Small sample corrections for Wald tests in latent variable models

Brice Ozenne,
*Copenhagen University Hospital and University of Copenhagen, Denmark*

Patrick M. Fisher
*Copenhagen University Hospital, Denmark*

and Esben Budtz-Jørgensen
*University of Copenhagen, Denmark*

[Received May 2018. Final revision March 2020]

**Summary.** Latent variable models are commonly used in psychology and increasingly used for analysing brain imaging data. Such studies typically involve a small number of participants ($n < 100$), where standard asymptotic results often fail to control the type 1 error appropriately. The paper presents two corrections improving the control of the type 1 error of Wald tests in latent variable models estimated by using maximum likelihood. First, we derive a correction for the bias of the maximum likelihood estimator of the variance parameters. This enables us to estimate corrected standard errors for model parameters and corrected Wald statistics. Second, we use a Student $t$-distribution instead of a Gaussian distribution to account for the variability of the variance estimator. The degrees of freedom of the Student $t$-distributions are estimated by using a Satterthwaite approximation. A simulation study based on data from two published brain imaging studies demonstrates that combining these two corrections provides superior control of the type 1 error rate compared with the uncorrected Wald test, despite being conservative for some parameters. The methods proposed are implemented in the R package lavaSearch2, which is available from https://cran.r-project.org/web/packages/lavaSearch2.

**Keywords:** Latent variable models; Maximum likelihood; Repeated measurements; Small sample inference; Wald test

1. **Introduction**

Understanding brain mechanisms is essential to improve prevention and treatment of brain disorders. For instance, it has been hypothesized that the serotonin system is a key factor in major depressive disorders and most antidepressants attempt to act on this system. Unfortunately, less than half of the patients respond to first-line antidepressant treatment. A deeper understanding of the serotonin brain system is therefore needed. Although it is not yet possible to measure the extracellular level of serotonin in the brain non-invasively, medical imaging enables us to visualize the brain structure (using magnetic resonance imaging) and to quantify the binding potential of certain serotonin receptors (by using positron emission tomography (PET)) simultaneously; for example see Fig. 1.

*Address for correspondence:* Brice Ozenne, Neurobiology Research Unit 6931, Rigshospitalet, Copenhagen University Hospital, 9 Blegdamsvej, Copenhagen 2100, Denmark.

E-mail: brice.ozenne@nru.dk
Fig. 1. Single-subject serotonin 4 receptor binding and five regions of interest—amygdala (red areas), caudate (blue), hippocampus (purple), neocortex (white) and putamen (black): (a) regions of interest overlaid on a brain image depicting voxel level serotonin 4 receptor binding (colour scale units, serotonin 4 receptor binding); (b) the same regions overlaid on a high resolution structural magnetic resonance image, for spatial orientation

Recently, latent variable models (LVMs) have been used to identify the brain serotonin level from the binding potentials measured in several brain regions and to relate this level to patient group, test performance or genotype status (Fisher et al., 2015, 2017; Stenbæk et al., 2017; Deen et al., 2017; Perfalk et al., 2017; da Cunha-Bang et al., 2018). These models were estimated by maximum likelihood (ML) and statistical inference was most often performed based on the asymptotic distribution of Wald statistics. However, the sample size in these studies is rather limited (respectively 68, 144, 24, 34, 41 and 43), especially in light of the number of parameters that are required to obtain a satisfying model fit (respectively 29, 29, 48, 31, 37 and 40). The application of asymptotic results is thus questionable; for example, one may be worried that the type 1 error is not at its nominal level. This has been shown by using simulation studies for global fit tests (i.e. the likelihood ratio test; Herzog et al. (2007)), for which corrections have been proposed (Bentler and Yuan, 1999; Jiang and Yuan, 2017; Maydeu-Olivares, 2017). To our knowledge, the small sample properties of the Wald test in LVMs has not been carefully studied and software packages for LVMs, such as the R package lavaan (Rosseel, 2012) or Mplus (Muthén and Muthén, 2017), implement several small sample corrections for global fit tests, but no corrections for Wald tests.

Current solutions for small sample inference include profile likelihood (Pek and Wu, 2015), the use of resampling procedures: the bootstrap, permutation, jackknife, or the use of Bayesian techniques such as Markov chain Monte Carlo methods. The main drawback of these methods is that they are computationally intensive. In addition, each method has specific pros and cons. For instance, the bootstrap and jackknife may not appropriately control the type 1 error rate
because they rely on asymptotic results, e.g. Parr (1983) and Carpenter and Bithell (2000). Although permutation procedures appropriately control the type 1 error rate, they can test only very specific combinations of the model parameters. McNeish (2016) has shown that Markov chain Monte Carlo sampling is highly sensitive to the specification of the prior distributions of the parameters in small samples and may, therefore, not be straightforward to use. In this paper we focus on LVMs estimated by ML and propose an analytical approach to approximate the distribution of the Wald statistics. This approach does not require any user input nor any additional model fit. It modifies the usual asymptotic distribution of the Wald statistic in two ways, similarly to the correction that was proposed by Kenward and Roger (1997) for mixed models:

(a) correcting the bias of the ML estimator for variance parameters and
(b) modelling the distribution of the Wald statistics by using Student’s $t$-distributions instead of Gaussian distributions.

The remainder of the paper is structured as follows: we formally introduce LVMs and discuss the validity of the traditional testing procedure in Section 2. In Section 3, we illustrate the use of LVMs and the inflated type 1 error rate of the traditional testing procedure in three applications. Our small sample correction is presented in Sections 4 and 5. They respectively extend strategies (a) and (b) when testing a single hypothesis in LVMs. Extension to multiple hypotheses and robust standard errors are discussed in Section 6. The control of the type 1 error rate after correction is assessed in Section 7 by using simulation studies inspired from the three applications. These are reanalysed with the proposed correction in Section 8. We end the paper with a discussion. The correction proposed is implemented in an R package called lavaSearch2, which is available from the Comprehensive R Archive Network (https://cran.r-project.org/web/packages/lavaSearch2). An overview of the functionalities and code examples can be found in the vignette of the package. The code that was used for the simulation studies and for the illustrations is available from https://github.com/bozenne/Article-lvm-small-sample-inference.

2. Inference in linear latent variable models

We consider a sample of $n$ independent and identically distributed replicates $(\mathbf{Y}_i, \mathbf{X}_i)_{i \in \{1, \ldots, n\}}$ generated by $m$ endogenous random variables $\mathbf{Y} = (Y_1, \ldots, Y_m)$ and $l$ exogenous random variables $\mathbf{X} = (X_1, \ldots, X_l)$. An LVM is defined by a measurement model linking the endogenous variables to a set of latent variables $\eta$ and to the exogenous variables,

$$\mathbf{Y}_i = \nu + \eta_i \Lambda + \mathbf{X}_i \mathbf{K} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \Sigma_\epsilon), \quad (1)$$

and by a structural model relating the latent variables to the exogenous variables,

$$\eta_i = \alpha + \eta_i \mathbf{B} + \mathbf{X}_i \mathbf{\Gamma} + \zeta_i, \quad \zeta_i \sim \mathcal{N}(0, \Sigma_\zeta), \quad (2)$$

where $\mathbf{B}$ is a matrix with 0 on its diagonal and such that $1 - \mathbf{B}$ is invertible. We denote by $p$ the number of parameters, by $\boldsymbol{\theta}$ the vector containing the model parameters (we use the notation bold to denote row vectors) and by $\mathcal{S}_\theta$ the set of model parameters. The conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ follows from equations (1) and (2):

$$\mathbf{Y}_i | \mathbf{X}_i \sim \mathcal{N}\{\boldsymbol{\mu}(\theta, \mathbf{X}_i), \Omega(\theta)\}$$

$$\boldsymbol{\mu}(\theta, \mathbf{X}_i) = \nu + \alpha(1 - \mathbf{B})^{-1} \Lambda + \mathbf{X}_i \mathbf{\Gamma}(1 - \mathbf{B})^{-1} \Lambda + \mathbf{X}_i \mathbf{K},$$

$$\Omega(\theta) = \Lambda^T (1 - \mathbf{B})^{-T} \Sigma_\zeta (1 - \mathbf{B})^{-1} \Lambda + \Sigma_\epsilon. \quad (3)$$

The parameters can either be involved in the conditional mean, both in the conditional mean and
variance or only in the conditional variance. Parameters of the first type, i.e. parameters \( \nu, \alpha, K \) and \( \Gamma \), will be called mean parameters and denoted \( \theta_\mu \). Parameters satisfying the latter type, i.e. parameters in \( \Sigma_\varepsilon \) and \( \Sigma_\zeta \), will be called variance parameters and denoted \( \theta_\Sigma \). Estimation can be carried out by using ML; see Holst and Budtz-Jørgensen (2013) for more details. ML is known to give asymptotically unbiased, efficient and normally distributed estimates. For a given parameter \( \theta \in S_\theta \), we can use the Wald statistic

\[
I_\theta = \hat{\theta} / \hat{\sigma}_\theta
\]

(4)
to assess whether \( \theta = 0 \). Under the null hypothesis, \( I_\theta \) is asymptotically normally distributed with mean 0 and variance 1. Here \( \hat{\theta} \) denotes the value of \( \theta \) estimated by using ML and \( \hat{\sigma}_\theta \) is the standard deviation of the estimator (the variance–covariance matrix of the estimator of \( \theta \) will be denoted \( \Sigma_\theta \)). In most applications \( \hat{\sigma}_\theta \) is not known but we can plug the ML estimate of the standard error, \( \hat{\sigma}_\theta \), into equation (4) to obtain a tractable test statistic:

\[
\hat{I}_\theta = \hat{\theta} / \hat{\sigma}_\theta
\]

(5)

This has two consequences in finite samples:

(a) the variance of \( \hat{I}_\theta \) will typically be greater than 1 and

(b) \( \hat{I}_\theta \) may not be normally distributed because of the variability of \( \hat{\sigma}_\theta \).

Indeed, if \( \hat{\sigma}_\theta^2 \) follows a \( \chi^2 \)-distribution and is independent of \( \hat{\theta} \), then \( I_\theta \) follows a Student \( t \)-distribution (up to a multiplicative factor). Regarding consequence (a), using a first-order Taylor series expansion and taking the expectation, we can express the first-order bias of \( \hat{\sigma}_\theta \) as

\[
E[\hat{\sigma}_\theta - \sigma_\theta] = E[\hat{\theta} - \theta] \frac{\partial \sigma_\theta^T}{\partial \theta} + o_p(n^{-1/2}).
\]

(6)

In correctly specified models, \( \sigma_\theta \) can be consistently estimated by using the appropriate element in the inverse of the expected information matrix (denoted \( I_\theta \)). As shown in the on-line supplementary material section A.3, the expected information relative to the parameters \( \theta \) and \( \theta' \) in an LVM can be expressed as

\[
\mathcal{I}(\theta, \theta') = \frac{n}{2} \text{tr} \left\{ \Omega(\theta)^{-1} \frac{\partial \Omega(\theta)}{\partial \theta'} \Omega(\theta)^{-1} \frac{\partial \Omega(\theta)}{\partial \theta} \right\} + \sum_{i=1}^{n} \frac{\partial \mu(\theta, X_i)}{\partial \theta} \Omega(\theta)^{-1} \frac{\partial \mu(\theta, X_i)^T}{\partial \theta'}.
\]

(7)

Since \( \sigma_\theta \) is an element of the inverse of \( \mathcal{I}(\theta) \), it depends on the variance parameters via \( \Omega(\theta) \) so \( \partial \sigma_\theta / \partial \theta \) is typically non-zero. In small samples, the ML estimator of the variance parameters is in general biased (Harville, 1977); it follows that the first term of equation (6) is non-zero and \( \hat{\sigma}_\theta \) is biased in finite samples. We expect that the ML estimator of the variance parameters will be biased downwardly and \( \sigma_\theta \) will increase with the variance parameters, so \( \hat{\sigma}_\theta \) will be biased downwardly.

### 3. Applications to real data

We consider three applications involving simple to complex LVMs (see Fig. 2 for the corresponding path diagrams). The first investigates the correspondance between mixed models and LVMs and the next two were chosen from the studies on the serotonin system that was mentioned in Section 1. The latter two applications are representative of scientific questions that are encountered in that field: can we correlate a genetic variable to indirect measurements of the brain serotonin system? Can we study the relationship between indirect measurements of the brain serotonin system and indirect measurements of the cognitive ability of a subject?
Fig. 2. Path diagram of the LVM associated with each application (the observed variables are displayed in blue (outcomes) and green (covariates) whereas the latent variable is displayed in red) (the unions of the full and dotted arrows represent the relationships between variables modelled by the LVM; the black dotted arrows denote the parameters of interest and symbols above the full arrows represent constraints; for example in application A the residual variances are constrained to have the same value and the loadings to be 1): (a) application A, growth of guinea pigs; (b) application B, serotonin 4 receptor and genetic polymorphisms (Fisher et al., 2015); (c) application C, serotonin 4 receptor and verbal memory recall (Stenbaek et al., 2017)

3.1. Growth of guinea pigs (application A)
The guinea pig growth application originates from a practical class on mixed models (http://publicifsv.sund.ku.dk/~jufo/RepeatedMeasures2017.html) where the students are asked to monitor the growth of two groups of guinea pigs over 7 weeks by using a random-intercept model. One group of pigs received vitamin E at the beginning of week 5, whereas the other group serves as a control (see the on-line supplementary material section G for a graphical display of the data) so the treatment effect is modelled only from week 5 to week 7 by using a different parameter for each week. The data set that is used is small: each group contains only five animals. The random-intercept model is equivalent to an LVM where the loadings are set to 1 (i.e. $\Lambda = (\lambda_1, \ldots, \lambda_6) = (1, \ldots, 1)$) and the residual variances are assumed constant over time:
Y_{it} = \begin{cases} 
\nu_t + \eta_i + \epsilon_{it}, & \text{if } t < 5, \\
\nu_t + \eta_i + \text{group}_i k_{t-4} + \epsilon_{it}, & \text{if } t \geq 5, 
\end{cases}

where \( \epsilon_{it} \sim \mathcal{N}(0, \sigma^2) \) and \( \eta_i \sim \mathcal{N}(0, \tau) \). The corresponding path diagram is shown in Fig. 2(a).

The interest lies in assessing whether vitamin E affects the growth of the pigs. This can be performed by testing whether all treatment parameters are 0 (i.e. \( k_1 = k_2 = k_3 = 0 \)) by using a Wald test. In a mixed model, we obtained a Wald statistic of 4.25 and a \( p \)-value of 0.0102 whereas the Wald statistic of the LVM was 5.02 and the corresponding \( p \)-value was 0.00176. The large discrepancy is because, in mixed models, restricted ML is preferred over ML when estimating the model because it corrects the bias of the ML estimator of the variance parameters. Kenward and Roger (1997) proposed an additional correction by modelling the distribution of the Wald statistic with a Student \( t \)-distribution and estimating its degrees of freedom by the method of moments. This has become the gold standard procedure in mixed models and is what has been used here. This correction is, however, not available for LVMs. To assess the type 1 error of the Wald test obtained with an LVM, we performed a simulation study using the estimated LVM with all treatment effects set to 0 as a generative model. We then fitted the LVM to the simulated data and observed that, in 10.2% of the simulations, the null hypothesis was rejected (at a 5% threshold), i.e. the type 1 error rate was inflated by 0.052.

3.2. Serotonin 4 receptor binding and genetic polymorphisms (application B)
Fisher et al. (2015) were interested in whether two genetic variants, i.e. BDNF val66met and 5-HTTLPR polymorphisms, were associated with brain serotonin levels. They collected genetic data and PET brain imaging data, using two scanners, from 68 healthy humans. The final data set contained 73 observations because five participants were scanned twice. The serotonin 4 receptor binding, a proxy for brain serotonin levels, was computed in five brain regions (amygdala, caudate, hippocampus, neocortex and putamen). Preliminary regression analyses suggested an association between the BDNF val66met genotype and serotonin 4 receptor levels in all brain regions, whereas the effect of the 5-HTTLPR polymorphisms was found to be specific to only the neocortex region. Data were subsequently analysed in an LVM where the regional serotonin 4 binding measurements were linked to a single latent variable, representing an unobservable brain serotonin level. This latent variable was affected by BDNF val66met genotype status to model a global effect (\( \Gamma \)-parameter), whereas the 5-HTTLPR effect was directly modelled on the neocortex region (\( K \)-parameter). Covariates that were assumed to affect the serotonin level were related to the latent variable. To remove systematic differences in serotonin measurements, a direct effect of the scanner on each brain region was modelled. The path diagram of the LVM is displayed in Fig. 2(b). The LVM had 29 parameters.

Inference was performed using cluster robust standard errors, i.e. the score was computed at the participant level when using the sandwich estimator. The effects of the BDNF val66met and 5-HTTLPR polymorphisms were estimated to be respectively \( \hat{\gamma}_2 = 0.074 \) (\( p \)-value 0.005) and \( \hat{k}_1 = -0.073 \) (\( p \)-value \( 7 \times 10^{-6} \)). In an LVM estimated under the constraint of no genetic effects as a generative model, a simulation study showed that the actual type 1 error rate was 0.074 for the effect of the BDNF val66met and 0.061 for the effect of the 5-HTTLPR polymorphisms.

3.3. Serotonin 4 receptor binding and verbal memory recall (application C)
Stenbæk et al. (2017) investigated the relationship between episodic memory and serotonin 4 receptor binding. They collected data from 24 healthy volunteers who underwent a PET scan and a verbal affective memory test. PET and memory measures were acquired proximal to
each other but not at the same time. In the verbal affective memory test, subjects are asked to remember words immediately after having learned them (immediate memory), 5 min after (short-term memory) and 30 min after (long-term memory). Words were divided into three categories: positive, negative and neutral valence words. The serotonin 4 receptor binding was measured in four brain regions that are known to be involved in affective processing and memory (amygdala, the anterior cingulate cortex, frontal cortex and hippocampus). Four latent variables were constructed: three that combine the immediate, short-term and long-term memory for each type of word and one that combines the serotonin 4 receptor binding across the brain regions (respectively $\eta_{\text{positive}}$, $\eta_{\text{neutral}}$, $\eta_{\text{negative}}$ and $\eta_{5HT4}$). Associations between the latent variables were adjusted for age. The path diagram of the LVM is shown in Fig. 2(c). The resulting LVM had 48 parameters.

With this LVM, they found an effect of the binding construct ($\eta_{5HT4}$) on the memory constructs ($\eta_{\text{positive}}, \eta_{\text{neutral}}, \eta_{\text{negative}}$) with respectively $b_1 = -7.3$ (p-value 0.0005), $b_2 = -6.7$ (p-value 0.004) and $b_3 = -3.7$ (p-value 0.07). A simulation study, using the estimated LVM as a generative model under the constraint of no relationship between memory and serotonin 4 receptor binding, led to a type 1 error rate of 0.063 for $b_1$, 0.085 for $b_2$ and 0.084 for $b_3$ (instead of 0.05) when using an (uncorrected) Wald test.

4. Bias correction for the maximum likelihood estimator of the variance parameters

We now develop a method to correct the small sample bias of the estimated variance parameters, $\hat{\theta}_\Sigma$. Denoting the observed residuals by $\xi_i(\hat{\theta}) = Y_i - \mu(\hat{\theta}, X_i)$, it is well known that in a standard linear model the variance of the observed residuals underestimates the (true) conditional variance of $Y$. We show in the on-line supplementary material section B that this result generalizes to LVMs. Indeed, given that $E[\xi_i(\theta)^T \xi_i(\theta)] = \Omega(\theta)$ and $\mathcal{I}(\theta)^{-1} = \Sigma_\theta$, the variance of the observed residuals can be decomposed into

$$E[\xi_i(\hat{\theta})^T \xi_i(\hat{\theta})] = \Omega(\theta) - \Psi_i + o_p(n^{-1})$$

$$\Psi_i = \frac{\partial \mu(\theta, X_i)^T}{\partial \theta} \Sigma_\theta \frac{\partial \mu(\theta, X_i)}{\partial \theta}.$$

Since the first-order bias $\Psi_i$ is positive definite, the variance of the observed residuals is a downwardly biased estimate of $\Omega(\theta)$. This result is similar to that found by Kauermann and Carroll (2001) for generalized estimating equation models. Denoting $\Psi = (1/n) \Sigma_{i=1}^n \Psi_i$, we obtain, by averaging over the samples,

$$E \left[ \frac{1}{n} \sum_{i=1}^n \xi_i(\hat{\theta})^T \xi_i(\hat{\theta}) \right] = \Omega(\theta) - \Psi + o_p(n^{-1}).$$

4.1. Example (standard linear model)

Consider the generative mechanism $Y_i = X_i \beta + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, where $Y$ is a univariate endogenous variable, $X$ contains $l$ exogenous variables and $\varepsilon_i$ are independent and identically normally distributed. As shown in the on-line supplementary material section C.1, formula (8) gives that $\Psi_i = \sigma^2 X_i (X^T X)^{-1} X_i^T$ and $\Psi = (p/n) \sigma^2$. Note that the first-order bias of the ML estimator can be removed by considering the estimator

$$(\hat{\sigma}^2)^2 = \hat{\sigma}^2 + \hat{\Psi} = \left(1 + \frac{p}{n}\right) \hat{\sigma}^2.$$
Section C.1 shows that the estimated bias of the residual variance at iteration \( k \) is 
\[
\hat{\Psi}^{(k)} = \left( \frac{p}{n} \right) \left( \hat{\sigma}^{(k-1)} \right)^2.
\]

The corresponding corrected residual variance is
\[
\left( \hat{\sigma}^{(k)} \right)^2 = \hat{\sigma}^2 \sum_{k=0}^{k} \frac{p^k}{n^k} \rightarrow \frac{n}{n-p} \sigma^2.
\]

The corrected residual variance tends towards the usual unbiased estimate of \( \sigma^2 \). The convergence is fast, especially when \( n \) is large, since there is a factor \( p/n \) between the contribution of the current iteration and the contribution of the next iteration. The same applies to \( \hat{\Psi}^{(k)} \). Note that \( \hat{\Psi}^{(k)} \) and \( \left( \hat{\sigma}^{(k)} \right)^2 \) are increasing sequences.

### 4.3. Example (mean–variance model)

We consider an LVM where no parameter appears in both the conditional mean and variance. This corresponds to common factor models where \( \alpha = 0 \) and \( \Gamma = 0 \), or mixed models where \( B = 0 \) and \( \Lambda \) is known (e.g. equals 1 for random-intercept models as in application A). In both

**Table 1. Algorithm 1**

| Initialize: \( I^{(0)} = \hat{I}(\theta) \) |
| for \( k = 1 \) to \( \infty \) do |
| \( \begin{align*}
& (a) \text{ compute for each subject } i \text{ the bias } \hat{\psi}_i^{(k)} \text{ by plugging } (I^{(k-1)})^{-1} \text{ and } \partial \mu(\hat{\theta}, X_i) / \partial \theta \text{ into equation (8); } \\
& (b) \text{ compute the average bias } \hat{\psi}^{(k)} = (1/n) \Sigma_i \hat{\psi}_i^{(k)}; \\
& (c) \text{ estimate } \hat{\theta}^{(k)} \text{ by } \hat{\theta}^{(k)} + \hat{\psi}^{(k)}; \\
& (d) \text{ find } \hat{\theta}_C^{(k)} \text{ satisfying } \hat{\sigma}^{(k)} = \hat{\Lambda}^{T}(1 - \hat{B})^{-T} \Sigma_\psi \hat{\theta}_C^{(k)} (1 - \hat{B})^{-1} \hat{\Lambda} + \Sigma_\psi \hat{\theta}_C^{(k)}; \\
& (e) \text{ compute the corrected derivatives } \partial \hat{\theta}_C^{(k)} / \partial \theta \text{ by using } \hat{\theta}_C^{(k)}; \\
& (f) \text{ compute } I^{(k)} \text{ by plugging } \hat{\theta}_C^{(k)}, \partial \mu(\hat{\theta}, X_i) / \partial \theta \text{ and } \partial \hat{\theta}_C^{(k)} / \partial \theta \text{ into equation (7)}.
\end{align*} \\end{aligned} |
| end |

a much smaller bias for the other parameters (for example, see on-line supplementary material section F.2). To do so, we assume that \( \Omega(\hat{\theta}) \) and \( \mathbb{E}[(1/n) \Sigma_i \xi_i(\hat{\theta})^T \xi_i(\hat{\theta})] \) have the same first-order bias. Then, given \( \Psi \), we can define a corrected ML estimator of \( \Omega: \hat{\Omega} = \Omega(\hat{\theta}) + \Psi \). From equation (3) and considering \( (\Lambda, B) \) fixed, we obtain that \( \Omega(\hat{\theta}) \) is linearly related to the variance parameters, so we can find a matrix \( Z \) (depending only on \( \Lambda \) and \( B \)) such that \( \vec{\vec{\Omega}} = Z \theta_C + r \). The conver-
cases, \( \partial \Omega(\theta) / \partial \theta \) and \( \partial \mu(\theta) / \partial \theta \) cannot be simultaneously non-zero for a given parameter, so we obtain from formula (7) that the information matrix is block diagonal. We show in the on-line supplementary material section D.1 that, if the number of mean parameters is smaller than \( n \), algorithm 1 converges and \( \Psi^{(k)} \) increases (in the sense of the spectral norm) over iterations.

The corrected estimates of the variance parameters that are obtained by algorithm 1 can be substituted in \( \hat{\theta} \) for the initial estimates to obtain \( \hat{\theta}^c \). As an important side product, we obtain also a corrected expected information matrix, denoted \( \hat{\Psi}^c \), that can be used to calculate a corrected Wald statistic.

### 4.4. Example (standard linear model)

In the standard linear model, the variance of the ML estimator of the regression parameters is \( \text{var}(\hat{\beta}) = \sigma^2 / (X'X) \). Using \( \hat{\Psi}^c \) instead of \( \hat{\Psi} \) is equivalent to plugging in \( \hat{\sigma}^2 \), an unbiased estimate of \( \sigma^2 \), instead of \( \hat{\sigma}^2 \), a downwardly biased estimate of \( \sigma^2 \). Whereas using algorithm 1 leads to a satisfactory estimator for \( \text{var}(\hat{\beta}) \), the estimator of \( \text{var}(\hat{\sigma}^2) \) can still be improved.

Indeed, the variance of \( \hat{\sigma}^2 \) equals \( 2\sigma^4 / (n - p) \) since

\[
\sum_{i=1}^{n} \frac{\hat{e}_i^2}{\sigma^2} \sim \chi^2_{n-p}.
\]

However, the variance estimator that is obtained after inverting \( \hat{\Psi}^c \) is \( 2(\hat{\sigma}^2)^4 / n \) which is downwardly biased in finite samples. We shall return to this problem in Section 5.3.

### 5. Modelling the distribution of the Wald statistics by using Student \( t \)-distributions

In this section we propose a method to account for the uncertainty in \( \hat{\sigma}^2 \) when deriving the distribution of the Wald statistic.

#### 5.1. Satterthwaite approximation

In a standard linear model, \( \hat{\sigma}^2 \) is known to be \( \chi^2 \)-distributed. Indeed \( \hat{\sigma}^2 \) is proportional to the residual variance, which can be expressed as the sum of the residuals squared. This motivates the use of a Student \( t \)-distribution instead of a Gaussian distribution for the Wald statistic. In multivariate models like LVMs, the distribution of \( \hat{\sigma}^2 \) is not generally known. However, it can still be approximated by using a \( \chi^2 \)-distribution by finding \( (\tau, \text{df}) \in (\mathbb{R}, \mathbb{R}^+ \) such that

\[
\tau \hat{\sigma}^2 \overset{\text{approx.}}{\sim} \chi^2(\text{df}).
\]

Here \( \tau \) and \( \text{df} \) can be identified from the method of moments: using that a \( \chi^2 \)-distribution with \( \text{df} \) degrees of freedom has expectation \( \text{df} \) and variance \( 2\text{df} \), we obtain that \( \mathbb{E}[\tau \hat{\sigma}^2] = \text{df} \) and \( \text{var}(\tau \hat{\sigma}^2) = 2\text{df} \). Therefore, \( \tau \) and \( \text{df} \) should satisfy

\[
\tau = \text{df} / \mathbb{E}[\hat{\sigma}^2],
\]

\[
\text{df} = 2\mathbb{E}[\hat{\sigma}^2]^2 / \text{var}(\hat{\sigma}^2)
\]
Denoting \( \theta^* = \hat{\theta} / \sigma_\theta \) and \( \sigma_{\theta}^2 = \text{df} \hat{\sigma}_\theta^2 / \sigma_\theta^2 \), we can rewrite the test statistic \( i_\theta \) as \( \theta^* / (\sigma_{\theta}^2 / \sqrt{\text{df}}) \).

Under the null hypothesis, \( \theta^* \sim \mathcal{N}(0, 1) \) and \( \sigma_{\theta}^2 \sim \chi^2(\text{df}) \), \( i_\theta \) follows a Student \( t \)-distribution with \( \text{df} \) degrees of freedom. This approximation is a classical technique in mixed models and it has been implemented in many software tools, e.g. SAS procedure \texttt{MIXED} or the R package \texttt{lmerTest} (Kuznetsova et al., 2017).

5.2. Application to latent variable models
Although we can directly substitute the estimate \( \hat{\sigma}_\theta^2 \) in equation (9) for \( \mathbb{E}[\hat{\sigma}_\theta^2] \), we need an estimator for \( \text{var}(\hat{\sigma}_\theta^2) \) to estimate \( \text{df} \). For a given \( \theta \in S_\theta \), \( \hat{\sigma}_\theta^2 = c_j I(\hat{\theta})^{-1} c_j^T \) where \( j \) is the index of \( \theta \) in \( S_\theta \) and \( c_j \) is a vector with a 1 at the \( j \)th position and 0 otherwise. From equation (7) we see that \( I(\theta) \) depends only on the model parameters (and on \( X_i \), which are fixed values). Using that \( \hat{\theta} \) is asymptotically normally distributed with variance \( \Sigma_\theta \), we can apply the multivariate delta method to obtain an estimator for \( \text{var}(\hat{\sigma}_\theta^2) \):

\[
n^{1/2}(\hat{\sigma}_\theta^2 - \sigma_\theta^2) \sim \mathcal{N}(0, \nabla_\theta \sigma_\theta^2 \Sigma_\theta \nabla_\theta \sigma_\theta^2 T^T).
\]

Here, \( \nabla_\theta \) denotes the vector of partial derivatives relative to each parameter in \( \theta \). Therefore \( \text{df} \) can be estimated by using the following procedure.

(a) For each \( k \in \{1, \ldots, p\} \), compute \( \partial I(\hat{\theta}) / \partial \theta_k \), the first derivative of the information matrix (see the on-line supplementary material section A.4).

(b) For each \( k \in \{1, \ldots, p\} \), compute

\[
\frac{\partial \hat{\sigma}_\theta^2}{\partial \theta_k} = -c_j I(\hat{\theta})^{-1} \frac{\partial I(\hat{\theta})}{\partial \theta_k} I(\hat{\theta})^{-1} c_j^T.
\]

Combining all the partial derivatives into a vector gives \( \nabla_\theta \hat{\sigma}_\theta^2 \).

(c) Estimate the degrees of freedom of \( \hat{\sigma}_\theta^2 \) as \( 2(\hat{\sigma}_\theta^2)^2 / (\nabla_\theta \hat{\sigma}_\theta^2 \Sigma_\theta \nabla_\theta \hat{\sigma}_\theta^2 T^T) \).

This approach can be generalized to any linear combination of parameters. To do so, the vector \( c_j \) in the expressions of \( \hat{\sigma}_\theta^2 \) and \( \partial \hat{\sigma}_\theta^2 / \partial \theta_k \) should be replaced by the vector defining the linear combination.

5.2.1. Example (standard linear model)
We denote by \( \sigma_\beta^2 = \sigma^2 / (X^T X) \) the variance of the estimated regression parameter. The variance of \( \hat{\sigma}_\beta^2 \) that is obtained with the delta method is \( 2\hat{\sigma}_\beta^4 / n \). The Satterthwaite approximation gives \( \text{df} = 2\hat{\sigma}_\beta^4 / (2\hat{\sigma}_\beta^4 / n) = n \) as the estimate of the degrees of freedom for the Wald statistic (online supplementary material section C.2). Although this approximation is better than using a standard normal distribution, it does not match the true value of \( n - p \) for the degrees of freedom. The estimator of the residual variance in the standard linear models is \( \chi^2 \) distributed with \( n - p \) degrees of freedom because the score equation induces \( p \) constraints between the observed residuals.

5.3. Effective sample size
So far, we have neglected the loss in degrees of freedom that is caused by the estimation of the parameters, i.e. using the actual sample size \( n \) is an upwardly biased estimator of the number of independent residuals. This number is used when computing the first term of the information matrix (equation (7)) and, as illustrated in the previous example, the bias also affects the
estimation of the degrees of freedom of the Wald statistic. We define the effective sample size as the sum of the dependence of each observed residual on the corresponding observation:

\[ n^c = \sum_{i=1}^{n} \frac{\partial \xi_i(\hat{\theta})}{\partial Y_i} = n - \sum_{i=1}^{n} \frac{\partial \mu(\hat{\theta}, X_i)}{\partial Y_i} \]  

(10)

where \( n^c = (n_1^c, \ldots, n_m^c) \) is the vector of effective sample sizes, with one element per endogenous variable, and \( n = (n_1, \ldots, n) \). If the observations would not affect the fit, then each element of \( n \) would equal \( n \). However, the constraints on the residuals reduce the variation of \( \xi_i(\hat{\theta}) \) relative to \( Y_i \), leading to each element of \( n^c \) being smaller than \( n \). We see that \( n^c \) depends on \( \partial \mu(\hat{\theta}, X_i)/\partial Y_i \), the generalized leverage, as defined by Wei et al. (1998).

5.3.1. Example (standard linear model)
We recover the standard result that the effective sample size is \( n^c = n - p \) (on-line supplementary material section C.3). The estimator of the variance of \( (\hat{\sigma}^c)^2 \) becomes \( 2\hat{\sigma}^2/(n - p) \) and the degrees of freedom of the Wald statistic obtained with the Satterthwaite approximation are \( n - p \): we now have unbiased estimators of the variance of \( (\hat{\sigma}^c)^2 \) and of the associated degrees of freedom.

5.3.2. Example (mean–variance model)
The effective sample size relative to the \( r \)th endogenous variable can be expressed as

\[ n - \sum_{i=1}^{n} \frac{\partial \mu(\hat{\theta}, X_i)}{\partial \mu} \left\{ \sum_{i=1}^{n} \frac{\partial \mu(\hat{\theta}, X_i)}{\partial \mu} \Omega(\theta)^{-1} \frac{\partial \mu(\hat{\theta}, X_i)}{\partial \mu} \right\}^{-1} \frac{\partial \mu(\hat{\theta}, X_i)}{\partial \mu} \Omega(\theta)^{-1} c_t \]

where \( c_t \) is an \( m \)-dimensional vector containing 1 at the \( t \)th position and 0 otherwise.

Algorithm 1 can be modified to compute the effective sample size and to obtain corrected degrees of freedom for the Wald statistics (on-line supplementary material section B.3).

6. Extensions
The Wald statistic that is defined in equation (4) can be modified to simultaneously testing several null hypotheses that can be defined via a non-singular contrast matrix \( C \) of rank \( Q \):

\[ F_C \theta = \frac{1}{Q} (C \theta)^T (C \Sigma_C C^T)^{-1} (C \theta). \]  

(11)

For instance, in application A, we test whether \( k_1 = k_2 = k_3 = 0 \), i.e. \( C \theta = 0 \) where \( C \) is a \( 3 \times p \) matrix containing 0 everywhere except in the first row (second and third), in the column corresponding to \( k_1 \) (respectively \( k_2 \) and \( k_3 \)) where it contains 1s. Asymptotically this statistic follows a \( \chi^2 \)-distribution with \( Q \) degrees of freedom, but in finite samples using a Fisher distribution may improve the type 1 error control. A Satterthwaite approximation can also be derived to estimate the degrees of freedom of the Fisher distribution (for example see the on-line supplementary material section E).

The sample that was collected in application B violates the assumption of independent and identically distributed replicates since some subjects were scanned twice: if \( i \) refers to the first scan and \( i' \) to the second scan of the same individual \( Y_i \) and \( Y_{i'} \) are likely to be positively correlated. However, we can find a partition of the observations into \( G \) clusters where the assumption of independent and identically distributed replicates is reasonable, by grouping the observations that were collected on the same individual. In that case \( G = 68 \) (while \( n = 73 \)) and we shall use \( G_1, \ldots, G_G \) to denote the clusters. Denoting \( U(\theta | Y_i, X_i) \) the individual score (see the
on-line supplementary material section A.2 for the mathematical expression), we can define the score of the $g$th cluster: $U_{g}(\theta) = \Sigma_{g} U(\theta|Y, X)$. When performing inference, White (1982) has shown that the robust estimator

$$
\Sigma_{\hat{\theta}}^{\text{robust}} = \mathcal{I}(\theta)^{-1} \left\{ \sum_{g=1}^{G} U_{g}(\theta)^{T} U_{g}(\theta) \right\} \mathcal{I}(\theta)^{-1}
$$

(12)
can consistently estimate the variance of $\hat{\theta}$, meaning that a Wald test based on a robust estimator of the variance will control the type 1 error (asymptotically). Such a test will be referred to as a robust Wald test. Note that the consistency of $\Sigma_{\hat{\theta}}^{\text{robust}}$ does not involve a distributional assumption on $Y|X$, which makes it robust to deviations to normality. In finite samples, however, we can expect that the robust Wald test suffers from the same limitations as the traditional Wald test. Fortunately, we can apply our bias correction to the estimator of $\Sigma_{\hat{\theta}}^{\text{robust}}$ by plugging the corrected score and information matrix in equation (12). As an approximation, we set the
degrees of freedom of the robust standard errors to be identical to those of the (non-robust) standard errors.

7. Simulation study

We performed three simulation studies to investigate the effect of the proposed corrections on the bias of the estimates and on the control of the type 1 error. The LVMs that were used in the simulation studies are simplified versions of the LVMs that were used in the real data applications (see Fig. 3 for the corresponding path diagrams). A simulation study was characterized by

(a) a generative model, i.e. the model defining the distribution that is used to simulate the data,
(b) an investigator model, i.e. the model fitted to the simulated data, and
(c) the set of null hypotheses, each testing whether one of the model parameters equals 0.

For each study, 20000 samples were generated by using the generative model. Each sample was used to estimate the parameters of the investigator model. Then, each null hypothesis was tested separately with a Wald test using

(a) the standard procedure (uncorrected information matrix; Gaussian distribution),
(b) the bias correction (corrected information matrix),
(c) the Satterthwaite approximation (Student’s t-distribution with degrees of freedom estimated according to Section 5) and
(d) the complete correction, i.e. bias correction and Satterthwaite approximation.

The type 1 error was then computed as the relative frequency of p-values lower than 0.05 and is displayed in Fig. 4 for a subset of the model parameters. Using the same simulation setting, we also investigated the type 1 error control of the robust Wald test (as defined in Section 6) when using the observations as clusters (i.e. \( n = G \)). The corresponding results are displayed in Fig. 5. The small sample bias was assessed for ML estimates and after application of the bias correction (ML-corrected estimates) and is displayed in Table S1 in the on-line supplementary material section F.2. When performing the simulation in small samples, the estimation algorithms did not always converge. The convergences issues and how they were handled are detailed in supplementary material section F.1.

7.1. Wald test in a mixed model (study A)

The first LVM is equivalent to a random-intercept model, where the endogenous variable is measured on three brain regions per subject. The resulting LVM has seven parameters. The first null hypothesis tests whether the conditional expectation of the endogenous variable is the same between the first and second repetition (within-subject parameter \( \nu_2 \)). The second null hypothesis tests whether there is an effect of genetic variant 1 on all repetitions of the endogenous variable (between-subject parameter \( \gamma_1 \)).

The ML estimates of the variance parameters showed a small bias that was removed by our correction (Table S1 in the on-line supplementary material section F.2). Without correction, a moderate inflation of the type 1 error rate was observed for the Wald tests, e.g. 0.015 for \( \nu_2 \) and 0.037 for \( \gamma_1 \) when \( n = 20 \) (Fig. 4(a)). The bias correction combined with the Satterthwaite approximation provided a satisfactory control of the type 1 error rate, e.g. 0.050 for \( \nu_2 \) and \( \gamma_1 \) when \( n = 20 \).

7.2. Robust Wald test in a single-factor model (study B)

Compared with the first simulation, the second LVM relaxes the assumption of a common vari-
Fig. 4. Type 1 error rate of the Wald test for study (a) A (mixed model), (b) B (single-factor model) and (c) C (two-latent-variable model) at various sample sizes (the type 1 error rates for the parameters that are not shown on the graph (e.g. testing $\lambda_2 = 1$) are similar to those shown (e.g. testing $\lambda_4 = 0$); at a given sample size, 20000 simulations were performed, meaning that the type 1 error rate is expected to fluctuate between 0.047 and 0.053): ■, standard procedure; ●, Satterthwaite approximation; △, bias correction; ★, Satterthwaite approximation with bias correction.
Fig. 5. Type 1 error rate of the robust Wald test for study (a) A (mixed model), B (single-factor model) and (c) C (two-latent-variable model) at various sample sizes (the type 1 error rates for the parameters that are not shown on the graph (e.g. testing $\lambda_2^D$) are similar to those shown (e.g. testing $\lambda_4^D$); at a given sample size, 20000 simulations were performed, meaning that the type 1 error rate is expected to fluctuate between 0.047 and 0.053): ■, standard procedure; ▲, Satterthwaite approximation; ●, bias correction; ○, Satterthwaite approximation with bias correction.
ance and covariance between the measurements, adds another polymorphism (genetic variant 2) and a fourth brain region. The resulting LVM had 15 parameters. Two additional null hypotheses are considered: whether the new region is correlated with the first ($\lambda_4$) and whether there is an effect of genetic variant 2 on the first region ($k_1$).

The simulation indicates that the ML estimator of the residual variance and of the variance of the latent variable were downwardly biased (Table S2 in the on-line supplementary material section F.2), i.e. the estimated variances are too small. For instance, for $n = 20$, the average bias was $-0.125$ for the residual variance and $-0.150$ for the variance of the latent variable ($1$ is the true value). The correction proposed was partially able to correct the bias; for example, for $n = 20$, the average bias became $-0.029$ for the residual variance and to $-0.018$ for the variance of the latent variable. Without correction, the inflation of the type 1 error rate in the robust Wald test was dependent on the coefficient (Fig. 5(b)): $0.038$ for $\nu_2$, $0.063$ for $\lambda_4$, $0.059$ for $\gamma_1$ and $0.067$ for $k_1$ when $n = 20$. The complete correction provided satisfactory control for $\nu_2$ and $\gamma_1$ (type 1 error $0.041$ and $0.056$ respectively for $n = 20$) but was slightly too liberal for $k_1$ (type 1 error $0.074$) and slightly too conservative for $\lambda_4$ (type 1 error $0.036$).

7.3. Wald test in a latent variable model with two latent variables (study C)

We now add a new latent variable ($\eta_{\text{memory}}$) and an additional measurement per latent variable. The resulting LVM had 36 parameters. The new null hypotheses test whether, conditionally on the latent variable, there is a residual covariance between the first two regions ($\sigma_{12}$) and whether the first latent variable influences the second ($b_1$).

The results were similar to those for study B: the covariance parameter $\sigma_{12}$ did not show any bias, whereas the correction reduced the bias for the other parameters (Table S3 in the on-line supplementary material section F.2). The inflation of the type 1 error was especially noticeable because the model involved more parameters (Fig. 4(c)). The bias correction alone reduced the inflation of the type 1 error rate by $0.01$ to $0.03$. When combined with the Satterthwaite correction, the resulting procedure was satisfactory for $\gamma_1$ (type 1 error $0.061$ when $n = 20$), conservative for $\nu_2$, $\lambda_4$, $b_1$ and $\sigma_{12}$ (type 1 error respectively $0.032$, $0.023$, $0.013$ and $0.024$ when $n = 20$) but still liberal for $k_1$ (type 1 error $0.078$ when $n = 20$).

7.4. Distribution of the Wald statistic in small samples

The simulation results show that the correction proposed does not always control the type 1 error rate exactly at the nominal level. This may be because the estimated Wald statistics are not $t$ distributed or that our estimators of the variance and degrees of freedom behave poorly in small samples. In the on-line supplementary material section F.3, we compare the distribution of the Wald statistic that was obtained after correction with the empirical distribution (obtained by simulation). We found that our estimators performed well in scenario A. However, in scenarios B and C, the corrected variance was slightly biased downwards and the Satterthwaite estimator of the degrees of freedom performed poorly for some parameters. The Student $t$-distribution appeared to be a good approximation of the empirical distribution except for two types of parameter: $b_1$ and $\sigma_{12}$.

8. Application of the small sample correction to real data

We now rerun the Wald tests that were presented at the end of Section 3 by using the small sample correction that was developed in Sections 4, 5 and 6. Table 2 gives an overview of the statistical tests that were performed with or without correction and the associated type 1 error (obtained by simulation).
Table 2. Statistical inference without (columns ‘ML’) and with (columns ‘ML with correction’) the proposed correction in the three applications (A, B and C) considered†

| Application | Parameter | Results for ML | Results for ML with correction |
|-------------|-----------|----------------|---------------------------------|
|             |           | σ   df  p-value | Type 1 error (%)      | σ   df  p-value | Type 1 error (%)      |
| A           | $k_1, k_2, k_3$ | $\infty$ 1.76×10^{-3} 10.22 | 49.1 1.23×10^{-2} 4.49 |
| B           | $\gamma_2$ | 0.026 $\infty$ 5.11×10^{-3} 7.41 | 0.027 65.4 9.03×10^{-3} 5.59 |
|             | $k_1$ | 0.016 $\infty$ 7.23×10^{-6} 6.14 | 0.017 67.5 4.80×10^{-5} 4.78 |
| C           | $b_1$ | 2.106 $\infty$ 5.45×10^{-4} 6.32 | 2.143 13.5 4.53×10^{-3} 2.02 |
|             | $b_2$ | 2.355 $\infty$ 4.21×10^{-3} 8.50 | 2.429 10.0 1.96×10^{-2} 3.71 |
|             | $b_3$ | 2.041 $\infty$ 7.20×10^{-2} 8.41 | 2.130 7.7 1.24×10^{-1} 3.42 |

†The parameter $\sigma$ refers to the standard error associated with the estimated parameter and df refers to the degrees of freedom of the Wald statistic. The type 1 error is computed by simulation under the null hypothesis.

Table 3. Comparison of the ML estimates, the proposed correction and the restricted ML estimates (using the Kenward and Roger (1997) correction) in application A

| Estimate            | Residual variance | Variance random intercept | Statistic | Degrees of freedom | p-value |
|---------------------|-------------------|---------------------------|-----------|--------------------|---------|
| ML                  | 0.148             | 0.349                     | 5.024     | $\infty$           | 0.0018  |
| ML with correction  | 0.177             | 0.389                     | 4.019     | 49.11              | 0.0123  |
| Restricted ML       | 0.176             | 0.389                     | 4.247     | 43.59              | 0.0102  |

8.1. Growth of guinea pigs (application A)

The correction that we proposed in this paper shares similarities with the correction of Kenward and Roger (1997): we attempt to correct the bias of the ML estimator when estimating the variance parameters and to use a Student $t$-distribution to model the distribution of the Wald statistic. The techniques that are used, however, differ; for example, our bias-corrected ML estimator is not identical to the restricted ML estimator. Nevertheless, the estimates of the residual variance and the variance of the random intercept that were obtained with our correction were similar to those obtained with restricted ML (relative difference less than 1%; Table 3). They were 19% and 11% larger than their corresponding ML estimates. The corrected test statistic was 4.02 with a corresponding $p$-value of 0.0123. Although our method does not replicate the results that were obtained with the correction of Kenward and Roger (1997) exactly, it gives estimates of the same order of magnitude. To validate our correction further, we performed the same simulation study as in the uncorrected case and found a type 1 error of 0.045 with the correction proposed.

8.2. Serotonin 4 receptor binding and genetic polymorphisms (application B)

Compared with the ML estimates of the variances parameters, the corrected (non-robust) variance estimates were larger by 23.9% (neocortex), 4.5% (caudate), 4.0% (putamen), 3.8% (hippocampus), 3.8% (amygdala) and 10.9% (latent variable). When using cluster robust standard
errors, the corrected \( p \)-values were 0.009 for the effect of the BDNF val66met (plus 76\% compared with the original \( p \)-value) and \( 4.8 \times 10^{-5} \) for the effect of 5-HTTLPR (plus 564\%). The same simulation study as in the uncorrected case gives, after correction, a type 1 error of 0.056 for the BDNF val66met effect and 0.048 for the 5-HTTLPR effect. Since the type 1 error that was estimated by simulations is close to the nominal level after correction and the corrected \( p \)-values are still below the critical level, this new analysis supports the conclusions of Fisher et al. (2015).

8.3. Serotonin 4 receptor binding and verbal memory recall (application C)

The bias correction increased the estimates of the variance parameters by a factor ranging between 4.8\% and 21.2\% (endogenous variables) and 9.1\% and 10.2\% (latent variables). The corrected \( p \)-values were 0.0045 for \( b_1 \) (plus 731\% compared with the original \( p \)-value), 0.02 for \( b_2 \) (plus 366\%) and 0.12 for \( b_3 \) (plus 72\%). The same simulation study as in the uncorrected case gives, after correction, a type 1 error of 0.02 for \( b_1 \), 0.037 for \( b_2 \) and 0.034 for \( b_3 \) after correction. This new analysis supports the existence of an association between serotonin 4 receptor binding and recall of positive and neutral words (\( b_1 \) and \( b_2 \)). The parameter \( b_3 \) did not reach significance before correction, where the testing procedure is liberal, so we should retain the null hypothesis for \( b_3 \).

In these applications, although the correction did not affect the conclusion of the statistical tests (when using a significance threshold of 0.05), the corrected \( p \)-values were better calibrated and therefore better reflected the strength of evidence against the null hypothesis.

9. Discussion

Concerns have been raised in the applied scientific literature about the lack of statistical power and the lack of reliability of studies involving small samples; for example, see Button et al. (2013) and Bakker et al. (2012). Compared with \( t \)-tests or linear regressions, multivariate approaches such as LVMs can be used to increase the power of testing procedures. They also provide a common framework to test the hypotheses of the investigator and to assess modelling assumptions. Although exact tests can be performed on univariate models, only approximate tests are tractable with LVMs. Using simulation studies, we performed a detailed investigation of control of the type 1 error rate when using Wald tests in LVMs with small samples. The overall conclusion from these simulations was that the type 1 error rate is inflated in small samples. For a sample size of 20, the type 1 error rate varied between 0.06 and 0.12 for the Wald test and between 0.07 and 0.14 for the robust Wald test. The nominal level of 0.05 was reached when the sample size reached 100–200, depending on the type of model parameter.

We proposed two corrections to obtain a better control of the type 1 error rate: a correction for the bias of the ML estimator of the variance parameters and the use of a Student \( t \)-distribution instead of a normal distribution to account for the uncertainty in the estimate of the variance of the model parameters. The corrections proposed have some desirable features:

(a) when combined, they match the traditional corrections performed in univariate linear models,
(b) they match the uncorrected ML inference in large samples,
(c) they are fast to compute, require no user input and can be applied to a large variety of models, and
(d) the first correction reduced the bias of the variance estimates and improved the control of the type 1 error in all studies.
Regarding feature (c), our implementation had a very reasonable run time: 75–475 ms in study A, 200–900 ms in study B and 1.5–6 s in study C. It converged in very few iterations, except with very small samples and complex LVMs. One drawback of the proposed testing procedure is that it is not parameterization invariant, meaning that different identifiability constraints (e.g. setting to 1 a loading in the measurement model or the variance of the latent variable) may lead to different $p$-values. This is a well-known issue when using Wald tests, and not specific to our corrections. Alternative test statistics (e.g. the likelihood ratio test) should be considered if this property is required (Larsen and Jupp, 2003).

A careful inspection of the simulation results showed that using a $t$-distribution to model the distribution of the Wald statistics was a good approximation for most parameters. We think that the inexact control of the type I error in small samples is mainly due to the poor performance of our estimator of the degrees of freedom. The estimation of the standard error could also be improved; indeed our bias correction does not completely remove the bias from the estimator of the variance parameter. A better bias correction may be achieved by using the formula of Cox and Snell (1968) for the small sample bias of the ML estimator. It gives an estimate of the small sample bias up to $o_p(n^{-1})$ but involves complex calculations (third-order derivative of the likelihood). The extension of the correction to robust standard errors could also be improved. Indeed the small sample correction has been derived assuming that the model was correctly specified (more precisely that $E[\xi_i(\theta)\xi_i(\theta)^\top] = \Omega(\theta)$ and $I(\theta)^{-1} = \Sigma_{\theta}$) and the current approximation for the degrees of freedom does not depend on the choice of the clusters $G_1, \ldots, G_G$. The estimation of the degree of freedom will perform poorly when the clusters contain many observations. We investigated other approximations (e.g. Pan and Wall (2002)) but did not obtain satisfying results. Finally, we note that the Satterthwaite correction was derived for the expected information matrix. In theory, a similar correction could be derived for the observed information matrix, but it would require more tedious derivations and complexify the software implementation. Nevertheless, this may be necessary in specific contexts; for example, see Savalei (2010) for a case where the expected information does not give consistent standard errors (missing data problems). In our software package implementing the proposed corrections, we provide a function called `calibrateType1` that can be used to assess the type I error of the corrected and uncorrected Wald tests via simulations—under the assumption that the investigator model is correctly specified. We hope that this will help to detect inflations in the type I error rate and improve the reliability of studies involving small samples.

As pointed out by a reviewer, alternative estimation techniques such as instrumental variables (IVs) (Bollen, 1996), generalized least squares (GLS) and weighted least squares (WLS) (Yuan and Bentler, 1997) could compare favourably with ML in small samples. Although a comprehensive comparison between these estimation techniques is beyond the scope of the present paper, we performed an additional simulation to compare ML, IV, GLS and WLS in study B under a correctly specified model (see the on-line supplementary material section F.4). We found that GLS and IVs showed an inflation of the type I error in small samples that is similar in magnitude to ML (uncorrected). WLS failed to estimate the model parameter for $n = 20$ and $n = 30$; it also had the worst control of the type I error. This poor behaviour of WLS in small samples is consistent with the existing literature (Olsson et al., 2000). Given the appealing properties of IVs (Bollen et al., 2007), it would be of interest to propose a small sample correction for IVs estimation.

**Acknowledgements**

BO was supported by the Lundbeck foundation (R231-2016-3236). This project has received
funding from the European Union's 'Horizon 2020' research and innovation programme under Marie Sklodowska-Curie grant agreement 746850.

References

Bakker, M., van Dijk, A. and Wicherts, J. M. (2012) The rules of the game called psychological science. Perspect. Psychol. Sci., 7, 543–554.

Bentler, P. M. and Yuan, K.-H. (1999) Structural equation modeling with small samples: test statistics. Multiv. Behav. Res., 34, 181–197.

Bollen, K. A. (1996) An alternative two stage least squares (2SLS) estimator for latent variable equations. Psychometrika, 61, 109–121.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M. and Chen, F. (2007) Latent variable models under misspecification: two-stage least squares (2SLS) and maximum likelihood (ML) estimators. Sociol. Meth. Res., 36, 48–86.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. and Munafò, M. R. (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neur., 14, no. 5, 365–376.

Carpenter, J. and Bithell, J. (2000) Bootstrap confidence intervals: when, which, what?: A practical guide for medical statisticians. Statist. Med., 19, 1141–1164.

Cox, D. R. and Snell, E. J. (1968) A general definition of residuals. J. R. Statist. Soc. B, 30, 248–275.

da Cunha-Bang, S., Fisher, P. M., Hjordt, L. V., Perfalk, E., Beliveau, V., Holst, K. and Knudsen, G. M. (2018) Men with high serotonin 1b receptor binding respond to provocations with heightened amygdala reactivity. Neuronlmage, 166, 79–85.

Deen, M., Hansen, H. D., Hougaard, A., da Cunha-Bang, S., Norgaard, M., Svarer, C., Keller, S. H., Thomsen, C., Ashima, M. and Knudsen, G. M. (2017) Low 5-htr1b receptor binding in the migraine brain: a PET study. Cephalalgia, 38, 519–527.

Fisher, P. M., Holst, K. K., Adriansen, D., Klein, A. B., Frokjaer, V. G., Jensen, P. S., Svarer, C., Gilllings, N., Baare, W. F., Mikkelson, J. D. and Knudsen, G. M. (2015) Bdnf val66met and 5-httlpr polymorphisms predict a human in vivo marker for brain serotonin levels. Hum. Brain Mapping, 36, 313–323.

Fisher, P., Ozene, B., Svarer, C., Adriansen, D., Lehel, S., Baaré, W., Jensen, P. and Knudsen, G. (2017) Bdnf val66met association with serotonin transporter binding in healthy humans. Transshl Psychiatr., 7, no. 2, article e1029.

Harville, D. A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Statist. Ass., 72, 320–338.

Herzog, W., Boomsma, A. and Reinecke, S. (2007) The model-size effect on traditional and modified tests of covariance structures. Struct. Equn Modlng, 14, 361–390.

Holst, K. K. and Budtz-Jørgensen, E. (2013) Linear latent variable models: the lava-package. Computnl Statist., 28, 1385–1452.

Jiang, G. and Yuan, K.-H. (2017) Four new corrected statistics for SEM with small samples and nonnormally distributed data. Struct. Equn Modlng, 24, 479–494.

Kauermann, G. and Carroll, R. J. (2001) A note on the efficiency of sandwich covariance matrix estimation. J. Am. Statist. Ass., 96, 1387–1396.

Kenward, M. G. and Roger, J. H. (1997) Small sample inference for fixed effects from restricted maximum likelihood. Biometrics, 53, 983–997.

Kuznetsova, A., Brockhoff, P. B. and Christensen, R. H. B. (2017) lmertest package: tests in linear mixed effects models. J. Statist. Softwr., 82, no. 13, 1–26.

Larsen, P. V. and Jupp, P. (2003) Parametrization-invariant Wald tests. Bernoulli, 9, 167–182.

Maydeu-Olivares, A. (2017) Maximum likelihood estimation of structural equation models for continuous data: standard errors and goodness of fit. Struct. Equn Modlng, 24, 383–394.

McNeish, D. (2016) On using bayesian methods to address small sample problems. Struct. Equn Modlng, 23, 750–773.

Muthén, L. K. and Muthén, B. O. (2017) Mplus User's Guide, 8th edn. Los Angeles: Muthén and Muthén.

Olsson, U. H., Foss, T., Troye, S. V. and Howell, R. D. (2000) The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. Struct. Equn Modlng, 7, 557–595.

Pan, W. and Wall, M. M. (2002) Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. Statist. Med., 21, 1429–1441.

Parr, W. C. (1983) A note on the jackknife, the bootstrap and the delta method estimators of bias and variance. Biometrics, 70, 719–722.

Pek, J. and Wu, H. (2015) Profile likelihood-based confidence intervals and regions for structural equation models. Psychometrika, 80, 1123–1145.

Perfalk, E., da Cunha-Bang, S., Holst, K. K., Keller, S., Svarer, C., Knudsen, G. M. and Frokjaer, V. G. (2017)
Testosterone levels in healthy men correlate negatively with serotonin 4 receptor binding. *Psychoneuroendocrinology*, 81, 22–28.

Rosseel, Y. (2012) lavaan: an R package for structural equation modeling. *J. Statist. Softwr.*, 48, no. 2, 1–36.

Savalei, V. (2010) Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychol. Meth.*, 15, no. 4, 352–367.

Stenbæk, D. S., Fisher, P. M., Ozenne, B., Andersen, E., Hjordt, L. V., McMahon, B., Hasselbalch, S. G., Frokjaer, V. G. and Knudsen, G. M. (2017) Brain serotonin 4 receptor binding is inversely associated with verbal memory recall. *Brain Behav.*, 7, no. 4, article e00674.

Wei, B.-C., Hu, Y.-Q. and Fung, W.-K. (1998) Generalized leverage and its applications. *Scand. J. Statist.*, 25, 25–37.

White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Yuan, K.-H. and Bentler, P. M. (1997) Mean and covariance structure analysis: theoretical and practical improvements. *J. Am. Statist. Ass.*, 92, 767–774.

**Supporting information**
Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplementary material for the article Small sample corrections for Wald tests in latent variable models’.