}^4) < \infty$. Now, the aim of ANCOVA is estimating and making inference across the components of the vectors $\boldsymbol{b}$ and $\boldsymbol{p}$, that is, testing the null hypotheses

$$H_0^{(b)} : \boldsymbol{Hb} = \boldsymbol{0} \quad \text{vs} \quad H_1^{(b)} : \boldsymbol{Hb} \neq \boldsymbol{0}, \text{ and} \tag{3}$$

$$H_0^{(p_\ell)} : p_\ell = 0 \quad \text{vs} \quad H_1^{(p_\ell)} : p_\ell \neq 0, \ell = 1, \dots, M. \tag{4}$$

Which hypothesis matrix $\boldsymbol{H}$ to chose depends on the actual model and research question of interest. For example, in a one-way design with a levels, $\mathbf{H} = \mathbf{P}_a$. In a two-way design involving a factor A with a levels and a factor B with b levels the null hypotheses of no main effect A, no main-effect B and no-interaction (AB) between A and B are tested using the contrast matrices

$$\begin{aligned}
H_0^{(b)}(A) &: \boldsymbol{H}_A\boldsymbol{b} = \left(\boldsymbol{P}_a \otimes \frac{1}{b}\boldsymbol{1}_b'\right)\boldsymbol{b} = \boldsymbol{0} \quad \text{or} \quad \overline{b}_{1\cdot} = \dots = \overline{b}_{a\cdot}, \quad \overline{b}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b} b_{ij}, \\
H_0^{(b)}(B) &: \boldsymbol{H}_B\boldsymbol{b} = \left(\frac{1}{a}\boldsymbol{1}_a' \otimes \boldsymbol{P}_b\right)\boldsymbol{b} = \boldsymbol{0} \quad \text{or} \quad \overline{b}_{\cdot 1} = \dots = \overline{b}_{\cdot b}, \quad \overline{b}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a} b_{ij}, \\
H_0^{(b)}(AB) &: \boldsymbol{H}_{AB}\boldsymbol{b} = (\boldsymbol{P}_b \otimes \boldsymbol{P}_b)\,\boldsymbol{b} = \boldsymbol{0} \quad \text{or} \quad \overline{b}_{ij} = \overline{b}_{i\cdot} + \overline{b}_{\cdot j} - \overline{b}_{\cdot\cdot}, \quad \overline{b}_{\cdot\cdot} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} b_{ij}
\end{aligned} \tag{5}$$

as known from linear model theory. Here, b_{ij} denotes the effect of cell (i,j). For more details we refer to, for example, Pauly et al.[4] We estimate the unknown model parameters $\boldsymbol{b}$ and $\boldsymbol{p}$ using methods of least squares and obtain

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{Y} - \boldsymbol{M}\hat{\boldsymbol{p}}) \quad \text{and} \quad \hat{\boldsymbol{p}} = (\boldsymbol{M}'\boldsymbol{Q}\boldsymbol{\Sigma}_*^{-1}\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{Q}\boldsymbol{\Sigma}_*^{-1}\boldsymbol{Y}. \tag{6}$$

Here, $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{P}$ denote the projection matrices and $\boldsymbol{\Sigma}_*$ denotes a suitable regular matrix, for example, ordinary least squares (OLS) estimators are obtained using $\boldsymbol{\Sigma}_* = \boldsymbol{I}_N$ and generalized least squares (GLS) estimators using $\boldsymbol{\Sigma}$ if it is known, or a consistent estimator thereof, respectively. Different estimation methods of $\boldsymbol{\Sigma}$ will be discussed in the next subsection.

Remark 1. The general model can be generalized to model interactions between the treatment effect and covariates. For instance, we aim to model an interaction term between group a and the ℓth covariate. In this case the regression part $\boldsymbol{Mp}$ is kept in the model as above whereas the fixed treatment effect $\boldsymbol{Xb}$ is written as a regression model for qualitative predictors by

$$\boldsymbol{Y} = \boldsymbol{X}^{\odot}\boldsymbol{b}^{\odot} + \boldsymbol{M}^{(\ell)}b_{a\ell} + \boldsymbol{Mp} + \boldsymbol{\epsilon}. \tag{7}$$

Here, $\boldsymbol{X}^{\odot}$ denotes the adjusted design matrix, $\boldsymbol{b}^{\odot} = (b_1, \dots, b_{a-1})'$ the vector of the treatment effects of groups $1, \dots, a-1$ and $b_{a\ell}$ denotes the interaction effect of treatment a and the ℓth covariate, respectively.

3.1 | Estimation of the variances

For the ease of representation, we rewrite the estimators $\widehat{\boldsymbol{b}}$ and $\widehat{\boldsymbol{p}}$ using appropriate generating matrices

$$\begin{aligned}\boldsymbol{A} &= (\boldsymbol{M}'\boldsymbol{Q}\boldsymbol{\Sigma}_*^{-1}\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{Q}\boldsymbol{\Sigma}_*^{-1} \quad \text{and}\\ \boldsymbol{D} &= (\boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1} - (\boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{M}\boldsymbol{A}\end{aligned} \tag{8}$$

as $\widehat{\boldsymbol{b}} = \boldsymbol{D}\boldsymbol{Y}$ and $\widehat{\boldsymbol{p}} = \boldsymbol{A}\boldsymbol{Y}$, respectively. For large sample sizes, that is, if $N \to \infty$ such that $N/n_i \to \lambda_i$ (and under some mild regulatory assumptions, see Zimmermann et al[7]), it can be shown that

$$\begin{aligned}\sqrt{N}(\widehat{\boldsymbol{b}} - \boldsymbol{b}) &\stackrel{\cdot}{\sim} N(\boldsymbol{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\Psi} = N\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}' \quad \text{and}\\ \sqrt{N}(\widehat{\boldsymbol{p}} - \boldsymbol{p}) &\stackrel{\cdot}{\sim} N(\boldsymbol{0}, \boldsymbol{\Xi}), \quad \boldsymbol{\Xi} = N\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}',\end{aligned}$$

where $\boldsymbol{\Sigma}$ is as given in (2). Both of these matrices are, however, unknown in practical applications and must be estimated from the data. Different methods for the estimation of $\boldsymbol{\Sigma}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Xi}$ are available and include bootstrap, maximum likelihood (ML and REML), MIVQUE0 as well as methods of moments.[16] Clearly, each estimation method has its own advantages and disadvantages, but beyond that, the resulting groupwise variance estimators depend on the outcomes in the other groups using any of them. For instance, in a two-sample design, the estimators of σ_1^2 and σ_2^2 would change when a third group (independent of the others) was included in the model (without changing any value in the initial groups). This is not the case when covariates are not included in the model. The variance components $\sigma_1^2, \dots, \sigma_a^2$ are model constants and we therefore prefer to estimate them on a group-specific level using methods of moments. We follow the idea from Cao et al[17] and estimate them using the corresponding submodels. Let $\boldsymbol{X}_i = \boldsymbol{1}_{n_i}$ denote the $n_i \times 1$ vector of 1's and let $\boldsymbol{M}_i$, $i = 1, \dots, a$ denote the matrices of the covariates for each group separately. Furthermore, let $\boldsymbol{B}_i = (\boldsymbol{X}_i \,\vdots\, \boldsymbol{M}_i)$ denote the a partitioned matrices of $\boldsymbol{X}_i$ and the corresponding covariates $\boldsymbol{M}_i$, and define the projection matrices

$$\boldsymbol{Q}_i = \boldsymbol{I}_{n_i} - \boldsymbol{B}_i(\boldsymbol{B}_i'\boldsymbol{B}_i)^{-1}\boldsymbol{B}_i'.$$

Then, unbiased and consistent estimators of the variances σ_i^2 are given by

$$\widehat{\sigma}_i^2 = \boldsymbol{Y}_i'\boldsymbol{Q}_i\boldsymbol{Y}_i/(n_i - 1 - rank(\boldsymbol{M}_i)), \quad i = 1, \dots, a. \tag{9}$$

Unbiased estimators of the matrices $\boldsymbol{\Sigma}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Xi}$ will be provided in the next corollary.

Corollary 1. *Let $\widehat{\sigma}_i^2$ as given in (9). Then, $\widehat{\boldsymbol{\Sigma}} = \bigoplus_{i=1}^a \widehat{\sigma}_i^2 \boldsymbol{I}_{n_i}$ is an unbiased and consistent estimator of $\boldsymbol{\Sigma}$. Furthermore, $\widehat{\boldsymbol{\Psi}} = N\boldsymbol{D}\widehat{\boldsymbol{\Sigma}}\boldsymbol{D}'$ and $\widehat{\boldsymbol{\Xi}} = N\boldsymbol{A}\widehat{\boldsymbol{\Sigma}}\boldsymbol{A}'$ are unbiased and consistent estimators of $\boldsymbol{\Psi}$ and $\boldsymbol{\Xi}$, respectively.*

We note that $\widehat{\sigma}_i^2$ is the ”classical” variance estimator for each group-specific submodel and therefore the result follows, see Cao et al.[17] Both the estimators $\widehat{\boldsymbol{b}}$ and $\widehat{\boldsymbol{p}}$ of the treatment effects as well as their consistent variance-covariance matrix estimators can now be used for the derivation of statistical procedures. This will be explained in the next section.

4 | THE ANCOVA-TYPE STATISTIC

In this section, test procedures for testing the null hypotheses $H_0^{(b)}$ as well as $H_0^{(p_\ell)}$ as given in (3) and (4) will be introduced. In general ANOVA and ANCOVA models, test procedures are mainly quadratic forms in the estimators (scaled with variance-covariance estimators). Let $\boldsymbol{H}$ denote a suitable hypothesis matrix for testing $H_0^{(b)} : \boldsymbol{H}\boldsymbol{b} = \boldsymbol{0}$ given in (3). For large sample sizes, a Wald-type statistic

$$Q_N(\boldsymbol{H}) = N\, \widehat{\boldsymbol{b}}'\boldsymbol{H}'\left(\boldsymbol{H}\widehat{\boldsymbol{\Psi}}\boldsymbol{H}'\right)^{-}\boldsymbol{H}\widehat{\boldsymbol{b}}$$

for testing $H_0^{(b)}$ is readily available. Here, $(\boldsymbol{B})^-$ denotes a generalized inverse of the matrix $\boldsymbol{B}$. For large sample sizes, $Q_N(\boldsymbol{H})$ follows a χ_r^2 distribution with $r = rank(\boldsymbol{H})$ degrees of freedom. The Wald-type statistic is also numerically available in SAS PROC MIXED using the CHISQ option within the model statement. Furthermore, F-tests are computed using estimated degrees of freedom. These statistics, however, should only be applied when sample sizes are large. In the present article, we focus on small sample sizes and we therefore emphasize the ANOVA-type statistic developed by Brunner et al.[5] In their original paper, the authors allow for mean comparisons only (ie, no covariates are involved in the model). However, SAS PROC MIXED implements an ANOVA-type statistic within the ANOVAF option. The statistics shall now be explained.

4.1 | ANOVAF Option in SAS PROC MIXED

A version of the ANOVA-type statistic for testing the aforementioned null hypotheses is numerically available in SAS PROC MIXED using the ANOVAF and MIVQUE0 options. Computational details are not provided, but the general procedure is explained in available handbooks, for example, on https://support.sas.com/kb/24/516.html. The following revisits its statements: "Let $\boldsymbol{H}$ denote the matrix of estimable functions for the hypothesis $H : \boldsymbol{Hb} = \boldsymbol{0}$ and let $\boldsymbol{T} = \boldsymbol{H}'(\boldsymbol{HH}')^-\boldsymbol{H}$ and let $\boldsymbol{C}$ denote the estimated variance-covariance matrix of $(\widehat{\boldsymbol{b}} - \boldsymbol{b})$ (see the section 'Statistical Properties' in the PROC MIXED documentation for the construction of $\boldsymbol{C}$). The ANOVAF F-statistics are computed as

$$F_A = \widehat{\boldsymbol{b}}'\boldsymbol{T}\widehat{\boldsymbol{b}}/t_1. \tag{10}$$

Note that this test is a modification of the usual F-statistic where $(\boldsymbol{HCH}')^-$ is replaced with $t_1 = tr(\boldsymbol{TC})$; see, for example Brunner, Domhof, and Langer (2002, sec. 5.4). The p-values for this statistic are computed from either an F_{ν_1,ν_2} or an $F_{\nu_1,\infty}$ distribution. The respective degrees of freedom are determined by the MIXED procedure as follows:

$$\nu_1 = \frac{t_1^2}{tr(\boldsymbol{TCTC})}, \quad \nu_2^* = \frac{2t_1^2}{\boldsymbol{g}'\boldsymbol{Ag}}, \quad \nu_2 = \begin{cases} \max\left\{\min\left\{\nu_2^*, df_\epsilon\right\}, 1\right\} & \boldsymbol{g}'\boldsymbol{Ag} > 1E3 \times MACEPS \\ 1 & otherwise \end{cases}$$

The term $\boldsymbol{g}'\boldsymbol{Ag}$ in the term ν_2^* for the denominator degrees of freedom is based on approximating $Var(tr(\boldsymbol{TC}))$ based on a first-order Taylor series about the true covariance parameters. This generalizes results in the appendix of Brunner, Dette, and Munk (1997) to a broader class of models. The vector $\boldsymbol{g} = (g_1, \dots, g_q)'$ contains the partial derivatives

$$tr\left(\boldsymbol{H}'(\boldsymbol{HH}')^-\boldsymbol{H}\frac{\partial \boldsymbol{C}}{\partial \theta_i}\right)$$

and $\boldsymbol{A}$ is the asymptotic variance-covariance matrix of the covariance parameter estimates (ASYCOV option in the PROC MIXED statement). PROC MIXED reports ν_1 and ν_2 as NumDF and DenDF under the ANOVA F heading in the output. The corresponding p-values are denoted as $Pr > F(DDF)$ for F_{ν_1,ν_2} and $Pr > F(infty)$ for $F_{\nu_1,\infty}$, respectively. P-values that are computed with the ANOVAF option can be identical to the nonparametric tests in Akritas, Arnold, and Brunner (1997) and in Brunner, Domhof, and Langer (2002), provided that the response data consists of properly created (and sorted) ranks and that the covariance parameters are estimated by MIVQUE0 in models with the REPEATED statement and properly chosen SUBJECT= and/or GROUP= effects."

Up to now, the statistical procedure is available in SAS only. Replicating the computational steps is a very challenging task, especially since some computational details are provided in a somewhat cryptic manner. Furthermore, as mentioned above, the group-specific variance estimators depend on outcomes in other groups. The detailed code for the computation of the statistic is as follows

```
PROC MIXED DATA =... METHOD = MIVQUE0 ANOVAF;
CLASS condition;
MODEL y = condition covariate1...covariateM;
REPEATED / TYPE=UN(1) GROUP=condition;
RUN;
```

The usage of the REPEATED statement is necessary for the groupwise estimation of the variances and should not be confused with "repeated measures". The repeated statement allows modeling of the structure of the variance-covariance matrix of the error term. When no covariates are available (ie, mean comparisons only), the ANOVA-type statistic is also implemented in the R-package GFD.[6]

4.2 | A novel approach

In the following, a novel ANCOVA-type statistic based on the point estimators $\widehat{\boldsymbol{b}}$ and $\widehat{\boldsymbol{p}}$ as well as their unbiased and consistent variance-covariance estimators $\widehat{\boldsymbol{\Psi}}$ and $\widehat{\boldsymbol{\Xi}}$ will be introduced. First, procedures for testing $H_0^{(b)}$ will be presented. Methods for testing the impact of the covariates will be obtained afterward in a similar way. Analogously to the ANOVAF realization in SAS, let $\boldsymbol{T} = \boldsymbol{H}'(\boldsymbol{H}\boldsymbol{H}')^{-}\boldsymbol{H}$ denote the projection onto the column space of $\boldsymbol{H}$ and consider the test statistic

$$F_N(\boldsymbol{T}) = \frac{N}{tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}})}\widehat{\boldsymbol{b}}'\boldsymbol{T}\widehat{\boldsymbol{b}}. \tag{11}$$

Note that $F_N(\boldsymbol{T})$ is very similar to F_A as given in (10), with the exception of using different variance estimators. The next proposition introduces an approximation of its asymptotic distribution.

Proposition 1. *Let* $\boldsymbol{D}$ *be as given in (8) and define the matrix* $\boldsymbol{K} = \boldsymbol{H}\boldsymbol{D} = [k_{ij}]_{i=1,\dots,a}^{j=1,\dots,N}$. *Furthermore, let* $\boldsymbol{D}_K = diag\{K_1, \dots, K_a\}$ *and* $\boldsymbol{D}_{\widehat{\sigma}} = diag\{\widehat{\sigma}_1^2, \dots, \widehat{\sigma}_a^2\}$ *denote the diagonal matrices with diagonal elements* $K_i = \sum_{j=1}^{n_i} k_{ij}^2$ *and* $\widehat{\sigma}_i^2$, $i = 1, \dots, a$, *respectively. Furthermore, let* $\boldsymbol{\Omega} = diag\{n_1 - 1 - rank(\boldsymbol{M}_1), \dots, n_a - 1 - rank(\boldsymbol{M}_a)\}$ *denote the diagonal matrix with elements obtained from the denominators of* $\widehat{\sigma}_i^2$. *Then, under* $H_0^{(b)}$, *the distribution of* $F_N(\boldsymbol{T})$ *as given in (11) can be approximated by a central* $F(\widehat{f}_1, \widehat{f}_2)$ *distribution, where*

$$\widehat{f}_1 = \frac{[tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}})]^2}{tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}}\boldsymbol{T}\widehat{\boldsymbol{\Psi}})} \quad \text{and} \quad \widehat{f}_2 = \frac{[tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}})]^2}{tr(\boldsymbol{D}_K^2\boldsymbol{D}_{\widehat{\sigma}}^2\boldsymbol{\Omega})}.$$

The derivation of the approximation is given in the appendix. Test procedures for testing the null hypothesis $H_0^{(p_\ell)}$: $p_\ell = 0$ (or any linear combination of the regression parameter as $H_0^{(p)}$: $\boldsymbol{c}'\boldsymbol{p} = 0$) are obtained in the same way. Here, $\boldsymbol{H}$ is replaced with the ℓth unit vector, $\widehat{\boldsymbol{\Psi}}$ is replaced with $\widehat{\boldsymbol{\Xi}}$, and the diagonal elements of $\boldsymbol{D}_K$ are obtained from the ℓth row of the generating matrix $\boldsymbol{A}$, respectively.

5 | SIMULATION RESULTS

All of the methods presented in this article are of asymptotic nature and we will therefore examine their behavior in small sample size situations with the help of extensive simulation studies (each with 10 000 simulation runs). We will use their control of the nominal type-1 error rate ($\alpha = 5\%$) as well as their powers to detect selected alternatives as quality criteria. Due to the abundance of possible general designs (numbers of factors), sample size/variance allocations, numbers of covariates, and their impact on the distribution of the response variable, we select a broad range of different designs to cover realistic practical scenarios, especially including the motivating example discussed in Section 2. Data have been generated from one- and two-way designs

$$Y_{ik} = b_i + \sum_{\ell=1}^{3} p_\ell M_{ik}^{(\ell)} + \sigma_i \cdot \frac{Z_{ik} - E(Z_{ik})}{\sqrt{Var(Z_{ik})}}, \quad i = 1, \dots, 4; k = 1, \dots, n_i, \tag{\textbf{Model I}}$$

$$Y_{ijk} = b_{ij} + \sum_{\ell=1}^{3} p_\ell M_{ijk}^{(\ell)} + \sigma_{ij} \cdot \frac{Z_{ijk} - E(Z_{ijk})}{\sqrt{Var(Z_{ijk})}}, \quad i = 1, 2; j = 1, \dots, 6; k = 1, \dots, n_{ij}, \tag{\textbf{Model II}}$$

each with $M = 3$ covariates drawn from discretized normal distributions with $\boldsymbol{p} = (0, \frac{1}{2}, 1)'$ and $b_i \equiv b_{ij} \equiv 10$. The random variables Z_{ik} and Z_{ijk} base the error terms and were generated from standard normal, t_3, Laplace, χ_{15}^2, χ_7^2, or exponential distributions, respectively. Thus, three differently tailed symmetric distributions as well as three skewed distributions with

TABLE 1 Overview of the different designs used in the simulation study for **Model I** and **Model II**

Setting	Model I	Model II	Meaning
1	$\boldsymbol{n}_{1,I} + m, \boldsymbol{\sigma}_{1,I}$	$\boldsymbol{n}_{1,II} + m, \boldsymbol{\sigma}_{1,II}$	Balanced/homoscedastic
2	$\boldsymbol{n}_{1,I} + m, \boldsymbol{\sigma}_{2,I}$	$\boldsymbol{n}_{1,II} + m, \boldsymbol{\sigma}_{2,II}$	Balanced/heteroscedastic
3	$\boldsymbol{n}_{2,I} + m, \boldsymbol{\sigma}_{1,I}$	$\boldsymbol{n}_{2,II} + m, \boldsymbol{\sigma}_{1,II}$	Unbalanced/homoscedastic
4	$\boldsymbol{n}_{2,I} + m, \boldsymbol{\sigma}_{2,I}$	$\boldsymbol{n}_{2,II} + m, \boldsymbol{\sigma}_{2,II}$	Positive pairing
5	$\boldsymbol{n}_{2,I} + m, \boldsymbol{\sigma}_{3,I}$	$\boldsymbol{n}_{1,II} + m, \boldsymbol{\sigma}_{3,II}$	Negative pairing
6	—	$\boldsymbol{n}_{2,II} + m, \boldsymbol{\sigma}_{4,II}$	Positive/negative pairing

different amount of skewness (mild, medium, and very) were chosen as error term distributions (standardized) in each of the two models. The constants σ_i and σ_{ij} ensure $Var(Y_{ik}) = \sigma_i^2$ and $Var(Y_{ijk}) = \sigma_{ij}^2$. As a major assessment criteria, we will exemplify the behavior of the methods in positive and negative pairings of sample sizes and variances. The resulting different settings of sample sizes and groupwise variance allocations are displayed in Table 1.

The initial parameters of **Model I** are $\boldsymbol{n}_{1,I} = (7, 7, 7, 7)'$, $\boldsymbol{n}_{2,I} = (7, 10, 13, 18)'$, $\boldsymbol{\sigma}_{1,I} = (1, 1, 1, 1)'$, $\boldsymbol{\sigma}_{2,I} = (1, \sqrt{2}, \sqrt{3}, 2)'$, and $\boldsymbol{\sigma}_{3,I} = (2, \sqrt{3}, \sqrt{2}, 1)'$, whereas the initial parameters of **Model II** were set as $\boldsymbol{n}_{1,II} = (\boldsymbol{n}'_{1,I}, \boldsymbol{n}'_{1,I}, \boldsymbol{n}'_{1,I})'$, $\boldsymbol{n}^{(1)}_{2,II} = (7, 10, 13, 15, 18, 20)'$, $\boldsymbol{n}_{2,II} = (\boldsymbol{n}^{(1)\prime}_{2,II}, \boldsymbol{n}^{(1)\prime}_{2,II})'$, and standard deviations $\boldsymbol{\sigma}_{1,II} = (\boldsymbol{\sigma}'_{1,I}, \boldsymbol{\sigma}'_{1,I}, \boldsymbol{\sigma}'_{1,I})'$, $\boldsymbol{\sigma}^{(1)}_{2,II} = (1, \sqrt{2}, \sqrt{3}, \sqrt{4}, \sqrt{5}, \sqrt{6})'$, $\boldsymbol{\sigma}_{2,II} = (\boldsymbol{\sigma}^{(1)\prime}_{2,II}, \boldsymbol{\sigma}^{(1)\prime}_{2,II})'$, $\boldsymbol{\sigma}^{(1)}_{3,II} = (\sqrt{6}, \sqrt{5}, \sqrt{4}, \sqrt{3}, \sqrt{2}, 1)'$, $\boldsymbol{\sigma}_{3,II} = (\boldsymbol{\sigma}^{(1)\prime}_{3,II}, \boldsymbol{\sigma}^{(1)\prime}_{3,II})'$, and $\boldsymbol{\sigma}_{4,II} = (\boldsymbol{\sigma}^{(1)\prime}_{3,II}, \boldsymbol{\sigma}^{(1)\prime}_{2,II})'$, respectively. In addition, $m \in \{0, 2, \dots, 10\}$ is a constant that is added to each initial sample size setting $\boldsymbol{n}_{\ell,I} = (n_1, \dots, n_4)'$ and $\boldsymbol{n}_{\ell,II} = (n_{11}, \dots, n_{12})'$, $\ell = 1, 2$. Settings 1 and 2 represent balanced cases with equal or unequal variances, while settings 3 to 6 cover unbalanced designs with homoscedastic and heteroscedastic variances. Each of the six different error distributions was simulated in all of the five (**Model I**) and six (**Model II**) scenarios with varying sample sizes and groupwise variances. We report the type-1 error ($\alpha = 5\%$) simulation results of the four main competitors: F_A in (10) implemented in `SAS PROC MIXED`, the wild-bootstrap test $W^*(\boldsymbol{T})$ from Zimmermann et al[7] (also see Equation (15)) (using $nboot = 10\,000$ bootstrap runs), the classical ANCOVA F-test as well as the novel ANCOVA-type test $F_N(\boldsymbol{T})$ in (11) for testing the null hypotheses

$$H_0^{(b)} : \boldsymbol{P}_4 \boldsymbol{b} = \boldsymbol{0} \qquad (\textbf{Model I}) \quad \text{and} \quad H_0^{(b)} : (\boldsymbol{P}_2 \otimes \boldsymbol{P}_6)\, \boldsymbol{b} = \boldsymbol{0} \qquad (\textbf{Model II})$$

in Figure 2 and in Figure 3, respectively.

First, it can readily be seen from Figure 2 that the two ANCOVA-type tests F_A (10) and $F_N(\boldsymbol{T})$ are of equal quality, and both procedures tend to control the nominal type-1 error rate quite accurate in all considered scenarios. When sample sizes are very small, they tend to be slightly conservative. It can also be seen that the quality of the approximation depends on distributional shapes. In case of symmetric distributions, the heavier the tail the more conservative the methods seem to be. This observation was made in each of the different settings. In case of skewed distributions, a similar pattern can be recognized. The higher the skewness the more conservative the tests behave. The classical F-test performs as expected and controls the nominal type-1 error rate very well under variance homoscedasticity, while it tends to be liberal or conservative under negative or positive pairings, respectively. The procedure is based on a pooled variance estimator (assuming equal variances), which is biased when the data actually have different variances in unbalanced designs. The wild-bootstrap method $W^*(\boldsymbol{T})$ (15) controls the type-1 error well in case of symmetric distributions. Under skewed distributions, the opposite as with the ANCOVA-type tests can be observed, namely the higher the skewness, the more liberal the test behaves and thus, over-rejects the null hypothesis. With increasing sample sizes, the conservatism of the ANCOVA-type tests as well as the liberality of $W^*(\boldsymbol{T})$ tend to decrease. Summing up all findings, the quality of each method depends on the shape of the underlying data distribution. In case of symmetric distributions, no major differences between the methods could be detected in the selected scenarios except for the t_3 distribution, where the wild-bootstrap test is closer to the nominal level than the somewhat conservative new ANCOVA-type tests. By contrast, the $F_N(\boldsymbol{T})$ seems more stable under skewed distributions than the wild-bootstrap method. Next, we will investigate the control of the type-1 error rate of the competing methods for testing the null hypothesis $H_0^{(p_1)} : p_1 = 0$. Since the results obtained under **Model II** and **Model I** were very similar, we omit the latter. The simulation results are displayed in Figure 4.

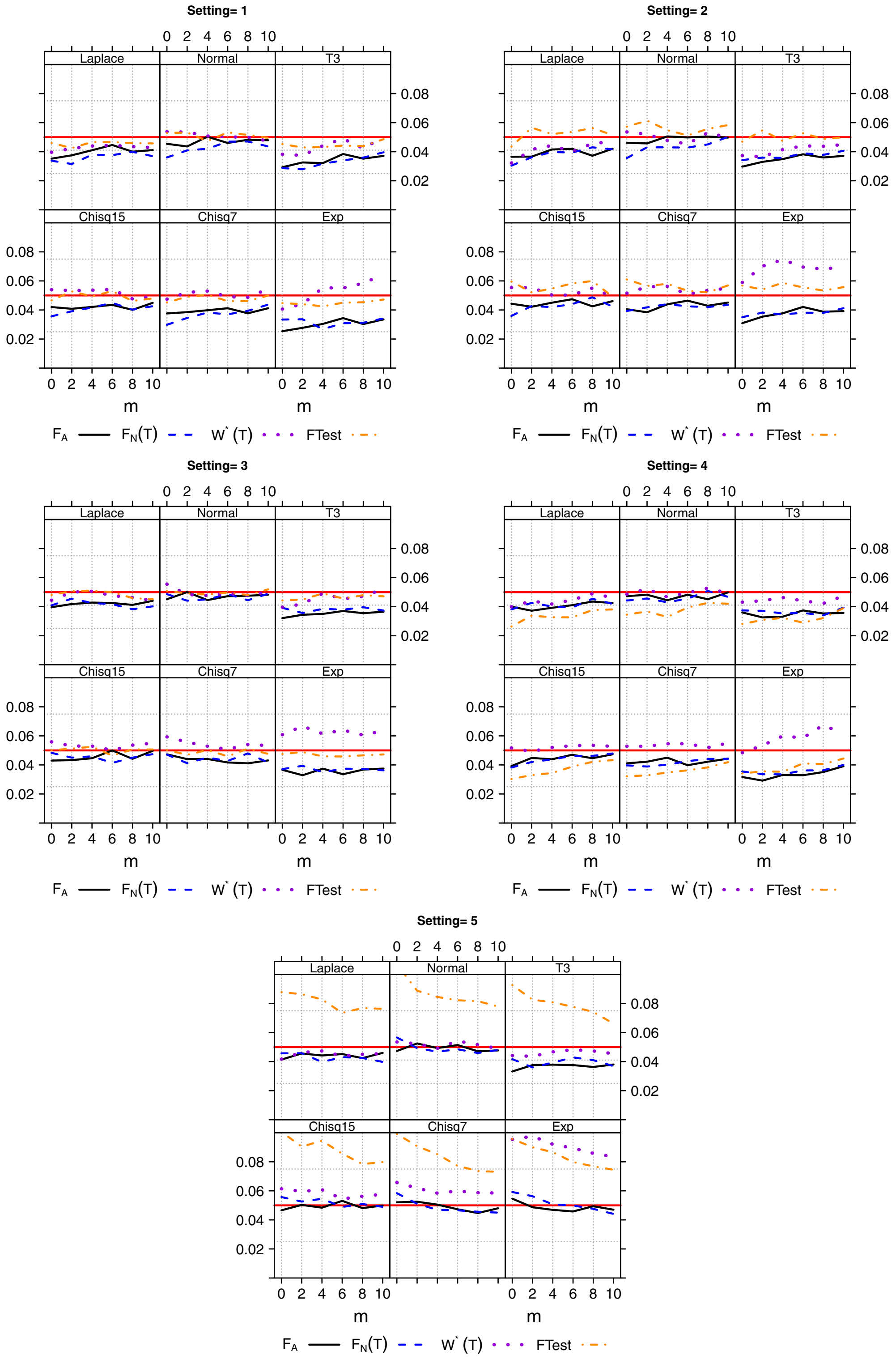

FIGURE 2 Type-1 error ($\alpha = 5\%$) simulation results of F_A in (10) implemented in SAS PROC MIXED, the wild-bootstrap test $W^*(\boldsymbol{T})$ in (15), the classical ANCOVA F-test, and the novel test $F_N(\boldsymbol{T})$ in (11) for testing $H_0^{(b)}$ in one-way designs (**Model I**) [Colour figure can be viewed at wileyonlinelibrary.com]

First, it can readily be seen that the ANCOVA-type test F_A implemented in SAS is pretty liberal and over-rejects the null hypothesis when sample sizes are either small or of medium size. This can be observed under all of the investigated data distributions and settings. With increasing sample sizes, the liberality vanishes, but, it still cannot be recommended when sample sizes are $n_i \approx 20$. This behavior of the test might occur because SAS uses the generalized least squares with the estimated variances as the point estimators. The estimation of the variances of the estimators might be unstable and not well tracked within the approximation procedure when sample sizes are as small as considered here. The novel

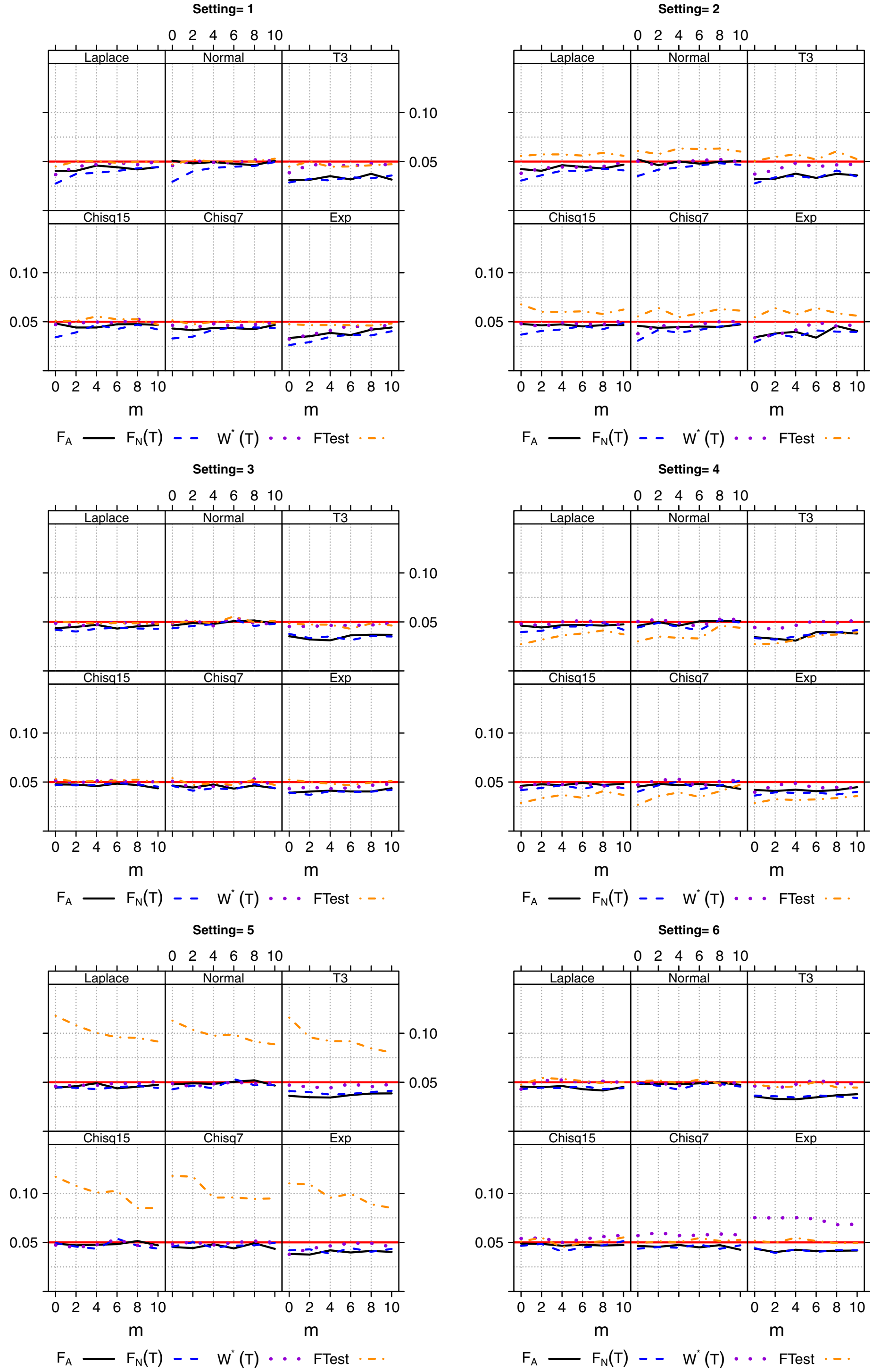

FIGURE 3 Type-1 error ($\alpha = 5\%$) simulation results of F_A in (10) implemented in `SAS PROC MIXED`, the wild-bootstrap test $W^*(\boldsymbol{T})$ in (15), the classical ANCOVA F-test, and the novel test $F_N(\boldsymbol{T})$ in (11) for testing $H_0^{(b)}$ of no interaction effect in two-way designs (**Model II**) [Colour figure can be viewed at wileyonlinelibrary.com]

ANCOVA-type test $F_N(\boldsymbol{T})$ as well as the resampling version $W^*(\boldsymbol{T})$ control the nominal type-1 error level very well in all of the considered scenarios.

5.1 | Power simulations

We conducted extensive simulation studies to compare the powers (at significance level $\alpha = 5\%$) of F_A in (10), $W^*(\boldsymbol{T})$ in (15) and $F_N(\boldsymbol{T})$ in (11) in various scenarios. Due to the abundance of different factorial designs, we compare the powers of the methods to detect the three selected alternatives

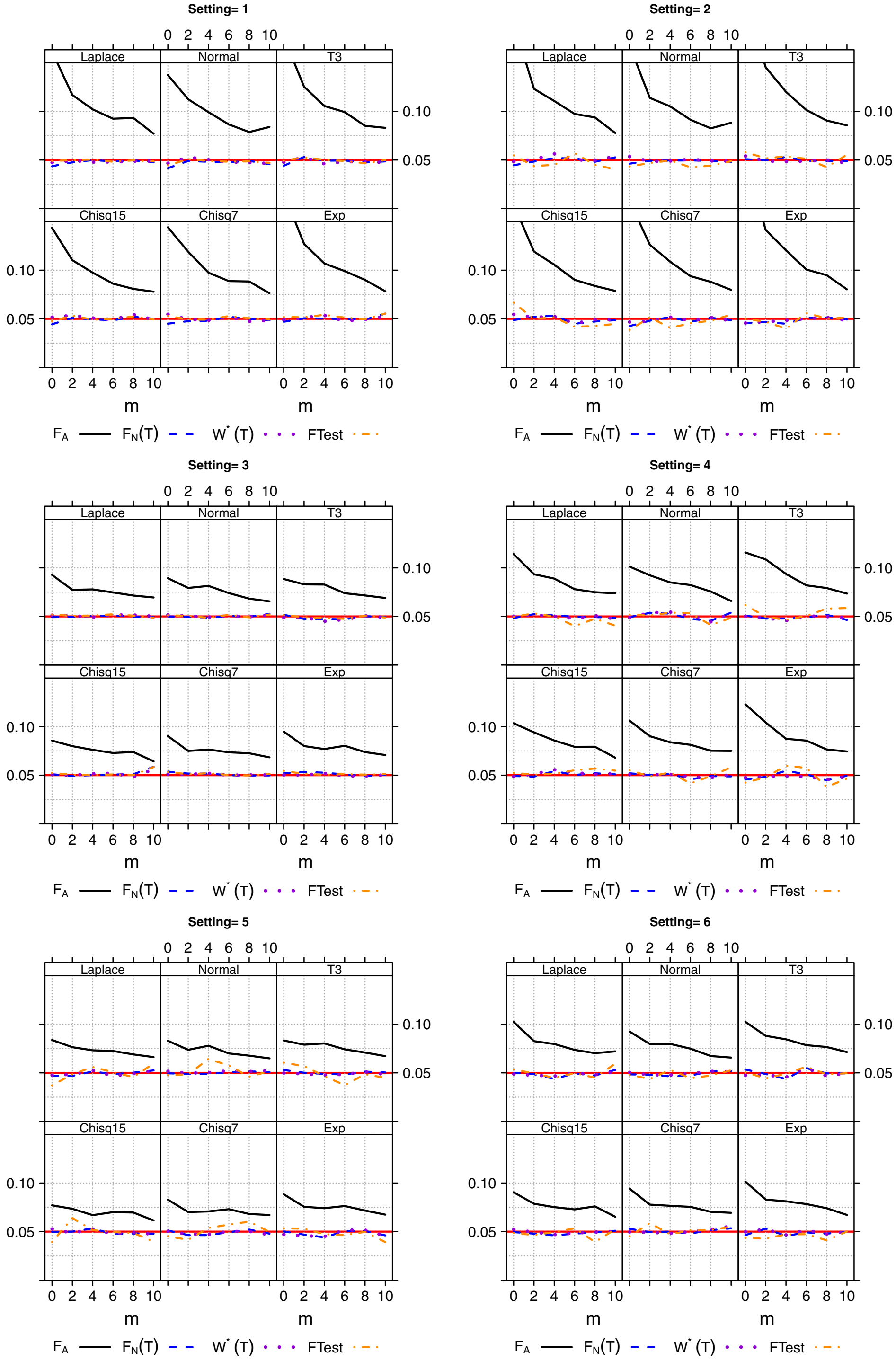

FIGURE 4 Type-1 error ($\alpha = 5\%$) simulation results of F_A in (10) implemented in `SAS PROC MIXED`, the wild-bootstrap test $W^*(\boldsymbol{T})$ in (15), the classical and novel ANCOVA F-test and $F_N(\boldsymbol{T})$ in (11) for testing $H_0^{(p_1)}$ in two-way designs (**Model II**) [Colour figure can be viewed at wileyonlinelibrary.com]

$$\begin{array}{lll}
\text{Alternative 1} & \text{Alternative 2} & \text{Alternative 3} \\
\boldsymbol{b} = \left(10-\delta, 10\boldsymbol{1}_3'\right)' & \boldsymbol{b} = \left(10-\delta, 10+\delta, 10\boldsymbol{1}_2'\right)' & \boldsymbol{b} = \left(10-\delta, 10+\delta, 10+\frac{1}{2}\delta, 10\right)' \\
\boldsymbol{b} = \left(10-\delta, 10\boldsymbol{1}_{11}'\right)' & \boldsymbol{b} = \left(10-\delta, 10\boldsymbol{1}_5', 10+\delta, 10\boldsymbol{1}_5'\right)' & \boldsymbol{b} = \left(10-\delta, 10+\delta, 10+\frac{1}{2}\delta, 10\boldsymbol{1}_3', 10+\delta, 10-\delta, 10-\frac{1}{2}\delta, 10\boldsymbol{1}_3'\right)'
\end{array} \tag{12}$$

in **Model I** (upper row) and **Model II** (lower row) with varying $\delta \in \{0, 0.1, \dots, 2\}$ in all of the different settings as described in Table 1. Throughout, we set $m = 0$ and thus do not compare the impact of increasing

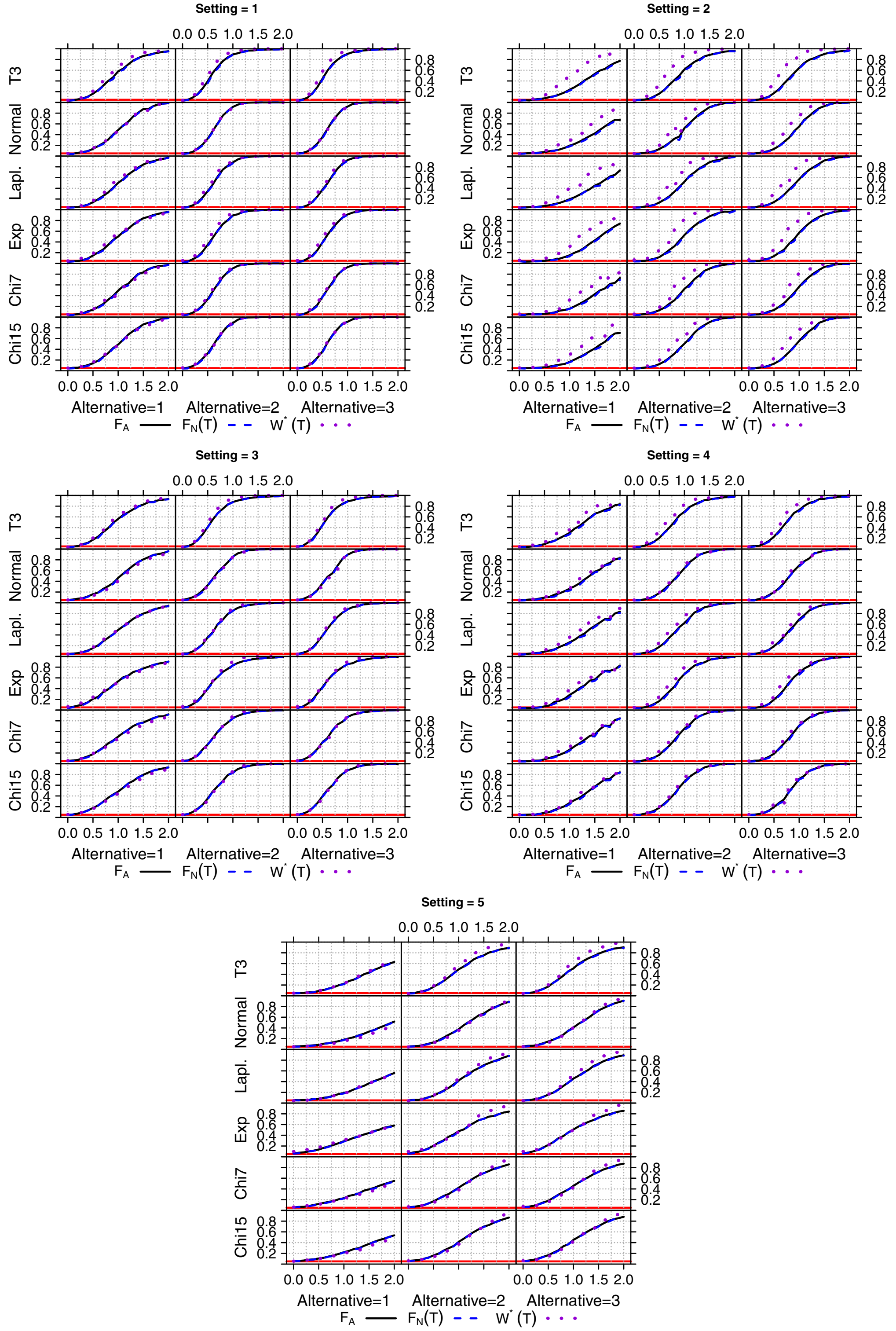

FIGURE 5 Power simulation results ($\alpha = 5\%$) to detect the three alternatives in (12) of F_A in (10) implemented in SAS PROC MIXED, the wild-bootstrap test $W^*(\boldsymbol{T})$ in (15) and novel ANCOVA F-test and $F_N(\boldsymbol{T})$ in (11) in one-way designs (**Model I**) [Colour figure can be viewed at wileyonlinelibrary.com]

sample sizes on the powers of the tests. Due to the liberal or conservative behavior of the classical F-test, we did not include the method in the comparison. The power curves are displayed in Figures 5 (**Model I**) and 6 (**Model II**), respectively.

We find that the wild-bootstrap method tends to have a higher power than the ANCOVA-tests in few scenarios (eg, Setting 2 in both **Model I** and **Model II**). A possible explanation could be that the method is a Wald-type statistic, while the others are approximate solutions. However, we also see the opposite in few scenarios (eg, Setting 3, Alternative 1). In general, the wild-bootstrap test $W^*(\boldsymbol{T})$ has a slightly higher power under variance heteroscedasticity,

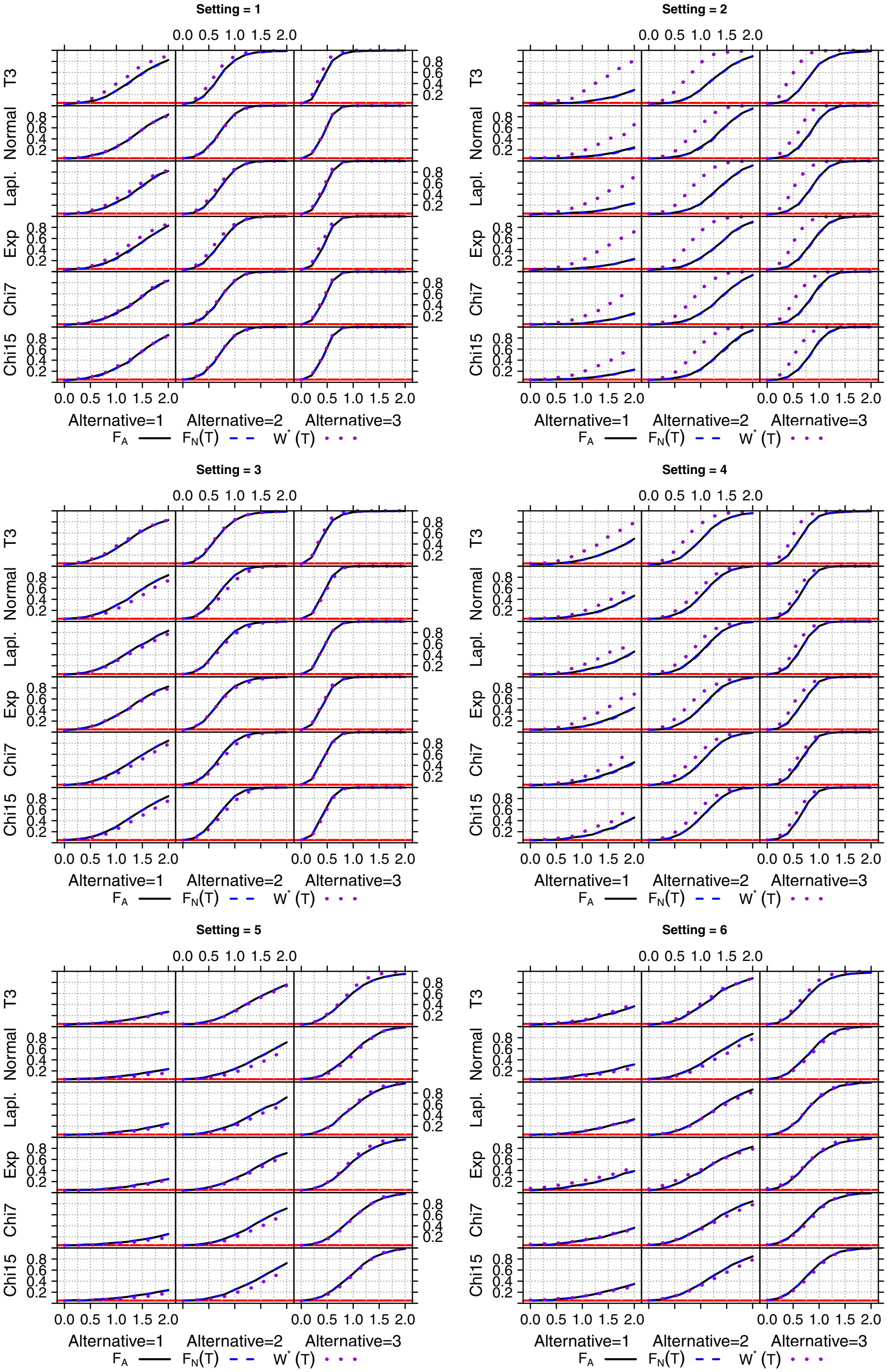

FIGURE 6 Power simulation results ($\alpha = 5\%$) to detect the three alternatives in (12) of F_A in (10) implemented in SAS PROC MIXED, the wild-bootstrap test $W^*(\boldsymbol{T})$ in (15) and novel ANCOVA F-test and $F_N(\boldsymbol{T})$ in (11) in two-way designs (**Model II**) [Colour figure can be viewed at wileyonlinelibrary.com]

depending on variance and sample size allocations. In case of equal variances, its power is slightly smaller than of its competitors. Furthermore, F_A in (10) also tends to have a slightly higher power than $F_N(\boldsymbol{T})$ in (11) (about 1% higher). The power increase might either result from using MIVQUE0 variance estimators and/or from using generalized least squares estimators. Overall, we see that the powers of the methods depend on various parameters, for instance the hypothesis of interest (contrast matrix), alternative pattern, sample sizes, degree of variance heteroscedasticity and especially their allocations. Therefore, finding a general conclusion is quasi impossible in the designs considered here. In general, none of the methods has a superior power, while there are situations in which any of the method outperforms the others.

TABLE 2 Group-specific point estimators of the treatment effects, their standard errors, and the variances of the two-way ANCOVA model

					Variance	
Gender	Pyridine con. (in ppm)	n_{ij}	Treatment effect ($\widehat{b}_{ij}$)	Standard error of $\widehat{b}_{ij}$	$\widehat{\sigma}^2_{ij}$	MIVQUE0
Female	0	10	22.58	0.26	7.12	6.86
	50	10	21.97	0.25	7.29	7.47
	100	10	23.02	0.25	8.56	13.40
	250	10	19.87	0.25	4.84	4.06
	500	10	21.43	0.26	3.72	4.94
	1000	8	24.88	0.26	14.71	13.52
Male	0	10	27.68	0.44	2.39	1.82
	50	10	25.77	0.45	4.01	4.79
	100	10	26.62	0.46	3.51	3.48
	250	10	26.59	0.45	5.48	4.68
	500	10	27.42	0.41	11.83	9.97
	1000	9	27.38	0.37	17.01	14.25

6 | DATA EVALUATIONS

The two-way factorial toxicological and carcinogenic study introduced in Section 2 can now be analyzed with the new methods. First, the data of this trial can be modeled by

$$Y_{ijk} = b_{ij} + \sum_{\ell=1}^{2} p_\ell M_{ijk}^{(\ell)} + \epsilon_{ijk}, \quad i = 1, \dots, 2; \quad j = 1, \dots, 6; \quad k = 1, \dots, n_{ij},$$

with Y_{ijk} denoting BUN (equivalently change of BUN), $M_{ijk}^{(1)}$ and $M_{ijk}^{(2)}$ BUN at baseline and change of bodyweight of rat k of gender i in dose level j, respectively. First, the estimated treatment effects ($\widehat{b}_{ij}$), standard errors of $\widehat{b}_{ij}$, and the variance estimators ($\widehat{\sigma}_{ij}$) for each factor-level combinations of the two factors: gender (A) and pyridine concentration (B) are listed in Table 2. For illustration, we also report the variance estimators obtained by `MIVQUE0` in `SAS PROC MIXED`.

It can readily be seen from Table 2 that the variance estimators from both estimation methods after adjusting for the two covariates are rather different for each combination of the two factors. Therefore, assuming homoscedastic variances in the classical ANCOVA model may lead to inaccurate statistical inferences.

In order to analyze the main effects and the interaction effect, we test each null hypothesis given in (5) using the novel ANCOVA-type test $F_N(\boldsymbol{T})$ in (11), the ANCOVA-type test F_A in (10) implemented in `SAS PROC MIXED`, the wild-bootstrap test $W^*(\boldsymbol{T})$ in (15), and the classical ANCOVA F-test. The results are summarized in Table 3 with the test statistics, degrees of freedom (DF), and the P-values.

It can readily be seen from Table 3 that the factor A (gender) has a significant impact on the endpoint BUN at 5% level of significance using the first three tests: $F_N(\boldsymbol{T})$, F_A, and $W^*(\boldsymbol{T})$, while an insignificant impact using the classical ANCOVA F-test. This difference in results might be because the test assumes homogenous variances. None of the tests detects a dose effect (pyridine concentration) at 5% level of significance. Furthermore, the bodyweight change does not seem to impact the response.

7 | DISCUSSION

Analysis of covariance is a fundamental inference method in statistical practice and allows estimation and testing of adjusted treatment effects in general factorial designs. For example, adjusting for baseline values is essential in a variety of trials. In addition, variance heteroscedasticity is a non-negligible impact which might complicate the statistical analysis, especially in small sample sizes. In the present article, we discussed available methods for the analysis of

TABLE 3 Test results for the two-way factorial toxicological and carcinogenic study. Here, two factors are the gender of rats (A) and the pyridine concentration (B)

Estimation method	Effect	Test statistic	DF Num	DF Den	P-value
Novel ANCOVA-type test: $F_N(\boldsymbol{T})$	A	4.16	1.00	54.58	.0462
	B	2.00	3.08	53.70	.1240
	$A \times B$	1.44	3.66	51.56	.2365
ANCOVA-type test: F_A in `SAS`	A	5.02	1.00	52.40	.0294
	B	2.29	3.53	84.20	.0740
	$A \times B$	1.45	3.95	76.90	.2265
Wild-bootstrap test: $W^*(\boldsymbol{T})$	A	5.77	—	—	.0267
	B	13.61	—	—	.0683
	$A \times B$	9.54	—	—	.1749
Classical ANCOVA F-test	A	2.51	1	103	.1161
	B	1.37	5	103	.2425
	$A \times B$	1.23	5	103	.3020

ANCOVA models under groupwise heteroscedasticity. In particular, we investigated the ANCOVA-type statistic F_A available in `SAS PROC MIXED`. In this spirit, we were able to derive its sibling statistic F_N, which can be seen as a generalization of the ANOVA-type statistic.[5] In comparison with F_A, F_N is numerically feasible and can be computed within few numerical steps. Extensive simulation studies show that the ANCOVA-type tests control the nominal type-1 error rate equally well and have comparable powers to detect alternatives. As further competitor, we investigated a wild-bootstrap approach.[7] Extensive simulation studies show that none of the methods has a superior power. In few situations depending on variance heteroscedasticity, type of hypothesis, sample sizes, and so forth, W^* has a higher power than its competitors. In general, the methods are applicable even in case of small sample sizes. As a rough recommendation, they are applicable when $n_i \geq 10$ (depending on the actual model under consideration). Furthermore, when confronted with small sample sizes, we recommend testing for no covariate effect(s) with the statistic F_N. All of the presented methods are, however, approximate solutions each with its own pros and cons. Besides evaluating data, statistical planning and sample size computations are of same importance. We will tackle them in future research projects. Furthermore, we investigated Box-type approximations only. The investigation of other methods, for example, the pretty popular Kenward-Roger approximation methods[18,19] (which are implemented in `SAS PROC MIXED` as well), will also be part of future research.

Extension: Completely heteroscedastic errors

The assumption of equal variances within groups (see Equation (2)) may be further relaxed, in order to cover also scenarios with subject-specific errors (ie, "completely heteroscedastic" settings), namely by only assuming

$$Var(\epsilon) = \bigoplus_{i=1}^{a} \bigoplus_{j=1}^{n_i} \sigma_{ij}^2. \tag{13}$$

Translating the ANCOVA-type approximation framework that has been considered in the present article to such a general setting is not straightforward. By contrast, a Wald-type approach is still applicable, because the covariance matrix given in (13) may be replaced by the estimator

$$\widehat{Var(\epsilon)} = \bigoplus_{i=1}^{a} \bigoplus_{j=1}^{n_i} \hat{\epsilon}_{ij}^2, \tag{14}$$

where $\hat{\epsilon}_{ij}$ denotes the ANCOVA residual of subject j in group i, $i \in \{1, \ldots, a\}$, $j \in \{1, \ldots, n_i\}$. In order to improve the finite-sample performance, the following wild-bootstrap Wald-type test has been proposed in Zimmermann et al.[7] Let $\boldsymbol{B} := (\boldsymbol{X}, \boldsymbol{M})$ denote the design matrix of the ANCOVA model. Independently from the data, we generate a sample of i.i.d. random variables $T_{11}, \ldots, T_{an_a}$ satisfying $P(T_1 = -1) = P(T_1 = 1) = 1/2$ and obtain the corresponding bootstrap observations $Y^*_{ij} = \hat{\epsilon}_{ij} T_{ij}(1 - p_{ij})^{-1/2}$, where p_{ij} denotes the diagonal element of the hat matrix $\boldsymbol{P_B} = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'$ corresponding to subject j within group i. Finally, we calculate the wild-bootstrap Wald-type statistic

$$W^*(\boldsymbol{T}) := (\boldsymbol{T}\hat{\boldsymbol{b}}^*)'(\boldsymbol{T}\hat{\boldsymbol{\Sigma}}^*\boldsymbol{T})^{-1}\boldsymbol{T}\hat{\boldsymbol{b}}^*. \tag{15}$$

Thereby, $\hat{\boldsymbol{b}}$ denotes the least squares estimator of $\boldsymbol{b}$ defined in (6), but with the original observations replaced by the bootstrap sample, and $\hat{\boldsymbol{\Sigma}}^*$ denotes the upper-left block of the 2×2 block matrix

$$(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'\widehat{Var^*(\boldsymbol{\epsilon})}\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1},$$

where $\widehat{Var^*(\boldsymbol{\epsilon})}$ is the bootstrap counterpart of the covariance matrix estimator from (14). Now, when repeating the procedure of drawing i.i.d. bootstrap samples a "large" number of times, we may use the empirical $(1 - \alpha)$ quantile of the conditional distribution of $W^*(\boldsymbol{T})$ as the critical value for testing $H_0^{(b)} : \boldsymbol{Tb} = \boldsymbol{0}$. Some limited simulation evidence indicates that the wild-bootstrap Wald-type test performs well even in the more general setting of unequal variances within groups, except for some tendency toward a conservative behavior in the log-normal case (see table 1 in Zimmermann et al[7]). In general, however, it should be kept in mind that in a well-designed study, one may expect at most slight heteroscedasticity within groups. Therefore, the ANCOVA-type approximation that has been considered in the present article might actually cover most practically relevant scenarios.

Another potential line of action might be to use alternative methods for estimating the variances of the estimated model coefficients instead of HCSE estimators.[8,9] For example, the Hadamard estimator considered in Dobriban and Su[20] may be less biased, and hence, improve the finite-sample performance of the wild-bootstrap ANCOVA. This idea needs to be further studied in future research. So far, we restricted our research to independent observations. Repeated measures designs and multivariate data will be part of future research. Implementations of the new methods in freely available software packages (eg, within the R-package GFD[6]) are envisioned.

ACKNOWLEDGEMENT

The work of Frank Konietschke and Asanka Gunawardana was supported by the Deutsche Forschungsgemeinschaft grant number DFG KO 4680/3-2, and the work of Georg Zimmermann by the WISS 2025 project IDA-Lab Salzburg (20204-WISS/225/197-2019 and 20102-F1901166-KZP).

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in https://manticore.niehs.nih.gov/cebssearch/test_article/110-86-1 (accessed November, 2020).

ORCID

Frank Konietschke https://orcid.org/0000-0002-5674-2076

REFERENCES

1. Huitema B. *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies*. Hoboken, NJ: John Wiley & Sons; 2011.
2. Sullivan LM, D'Agostino RB Sr. Robustness and power of analysis of covariance applied to ordinal scaled data as arising in randomized controlled trials. *Stat Med*. 2003;22(8):1317-1334.
3. Acitas S, Senoglu B. Robust factorial ANCOVA with LTS error distributions. *Hacettepe J Math Stat*. 2018;47(2):347-363.
4. Pauly M, Brunner E, Konietschke F. Asymptotic permutation tests in general factorial designs. *J Royal Stat Soc Ser B Stat Methodol*. 2015;77(2):461-473.
5. Brunner E, Dette H, Munk A. Box-type approximations in nonparametric factorial designs. *J Am Stat Assoc*. 1997;92(440):1494-1502.
6. Friedrich S, Konietschke F, Pauly M. GFD: An R package for the analysis of general factorial designs. *J Stat Softw*. 2017;79(1):1-18.

7. Zimmermann G, Pauly M, Bathke AC. Small-sample performance and underlying assumptions of a bootstrap-based inference method for a general analysis of covariance model with possibly heteroskedastic and nonnormal errors. *Stat Methods Med Res*. 2019;28(12):3808-3821.
8. Eicker F. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Ann Stat*. 1963;34(2):447-456.
9. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 1980;48(4):817-838.
10. Liu RY. Bootstrap procedures under some non-iid models. *Ann Stat*. 1988;16(4):1696-1708.
11. Wu CFJ. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann Stat*. 1986;14(4):1261-1295.
12. Rao CR. Minimum variance quadratic unbiased estimation of variance components. *J Multivar Anal*. 1971;1(4):445-456.
13. Box GE. 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. *Ann Math Stat*. 1954;25(2):290-302.
14. Mathai AM, Provost SB. *Quadratic Forms in Random Variables: Theory and Applications*. New York, NY: Marcel Dekker; 1992.
15. NTP. *TR-470: Pyridine (110-86-1). Chemical Effects in Biological Systems (CEBS)*. Research Triangle Park, NC: National Toxicology Program (NTP); 2000.
16. Wiesner AJ. *Comparing Bootstrap, MIVQUE0 and REML Estimates of Variance Components of a Two-Level Random Coefficient Model with Non-normal Errors: A Simulation Study* [Doctoral dissertation]. University of Pittsburgh; 2004.
17. Cao C, Pauly M, Konietschke F. The Behrens-Fisher problem with covariates and baseline adjustments. *Metrika*. 2020;83(2):197-215.
18. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*. 1997;53:983-997.
19. Kenward MG, Roger JH. An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Comput Stat Data Anal*. 2009;53(7):2583-2595.
20. Dobriban E, Su WJ. Robust inference under heteroskedasticity via the hadamard estimator; 2018. arXiv preprint arXiv:1807.00347.

How to cite this article: Konietschke F, Cao C, Gunawardana A, Zimmermann G. Analysis of covariance under variance heteroscedasticity in general factorial designs. *Statistics in Medicine*. 2021;40:4732–4749. https://doi.org/10.1002/sim.9092

APPENDIX. DERIVATION OF THE APPROXIMATION PROCEDURE

We will prove Proposition 1 in the following. It follows from the (asymptotic) multivariate normality of $\sqrt{N}(\widehat{\boldsymbol{b}} - \boldsymbol{b})$ and the representation theorem of a quadratic form in Mathai[14] that the (asymptotic) distribution of $\tilde{B}_N(\boldsymbol{T})$ can be approximated by a weighted sum of independent χ_1^2 random variables. Following the idea of Box[13] and Brunner et al,[5] we approximate its distribution by a scaled $g \cdot \chi_{f_1}^2$ distribution, equate the first two moments, and obtain

$$B_N(\boldsymbol{T}) = N\widehat{\boldsymbol{b}}'\boldsymbol{T}\widehat{\boldsymbol{b}} \overset{\cdot}{\sim} \sum_{i=1}^{a} \kappa_i \chi_1^2 \overset{!}{\sim} g \cdot \chi_{f_1}^2 \text{ with}$$

$$E\left(B_N(\boldsymbol{T})\right) = \sum_{i=1}^{a} \kappa_i = tr(\boldsymbol{T\Psi}) = g \cdot f_1 \text{ and}$$

$$Var\left(B_N(\boldsymbol{T})\right) = 2\sum_{i=1}^{a} \kappa_i^2 = 2tr(\boldsymbol{T\Psi T\Psi}) = 2g^2 \cdot f_1.$$

It follows that

$$g \cdot f_1 = tr(\boldsymbol{T\Psi}) \quad \text{and} \quad f_1 = \frac{[tr(\boldsymbol{T\Psi})]^2}{tr(\boldsymbol{T\Psi T\Psi})}.$$

Hence, under $H_0^{(b)}$, we obtain the approximation

$$\tilde{F}_N(\boldsymbol{T}) = \frac{N}{g \cdot f_1}\widehat{\boldsymbol{b}}'\boldsymbol{T}\widehat{\boldsymbol{b}} = \frac{N}{tr(\boldsymbol{T\Psi})}\widehat{\boldsymbol{b}}'\boldsymbol{T}\widehat{\boldsymbol{b}} \overset{\cdot}{\sim} \chi_{f_1}^2/f_1.$$

Since $tr(\boldsymbol{T\Psi})$ depends on unknown parameters and thus is unknown in practical applications, we replace it by its empirical counterpart $tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}})$. However, we observe that we can write the latter as a quadratic form and approximate its distribution in the same way as above

$$
\begin{aligned}
tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}}) &= tr(\underbrace{\boldsymbol{TD}}_{=:\boldsymbol{K}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{D}'\boldsymbol{T}') = tr(\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}\boldsymbol{K}') \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n_i} k_{ij}^2\widehat{\sigma}_i^2 = \sum_{i=1}^{a}\widehat{\sigma}_i^2 \underbrace{\sum_{j=1}^{n_i} k_{ij}^2}_{K_i} = \sum_{i=1}^{a} K_i\widehat{\sigma}_i^2.
\end{aligned}
$$

Thus, it follows that

$$
tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}}) \sim N\sum_{i=1}^{a}\frac{K_i\sigma_i^2}{n_i-1-rank(\boldsymbol{M}_i)}V_i, V_i \sim \chi^2_{n_i-1-rank(\boldsymbol{M}_i)}.
$$

In the last step we assumed $\frac{n_i-1-rank(\boldsymbol{M}_i)}{\sigma_i^2}\widehat{\sigma}_i^2 \sim \chi^2_{n_i-1-rank(\boldsymbol{M}_i)}$ (at least approximately). Following the ideas of Box[13] and Brunner et al[5] again, we approximate the distribution of $tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}})$ by a scaled $g_2\chi^2_{f_2}/f_2$ distribution such that the first two moments coincide and obtain

$$
\begin{aligned}
E\left(tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}})\right) &= tr\,(\boldsymbol{T}\boldsymbol{\Psi}) = E(g_2\chi^2_{f_2}/f_2) = g_2, \\
Var\left(tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}})\right) &= 2N^2\sum_{i=1}^{a}\frac{K_i^2\sigma_i^4}{n_i-1-rank(\boldsymbol{M}_i)} = 2g_2^2/f_2
\end{aligned}
$$

which yields

$$
f_2 = \frac{[tr(\boldsymbol{T}\boldsymbol{\Psi})]^2}{\sum_{i=1}^{a}\frac{K_i^2\sigma_i^4}{n_i-1-rank(\boldsymbol{M}_i)}} = \frac{[tr(\boldsymbol{T}\boldsymbol{\Psi})]^2}{tr(\boldsymbol{D}_K^2\boldsymbol{D}_{\boldsymbol{\sigma}}^2\boldsymbol{\Omega})}.
$$

Therefore, we obtain the approximation

$$
F_2(\boldsymbol{T}) = \frac{tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}})}{tr(\boldsymbol{T}\boldsymbol{\Psi})} \overset{.}{\sim} \chi^2_{f_2}/f_2.
$$

The quadratic forms $\tilde{F}_N(\boldsymbol{T})$ and $F_2(\boldsymbol{T})$ are independent which motivates as approximation

$$
F_N^*(\boldsymbol{T}) = \frac{\tilde{F}_N(\boldsymbol{T})}{F_2(\boldsymbol{T})} = \frac{N}{tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}})}\widehat{\boldsymbol{b}}'\boldsymbol{T}\widehat{\boldsymbol{b}} \overset{.}{\sim} \frac{\chi^2_{f_1}/f_1}{\chi^2_{f_2}/f_2} = F(f_1, f_2).
$$

Finally, since f_1 and f_2 depend on unknown parameters, we replace them with their consistent estimators and obtain

$$
F_N(\boldsymbol{T}) = \frac{N}{tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}})}\widehat{\boldsymbol{b}}'\boldsymbol{T}\widehat{\boldsymbol{b}} \overset{.}{\sim} F(\widehat{f}_1, \widehat{f}_2),
$$

where

$$
\widehat{f}_1 = \frac{[tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}})]^2}{tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}}\boldsymbol{T}\widehat{\boldsymbol{\Psi}})} \quad \text{and} \quad \widehat{f}_2 = \frac{[tr(\boldsymbol{T}\widehat{\boldsymbol{\Psi}})]^2}{tr(\boldsymbol{D}_K^2\boldsymbol{D}_{\widehat{\boldsymbol{\sigma}}}^2\boldsymbol{\Omega})},
$$

respectively.