Psicothema

Psicothema 2018, Vol. 30, No. 4, 434-441
doi: 10.7334/psicothema2018.192

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Data analysis of incomplete repeated measures using a multivariate extension of the Brown-Forsythe procedure

Guillermo Vallejo Seco[1], María Paula Fernández García[1], Pablo Esteban Livacic Rojas[2], and Ellián Tuero Herrero[1]
[1] Universidad de Oviedo and [2] Universidad de Santiago de Chile

Abstract

Background: A multivariate extension of the Brown-Forsythe (MBF) procedure can be used for the analysis of partially repeated measure designs (PRMD) when the covariance matrices are arbitrary. However, the MBF procedure requires complete data over time for each subject, which is a significant limitation of this procedure. This article provides the rules for pooling the results obtained after applying the same MBF analysis to each of the imputed datasets of a PRMD. **Method:** Montecarlo methods are used to evaluate the proposed solution (MI-MBF), in terms of control of Type I and Type II errors. For comparative purposes, the MBF analysis based on the complete original dataset (OD-MBF) and the covariance pattern model based on an unstructured matrix (CPM-UN) were studied. **Results:** Robustness and power results showed that the MI-MBF method performed slightly worse than tests based on CPM-UN when the homogeneity assumption was met, but slightly better when that assumption was not met. We also note that without assuming equality of covariance matrices, little power was sacrificed by using the MI-MBF method in place of the OD-MBF method. **Conclusions:** The results of this study suggest that the MI-MBF method performs well and could be of practical use.

Keywords: attrition, multiple imputation, MBF procedure, covariance pattern model, heterogeneous covariance matrices.

Resumen

Análisis de datos de medidas repetidas incompletas usando una extensión multivariante del enfoque de Brown-Forsythe. **Antecedentes:** para analizar diseños de medidas parcialmente repetidas (DMPR) con matrices de covarianza arbitrarias se puede usar una extensión multivariante del enfoque de Brown-Forsythe (MBF). Una importante limitación de este enfoque es que requiere datos completos para cada sujeto. Este artículo proporciona las reglas para agrupar los resultados obtenidos tras aplicar el análisis MBF a los diferentes conjuntos de datos imputados de un DMPR. **Método:** se aplican técnicas de Montecarlo para evaluar la solución propuesta (IM-MBF), en términos de control de los errores Tipo I y Tipo II. Con fines comparativos, también se evalúan los resultados obtenidos con el enfoque MBF basado en los datos originales (DO-MBF), así como con el modelo de patrones de covarianza basado en asumir una matriz no estructurada (MPC-NE). **Resultados:** cuando se cumple el supuesto de homogeneidad, el desempeño de la prueba IM-MBF es ligeramente inferior al obtenido con la prueba MPC-NE, mientras que sucede lo contrario cuando se incumple dicho supuesto. También encontramos que se pierde poca potencia usando el enfoque MI-MBF, en lugar del enfoque DO-MBF, cuando las matrices de covarianza son heterogéneas. **Conclusiones:** los resultados sugieren que el enfoque MI-MBF funciona bien y podría ser de uso práctico.

Palabras clave: abandono del estudio, imputación múltiple, enfoque MBF, modelo de patrón de covarianza, heterogeneidad de las matrices de covarianza.

Data from partially repeated-measure designs (PRMD), or other structurally similar designs (e.g., split-plot and variations of the Latin square), are usually analyzed by using univariate analysis of variance (ANOVA) or multivariate analysis of variance (MANOVA), despite being highly sensitive to deviations from the multisample sphericity assumption (i.e., equality of covariance matrices across groups and sphericity for the common covariance matrix) when group sizes are not reasonably balanced. From at the end of the 20th century until today, several analytical strategies have been proposed for testing hypotheses about means without assuming homoscedasticity, including bootstrapping and permutation resampling methods (Friedrich, Konietschke, & Pauly, 2017; Xu, 2015), generalized *p*-values (Weerahandi, 2004), and aligned ranks (Brunner, Munzel, & Puri, 2002; Kowalchuk, Keselman, & Algina, 2003; Oliver-Rodríguez & Wang, 2015), in addition to many other procedures proposed to approximate the degrees of freedom for the classical univariate and multivariate general linear models (Bathke, Schabenberger, Tobias, & Madden, 2009; Lix, Algina, & Keselman, 2003; Vallejo & Ato, 2006).

The degrees of freedom that will most likely result in the best approximation have been reported by Vallejo and Ato (2006) for the multivariate Brown-Forsythe (MBF) approach (see Vallejo, Moris, & Conejo, 2006, for a SAS/IML® implementation). Simulation studies carried out by Vallejo, Arnau, and Ato (2007) and Livacic-Rojas, Vallejo, Fernández, and Tuero-Herrero (2017) in the context

Received: May 5, 2018 • Accepted: September 6, 2018
Corresponding author: Guillermo Vallejo Seco
Facultad de Psicología
Universidad de Oviedo
33003 Oviedo (Spain)
e-mail: gvallejo@uniovi.es

of multivariate analysis of data collected in PRMD, suggest that the MBF method performs very well and could be practical for use when sample sizes are moderate to small (i.e., the ratio of the smallest group size to the number of repeated measurements could be approximately 3 to 2, to obtain a robust MBF test). Vallejo and Ato (2012) arrive at this same conclusion using cross-sectional multivariate designs. It should be noted that the MBF procedure assumes complete measurements for all subjects, which represents an important limitation of this procedure, given that missing data is a relatively common problem in almost all research studies (Paniagua, Amor, Echeburúa, & Abad, 2017).

In many studies, however, researchers who do not have complete measurements on all subjects across time can use the MBF procedure, excluding the incomplete vectors or by using the multiple imputation procedure proposed by Rubin (MI; 1987). The first strategy involves discarding any participant who has missing values in any of the variables selected and proceeding with the complete-case (CC) analysis using standard methods. If the data are missing completely at random (MCAR), then the results of the CC analysis will be unbiased, although there will often be less power for testing hypotheses. When data are not MCAR and/or the proportion of cases lost is large, not only can CC analysis reduce the sensitivity of the tests but it can also produce biased results. The second strategy involves replacing the missing values with two or more imputed values in order to reflect uncertainty about which value to impute. Rubin (1987) recommends drawing imputations from conditional (predictive) distribution of the missing data given the observed data to ensure that imputations are proper. In this paper, we focus on the MI method for dealing with monotone missing data. The framework that we will follow involves creating multiple complete datasets, separately for each treatment group, using the propensity score method available in SAS PROC MI (SAS Institute, 2017). There are other methods for imputing the missing values such as regression model which can be used in this particular context.

Several combining rules have been developed for obtaining MI inference with a variety of statistical quantities. These rules have been developed, for example, for pooling the point estimates of the parameters of interest, such as a single regression coefficient or the difference between two sample means (Rubin & Schenker, 1986), multivariate tests of hypotheses (Li, Raghunathan, & Rubin, 1991), *p*-values (Li, Meng, Raghunathan, & Rubin, 1991), likelihood ratio tests (Meng & Rubin, 1992; Shaffer, 1997) or ANOVA sums of squares, *F*-tests, and the like (Grund, Lüdtke, & Robitzsch, 2016; Raghunathan, 2016; Raghunathan & Dong, 2011; Van Ginkel & Kroonenberg, 2014). In order to facilitate the MI inference process, a variety of software programs, both generalist and specific, incorporate functions that combine the results of multiple analyses into a single multiple-imputation result. However, the widespread application of the MI technique poses new challenges for researchers, since little or no work has been done with it in the context of ANOVA and MANOVA models when dispersion matrices are heterogeneous.

Thus, our objective in this paper is twofold. Firstly, this paper shows the rules for pooling the results obtained after applying the same MBF analysis to each of the imputed datasets of a PRMD, when the assumption of homogeneous covariance matrices is not met. The paper's second aim is to examine the operating characteristics of the MBF analysis based on the data completed by multiple imputation (referred to hereafter as MI-MBF) for testing interaction effects when the covariance homogeneity assumption is violated. In particular, we investigated if MI-MBF procedure to perform reasonably well in terms of controlling Type I error probabilities (does not claim treatment effects when none are present) or power (does not fail to detect effects when they are present). For comparative purposes, the original data analysis based MBF approach (OD-MBF), the complete-case analysis based on the MBF approach (CC-MBF), and the covariance pattern model adjusted by the Kenward-Roger solution and based on a unstructured covariance matrix (CPM-UN) were also studied. An advantage of the CPM-UN approach with covariance parameters estimated by restricted maximum likelihood (REML) estimation as implemented through the SAS PROC MIXED module, is that it can easily accommodate incomplete data (uses all of the available information for each case) and will tend to produce correct analyses provided the data are missing at random (MAR) and the distributional assumptions are met (Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2011a); in addition, it has been found that to be generally robust under similar conditions to those investigated in our article (Vallejo et al., 2011a; Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2011b).

Method

Design

To examine the operating characteristics (i.e., Type I and Type II error rates) of the tests described in the previous section, we carried out a simulation study using a completely randomized design in which n subjects were randomly distributed into two treatment groups with t equally spaced measurements from each subject. Here, we simulate data according to the following regression model for the mean response: $E(y_{ijk}) = \beta_0 + \beta_1 Trt_{ij} + \beta_2 Time_{ik} + \beta_3 Trt_{ij} \times Time_{ik}$, where Trt_{ij} denotes an indicator variable for subject i in treatment group j (i.e., 0 or 1), and $Time_{ik}$ was coded 0 to t time points. The selected set of regression coefficients under the null hypothesis represents a situation in which the patterns of change in the mean response over time are the same in both groups and their mean response profiles are equal. While the selected set of regression coefficients under the alternative hypothesis represents a situation in which the rate of change is greater in the treatment group than in the control group, at the first time point however the mean response is independent of treatment assignment. We selected regression coefficients that provided nominal powers of .80 for the MBF procedure when covariance matrices were homogeneous across groups and completely balanced design (i.e., equal sample sizes and complete data).

Procedure

Firstly, we defined the pooling rules for obtaining MI inference with the MBF approach. Secondly, we compared the robustness of the OD-MBF, CC-MBF, MI-MBF, and CPM-UN procedures when the homogeneity assumption was violated. Thirdly, we compared the power of these approaches to detect a differential pattern of change over time between treatment groups, under conditions where the tests reasonably controlled the Type I error rates. Comparisons of the performance of different methods for testing the interaction effects were made in the complete dataset before data absence was introduced and the time-related dropout

of the incomplete datasets was derived from the complete dataset. Preliminary simulations suggest that little would be learned beyond the conclusions reached using the interaction effect, commonly the most interesting issue for researchers (Fitzmaurice, Laird, & Ware, 2011; Vallejo, Fernández, Herrero, & Conejo, 2004; Vallejo, Ato, Fernández, & Livacic-Rojas, in press), so the performance for the repeated measures main effect was not examined.

Study variables

The following variables were examined in the simulation study:

1. *Number of repeated measurements*. A two-group parallel design containing either $t = 4, 6$, or 8 repeated measures per subject was considered.
2. *Homogeneity of covariance matrices*. The relative performance of the procedures was studied when covariance structures were homogeneous and heterogeneous. In the latter case, the unequal group covariance matrices were in the ratio of 1:2.
3. *Relationship between the group sizes and unequal covariance matrices*. Previous studies have shown that the relationships between the group sizes and the covariance matrices can have differing effects on test statistics. Therefore, null, positive and negative pairings of group sizes and covariance matrices were studied. A null pairing refers to the case in which the design is balanced, that is, the size of the element values of the covariance matrices is not related to the group size because all groups are of equal size. A positive pairing refers to the case in which the largest n_j is associated with the covariance matrix containing the largest element values; a negative pairing refers to the case in which the largest n_j is associated with the covariance matrix containing the smallest element values. The group sizes were respectively equal to: (a) $n_1 = 50$, $n_2 = 50$; (b) $n_1 = 40$, $n_2 = 60$; and (c) $n_1 = 60$, $n_2 = 40$. Thus, the value of the coefficient of sample size variation, Δ, was set equal to 0.20 (i.e., a moderate degree of inequality).
4. *Number of imputations*. Here, we will focus on examining the performance of the MI-MBF approach after performing 5, 10 and 20 imputations for the case in which the missing data pattern is monotone under MAR. Although the optimal number of imputations depends on the percentage of missing information, current practice favors carrying out between 5 and 20 imputations in the case of having a maximum of 20% of missing values and up to 50 imputations in the case of having a higher percentage of unobserved data (Kenward & Carpenter, 2007).

The method used to generate the imputations (i.e., the propensity score method as implemented through the SAS PROC MI module), the degree of heterogeneity between the covariance matrices (i.e., $\Sigma_2 = 2\Sigma_1$), the total sample size (i.e., $n = 100$), the shape of the distribution of the measure variable (i.e., $\gamma_1 = 0$; $\gamma_2 = 0$), the missing data mechanisms (i.e., MAR) and the patterns of missingness (i.e., monotone) were kept constant in the study. According to this missingness mechanism, the data point for subject i was missing at time k and the subsequent times if $U_{ik} < \Phi[\lambda_k + Y_{i(k-1)}]$, where U_{ik} is a uniform random variable and Φ is the cumulative normal distribution function. The values of λ_k in the above mechanisms were chosen to yield time-related dropout rates of 0%, 10%, 19% and 27% for the four respective occasions, of 0%, 10%, 19%, 27%, 34% 41% for the six respective occasions, and of 0%, 10%, 19%, 27%, 34%, 41%, 47% and 52% for the eight respective occasions.

Data analysis

Data Generation. In each treatment group, Gaussian continuous longitudinal data were simulated using the method of Ripley (1987). This procedure involves the following two steps:

1. Generate pseudorandom observation vectors $\mathbf{z}_{ij}$ with $E(\mathbf{z}_{ij}) = \mathbf{0}$ and $Cov(\mathbf{z}_{ij}) = \mathbf{I}$ from a t-variate normal distribution, where $\mathbf{I}$ is the identity matrix. These vectors were obtained using the RANNOR function in SAS.
2. Create complete datasets $\mathbf{y}_{ij}$ by multiplying the vector $\mathbf{z}_{ij}$ by the Cholesky decomposition $\mathbf{L}_l$, that is, $\mathbf{y}_{ij} = \boldsymbol{\beta}_j + \mathbf{L}_l\mathbf{z}_{ij}$, where $\mathbf{y}_{ij}$ is a vector of length t for the (i, j) th subject, $\boldsymbol{\beta}_j$ is a p-dimensional vector containing the population fixed effects, and $\mathbf{L}_l$ is a lower triangular matrix of dimension t satisfying $\boldsymbol{\Sigma}_l = \mathbf{L}_l\mathbf{L}'_l$, l =1,...,4, 6, and 8.

Five thousand replications of each condition were performed using a 0.05 significance level.

Rules for combining MBF results from multiple imputed datasets. Once the data have been multiply imputed and the MBF procedure has been used to analyze the completed datasets repeatedly, the problem lies in knowing how to combine the multiple values of Wilks' Lambda test to yield a single inference. Fortunately, Raghunathan (2016) and Raghunathan and Dong (2011) have developed a theoretical framework to combine random variables with Snedecor's F distribution in the context of ANOVA models. It is worth noting that Vallejo and Ato (2006) use only Wilks' (perhaps the most widely used criterion) approach to MANOVA to develop the MBF test, however, nothing impedes the development of tests based on other popular multivariate criteria.

Using the theory developed by Raghunathan (2016) and Raghunathan and Dong (2011) and relating Wilks' Lambda statistic, or criterion, to the F-test, it is possible to derive the formulas that facilitate combining the estimates obtained in the analyses of each complete dataset with the MBF approach. Under the null hypothesis (H_0), the pivotal statistic associated with an F-distribution arises considering the quotient $F = (H\nu_d)/(E\nu_n)$, where H and E are two independent chi-squared random variables with ν_n and ν_d degrees of freedom, respectively. If we consider the distribution $\Lambda = |\mathbf{E}| / |\mathbf{E} + \mathbf{H}|$. the relationship between Λ and F is

$$F = \frac{1-\Lambda^{1/s}}{\Lambda^{1/s}} \cdot \frac{\nu_d}{\nu_n}$$

where $|\boldsymbol{H}|$ and $|\mathbf{E}|$ are determinants of the hypothesis and error sum of squares and cross-product matrices and $s = [(l^2\nu^2_h - 4) / (l^2 + \nu^2_h - 5)]^{1/2}$, with l equal to the dimension of $\mathbf{E}$ and $\nu_\mathbf{h}$ equal to the degrees of freedom of $\mathbf{H}$. To obtain a valid frequentist hypothesis test when the assumption of homogeneous covariance matrices is not met, Vallejo and Ato (2006) suggest approximating the

degrees of freedom of the **H** and **E** matrices using what is known as multivariate Satterthwaite's approximation.

As mentioned previously, after generating M (≥2) completed datasets and performing the same analysis on each of the imputed datasets, the results obtained from the analyses are combined into a single pooled result. In our concrete case, the $Q_n^{(m)}[= (1\text{-}\Lambda^{1/s})^{(m)}\nu_d^{(m)}]$ and $Q_d^{(m)}[= (\Lambda^{1/s})^{(m)}\nu_n^{(m)}]$ quantities are associated with the numerator and denominator of the transformation of Wilks' lambda criterion to F statistic. With regard to the numerator of the F-ratio, the aggregation process necessary to obtain MI inference is done by computing the following quantities:

$$\bar{Q}_n = \frac{1}{M}\sum_{m=1}^{M}(Q_n^{(m)})^{-1}$$

$$\bar{U}_n = \frac{1}{M}\sum_{m=1}^{M}\left(\nu_n^{(m)}Q_n^{2(m)}\right)^{-1}$$

$$B_n = \frac{1}{M-1}\sum_{m=1}^{M}\left(Q_n^{(m)} - \bar{Q}_n\right)^2$$

where $\bar{Q}_n$ (point estimate of the parameter being studied) is the harmonic average of $Q_n^{(m)}$ over the M imputed datasets, $\bar{U}_n$ (within-imputation variance) is the average of the squared $Q_n^{(m)}$ over the analyses from the M imputed datasets, and B_n (between-imputation variance) is the sample variance of $Q_n^{(m)}$ over the M imputed datasets or the uncertainty that is due to missing data. The quantities referring to the denominator of the F-ratio are defined in a similar way to those used to describe the numerator, namely:

$$\bar{Q}_d = \frac{1}{M}\sum_{m=1}^{M}\left(Q_d^{(m)}\right)^{-1}$$

$$\bar{U}_d = \frac{1}{m}\sum_{m=1}^{M}\left(\nu_d^{(m)}Q_d^{2(m)}\right)^{-1}$$

$$B_d = \frac{1}{M-1}\sum_{m=1}^{M}\left(Q_d^{(m)} - \bar{Q}_d\right)^2$$

Lastly, the H_0 of primary interest when we fit the PRMD with a between-subjects factor and a within-subjects factor can be tested by using the multiple imputation F-statistic that follows:

$$F_{MI} = \frac{\bar{Q}_d}{\bar{Q}_n}$$

As the reference distribution for F_{MI} it is recommended to use an F-distribution, F_{df_n, df_d} with $df_n = 2\bar{Q}_n^2 / T_n$ degrees of freedom in the numerator and $df_d = 2\bar{Q}_d^2 / T_d$ degrees of freedom in the denominator, where $T_n = 2\bar{U}_n + (1+1/M)B_n$ and $T_d = 2\bar{U}_d + (1+1/M)B_d$ estimate the posterior variance of Q_n^{-1} and Q_d^{-1} given the observed data Y_{obs}, as discussed by Rubin (1987). When $M = 1$, the ratio F_{MI} and numerator and denominator degrees of freedom associated with the terms of the ratio (i.e., df_n, df_d) coincide with the quantities obtained using the MBF method when there are no missing data.

To derive df_n it is assumed that $Q_n^{-1} \mid Y_{obs}$ is approximately distributed as $c_n\chi^2_{f_n}$, where c_n is a constant and $\chi^2_{f_n}$ is a central chi-square with f_n degrees of freedom, and c_n and f_n are determined by matching the first two moments of $Q_n^{-1} \mid Y_{obs}$ to those of $c_n\chi^2_{f_n}$, respectively. In particular, the numerator df_n is obtained by solving simultaneously the equations:

$$\bar{Q}_n = E(c_n\chi^2_{f_n}) = c_n f_n$$
$$T_n = V(c_n\chi^2_{f_n}) = 2c_n^2 f_n$$

i.e.,

$$df_n = \frac{2\bar{Q}_n^2}{2\bar{U}_n + (1+1/M)B_n}$$

.

The degrees of freedom for the denominator are derived in the same way as the numerator.

Statistical analysis of simulated dataset. The proportion of rejections in 5,000 simulations was taken as a measure of Type I error in null conditions and power in nonnull conditions.

Results

Type I Error Rates

In order to help identify the conditions when the tests are robust and when they are not, a standard test based on the normal approximation to the binomial distribution can be used (see Austin, 2009, or Brunner et al., 2002). Thus, according to this criterion, any empirical Type I rate that is higher than 0.0562 or lower than 0.0438 would be significantly different from 0.05. In the tables, values that differ significantly from the nominal alpha level of 0.05 are in boldface type. Tests with empirical estimates that are significantly lower than the nominal level are referred to as conservative, while those whose rates are significantly higher are referred to as liberal. With regards to the method of identifying a non-robust procedure, it should be pointed out that several standards have been used by researchers to identity non-robust procedures. Therefore, it should be noted that with other standards different interpretations of the results are possible.

Homogeneous data. Estimates of actual Type I error rate and power (percent) for the interaction groups by trials when the number of levels of the within-subjects factor was equal to four, six and eight (t = 4, 6, and 8) are shown in Tables 1-3 (top panel). As seen from the tables, all error rates were close to the nominal 5% level when the OD-MBF, CC-MBF, and CPM-UN methods were used. However, MI-MBF rates were conservative (ranging from 3.12 to 4.22%) in 18 out of the 27 investigated conditions. It is important to note that the tendency of the MI-MBF test to be conservative was stronger when t = 4, but declined as the number of repeated measurements increased.

Heterogeneous data. As seen in Tables 1-3 (bottom panel), all error rates were close to the nominal 5% level when the OD-MBF procedure was used. The CC-MBF analysis was liberal (ranging from 5.92 to 6.58%) in 21 out of the 27 investigated conditions. Also, the CPM-UN method generally resulted in liberal rates of error; in fact, it was liberal (ranging from 6.32 to 9.52%) in 18 out of the 27 investigated conditions. The degree of liberalism of both approaches increased as the number of repeated measurements increased. On the other hand, error rates for the MI-MBF method were conservative (ranging from 2.72 to 4.22%) in three out of the 27 investigated conditions and liberal (ranging from 5.72 to 6.30%) in three others.

Table 1
Empirical Power and Type I Error Rates (percent) for the Interaction Term of a PRMD ($t = 4$, $\alpha = 5\%$ and $\beta = 20\%$)

		OD-MBF analysis		CC-MBF analysis		MI-MBF analysis		CPM-UN analysis	
M	**Pairing**	**Type I Error**	**Power**	**Type I Error**	**Power**	**Type I Error**	**Power**	**Type I Error**	**Power**
					Homogeneity				
	Null	5.22	80.78	5.62	62.90	**3.80**^{C}	67.76	5.22	73.90
5	+	5.40	78.26	5.70	60.22	**4.22**^{C}	66.10	4.98	72.14
	–	5.10	76.80	5.44	61.08	**3.74**^{C}	64.24	5.04	71.10
	Null	5.22	80.78	5.62	62.90	**3.87**^{C}	68.02	5.22	73.90
10	+	5.40	78.26	5.70	60.22	**3.88**^{C}	65.46	4.98	72.14
	–	5.10	76.80	5.44	61.08	**3.87**^{C}	65.24	5.04	71.10
	Null	5.22	80.78	5.62	62.90	**3.29**^{C}	68.28	5.22	73.90
20	+	5.40	78.26	5.70	60.22	**3.32**^{C}	65.46	4.98	72.14
	–	5.10	76.80	5.44	61.08	**3.12**^{C}	65.42	5.04	71.10
					Heterogeneity				
	Null	4.86	61.26	5.54	55.88	4.46	56.92	4.88	55.38
5	+	5.58	62.38	**5.92**^{L}	**55.92***	5.16	57.44	5.38	57.54
	–	4.92	53.96	5.62	49.18	**3.92**^{C}	50.20	5.26	49.86
	Null	4.86	61.26	5.54	55.88	**3.36**^{C}	56.64	4.88	55.38
10	+	5.58	62.38	**5.92**^{L}	**55.92***	**3.80**^{C}	57.86	5.38	57.54
	–	4.92	53.96	5.62	49.18	**3.28**^{C}	51.25	5.26	49.86
	Null	4.86	61.26	5.54	55.88	**3.22**^{C}	56.08	4.88	55.38
20	+	5.58	62.38	**5.92**^{L}	**55.92***	**3.78**^{C}	57.26	5.38	57.54
	–	4.92	53.96	5.62	49.18	**2.72**^{C}	51.06	5.26	49.86

Note: OD-MBF = MBF based on original data; CC-MBF = MBF based on complete case; MI-MBF = MBF based on imputed data; CPM-UN = covariance pattern model based on an unstructured matrix; *M* = number of imputations performed with MI; *t* = number of levels of the within-subjects factor; Null pair = null pairing of group sizes and covariance matrices; + pair = positive pairing of group sizes and covariance matrices; – pair = negative pairing of group sizes and covariance matrices; *C* = empirical estimates significantly lower than the nominal alpha level; *L* = empirical estimates significantly higher than the nominal alpha level. * The asterisks correspond to the values associated with the conditions under which the approaches yielded empirical Type I error rates above the upper bound criterion (i.e., 5.70%). PRMD = partially repeated mesures design

Table 2
Empirical Power and Type I Error Rates (percent) for the Interaction Term of a PRMD ($t = 6$, $\alpha = 5\%$ and $\beta = 20\%$)

		OD-MBF analysis		CC-MBF analysis		MI-MBF analysis		CPM-UN analysis	
M	**Pairing**	**Type I Error**	**Power**	**Type I Error**	**Power**	**Type I Error**	**Power**	**Type I Error**	**Power**
					Homogeneity				
	Null	4.82	80.52	4.86	49.52	5.40	66.20	5.40	72.10
5	+	5.04	77.78	4.78	44.76	5.62	65.92	4.76	70.16
	–	5.10	78.44	5.08	48.92	4.86	62.18	5.08	69.96
	Null	4.82	80.52	4.86	49.52	4.38	67.72	5.40	72.10
10	+	5.04	77.78	4.78	44.76	4.44	65.58	4.76	70.16
	–	5.10	78.44	5.08	48.92	4.30	64.24	5.08	69.96
	Null	4.82	80.52	4.86	49.52	**3.96**^{C}	66.54	5.40	72.10
20	+	5.04	77.78	4.78	44.76	**4.18**^{C}	65.00	4.76	70.16
	–	5.10	78.44	5.08	48.92	**3.82**^{C}	63.20	5.08	69.96
					Heterogeneity				
	Null	4.80	62.04	**5.96**^{L}	**50.28***	5.32	60.16	**6.56**^{L}	**55.40***
5	+	4.96	61.58	**6.14**^{L}	**48.50***	5.46	57.64	**6.32**^{L}	**53.72***
	–	5.20	54.72	**6.08**^{L}	**45.64***	**4.22**^{C}	54.32	**6.84**^{L}	**50.88***
	Null	4.80	62.04	**5.96**^{L}	**50.28***	5.06	59.08	**6.56**^{L}	**55.40***
10	+	4.96	61.58	**6.14**^{L}	**48.50***	4.66	59.18	**6.32**^{L}	**53.72***
	–	5.20	54.72	**6.08**^{L}	**45.64***	**3.74**^{C}	53.56	**6.84**^{L}	**50.88***
	Null	4.80	62.04	**5.96**^{L}	**50.28***	**4.16**^{C}	59.50	**6.56**^{L}	**55.40***
20	+	4.96	61.58	**6.14**^{L}	**48.50***	**3.84**^{C}	56.80	**6.32**^{L}	**53.72***
	–	5.20	54.72	**6.08**^{L}	**45.64***	**3.96**^{C}	52.66	**6.84**^{L}	**50.88***

Note: See the note in Table 1

Table 3
Empirical Power and Type I Error Rates (percent) for the Interaction Term of a PRMD ($t = 8$, $\alpha = 5\%$ and $\beta = 20\%$)

		OD-MBF analysis		CC-MBF analysis		MI-MBF analysis		CPM-UN analysis	
M	**Pairing**	**Type I Error**	**Power**	**Type I Error**	**Power**	**Type I Error**	**Power**	**Type I Error**	**Power**
					Homogeneity				
	Null	4.72	80.20	4.74	37.36	5.50	67.64	5.00	68.72
5	+	5.04	78.52	4.64	33.50	5.62	67.56	4.38	67.64
	–	4.94	78.40	4.96	37.48	4.68	62.66	4.46	66.42
	Null	4.72	80.20	4.74	37.36	5.44	67.08	5.00	68.72
10	+	5.04	78.52	4.64	33.50	4.20	65.70	4.38	67.64
	–	4.94	78.40	4.96	37.48	**4.26**^{C}	61.52	4.46	66.42
	Null	4.72	80.20	4.74	37.36	**4.16**^{C}	66.02	5.00	68.72
20	+	5.04	78.52	4.64	33.50	**4.06**^{C}	63.86	4.38	67.64
	–	4.94	78.40	4.96	37.48	**3.92**^{C}	61.74	4.46	66.42
					Heterogeneity				
	Null	4.82	59.24	**6.58**^{L}	**44.40***	**6.20**^{L}	**58.96***	**9.52**^{L}	**55.88***
5	+	4.98	62.62	**6.36**^{L}	**43.22***	**6.30**^{L}	**58.84***	**8.78**^{L}	**54.36***
	–	4.80	56.38	**6.24**^{L}	**42.40***	5.18	56.18	**8.94**^{L}	**52.30***
	Null	4.82	59.24	**6.58**^{L}	**44.40***	5.24	59.54	**9.52**^{L}	**55.88***
10	+	4.98	62.62	**6.36**^{L}	**43.22***	**5.72**^{L}	**59.58***	**8.78**^{L}	**54.36***
	–	4.80	56.38	**6.24**^{L}	**42.40***	**4.28**^{C}	55.60	**8.94**^{L}	**52.30***
	Null	4.82	59.24	**6.58**^{L}	**44.40***	**4.28**^{C}	58.28	**9.52**^{L}	**55.88***
20	+	4.98	62.62	**6.36**^{L}	**43.22***	5.10	58.92	**8.78**^{L}	**54.36***
	–	4.80	56.38	**6.24**^{L}	**42.40***	**4.26**^{C}	54.62	**8.94**^{L}	**52.30***

Note: See the note in Table 1

Power Rates

The homoscedastic and heteroscedastic power values (%) are also presented in Tables 1-3. Because power comparisons can be made between methods that give comparable Type I error control, only the power values for the procedures that provide a similar degree of Type I error control are reported in tables without additional information. The power values with asterisks correspond to the values associated with the conditions under which the approaches yielded empirical Type I error rates above the upper bound criterion (i.e., 5.70%).

Homogeneous data. As would be expected, the power values (percent) in Tables 1-3 (top panel) indicate that the OD-MBF procedure has more power than the remaining methods over the whole range of conditions that we have considered. At the same time, the results show that the CPM-UN outperformed the MI-MBF and the CC-MBF methods. In particular, the power rates, averaged over all conditions studied, were 78.92%, 48.42%, 65.40% and 70.24%, for the OD-MBF, CC-MBF, MI-MBF and CPM-UN methods, respectively.

Heterogeneous data. The results in Tables 1-3 (bottom panel) indicate that the OD-MBF procedure has substantially more power than the CC-MBF approach and slightly more than the CPM-UN method over the whole range of conditions that we have considered. This was despite the fact that this difference is counteracted by the inflated estimates of Type I error rate mentioned earlier for the last two procedures. On the other hand, when the homogeneity assumption was not met, the power rates of the MI-MBF method were comparable to that of the OD-MBF. Specifically, the power rates, averaged over all conditions studied, were 59.36% and 56.6%, for MBF based on original data and MI-MBF methods, respectively.

Discussion

The primary aim of the present article was to provide the rules for combining the results obtained after applying a complete data approach to each of the different *M* imputed datasets of a PRMD when the assumption of homogeneity of variance-covariance matrices is violated. The proposed approach is the direct extension of the work by Vallejo and Ato (2006) and Raghunathan (2016) to analyze imputed values with the MBF procedure. As noted earlier, Vallejo and Ato (2006) restricted their approach to the analysis of complete repeated measures data without assuming equality of covariance matrices while Raghunathan (2016) focused on adapting the approach commonly known as Rubin's rules (Rubin, 1987) to sums of squares ANOVA, F-tests, and the like.

Our simulation study suggest that the MI-MBF approach (about 10 imputations are sufficient for good results) constitutes a viable alternative for testing the repeated measures interaction effects when the data are MAR and dispersion matrices vary across groups. However, it should be noted that there is no universally best statistical method available for the analysis of incomplete repeated measures data. On the whole, robustness and power results revealed that the MI-MBF approach had a slightly worse performance than tests based on the CPM-UN when the homogeneity assumption was met. In contrast, the MI-MBF approach outperformed that of the CPM-UN when the homogeneity assumption was not met. Most importantly for the results of the simulation study, without assuming equality of covariance matrices, the OD-MBF method

(i.e., based on complete dataset before introducing data absence) provides a modest power advantage over the MI-MBF approach. This is despite the fact that for the errors between the estimated Type I error rate and the nominal level 0.05, it is clear that the MI-MBF method gives the largest negative errors (tends to yield many conservative values) while the OD-MBF method has a much better performance in achieving the nominal level.

Another important finding from this study is the profound effect of heterogeneity of covariance matrices on the power of all examined procedures, including the OD-MBF procedure. Compared to the case of equal covariance matrices and complete data, in the current study it was observed (these results are not reported to save space, but available from the first author on request) that with unequal covariance matrices the sample size is required to increase by 25-30% in order to reach a specific statistical power of 80%. Obviously, with dropout rates of 10% at every time point, the sample size must be increased by a substantially higher percentage. However, longitudinal research often involves small samples (Hertzog, Lindenberger, Ghisletta, & von Oertzen, 2008). Therefore, researchers interested in carrying out studies that have sufficient power to reject the null hypothesis should avoid planning to use small sample sizes whenever possible.

On the other hand, the simulation study covered in this paper also reveal that the MI-MBF and CPM-UN methods are clearly a better choice than using the CC-MBF analysis, given that both approaches are superior in controlling Type I and Type II error rates. In fact, MI-MBF and CPM-UN methods are always at least as good as the CC-MBF analysis, and usually MI-MBF and CPM-UN methods are better or substantially better. The CC-MBF only tends to perform quite well, compared to MI-MBF and CPM-UN analyses, when the number of repeated measurements is low (t = 4). There is a simple reason for this: keeping the percentage of missing values constant, the number of cases to discard increases as does the number of repeated measures.

Finally, readers should note that our results and recommendations are based on assuming normality for the continuous outcome variable. The effect of non-normality on the Type I and Type II error rates will not be of much consequence in the case of near-normal populations. However, the presence of a fair degree of skewness and/or kurtosis, as is not uncommon in educational, health, and psychological studies (see, e.g., Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013; Bono, Blanca, Arnau, & Gómez-Benito, 2017; Cain, Zhang, & Yuan, 2017), leads to a more conservative alpha level and, thus, to more demanding sample size requirements. Another limitation of our work is that the attrition rate is always assumed to be equal between two groups across the study. In other words, the missingness is only allowed to vary across time, but not by group by time. However, in longitudinal studies one may reasonably expect that selective attrition occurs. Therefore, before proceeding with the generalization of the two-stage approach developed in the current work, in future research it would be very useful to examine its performance under some of the conditions maintained constant due to not being the focus of this article (e.g., missing data mechanisms, patterns of missingness, number of groups, shape of distribution, or group by time related dropout rates).

Acknowledgments

We thank the three anonymous reviewers for valuable comments and suggestions. This research was supported by grant PSI-2015-67630-PSIC (AEI/FEDER, UE) from the Spanish Ministry of Economy and Competitiveness, and by grant 1170642 (FONDECYT) from the Chilean National Fund for Scientific and Technological Development.

References

Austin, P.C. (2009). Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The International Journal of Biostatistics*, *5*, 13-38. doi.org/10.2202/1557-4679.1146

Bathke, A.C., Schabenberger, O., Tobias, R.D., & Madden, L.V. (2009). Greenhouse-Geisser adjustment and the ANOVA-type statistic: Cousins or twins? *The American Statistician, 63,* 239-246. doi.org/10.1198/tast.2009.08187

Bono, R., Blanca, M.J., Arnau, J., & Gómez-Benito, J. (2017). Non-normal distributions commonly used in health, education, and social sciences: A systematic review. *Frontiers in Psychology*, *8*. doi.org/10.3389/fpsyg.2017.01602

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology, 9*, 78-84. doi.org/10.1027/1614-2241/a000057

Brunner, E., Munzel, U., & Puri, M.L. (2002). The multivariate nonparametric Behrens-Fisher problem. *Journal of Statistical Planning and Inference*, *108*, 37-53. doi.org/10.1191/0962280205sm392oa

Cain, M.K., Zhang, Z., & Yuan, K.H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*, 1716-1735. doi.org/10.3758/s13428-016-0814-1

Fitzmaurice, G., Laird, N., & Ware, J. (2011). *Applied longitudinal analysis* (2nd edition). Hoboken, NJ: Wiley.

Friedrich, S., Konietschke, F., & Pauly, M. (2017). A wild bootstrap approach for nonparametric repeated measurements. *Computational Statistics & Data Analysis*, *113*, 38-52. doi.org/10.1016/j.csda.2016.06.016

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Pooling ANOVA results from multiply imputed datasets: A simulation study. *Methodology*, *12*, 75-88. doi.org/10.1027/ 1614-2241/a000111

Hertzog, C., Lindenberger, U., Ghisletta, P., & von Oertzen, T. (2008). Evaluating the power of latent growth curve models to detect individual differences in change. *Structural Equation Modeling*, *15*, 541-563. doi.org/10.1080/10705510802338983

Kenward, M., & Carpenter, J. (2007). Multiple imputation: Current perspectives. *Statistical Methods in Medical Research*, *16,* 199-218. doi.org/10.1177/0962280206075304

Kowalchuk, R.K., Keselman, H.J., & Algina, J. (2003). Repeated measures interaction test with aligned ranks. *Multivariate Behavioral Research*, *38*, 433-461.doi.org/10.1207/ s15327906mbr3804_2

Li, K.H., Meng, X L., Raghunathan, T.E., & Rubin, D.B. (1991). Significance levels from repeated p-values with multiply imputed data. *Statistica Sinica*, *1*, 65-92.

Li, K.H., Raghunathan, T.E., & Rubin, D.B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistic and an F-reference distribution. *Journal of the American Statistical Association*, *86*, 1065-73. doi.org/10.1080/ 01621459.1991.10475152

Livacic-Rojas, P.E., Vallejo, G. Fernández, P., & Tuero-Herrero, E. (2017). Power of modified Brown-Forsythe and mixed-model approaches in split-plot designs. *Methodology*, *13*, 9-22. doi.org/10.1027/1614-2241/a000124

Lix, L.M., Algina, J., & Keselman, H.J. (2003). Analyzing multivariate repeated measures designs: A comparison of two approximate degrees

of freedom procedures. *Multivariate Behavioral Research, 38*, 403-431. doi.org/10.1207/s15327906mbr3804_1

Meng, X.L., & Rubin, D.B. (1992). Performing likelihood ratio tests with multiply imputed data set. *Biometrika*, *79*, 103-111. doi.org/10.1093/biomet/79.1.103

Oliver-Rodríguez, J.C., & Wang, X.T. (2015). Non-parametric three-way mixed ANOVA with aligned rank tests. *British Journal of Mathematical and Statistical Psychology*, *68*, 23-42. doi.org/10.1111/bmsp.12031

Paniagua, D., Amor, P.J., Acheburúa, E., & Abad, F.J. (2017). Comparison of methods for dealing with missing values in the EPV-R. *Psicothema*, *29*, 384-389. doi.org/10.7334/psicothema2016.75

Raghunathan, T. (2016). *Missing Data Analysis in Practice*. Boca Raton, FL: Chapman and Hall/CRC.

Raghunathan, T., & Dong, Q. (2011). Analysis of variance from multiply imputed data sets. Unpublished manuscript, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan. Retrieved from http://www-personal.umich.edu/_teraghu/Raghunathan-Dong.pdf.

Ripley, B E. (1987). *Stochastic Simulation*. New York: Wiley.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D.B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, *81*, 366-374. doi.org/10.1080/01621459.1986.10478280

SAS Institute, Inc (2017). *SAS/STAT® 14.3 user's guide*. Cary, NC: SAS Institute, Inc.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Vallejo, G., & Ato, M. (2006). Modified Brown-Forsythe procedure for testing interaction effects in split-plot designs. *Multivariate Behavioral Research*, *41*, 549-578.

Vallejo, G., & Ato, M. (2012). Robust tests for multivariate factorial designs under heteroscedasticity. *Behavior Research Methods*, *44*, 471-489. doi.org/10.1207/s15327906mbr4104_6

Vallejo, G., Arnau, J., & Ato, M. (2007). Comparative robustness of recent methods for analyzing multivariate repeated measures. *Educational & Psychological Measurement*, *67*, 1-27. doi.org/10.1177/0013164406294777

Vallejo, G., Fernández, P., Herrero, F.J., & Conejo, N. M. (2004). Alternative procedures for testing fixed effects in repeated measures designs when assumptions are violated. *Psicothema, 16*, 498-508.

Vallejo, G., Ato, M., Fernández, M. P., & Livacic-Rojas, P.E. (in press). Sample size estimation for heterogeneous growth curve models with attrition. *Behavior Research Method*. doi.org/10.3758/s13428-018-1059-y

Vallejo, G., Fernández, M.P., Livacic-Rojas, P.E., & Tuero-Herrero, E. (2011a). Comparison of modern methods for analyzing unbalanced repeated measures data with missing values. *Multivariate Behavioral Research*, *46*, 900-937. doi.org/10.1080/00273171.2011.625320

Vallejo, G., Fernández, M.P., Livacic-Rojas, P.E., & Tuero-Herrero, E. (2011b). Selecting the best unbalanced repeated measures model. *Behavior Research Methods, 43*, 18-36. doi.org/10.3758/s13428-010-0040-1

Vallejo, G., Moris, J., & Conejo, N. (2006). A SAS/IML program for implementing the modified Brown-Forsythe procedure in repeated measures designs. *Computer Methods & Programs in Biomedicine, 83,* 169-177. doi.org/10.1016/j.cmpb.2006.06.006

Van Ginkel, J.R., & Kroonenberg, P.M. (2014). Analysis of variance of multiply imputed data. *Multivariate Behavioral Research, 49,* 78-91. doi.org/10.1080/00273171.2013.855890

Weerahandi, S. (2004). *Generalized Inference in Repeated Measures*: *Exact Methods in MANOVA and Mixed Models*. New Jersey, NJ: John Wiley & Sons.

Xu, L.W. (2015). Parametric bootstrap approaches for two-way MANOVA with unequal cell sizes and unequal cell covariance matrices. *Journal of Multivariate Analysis*, *133*, 291-303. doi.org/10.1016/j.jmva.2014.09.015