Computational Statistics (2023) 38:1–23
https://doi.org/10.1007/s00180-022-01214-7

ORIGINAL PAPER

Controlling the familywise error rate when performing multiple comparisons in a linear latent variable model

Brice Ozenne[1,2] · **Esben Budtz-Jørgensen**[1] · **Sebastian Elgaard Ebert**[2]

Received: 8 May 2021 / Accepted: 15 February 2022 / Published online: 25 March 2022

Abstract

In latent variable models (LVMs) it is possible to analyze multiple outcomes and to relate them to several explanatory variables. In this context many parameters are estimated and it is common to perform multiple tests, e.g. to investigate outcome-specific effects using Wald tests or to check the correct specification of the modeled mean and variance using a forward stepwise selection (FSS) procedure based on Score tests. Controlling the family-wise error rate (FWER) at its nominal level involves adjustment of the *p*-values for multiple testing. Because of the correlation between test statistics, the Bonferroni procedure is often too conservative. In this article, we extend the max-test procedure to the LVM framework for Wald and Score tests. Depending on the correlation between the test statistics, the max-test procedure is equivalent or more powerful than the Bonferroni procedure while also providing, asymptotically, a strong control of the FWER for non-iterative procedures. Using simulation studies, we assess the finite sample behavior of the max-test procedure for Wald and Score tests in LVMs. We apply our procedure to quantify the neuroinflammatory response to mild traumatic brain injury in nine brain regions.

Keywords Latent variable model · Multiple comparisons · Max-test procedure · Familywise error rate

B.O. has received funding from the Lundbeck foundation (R231-2016-3236) and the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 746850.

✉ Brice Ozenne
brice.ozenne@nru.dk

[1] Section of Biostatistics, University of Copenhagen, 5 Øster Farimagsgade, 1014 Copenhagen, Denmark

[2] Neurobiology Research Unit, Rigshospitalet and University of Copenhagen, 6-8 Inge Lehmanns Vej, 2100 Copenhagen, Denmark

1 Introduction

Latent variable models (LVM) are an attractive tool for studying systems of variables where an exposure, e.g. a treatment or a disease, is to be related to several outcome variables, e.g. the concentration of a specific protein in various brain regions. They are able to jointly analyze several dependent variables, relate them to exogenous factors, and to investigate shared correlation structures. They encompass linear regressions, probit models, and mixed models as specific cases (Holst and Budtz-Jørgensen 2013). They also admit a graphical representation called path diagram (Fig. 1).

In applications, LVMs include a high number of parameters and several of them (or combinations) are of interest. Global tests, such as likelihood ratio tests, enable to simultaneously test several null hypotheses but provide no guidance as to which null hypotheses are false. However, this is of critical importance, e.g. when investigating the effect of a disease on several brain regions, and motivate the use of separate tests. Traditional adjustments for handling multiple testing like Bonferroni ignore the correlation between test statistics. If the correlation is strong, e.g. above 0.7 in our real data application, power is lost which is problematic in fields like neuroscience where the inclusion of many subjects is expensive.

Another situation where multiple testing arises is model checking. Traditionally, practitioners specify LVMs drawing a path diagram based on a priori information, fit the model, and assess its goodness of fit. In absence of a priori information, one often considers a parsimonious structure, e.g. a single latent variable is sufficient to capture the covariance structure (i.e. no double-headed arrows). Searching for local dependencies (i.e. conditional dependency between two variables that are not connected in the path diagram) is then a recommended practice (Ropovik 2015) and identifying omitted local dependencies will raise doubts about the validity of the LVM. A possible remedy is to sequentially include the omitted local dependencies until the fit of the model is considered satisfying. At each step, the most relevant local dependency is identified using Score tests over the set of possible additional local dependencies. While widely used, such forward stepwise selection (FSS) procedures have been criticized due to low reproducibility (MacCallum et al. 1992) and inflation of Type 1 error (Ropovik 2015).

Efficient methods for handling multiple comparisons have been developed for a number of years (see Dmitrienko and D'Agostino 2013 for an introduction in the context of clinical trials), but they are not often used in conjuction with LVMs. For instance, the max-test procedure is not implemented in statistical softwares specialized in LVM such as M-plus (Muthén and Muthén 2017), the PROC CALIS in SAS, the R packages lava (Holst and Budtz-Jørgensen 2013) or lavaan (Rosseel 2012). A few articles have stressed the importance of controlling the FWER in the LVM literature, promoting the Bonferroni procedure (Cribbie 2000; Cudeck and O'dell 1994) or a procedure similar to Bonferroni–Holm when performing backward stepwise selection (Green et al. 2001). Interestingly, Smith and Cribbie (2013) proposed a modified Bonferroni procedure to account for the correlation between the test statistics. This is performed in an ad-hoc way by correcting the number of tests to adjust for using the average absolute correlation between estimated coefficients. In particular, there is no guarantee that the FWER is appropriately controlled (it is not difficult to construct

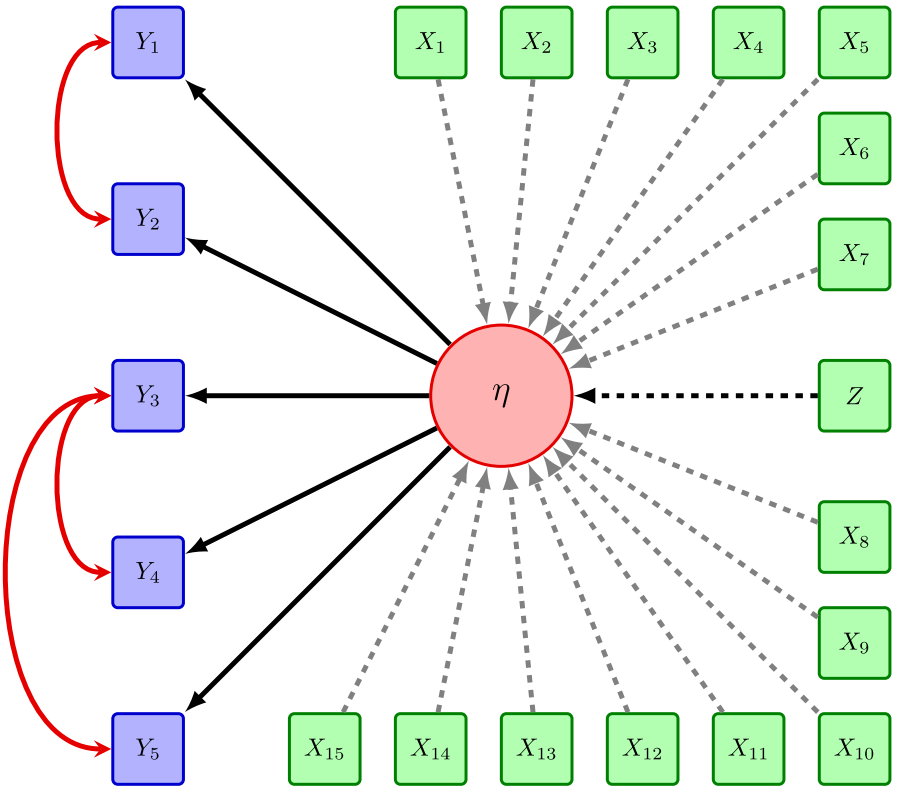

Fig. 1 Path diagram of the generative LVM used for the simulations in Sect. 5.2. The outcomes are represented in blue, the latent variable in red, and the covariates in green. Regression links are indicated with black single-headed arrows, covariance links are indicated with red double-headed arrow, and the absence of arrows or a gray arrow indicates conditional independence between two variables. Dashed arrows indicate the links that are tested in the FSS (not displayed in a traditional path diagram)

examples where it is not the case). In comparison a max-test procedure can be shown to control the FWER while having a power advantage over the Bonferroni procedure. It can be carried in a parametric way for normally (or Student's t) distributed test statistics (Hothorn et al. 2008). Westfall and Troendle (2008) show that the procedure generalizes to other distributions provided that one can compute cumulative distribution function (cdf) of the maximum. However, for Score tests, the cdf of the maximum of χ^2 variables is difficult to calculate.

In this article, we extend the parametric max-test procedure to LVMs (i) when testing multiple parameters using Wald statistics and (ii) when using a Score test. To achieve (i), we apply the max-test procedure proposed by Hothorn et al. (2008) in conjunction with a modification of the classical Wald statistics (Ozenne et al. 2020) to obtain a max-test procedure for LVMs that is valid in small samples (e.g. n=36 in our real data application). To achieve (ii), we have developed two novel procedures to approximate the max-distribution for χ^2 distributed variables. The max-test procedures are implemented in a package for the R software called `lavaSearch2`, available on CRAN

(https://cran.r-project.org/web/packages/lavaSearch2/index.html). The code used for the simulation studies and for the data analysis is available at https://github.com/bozenne/Article-lvm-multiple-comparisons.

2 Latent variable model (LVM) framework

Let us consider a vector of outcome variables $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ and a vector of covariates $\boldsymbol{X} = (X_1, \ldots, X_l)$ with arbitrary distribution. We observe a sample $(\mathcal{X}_i)_{i\in\{1,\ldots,n\}} = (\boldsymbol{y}_i, \boldsymbol{x}_i)_{i\in\{1,\ldots,n\}}$ of n replications of $\mathcal{X} = (\boldsymbol{Y}, \boldsymbol{X})$. We assume that the sample contains independent and identically distributed (iid) replicates. For this we consider a LVM, denoted $\mathcal{M}(\boldsymbol{\Theta})$, which models the conditional distribution of the outcomes as a function of the covariates and the vector of model parameters $\boldsymbol{\Theta}$ through a normal distribution:

$$\boldsymbol{Y}|\boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{\mu}(\boldsymbol{\Theta}, \boldsymbol{X}), \Omega(\boldsymbol{\Theta})\right)$$

To express the conditional mean $\boldsymbol{\mu}(\boldsymbol{\Theta}, \boldsymbol{X})$ and the conditional variance $\Omega(\boldsymbol{\Theta})$, we introduce a set of latent variables $\boldsymbol{\eta}$ and relate them to the observed variables via the measurement model:

$$\boldsymbol{Y} = \boldsymbol{\nu} + \Lambda\boldsymbol{\eta} + K\boldsymbol{X} + \boldsymbol{\varepsilon}, \text{ where } \boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \Sigma_\varepsilon\right)$$

and the structural model:

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + B\boldsymbol{\eta} + \Gamma\boldsymbol{X} + \boldsymbol{\zeta}, \text{ where } \boldsymbol{\zeta} \sim \mathcal{N}\left(0, \Sigma_\zeta\right)$$

where $\boldsymbol{\varepsilon} \perp\!\!\!\perp \boldsymbol{\zeta}$ ($\perp\!\!\!\perp$ denotes stochastic independence) and B is a matrix with 0 on its diagonal and such that $1 - B$ is invertible. We also impose constraints on the parameters $\boldsymbol{\alpha}$ and Λ, typically that their first element is, respectively, 0 and 1, to ensure identifiability of the model. Thus $\boldsymbol{\Theta}$ contains the unconstrained parameters of $\boldsymbol{\nu}, \lambda, K, \Sigma_{\boldsymbol{\varepsilon}}, \boldsymbol{\alpha}, B, \Sigma_\zeta$. See supplementary material E for an example. In this model, the conditional mean and variance are:

$$\begin{aligned}\boldsymbol{\mu}(\boldsymbol{\Theta}, \boldsymbol{X}) &= \boldsymbol{\nu} + \Lambda(1 - B)^{-1}\boldsymbol{\alpha} + \left[\Lambda(1 - B)^{-1}\Gamma + K\right]\boldsymbol{X} \\ \Omega(\boldsymbol{\Theta}) &= \Lambda(1 - B)^{-1}\Sigma_\zeta(1 - B)^{-\intercal}\Lambda^\intercal + \Sigma_\varepsilon\end{aligned}$$

and the log-likelihood is:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{Y}, \boldsymbol{X}) = \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{Y}_i, \boldsymbol{X}_i) = \sum_{i=1}^{n} -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\Omega(\boldsymbol{\Theta})| \\ - \frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{\mu}(\boldsymbol{\Theta}, \boldsymbol{X}_i))\Omega(\boldsymbol{\Theta})^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}(\boldsymbol{\Theta}, \boldsymbol{X}_i))^\intercal\end{aligned}$$

Modification of this model to handle non-linear effects of some covariates, clustered data, binary data, or censored data can be found in Holst and Budtz-Jørgensen (2013). Note that in the following, we will assume that four regularity conditions are satisfied, at least in a neighborhood of $\boldsymbol{\Theta}_0$: (i) $\boldsymbol{\Theta}_0$ is interior to the space of possible parameter values, (ii) distinct $\boldsymbol{\Theta}$ values represent distinct distributions, (iii) Σ_ε and Σ_ζ, are positive definite, and (iv) $\frac{\partial \boldsymbol{\mu}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}$ and $\frac{\partial \Omega(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}$ are of full column rank. We denote by $\widehat{\boldsymbol{\Theta}}$ the estimate of $\boldsymbol{\Theta}$ obtained by maximum likelihood (ML) estimation. The estimation can be carried out using the Newton-Raphson and iteratively computing the vector of scores:

$$\mathcal{S}(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \mathcal{S}_i(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \frac{\partial \log(\mathcal{L}(\boldsymbol{\Theta}|Y_i, X_i))}{\partial \boldsymbol{\Theta}}$$

and the expected information matrix:

$$\mathcal{I}(\boldsymbol{\Theta}) = -\sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial^2 \log(\mathcal{L}(\boldsymbol{\Theta}|Y_i, X_i))}{\partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}^T}\right] = n\mathcal{I}_1(\boldsymbol{\Theta})$$

for updating $\widehat{\boldsymbol{\Theta}}$ until convergence. Explicit expression for $\mathcal{S}$ and $\mathcal{I}$ in LVMs can be found in supplementary material A. They can be used to show that $\mathcal{I}$ is invertible under (i-iv). We can then obtain an estimate $\hat{\Sigma}_{\widehat{\boldsymbol{\Theta}}}$ of the variance-covariance matrix of $\widehat{\boldsymbol{\Theta}}$ (denoted $\Sigma_{\widehat{\boldsymbol{\Theta}}}$) in two ways: by estimating model-based variance-covariance matrix, $\Sigma_{m,\widehat{\boldsymbol{\Theta}}} = \mathcal{I}(\boldsymbol{\Theta})^{-1}$, or using the robust variance-covariance matrix, $\Sigma_{r,\widehat{\boldsymbol{\Theta}}} = \mathcal{I}(\boldsymbol{\Theta})^{-1}\mathcal{S}(\boldsymbol{\Theta})^\intercal\mathcal{S}(\boldsymbol{\Theta})\mathcal{I}(\boldsymbol{\Theta})^{-1}$. In addition, from the theory of M-estimators (e.g. see Tsiatis 2006, section 3.2), we get that ML estimators for LVM are asymptotically linear. This means that there exists a function $\psi_{\boldsymbol{\Theta}}$, called the influence function, such that:

$$\sqrt{n}\left(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_0, \mathcal{X}_i) + o_p(1) \tag{1}$$

where $\boldsymbol{\Theta}_0$ denotes the true value of $\boldsymbol{\Theta}$ and $\psi_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_0, \mathcal{X}_i) = \mathcal{S}_i(\boldsymbol{\Theta}_0)\mathcal{I}_1(\boldsymbol{\Theta}_0)^{-1}$ are iid terms.

Hypothesis testing in LVMs estimated by ML can be done using classical tests such as likelihood ratio tests, Score tests, or Wald tests. When testing parameters or combinations of parameters, Wald tests are the privileged approach since confidence intervals can easily be obtained. For model building, e.g. when deciding whether to include a new parameter, Score tests are often preferred for computational reasons.

3 Multiple inference using Wald tests

In this section, we consider a LVM $\mathcal{M}(\boldsymbol{\Theta})$ and assume that it is correctly specified. This means that there exists a vector $\boldsymbol{\Theta}_0$ such that $\mathcal{M}(\boldsymbol{\Theta}_0)$ is the distribution that

has generated $(\boldsymbol{y}_i)_{i\in\{1,\dots,n\}}$ given $(\boldsymbol{x}_i)_{i\in\{1,\dots,n\}}$. We also denote by $\Sigma_{\widehat{\boldsymbol{\Theta}},0}$ the true value of $\Sigma_{\widehat{\boldsymbol{\Theta}}}$; in a univariate linear model $Y = X\beta + \varepsilon$ where $\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$ we have $\Sigma_{\hat{\beta},0} = \sigma^2(X^\intercal X)^{-1}$. We are interested in testing c null hypotheses:

$$\left\{\begin{array}{c} \mathcal{H}_0^{(1)} : \beta_{1,0} = 0 \text{ vs. } \mathcal{H}_1^{(1)} : \beta_{1,0} \neq 0 \\ \dots \\ \mathcal{H}_0^{(c)} : \beta_{c,0} = 0 \text{ vs. } \mathcal{H}_1^{(c)} : \beta_{c,0} \neq 0 \end{array}\right\}$$

where $(\beta_1, \dots, \beta_c)$ are distinct elements of $\boldsymbol{\Theta}$ and $\beta_{1,0}$ denotes the true value of β_1, e.g. if $\beta_1 = \Theta_1$ then $\beta_{1,0} = \Theta_{1,0}$. This set of null hypotheses can also be written in a matrix form: $\boldsymbol{\beta}_0 = \mathbf{0}$ or, more generally, $C\boldsymbol{\Theta} = \mathbf{b}$ where C is any full rank matrix (often called contrast matrix) and $\mathbf{b}$ is a vector. From maximum likelihood theory (e.g. Van der Vaart (2000), section 5.5), we know that:

$$\sqrt{n}\left(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\right) \overset{d}{\sim} \mathcal{N}\left(\mathbf{0}, \mathcal{I}_1^*(\boldsymbol{\Theta})^{-1}\right)$$

where $\overset{d}{\sim}$ denotes convergence in distribution as $n \to \infty$ and $\mathcal{I}_1^*(\boldsymbol{\Theta})$ that large sample limit of $\mathcal{I}_1(\boldsymbol{\Theta})$. We introduce $D_{\widehat{\boldsymbol{\beta}}}$ the diagonal matrix containing the diagonal elements of $C\mathcal{I}(\boldsymbol{\Theta})^{-1}C^\intercal$. Defining the vector of Wald statistics by $\mathbf{T} = D_{\widehat{\boldsymbol{\beta}}}^{-\frac{1}{2}}\left(C\widehat{\boldsymbol{\Theta}} - \mathbf{b}\right)$, we have under the null:

$$\sqrt{n}\, D_{\widehat{\boldsymbol{\beta}}}^{\frac{1}{2}}\mathbf{T} \overset{d}{\sim} \mathcal{N}\left(\mathbf{0}, C\mathcal{I}_1^*(\boldsymbol{\Theta})^{-1}C^\intercal\right)$$

When testing $\mathcal{H}_0^{(1)}, \dots, \mathcal{H}_0^{(c)}$, C is the identity matrix and $\mathbf{T}$ is simply $\left(\frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_1}}, \dots, \frac{\hat{\beta}_c}{\sigma_{\hat{\beta}_c}}\right)$, where $\sigma_{\hat{\beta}_j}$ is the square root of the j-th diagonal element of $\mathcal{I}(\boldsymbol{\Theta})^{-1}$. Since $\left(n\hat{\mathcal{I}}_1(\boldsymbol{\Theta})\right)^{-1}$ converges in probability toward $\Sigma_{\widehat{\boldsymbol{\Theta}},0}$ and using the same arguments as in Hothorn et al. (2008), we can approximate (under the null hypothesis) the distribution of $\mathbf{T}$ by a normal distribution with mean 0 and variance-covariance matrix $\hat{\Sigma}_{\mathbf{T}} = \hat{D}_{\widehat{\boldsymbol{\beta}}}^{-\frac{1}{2}} C\,\hat{\Sigma}_{\widehat{\boldsymbol{\Theta}}} C^\intercal \hat{D}_{\widehat{\boldsymbol{\beta}}}^{-\frac{1}{2}}$. Defining $|T|_{\max} = \max_{j\in\{1,\dots,c\}} |T_j|$, where T_j is the j-th element of $\mathbf{T}$, an adjusted p-value for the j-th statistic can be obtained by computing $1 - \mathbb{P}\left[|T|_{\max} \leq |t_j|\right]$ where:

$$\mathbb{P}\left[|T|_{\max} \leq |t|\right] = \int_{\mathbf{u}\in[-t;t]^c} f_{\mathbf{T}}(\mathbf{u})d\mathbf{u}. \tag{2}$$

Here $f_{\mathbf{T}}$ denotes the joint density of $\mathbf{T}$ and $[-t; t]^c$ the cartesian product of c intervals $[-t; t]$. Asymptotically $f_{\mathbf{T}}$ equals the density of a multivariate Gaussian distribution with $\mathbf{0}$ mean and variance-covariance matrix $\Sigma_{\mathbf{T}}$. In finite samples, one can use the asymptotic distribution as an approximation for the distribution $f_{\mathbf{T}}$. This procedure is called a max-z test.

The previous derivations do not account for the fact that, in practice, $D_{\widehat{\boldsymbol{\beta}}}$ is estimated and plugged-in to estimate $\mathbf{T}$ nor for the small sample bias of the ML estimator. As a result, Wald tests based on the asymptotic ML theory generally show inflated type 1 error rates in small samples. For linear models and linear mixed models it is recommended to model the distribution of the Wald statistic using a Student's t-distribution and to estimate the variance parameters using restricted maximum likelihood (REML). Because REML has not been developed for LVMs, Ozenne et al. (2020) recently proposed a procedure, called hereafter "small sample correction", to correct the finite sample bias of the ML estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}$ and used a Satterwaithwaite approximation to estimate the degrees of freedom corresponding to the Student's t-distribution. This enables us to use a multivariate Student's distribution in equation (2) instead of a Gaussian distribution, and will be referred to as a max-t test. As an approximation, the degrees of freedom of the multivariate Student's t-distribution is computed as the average of the Wald's degrees of freedom. We summarize the multiple testing procedure in the following definition:

Definition 1 (Single step max-test procedure for Wald tests in LVM) Given a LVM estimated by ML with estimated parameter $\widehat{\boldsymbol{\Theta}}$ and estimated variance-covariance matrix $\hat{\Sigma}_{\widehat{\boldsymbol{\Theta}}}$, the max-test procedure for testing the set of null hypotheses $\left(\mathcal{H}_0^{(j)}\right)_{j\in\{1,\dots,c\}}$ using Wald tests is:

1. Extract $\widehat{\boldsymbol{\beta}} = C\widehat{\boldsymbol{\Theta}}$ the estimated parameters relative to each null hypothesis from $\widehat{\boldsymbol{\Theta}}$.
2. Extract $\hat{\Sigma}_{\widehat{\boldsymbol{\beta}}} = C\hat{\Sigma}_{\widehat{\boldsymbol{\Theta}}}C^\intercal$ the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ from $\hat{\Sigma}_{\widehat{\boldsymbol{\Theta}}}$ and denote $\hat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}$ its diagonal elements. Create $\hat{D}_{\widehat{\boldsymbol{\beta}}}$ the diagonal matrix containing $\hat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}$. When using a max-t test, extract $df_{\widehat{\boldsymbol{\beta}}}$ the degrees of freedom relative to $\hat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}$ and use the bias-corrected estimate of the variance-covariance matrix obtained from the small sample correction for $\hat{\Sigma}_{\widehat{\boldsymbol{\Theta}}}$.
3. Compute the Wald statistics as $\hat{\mathbf{T}} = \hat{D}_{\widehat{\boldsymbol{\beta}}}^{-\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \mathbf{b})$.
4. Compute p-values using formula (2) with $f_{\mathbf{T}}$ being the density of a multivariate normal distribution (max-z test) or of a multivariate Student's t-distribution (max-t test) with variance-covariance matrix $\hat{\Sigma}_{\widehat{\boldsymbol{\beta}}}$. When using a max-t test, estimate the degree of freedom of the Student's t-distribution by $\frac{1}{c}\sum_{j=1}^{c} df_{\widehat{\beta}_j}$
5. Compute confidence intervals using $\left[\widehat{\boldsymbol{\beta}} - q_\alpha\hat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}; \widehat{\boldsymbol{\beta}} + q_\alpha\hat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}\right]$ where q_α is the $1-\frac{\alpha}{2}$ equicoordinate quantile of $f_{\mathbf{T}}$.

Note that at step 4, resampling technics (e.g. Chernozhukov et al. (2013)) could also be used to obtain a non-parametric estimator of $\mathbb{P}\left[|T|_{\max} \leq |t|\right]$ based the iid decomposition of equation (1) instead of assuming a multivariate normal or Student's t-distribution and performing numerical integration to compute the right hand side of equation (2).

The max-test procedure enjoys several desirable properties: asymptotically, it provides a strong control of the FWER, i.e. the probability of incorrectly rejecting at least

one hypothesis is at most 5%. It is asymptotically exact in the sense that the FWER tends to 5% as the sample size tends to infinity (the Bonferroni correction does not have this property with correlated test statistics). Consequently, the max-test procedure will be equally or more powerful than tests adjusted with the Bonferroni procedure. It is also known to be a coherent and consonant procedure (Bretz et al. (2011), section 2.1.2) leading to decision patterns that are logical and simple to communicate: rejection of any null hypotheses implies rejection the global null hypothesis (intersection of the c null hypotheses) and rejection of the global null hypothesis implies rejection of at least one null hypothesis. The power of the procedure could be further improved, e.g., by considering step-down or step-up max-test procedures. However, this complicates the definition of the confidence intervals.

4 Multiple inference using Score tests

To motivate the use of Score tests, we will consider a LVM $\mathcal{M}_p(\boldsymbol{\theta})$ with p parameters, defined blinded to the data. We also consider the set of LVMs with $p+1$ parameters (referred to as the extended models), containing $\mathcal{M}_p$ as a submodel, and that are identifiable. For instance if $\mathcal{M}_p$ is the LVM defined by Fig. 1, $\mathcal{M}_{p+1}^{(1)}$ may add a regression parameter between X_1 and η, while $\mathcal{M}_{p+1}^{(2)}$ may add, instead, a regression parameter between X_2 and η (and so on). We denote by c the number of extended LVMs and by $\boldsymbol{\Theta}_j = (\boldsymbol{\theta}, \beta_j)$ the parameters in $\mathcal{M}_{p+1}^{(j)}$, the j-th extended LVM. As a diagnostic test, the practitioner would like to know whether any of the extended LVM has a significantly better fit compared to the original LVM. We therefore test, for each extended model $\mathcal{M}_{p+1}^{(j)}$, the null hypothesis $\mathcal{H}_0^{(j)} : \beta_{j,0} = 0$ vs. $\mathcal{H}_1^{(j)} : \beta_{j,0} \neq 0$. Here $\beta_{j,0}$ denotes the true parameter value in model $\mathcal{M}_{p+1}^{(j)}$. Traditionally, Score tests are used since the Score test statistic:

$$U^{\mathcal{M}_{p+1}^{(j)}} = \mathcal{S}^{\mathcal{M}_{p+1}^{(j)}}\left(\widetilde{\boldsymbol{\Theta}}\right)\left[\mathcal{I}^{\mathcal{M}_{p+1}^{(j)}}\left(\widetilde{\boldsymbol{\Theta}}\right)\right]^{-1}\mathcal{S}^{\mathcal{M}_{p+1}^{(j)}}\left(\widetilde{\boldsymbol{\Theta}}\right)^{\intercal} \tag{3}$$

do not require to fit any additional model; only to compute the score and the expected information matrix of the extended model (respectively denoted $\mathcal{S}^{\mathcal{M}_{p+1}^{(j)}}$ and $\mathcal{I}^{\mathcal{M}_{p+1}^{(j)}}$). In the equation above, $\widetilde{\boldsymbol{\Theta}} = (\widetilde{\boldsymbol{\theta}}, 0)$ denotes the ML estimator in any extended model under the null, i.e. under the constraint that $\widetilde{\beta}_j = 0$. In contrast, we denote by $\widehat{\boldsymbol{\Theta}}_j = (\widehat{\boldsymbol{\theta}}, \widehat{\beta}_j)$ the (unconstrained) ML estimator of the parameters in $\mathcal{M}_{p+1}^{(j)}$. It is a classical result from Maximum Likelihood theory that, under the null hypothesis, $U^{\mathcal{M}_{p+1}^{(j)}} \overset{d}{\sim} \chi_1^2$, where χ_1^2 denotes the χ^2 distribution with one degree of freedom. Using the same reasoning as in the previous section, we are once more interested in a max statistic: $U_{\max} = \max_{j\in\{1,\dots,c\}}\left(U^{\mathcal{M}_{p+1}^{(j)}}\right)$. If the test statistics were independent then up to a linear transformation, $U_{\max}$ is known to converge toward a Gumbel distribution when c tends to infinity (Gasull et al. 2015). In practice, the number of tests can be

small (e.g. $c < 10$) and the test statistics in LVMs are typically correlated; we therefore need an alternative approach.

4.1 Resampling of the score under the null hypothesis

The principle of this approach is first to identify the joint distribution of the scores across all extended LVMs, under the global null hypothesis. Then we resample from this distribution, compute the Score statistics for each extended LVM, take their maximum, and therefore obtain iid realizations of $U_{\max}$ under the global null hypothesis. The p-value can then be computed as the frequency at which the sampled realizations of $U_{\max}$ are more extreme than the realization of $U_{\max}$ obtained from the data. If we would have an iid decomposition of the scores in each extended model, we could use the same approach as in Pipper et al. (2012) to identify the joint distribution of the scores: stack the iid decompositions across models and use the multivariate central limit theorem to show that the scores are asymptotically jointly normally distributed. A consistent estimator of the variance-covariance matrix of this distribution could then be deduced from the iid decomposition.

An intuitive idea for the iid decomposition would be to use that $\mathcal{S}^{\mathcal{M}^{(j)}_{p+1}}(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \mathcal{S}_i^{\mathcal{M}^{(j)}_{p+1}}(\boldsymbol{\Theta})$. While this is a valid iid decomposition at $\boldsymbol{\Theta}_0$ this is not the case at $\widehat{\boldsymbol{\Theta}}_j$ or at $\widetilde{\boldsymbol{\Theta}}$ due to the constraints on the score implied by the ML estimation. For instance $\mathcal{S}^{\mathcal{M}^{(j)}_{p+1}}(\widehat{\boldsymbol{\Theta}}_j) = \mathbf{0}$ and has variance 0, while the individual terms $\mathcal{S}_i^{\mathcal{M}^{(j)}_{p+1}}(\widehat{\boldsymbol{\Theta}}_j)$ have non-0 variance. We therefore developed another decomposition (see supplementary material B for details) by first expressing $\widehat{\boldsymbol{\Theta}}_j$ as a function of $\widetilde{\boldsymbol{\Theta}}_j$ and then using a Talyor expansion of the score around $\widehat{\boldsymbol{\Theta}}_j$. We obtain that the first p components of the score evaluated at $\widetilde{\boldsymbol{\Theta}}_j$ are 0 and the last component (corresponding to β_j) is $\hat{\beta}_j \left(\Sigma^{\mathcal{M}^{(j)}_{p+1}}_{\beta_j,\beta_j}(\widetilde{\boldsymbol{\Theta}}_j) \right)^{-1}$. Therefore, denoting by $\psi_{\beta_j}(\boldsymbol{\Theta}_0, \mathcal{X}_i)$ the contribution of the i-th observation to the iid decomposition of $\hat{\beta}_j$ (equation (1)), we can introduce the normalized score $\mathcal{U}^{\mathcal{M}^{(j)}_{p+1}}$, a vector of length $p+1$ with iid decomposition:

$$\mathcal{U}^{\mathcal{M}^{(j)}_{p+1}}\left(\widetilde{\boldsymbol{\Theta}}\right) = \frac{1}{n}\sum_{i=1}^{n} \psi_{\mathcal{U}}^{\mathcal{M}^{(j)}_{p+1}}(\widetilde{\boldsymbol{\Theta}}, \mathcal{X}_i)$$

$$\text{where } \psi_{\mathcal{U}}^{\mathcal{M}^{(j)}_{p+1}}(\widetilde{\boldsymbol{\Theta}}, \mathcal{X}_i) = \psi_{\beta_j}(\widetilde{\boldsymbol{\Theta}}, \mathcal{X}_i) \left[\mathbf{0} \left(\Sigma^{\mathcal{M}^{(j)}_{p+1}}_{\beta_j,\beta_j}(\widetilde{\boldsymbol{\Theta}}_j) \right)^{-1} \right] \left(\mathcal{I}^{\mathcal{M}^{(j)}_{p+1}} \right)^{-\frac{1}{2}} (\widetilde{\boldsymbol{\Theta}}) \tag{4}$$

Once squared, $\mathcal{U}^{\mathcal{M}^{(j)}_{p+1}}\left(\widetilde{\boldsymbol{\Theta}}\right)$ is equivalent to the Score statistic: $U^{\mathcal{M}^{(j)}_{p+1}} = \mathcal{U}^{\mathcal{M}^{(j)}_{p+1}}\left(\widetilde{\boldsymbol{\Theta}}\right) \mathcal{U}^{\mathcal{M}^{(j)}_{p+1}}\left(\widetilde{\boldsymbol{\Theta}}\right)^{\mathsf{T}} + o_p(n^{-\frac{1}{2}})$. This leads to the following procedure:

Definition 2 (Single step max-test procedure for Score tests in LVM) Given a LVM estimated by ML, the max-test procedure for testing the set of null hypotheses $\left(\mathcal{H}_0^{(j)}\right)_{j\in\{1,\dots,c\}}$ using Score tests is:

1. Evaluate the score function and the expected information matrix at $\widetilde{\boldsymbol{\theta}}$ for each extended model.
2. Calculate the Score statistics $\left(U^{\mathcal{M}_{p+1}^{(j)}}\right)_{j\in\{1,\dots,c\}}$ using equation (3) for each extended models.
3. Estimate $\psi_{\mathcal{U}} = \left(\psi_{\mathcal{U}}^{\mathcal{M}_{p+1}^{(j)}}\right)_{j\in\{1,\dots,c\}}$ the iid decomposition of the score using equation (4) for each extended model. Estimate the covariance matrix of the normalized score $\Sigma_{\mathcal{U}} = \frac{1}{n}\sum_{i=1}^{n} \psi_{\mathcal{U}}\psi_{\mathcal{U}}^{\intercal}$.
4. Resample the normalized score under the null hypothesis, either:
   (a) sampling in a multivariate normal distribution with covariance matrix $\Sigma_{\mathcal{U}}$.
   (b) using wild bootstrap, i.e., weight the individual iid terms with individual specific weight sampled from a standard normal distribution and sum the iid terms over the individuals to obtain a sample of the normalized score.
5. For each sample: compute the Score statistic for each extended model, take the maximum over the extended models. Estimate the p-value using the relative frequency of the event that the sampled maximum is greater than the observe Score statistic.

This procedure can be limited in practice by step 4(a) or step 4(b). Step 4(a) involves sampling in a Gaussian distribution of dimension $(p+1)c$ which can be very large for complex LVMs. For instance in our illustration, this dimension will be (45+1)*36=1656 which is already quite large for a rather simple LVM where we limit the search to covariance links. Using step 4(b) may then be more tractable. Also, step 4(b) does not rely on the assumption that $\mathcal{U}$ follows a normal distribution, which may not be valid in small samples. However, it involves sampling a weight for each individual so step 4(b) may be slow for very large n. In the following subsection, we propose a procedure that is numerically more efficient.

4.2 Approximation of the max-distribution using latent Gaussian variables

Krishnaiah and Armitage (1965) proposed to compute the distribution of the maximum of χ^2 distributions using the joint distribution of the underlying Gaussian variables: given a vector $\left(T_j\right)_{j\in\{1,\dots,c\}}$ such that $T_j^2 = U^{\mathcal{M}_{p+1}^{(j)}}$ then $\mathbb{P}\left[U_{\max} \leq u\right] = \mathbb{P}\left[|T|_{\max} \leq \sqrt{u}\right]$ where $|T|_{\max} = \max_{j\in\{1,\dots,c\}}\left(\left|T_j\right|\right)$. Since each $U^{\mathcal{M}_{p+1}^{(j)}}$ is χ^2 distributed with one degree of freedom, the marginal distribution of each T_j is a standard

normal distribution. So it only remains to identify $R_{\mathbf{T}}$, the correlation matrix of $\left(T_j\right)_{j\in\{1,\ldots,c\}}$. Introducing:

$$\varphi_i^{\mathcal{M}_{p+1}^{(j)}} = \mathcal{S}_i^{\mathcal{M}_{p+1}^{(j)}}\left(\widetilde{\boldsymbol{\Theta}}\right)\left[\mathcal{I}^{\mathcal{M}_{p+1}^{(j)}}\left(\widetilde{\boldsymbol{\Theta}}\right)\right]^{-1}\mathcal{S}^{\mathcal{M}_{p+1}^{(j)}}\left(\widetilde{\boldsymbol{\Theta}}\right)^{\mathsf{T}} \tag{5}$$

where the difference with Eq. (3) is that the first element on the right-hand side is the individual score instead of the total score. We can then express the Score statistic as $U^{\mathcal{M}_{p+1}^{(j)}} = \sum_{i=1}^{n}\varphi_i^{\mathcal{M}_{p+1}^{(j)}}$. We therefore propose the following estimator for the pairwise correlation between two underlying Gaussian variables j and j':

$$\widehat{\mathbb{C}or}(T_j, T_{j'}) = \sum_{i=1}^{n}\frac{\left(\varphi_i^{\mathcal{M}_{p+1}^{(j)}} - m_{\varphi}^{\mathcal{M}_{p+1}^{(j)}}\right)\left(\varphi_i^{\mathcal{M}_{p+1}^{(j')}} - m_{\varphi}^{\mathcal{M}_{p+1}^{(j')}}\right)}{\sqrt{s_{\varphi}^{\mathcal{M}_{p+1}^{(j)}} s_{\varphi}^{\mathcal{M}_{p+1}^{(j')}}}} \tag{6}$$

where $m_{\varphi}^{\mathcal{M}_{p+1}^{(j)}}$ and $s_{\varphi}^{\mathcal{M}_{p+1}^{(j)}}$ denote the empirical mean and variance of $\left(\varphi_i^{\mathcal{M}_{p+1}^{(j)}}\right)_{i\in\{1,\ldots,n\}}$. As illustrated in the following example, the estimator defined in equation (6) may not always be consistent but can provide a reasonable approximation of the magnitude of the correlation.

EXAMPLE: consider for $\mathcal{M}_p$ the univariate model $Y_i = \nu + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and for the alternative models $Y_i = \nu + K_j X_{ij} + \varepsilon_i$ for $j \in \{1, \ldots, q\}$. To simplify we will assume that X_j has mean 0 and variance 1. Then the score vector for the parameters ν, σ^2, and K_j is $\mathcal{S}^{\mathcal{M}_{p+1}^{(j)}}\left(\widetilde{\boldsymbol{\Theta}}\right) = (0, 0, s_j)$ where $s_j = \sum_{i=1}^{n} s_{ij} = \sum_{i=1}^{n}\frac{X_{ij}(Y_{ij}-\hat{\nu})}{\hat{\sigma}^2}$ is non-0 in finite samples. So $U^{\mathcal{M}_{p+1}^{(j)}} = \frac{\sigma^2 s_j^2}{n}$ and we can identify the latent Gaussian variable $T_j = \frac{\sigma s_j}{\sqrt{n}}$. For $(j, j') \in \{1, \ldots, q\}^2$, $\mathbb{C}or(T_j, T_{j'}) = \mathbb{C}or(s_j, s_{j'})$ which can be consistently estimated by computing the correlation between the vectors $\left(s_{ij}\right)_{i\in\{1,\ldots,n\}}$ and $\left(s_{ij'}\right)_{i\in\{1,\ldots,n\}}$. However, the empirical correlation between $\left(\varphi_i^{\mathcal{M}_{p+1}^{(j)}}\right)_{i\in\{1,\ldots,n\}}$ and $\left(\varphi_i^{\mathcal{M}_{p+1}^{(j')}}\right)_{i\in\{1,\ldots,n\}}$ estimates $sign(s_j)sign(s'_j)\mathbb{C}or(s_j, s_{j'})$, where $sign$ denotes the function returning -1 for negative numbers and +1 for positive numbers. Therefore $\widehat{\mathbb{C}or}(T_j, T_{j'}) = \mathbb{C}or(s_j, s_{j'})sign\left(s_j s_{j'}\right)$.

We obtain the following procedure for handling multiple testing:

Definition 3 (Approximate single step max-test procedure for Score tests in LVM) Given a LVM estimated by ML, the approximate max-test procedure for testing the set of null hypotheses $\left(\mathcal{H}_0^{(j)}\right)_{j\in\{1,\dots,c\}}$ using Score tests is:

1. Define the score function and the expected information matrix at $\widetilde{\boldsymbol{\theta}}$ for each extended model.
2. Compute the Score statistics $\left(U^{\mathcal{M}_{p+1}^{(j)}}\right)_{j\in\{1,\dots,c\}}$ using equation (3) for each extended models. Denote $\left(T_j\right)_{j\in\{1,\dots,c\}}$ their square root value.
3. Compute $\left(\varphi^{\mathcal{M}_{p+1}^{(j)}}\right)_{j\in\{1,\dots,c\}}$ using equation (5).
4. Estimate the correlation matrix $R_{\mathbf{T}}$ using equation (6).
5. Compute p-values applying formula (2) to $\left(T_j\right)_{j\in\{1,\dots,c\}}$ where $f_{\mathbf{T}}$ is the density of a multivariate normal distribution with mean zero and variance-covariance matrix $R_{\mathbf{T}}$.

This procedure is expected to be more numerically efficient than the resampling procedure proposed in the previous subsection. Indeed, equation (5) gives a univariate influence function, compared to formula (4) where it is p+1 dimensional, making step 5 reasonably fast. However, the validity of the approximation performed with this procedure is unclear and will be empirically assessed in simulation studies (see section 5.2).

5 Control of the FWER of the max-test procedure in finite samples

5.1 Multiple comparisons using Wald tests

We consider a latent factor model with 9 outcomes and two binary covariates called group and gene. In the generative model all the outcomes are equally correlated, normally distributed, independent of the group variable but dependent of the gene variable. This corresponds to the following measurement and structural models:

$$Y_{i,j} = \nu_j + \lambda_j \eta_i + K_{1,j} Group_i + K_{2,j} Gene_i + \varepsilon_{i,j}, \text{ where } \varepsilon_{i,j} \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon,j}\right) \tag{7}$$

$$\eta_i = \alpha + \zeta_i, \text{ where } \zeta_i \sim \mathcal{N}\left(0, \sigma^2_{\zeta}\right) \tag{8}$$

$$\forall j \in \{1,\dots,9\},\ \zeta_i \perp\!\!\!\perp \varepsilon_{i,j} \text{ and } \forall (j,j') \in \{1,\dots,9\}^2,\ j \neq j',\ \varepsilon_{i,j} \perp\!\!\!\perp \varepsilon_{i,j'} \tag{9}$$

$$\nu_1 = 0 \tag{10}$$

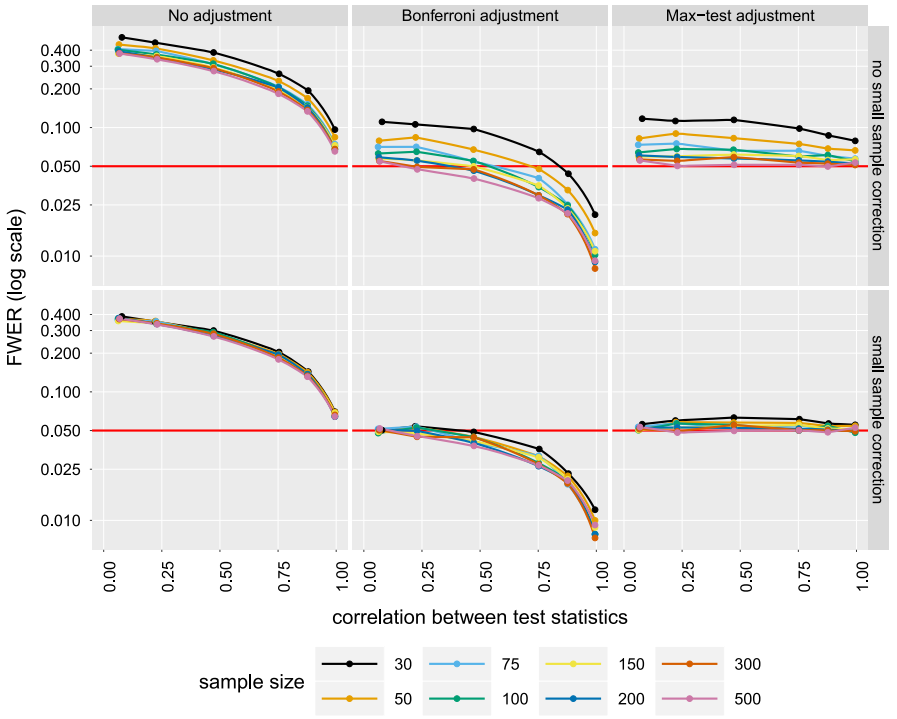

Fig. 2 FWER when testing 9 null hypotheses with Wald tests. using different procedures to adjust the p-values for multiple comparisons (columns). The rows indicate whether a small sample correction is used. For instance in the third column, the upper row uses a max-z test while the lower row uses a max-t test. The correlation reported in the x-axis is the median Pearson correlation between the test statistics computed over the repetitions. For the y-axis, a logarithmic scale was used

in the special case where we constrain (i) $\forall j \in \{1, \ldots, 9\}\, \lambda_r = \lambda_1$ and (ii) $\forall j \in \{1, \ldots, 9\}\, K_{1,j} = 0$. The values of the remaining parameters were obtained fitting the unconstrained model to the data used for the illustration (see section 6). To assess the control of the FWER of the max-test procedure, we consider the LVM defined by the equations (7), (8), (9), and (10) under the constraint $\lambda_1 = 1$. This model will be referred as the investigator model thereafter. The set of null hypotheses that is being tested is $\left\{K_{1,j} = 0\right\}_{j \in \{1,\ldots,9\}}$, i.e. the group effects are zero. We consider several scenarios where we varied the sample size: $n \in \{30, 50, 75, 100, 150, 200, 300, 500\}$ and the covariance between the outcomes: $\lambda_1 \in \{0.1, 0.2, 0.35, 0.65, 1, 5\}$. We generated 10 000 datasets, analyzed them using the investigator model, and computed the p-value for the global null hypothesis using no adjustment for multiple comparison, the Bonferroni procedure, or the max-test procedure with the model-based variance-covariance matrix. To improve the control of the FWER in finite samples, the p-values were also computed after application of the small sample correction.

The upper panel of Fig. 2 shows the FWER in absence of adjustment, when using the Bonferroni procedure or the max-test procedure. For large sample sizes (e.g. n=500), the FWER was above its nominal level in absence of adjustment (except when the

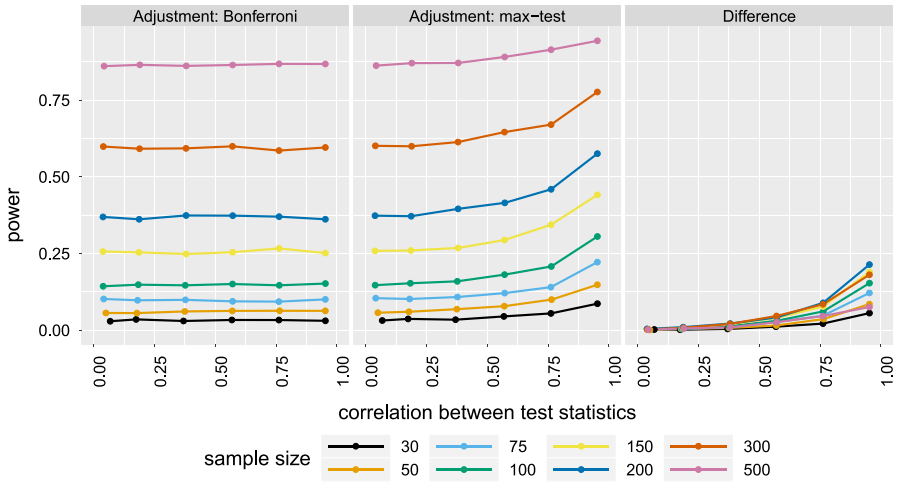

Fig. 3 Power when testing 9 null hypotheses with Wald tests using the Bonferroni or the max-test procedure to adjust the p-values for multiple comparisons. The last column displays the difference in power between the two procedures

test statistics were perfectly correlated) while below its nominal level when using the Bonferroni procedure (except when the test statistics were independent). The max-test procedure managed to keep the FWER at its nominal level regardless the correlation. Without small sample size correction, the FWER increased when the sample size decreased, e.g. the max-test procedure had a FWER of approximately 0.1 for n=30. This was corrected when using the small sample correction (Fig. 2 lower panel).

To assess the gain in power when using the max-test procedure instead of the Bonferroni procedure, we simplified the generative used in the previous simulation. We set the intercepts to 0 and the other coefficients to 1, except for the group effects where the first was set to 0.4 and the others to 0 (i.e $K_{1,1} = 0.4$ and $\forall j \in \{2, \dots, 9\}\, K_{1,j} = 0$), the loadings $(\lambda_j)_{j\in\{1,\dots,9\}}$ were all set to a value a, and the residual variances $(\sigma^2_{\varepsilon,j})_{j\in\{1,\dots,9\}}$ were all set to $5.25 - \sqrt{a}$, to vary the degree of correlation between the outcomes. This ensured that the conditional variance of the outcomes ($\mathbb{V}ar[Y_j|Group, Gene] = \lambda_j^2 + \sigma^2_{\varepsilon,j}$) remained constant when we varied $a \in \{0.25, 1, 2, 3, 4, 5\}$. We generated 5000 datasets for each configuration and the power was computed as the frequency at which the p-value for the global null hypothesis was lower than 0.05. Figure 3 shows that the max-test procedure was always more powerful than the Bonferroni procedure with a gain in power that ranged between 0% and 22%, when the correlation between the test statistics was respectively low and high. We also compared the max-test procedure to step-up procedures (Hochberg and Hommel, see figure A in supplementary material C) and found that the power improvement obtained using a step-up procedure was neglectable.

5.2 Multiple comparisons using Score tests

We now assess the control of the FWER when testing multiple hypotheses using Score tests. The generative model is a latent factor model with 5 outcomes loading on a single latent variable η. The latent variable was correlated with a single variable called treatment. 15 other covariates were simulated ($X_1, \ldots, X_{15}$) and the investigator aimed to assess whether they had an effect on the latent variable. The covariates were simulated with a common pairwise covariance that was varied: $a \in \{0, 0.6, 1, 1.5, 2.5, 5\}$. In each scenario, the 15 possible extended models were formed and the Score statistics computed. The corresponding p-values were calculated using no adjustment, the Bonferroni procedure, the max-test procedure with resampling (i.e. definition 2 with step 4(a)), or the approximate max-test procedure (i.e. definition 3). For each sample size and covariance value 10000 datasets were simulated and analyzed. The FWER was computed as the relative frequency at which the smallest p-value was below 5%. Figure 4 displays the FWER relative to each procedure. Results are similar to those of the previous simulation, the max-test correction providing a good control of the FWER regardless to the correlation, while the Bonferroni procedure was too conservative for correlated test statistics. In small samples, the approximate max-test procedure appeared to provide a better FWER compared to the use of resampling. In 0.01% of the datasets the p-value adjusted using the approximate max-test procedure were greater than for Bonferroni. This only occurred when the non-adjusted p-value was small (< 0.001) and we think this is due to inaccuracies in the numerical integration procedure required to compute the integral in Eq. (2).

In term of computation time, the approximate max-test procedure was similar to the bootstrap in small samples, e.g. for $n = 50$ the median [5% quantile; 95% quantile] computation time in seconds was 1.24 [1.12;2.46] vs. 1.90 [1.71;3.4]. However it scaled better with n, e.g. 2.67 [2.48;3.13] vs. 11.5 [9.09;15.3] for $n = 500$.

We also repeated this simulation including an additional variable Z in the generative model. This variable had an effect on the outcomes through the latent variable (in the structural model $\eta_i = \alpha + \gamma Z_i + \zeta$, the regression coefficient γ was set to 0.25) but was independent of the covariates $X_1, \ldots, X_{15}$ (Fig. 1, the correlation between the variables $X_1, \ldots, X_{15}$ is omitted for readability). The investigator aimed to assess whether Z or any of the other 15 covariates had an effect on the latent variable. This would correspond to the first step of FSS: the variable Z is selected if the p-value of relative to the parameter γ is significant and if it has the greatest test statistic (in absolute value). The p-values were adjusted for multiple testing using either the Bonferroni procedure or the approximate max-test procedure. The power was then defined as the proportion of simulations where Z was selected. 5000 datasets were generated for each scenario. The upper panels of Fig. 5 display the relative frequency at which the effect of Z reached the significance level. As expected, when the test statistics were correlated the effect Z reaches significance more often with the max-test procedure than with the Bonferroni procedure. The observed increase in frequency varied between 0% and 20%, depending on the sample size and the correlation. A similar pattern was observed when looking at the empirical power, i.e. when the effect

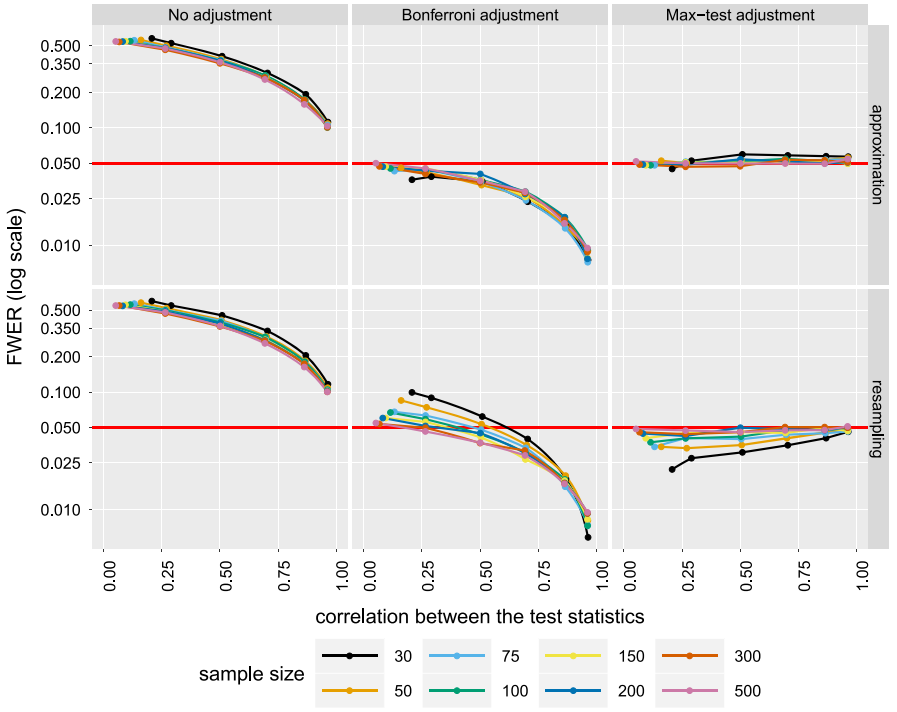

Fig. 4 FWER when testing 15 null hypotheses with Score tests using different procedures to adjust the p-values for multiple comparisons (columns). The rows indicate correspond to the use of analytic formula or resampling to compute the p-values. The correlation reported in the x-axis is the median Pearson correlation between the test statistics computed over the datasets. For the y-axis, a logarithmic scale was used

of Z reached the significance level and had the greatest test statistic among all the tested effects (lower row of panels of Fig. 5).

6 Illustration

Mild traumatic brain injury (mTBI) is an injury to the head inducing disruption of brain function, e.g. loss of consciousness not exceeding 30 min and dysfunction of memory around the time of injury not exceeding 24 h. Because the pathogenesis of symptoms following mTBI is poorly understood and no evidence-based treatments are available for patients with bad recovery, there is a medical interest in a more precise and objective characterization of mTBI using medical imaging. It has been hypothesized that mTBI induces a neuroinflammatory response that could act as a therapeutic target. The neuroinflammatory response is expected to vary over the brain dependent on the trauma mechanism in the individual subject, but with deeper lying brain regions generally being more vulnerable due to local concentration of shock waves. Neuroinflammation can be measured indirectly using single-photon emission computed tomography (SPECT) and injection of the radioligand [123I]-CLINDE which

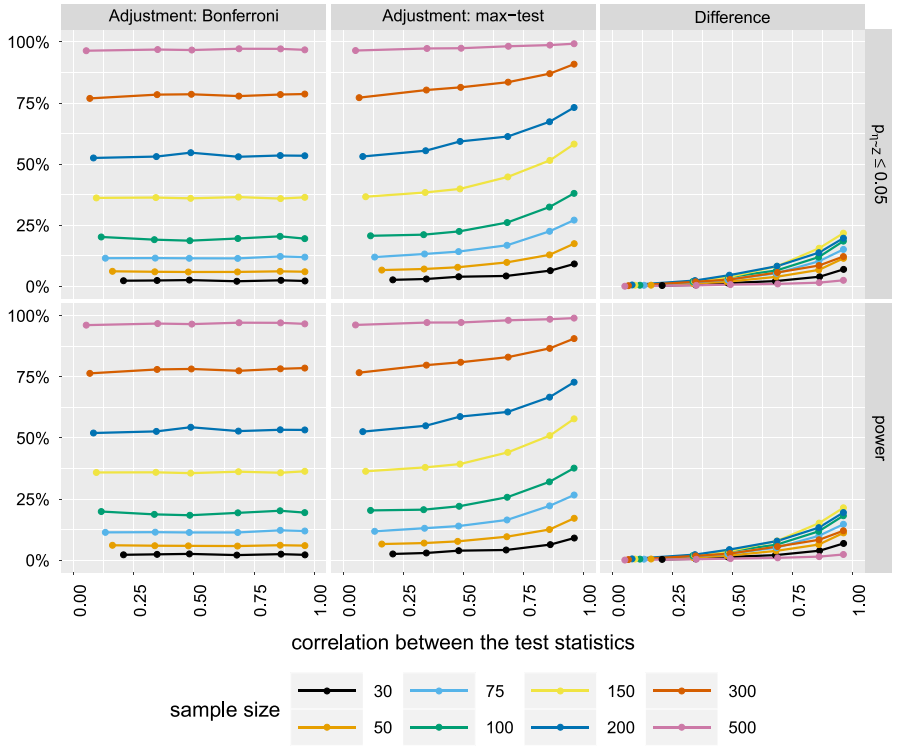

Fig. 5 Percentage of simulations where the p-value of the Score test $\gamma = 0$ was significant (upper panels) and power (lower panels). The third column displays the difference between the values obtained using the approximate max-test procedure (second column) vs. Bonferroni (first column). The correlation reported in the x-axis is the median Pearson correlation between the test statistics computed over the datasets

visualizes the translocator protein (TSPO); a protein upregulated in active immune cells. A genetic polymorphism of TSPO is known to affect [123I]-CLINDE binding to TSPO and partially explain interindividual variability.

Clinical, genetic and [123I]-CLINDE-SPECT data of 14 patients with mTBI and 22 healthy controls were collected. Patients were scanned first one to two weeks after the injury, and a second time 3 to 4 months after the injury (this second scan will be ignored in the following). Healthy controls were only scanned once. One of the aims of the study (Ebert et al. 2019) was to compare [123I]-CLINDE binding to TSPO between the two groups in 9 brain regions: thalamus, pallidostratum, impact region (patients) or neocortex (healthy controls), midbrain, pons, cingylate gyrus, hippocampus, supramarginal gyrus, corpus callosum. To quantify [123I]-CLINDE binding to TSPO, regional distribution volumes of [123I]-CLINDE were calculated with a two-tissue compartment model using arterial plasma as the input function. Distribution volume is the ratio between radiotracer concentration in the brain and in the blood at equilibrium. In the following we ignore the uncertainty related to this method of quantification and treat the distribution volumes as if they were directly measured.

The LVM defined with the investigator included the log of the TSPO distribution volumes of [123I]-CLINDE in the 9 regions as outcomes, a single latent variable to account for the covariance between the outcomes, region-specific genetic effects and group effects. This corresponds to the LVM defined previously (Eqs. (7), (8), (9), (10) under the constraint $\lambda_1 = 1$) which contains 45 parameters: 8 parameters from $\boldsymbol{\nu}$, 8 parameters from Λ, 18 parameter from K, 9 parameters from $\Sigma_{\boldsymbol{\varepsilon}}$, 1 parameter from $\boldsymbol{\alpha}$, no parameter from B or Γ (since here B and Γ are null) and one parameter from Σ_{ζ}. Their value are reported in supplementary material E. With this model, the χ^2 statistic testing whether the modeled variance-covariance matrix differs from the observed covariance matrix was found significant (p=0.0016). To assess which covariance parameters should be added, we use a FSS on the 36 possible covariance links. The Score statistics were weakly correlated, with a median absolute pairwise Pearson correlation of 0.17 (2.5% quantile : 0.013, 97.5% quantile 0.59). Table 1 contains the results of the FSS using no adjustment, or the Bonferroni procedure, or the approximate max-test procedure. Without adjustment two covariance parameters would be added while when adjusting the p-values with the Bonferroni or max-test approach one covariance parameter would be added. The max-test approach multiplied the unadjusted p-value by a factor of 18 (first step), 26 (second step) and 13 (third step) instead of 36, 35 and 34 for the Bonferroni approach. Including the first covariance parameter enables to obtain a non-significant χ^2 test (p=0.31) and this model will be the one retained for performing inference. The qqplots of the residuals of the measurement models did not show any clear violation of the normality assumption.

Using an F-test for testing the global null hypothesis of no effect of mTBI on the distribution volume in any region gives a p-value of 0.011. While the p-value supports that mTBI induces neuro-inflammation, it does not inform on which brain region is affected which limits its practical value. This motivates the use of the max-test procedure when assessing the significance of the region-specific effects. The 9 Wald statistics were highly correlated with a median correlation of 0.831 (min 0.702, max 0.960). The p-values adjusted with the max-test approach were only between 1.17 to 3.3 times larger than the unadjusted p-values (instead of 9 times larger for the Bonferroni approach, see Table 2). Corpus Callosum was the region with the largest effect and test statistic. The unadjusted p-value was 0.026, the Bonferroni adjusted was 0.234 and the max-t adjusted was 0.086. The next region was cingulate gyrus (max-t adjusted p=0.108) while the remaining regions had a large adjusted p-value ($>$0.3).

In this data the smallest adjusted p-value was higher than for the F-test. This may be happen when the effect is similar across the regions and the opposite happens when the effect is only present in a few regions.

7 Concluding remarks

When dealing with complex systems of variables, LVMs are a convenient modeling tool that provides, under some assumptions, interpretable and efficient estimates. This enables the investigator to translate her hypotheses into a function of the parameters and test whether this function evaluated at the estimated parameters equals a particular

Table 1 Result of the first three steps of FSS for local dependencies. Each row corresponds to a step

Covariance parameter	Number of tests	Max statistic	*p*-value		
			No adjustment	Bonferroni	Max-test
(neocortex, supramarginalGyrus)	36	23.27	0.000001	0.000051	0.000028
(pallidostriatum, midbrain)	35	7.74	0.005387	0.188535	0.140737
(thalamus, pallidostriatum)	34	3.78	0.051955	1	0.68033

Table 2 Test of the null hypothesis of equal average distribution volume in the healthy group vs. the mTBI group. The column mTBI effect indicates the estimated percentage increase in distribution volume after mTBI

Region	mTBI effect (%)	Statistic	p-value			
			No adjustment	Bonferroni	Max-t-test	Step-down max-t-test
thalamus	12.75	1.22	0.23	1	0.534	0.395
pallidostriatum	12.03	1.37	0.177	1	0.435	0.358
neocortex	4.38	0.53	0.601	1	0.969	0.753
midbrain	10.4	1.25	0.219	1	0.514	0.395
pons	1.56	0.18	0.858	1	1	0.858
cingulateGyrus	17.28	2.18	0.034	0.304	0.108	0.096
hippocampus	12.64	1.5	0.139	1	0.358	0.317
supramarginalGyrus	5.22	0.61	0.547	1	0.94	0.736
corpusCallosum	19.02	2.3	0.026	0.234	0.086	0.086

value. Like in many other models, the statistical testing framework in LVM is well established for a single statistical test. However, investigators are often interested in testing multiple clinical hypotheses (using Wald tests) or performing multiple diagnostic tests (using Score tests). In this article, we present adjustments for multiple comparisons applicable to both Wald and Score tests that appropriately control the FWER without sacrificing statistical power. While both procedures rely on asymptotic results, we found via simulation studies that they had a satisfying behavior in finite samples. The procedures are implemented in a freely available R package (`lavaSearch2`). Our implementation of the max-test procedure rely on numerical integration (Genz et al. 2018) to compute tail probabilities of the multivariate Gaussian or Student's t-distributions, restricting its applicability to low dimensional problems.

The power of the max-test procedure can be further increased using a step-down max-test procedure (analogue to a Bonferroni-Holm procedure but accounting for the correlation between the test statistics). While the most significant p-value is not affected, the other p-values can sometimes be substantially reduced (e.g. compare the last two columns of Table 2). The power of the proposed procedure could also be improved by taking advantage of logical restrictions between null hypotheses (Westfall and Tobias 2007). While none were present in our simulation study and illustration, they typically arise when considering all pairwise differences between exposures (A vs. B, A vs. C, B vs. C). One common limitation of these improved procedures is that it is difficult to obtain simultaneous confidence intervals that are informative (i.e. provide information additional to the rejection of the null hypothesis) and consistent with the adjusted p-values. This is also why we focused on the single-step max-test procedure in the article. We refer to Strassburger and Bretz (2008) and Brannath and Schmidt (2014) for a more detailed discussion on simultaneous confidence intervals. Another possible improvement would be to handle sequential hypothesis testing. For instance, in our illustration, we first performed several Score test until finding a satisfying model and then tested the clinical hypothesis based on the retained model. Based on a simulation study (supplementary material D.3), the type 1 error appeared to be properly controled in that example. However this is likely not to be the case if the model misspecifications are directly related to the clinical hypothesis (e.g. region-specific group effects). Resampling procedures (e.g. supplementary material D.2) being generaly too computer-intensive to be used, more efficient post-selection procedures would be beneficial.

As suggested by a reviewer post-selection methods could also be used to avoid multiple comparisons, e.g. by using part of the data (Cox 1975; DiCiccio et al. 2020) to identify the most promising region and another part to assess its statistical significance. In the present application, this lead to a median p-value of 0.066 for a critical threshold of 0.025 so no apparent gain in power. We believe that this approach is mostly relevant when testing a large number of hypotheses and there is no interest in assessing the individual null hypotheses (i.e. here identifying which brain regions were subject to inflammation).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00180-022-01214-7.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Brannath W, Schmidt S (2014) A new class of powerful and informative simultaneous confidence intervals. Stat Med 33(19):3365–3386

Bretz F, Hothorn T, Westfall P (2011) Multiple comparisons using R. CRC Press, Cambridge

Chernozhukov V, Chetverikov D, Kato K et al (2013) Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. Ann. Stat. 41(6):2786–2819

Cox DR (1975) A note on data-splitting for the evaluation of significance levels. Biometrika 62(2):441–444

Cribbie RA (2000) Evaluating the importance of individual parameters in structural equation modeling: the need for type I error control. Personal. Indiv. Diff. 29(3):567–577

Cudeck R, O'dell LL (1994) Applications of standard error estimates in unrestricted factor analysis: significance tests for factor loadings and correlations. Psychol. Bull. 115(3):475

DiCiccio CJ, DiCiccio TJ, Romano JP (2020) Exact tests via multiple data splitting. Stat Probab. Lett. 166:108865

Dmitrienko A, D'Agostino R Sr (2013) Traditional multiplicity adjustment methods in clinical trials. Stat. Med. 32(29):5172–5218

Ebert SE, Jensen P, Ozenne B, Armand S, Svarer C, Stenbaek DS, Moeller K, Dyssegaard A, Thomsen G, Steinmetz J, et al. (2019). Molecular imaging of neuroinflammation in patients after mild traumatic brain injury: a longitudinal 123i-clinde single photon emission computed tomography study. Eur J Neurol

Gasull A, López-Salcedo JA, Utzet F (2015) Maxima of gamma random variables and other weibull-like distributions and the lambert w function. Test 24(4):714–733

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2018) mvtnorm: Multivariate normal and t distributions. CRAN. R package version 1.0-8

Green SB, Thompson MS, Poirer J (2001) An adjusted Bonferroni method for elimination of parameters in specification addition searches. Struct Equ Model 8(1):18–39

Holst KK, Budtz-Jørgensen E (2013) Linear latent variable models: the lava-package. Comput Stat 28(4):1385–1452

Hothorn T, Bretz F, Westfall P (2008) Simultaneous inference in general parametric models. Biometr J 50(3):346–363

Krishnaiah PR, Armitage JV (1965) Tables for the distribution of the maximum of correlated chi-square variates with one degree of freedom. Trabajos de estadística y de investigación operativa 16(2–3):91–115

MacCallum RC, Roznowski M, Necowitz LB (1992) Model modifications in covariance structure analysis: the problem of capitalization on chance. Psychol Bull 111(3):490

Muthén LK, Muthén BO (2017) Mplus user's guide, 8th edn. Muthén & Muthén, Los Angeles, CA

Ozenne B, Fisher PM, Budtz-Jørgensen E (2020) Small sample corrections for wald tests in latent variable models. arXiv preprint arXiv:2002.02272

Pipper CB, Ritz C, Bisgaard H (2012) A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. J R Stat Soc Ser C Appl Stat 61(2):315–326
Ropovik I (2015) A cautionary note on testing latent variable models. Front Psychol, 6
Rosseel Y (2012) lavaan: an R package for structural equation modeling. J Stat. Softw 48(2):1–36
Smith CE, Cribbie RA (2013) Multiplicity control in structural equation modeling: incorporating parameter dependencies. Struct Equ Model Multidiscip J 20(1):79–85
Strassburger K, Bretz F (2008) Compatible simultaneous lower confidence bounds for the holm procedure and other Bonferroni-based closed tests. Stat Med 27(24):4914–4927
Tsiatis AA (2006) Semiparametric Theory and Missing Data. Springer Series in Statistics
Van der Vaart AW (2000) Asympt Stat. Cambridge University Press, Cambridge
Westfall PH, Tobias RD (2007) Multiple testing of general contrasts: truncated closure and the extended Shaffer–Royen method. J Am Stat Assoc 102(478):487–494
Westfall PH, Troendle JF (2008) Multiple testing with minimal assumptions. Biometr J J Math Methods Biosci 50(5):745–755