Behavior Genetics (2019) 49:99–111
https://doi.org/10.1007/s10519-018-9942-y

ORIGINAL RESEARCH

Type I Error Rates and Parameter Bias in Multivariate Behavioral Genetic Models

Brad Verhulst[1] · Elizabeth Prom-Wormley[2] · Matthew Keller[3] · Sarah Medland[4] · Michael C. Neale[5]

Received: 26 June 2018 / Accepted: 27 November 2018 / Published online: 20 December 2018

Abstract

For many multivariate twin models, the numerical Type I error rates are lower than theoretically expected rates using a likelihood ratio test (LRT), which implies that the significance threshold for statistical hypothesis tests is more conservative than most twin researchers realize. This makes the numerical Type II error rates higher than theoretically expected. Furthermore, the discrepancy between the observed and expected error rates increases as more variables are included in the analysis and can have profound implications for hypothesis testing and statistical inference. In two simulation studies, we examine the Type I error rates for the Cholesky decomposition and Correlated Factors models. Both show markedly lower than nominal Type I error rates under the null hypothesis, a discrepancy that increases with the number of variables in the model. In addition, we observe slightly biased parameter estimates for the Cholesky decomposition and Correlated Factors models. By contrast, if the variance–covariance matrices for variance components are estimated directly (without constraints), the numerical Type I error rates are consistent with theoretical expectations and there is no bias in the parameter estimates regardless of the number of variables analyzed. We call this the direct symmetric approach. It appears that each model-implied boundary, whether explicit or implicit, increases the discrepancy between the numerical and theoretical Type I error rates by truncating the sampling distributions of the variance components and inducing bias in the parameters. The direct symmetric approach has several advantages over other multivariate twin models as it corrects the Type I error rate and parameter bias issues, is easy to implement in current software, and has fewer optimization problems. Implications for past and future research, and potential limitations associated with direct estimation of genetic and environmental covariance matrices are discussed.

Keywords Twin models · Type I error · Cholesky decomposition · Correlated factors model · Direct symmetrical matrix

Introduction

The classic twin study design compares the covariances of monozygotic (MZ) and dizygotic (DZ) twin pairs reared by their biological parents in the same home. This approach has been popular since the 1960s and is the basis for most of the heritability estimates that have been obtained without direct measurement of genomic similarity. This approach estimates the degree to which genetic and environmental factors influence a phenotype using the path coefficients *a, c* and *e* (Fig. 1a). These path coefficients are regression weights of the phenotype of interest on latent genetic (*A*), common environmental (*C*) and unique environmental (*E*) factors (Neale and Cardon 1992). This approach is typically referred to as the "univariate ACE model". Heritability (h^2) is estimated by dividing the squared additive genetic path coefficient (a^2) by the total phenotypic variance ($a^2+c^2+e^2$).

Edited by Stacey Cherny.

✉ Brad Verhulst
brad.verhulst@gmail.com

1 Department of Psychology, Michigan State University, East Lansing, USA

2 Family Medicine and Population Health, Virginia Commonwealth University, Richmond, USA

3 Department of Psychology and Neuroscience, University of Colorado, Boulder, USA

4 Psychiatric Genetics Laboratory, QIMR Berghofer Medical Research Institute, Brisbane, Australia

5 Department of Psychiatry and Human Genetics, Virginia Commonwealth University, Richmond, USA

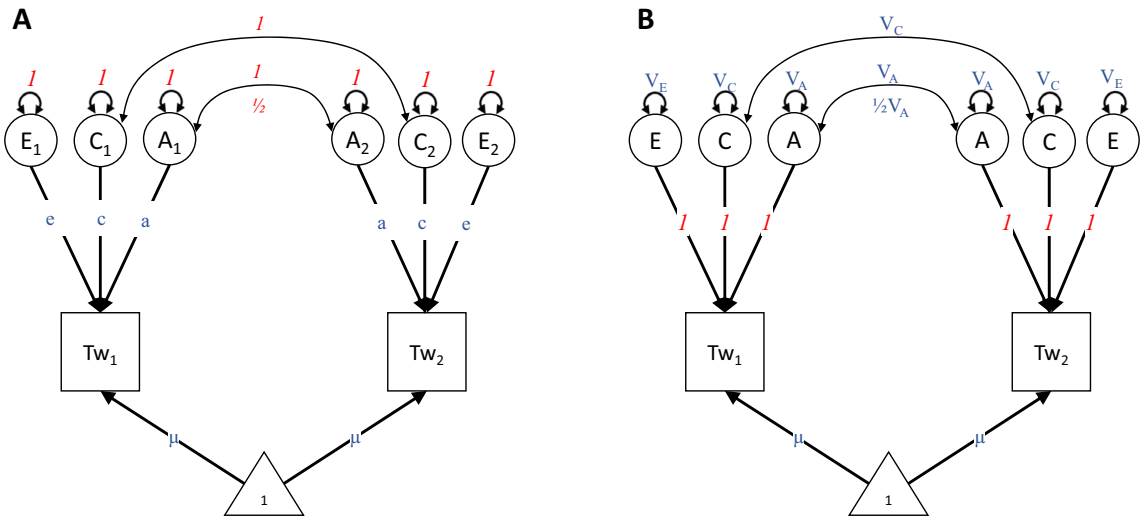

Fig. 1 Alternative parameterizations of the univariate ACE model for a pair of twins for **a** the standard path specification and **b** the direct variance specification. *Note* T_1 and T_2 represent the observed phenotypes for twin 1 and twin 2, respectively. The latent variables, depicted as circles representing the effects of additive genetic (A), common environment (C) and specific environment (E) variation generate phenotypic variation. Path labels in blue are estimated parameters and paths labels in red italics are fixed at the specified values. In (**a**), regression path coefficients a, c and e capture the relationship between the latent variable and the phenotypes, which are squared and summed to the phenotypic variance in the twin model. In (**b**), variance components V_A, V_C and V_E are specified as variances which can be directly summed to the phenotypic variation. The phenotypic means are represented by μ. (Color figure online)

The statistical significance of the parameters from a univariate ACE model is often assessed using a likelihood ratio test, where the −2 log-likelihood of a model with a freely estimated parameter (e.g. *c* or the shared environment) is compared with −2 log-likelihood of a nested model where the parameter is fixed to zero. Under certain regularity conditions (Steiger 1980) this statistic is asymptotically distributed as chi-squared with one degree of freedom. Unfortunately, these regularity conditions are not met when models have either implicit or explicit bounds. Using standard methods to estimate variance components in twin models, under the null hypothesis that a variance component is zero, this test is distributed as a 50:50 distribution of zero and Chi square with one degree of freedom. This parameterization is entirely logical as variance is a squared quantity and we measure individual differences in real (not imaginary) number values and our statistical models have been parameterized accordingly (Fig. 1a). This approach, however, has adverse and unintended consequences.

Type I error rates in univariate genetic models

Squaring the path coefficients in a univariate twin model prevents negative estimates of the variance components which places an implicit, artificial boundary on the parameter space. The boundary truncates the sampling distribution of the variance components and leads to lower than nominal Type I error rates (Carey 2005; Dominicus et al. 2006; Visscher 2006).

In addition to affecting the Type I error rate, forcing *A, C*, and *E* variances to be positive also results in biased parameter estimates. For example, suppose that the null hypothesis of no common environment is true and consider the expected distribution of MZ and DZ correlations for a trait where $a^2 = 0.5$, $c^2 = 0$ and $e^2 = 0.5$. With finite sample sizes, both the MZ and DZ correlations will vary around their expected values of 0.5 and 0.25, respectively. We would expect that 50% of the time the MZ correlation (r_{MZ}) will be greater than twice the DZ correlation (r_{DZ}), and 50% of the time the opposite will hold. When $r_{MZ} < 2r_{DZ}$, a non-zero estimate of c^2 will be obtained, and when $r_{MZ} > 2r_{DZ}$ an estimate of c^2 at its lower bound of zero will be obtained. Thus, the true value of c^2 is zero will be obtained 50% of the time, albeit because it falls exactly on the implicit boundary, but a positive estimate of c^2 *would be returned otherwise*. Consequently, on average the estimate of c^2 would be positive, which is biased. The same thought experiment may be conducted for the case where the null hypothesis of $a^2 = 0$ is true. For example, *with* $a^2 = 0$, $c^2 = 0.5$ and $e^2 = 0.5$, the expected distributions of r_{MZ} and r_{DZ} would both vary around 0.5, with a 50:50 split of $r_{MZ} > r_{DZ}$ and $r_{MZ} < r_{DZ}$. In the former case, a positive estimate of a^2 would be obtained, whereas in the latter the lower bound of $a^2 = 0$ would be estimated. Thus, both familial variance components, a^2 and c^2, are expected to be biased under the null.

Adjusting Type I error rates in the univariate ace model

One method to address the overly conservative Type I error rate for the univariate ACE model only requires the simple adjustment of dividing the observed p-value by two. However, this solution does not address the underlying issue of boundary constraints and the resulting bias in the estimates of the path coefficients *a, c*, and *e*. Therefore, the source of divergence in Type I error rates in the standard approach to the ACE model is that variance components have a lower bound of zero (Dominicus et al. 2006; Visscher 2006; Wu and Neale 2012). As an alternative, we propose the model parameterized in Fig. 1b, in which the quantities V_A, V_C and V_E are estimated directly as unbounded free parameters (which therefore may take negative values). This model has neither implicit (due to squaring) nor explicit (imposed by software during optimization) boundaries.

Type I error rates in multivariate genetic models

Most current variance components modeling of multivariate twin data involves estimating matrices that can be algebraically transformed into *A, C* and *E* covariance matrices with variances on the diagonal elements, and covariances in the off-diagonal elements. To date, there are two popular parameterizations for estimating multivariate genetic and environmental covariance matrices: the Cholesky or triangular decomposition (Neale and Cardon 1992) and the Correlated Factors model (Neale et al. 2006).

In the Cholesky decomposition approach, the lower triangular and diagonal elements of the matrix are freely estimated, while the upper triangular (above the diagonal) are fixed at zero. The predicted variance–covariance matrix is obtained by post-multiplying the matrix by its transpose, e.g., $\mathbf{A} = \mathbf{aa}^{\mathbf{T}}$ where $\mathbf{a}$ is the lower triangular matrix containing the genetic path coefficients. This may be thought of as a multivariate analog to the square root of a variance and is exactly this in the single variable case. The "Triple Cholesky" model for twin data estimates separate Cholesky matrices for all three variance components (*A, C* and *E*). Importantly, Carey (2005) pointed out that the Cholesky decomposition imposes two important constraints on the parameter space: (1) an implicit lower bound of zero for the variance of each variable, and (2) that each variance component is non-negative definite. In a Cholesky decomposition, both nonsensical variances less than 0 or nonsensical correlations outside the range of -1 to $+1$ cannot be estimated. Both of these features of the Cholesky decomposition can influence the Type I error rate. The attraction of not having to explain such nonsensical estimates has, to some extent, contributed to the popularity of the approach for the past quarter century.

A second approach to parameterizing genetic and environmental covariance matrices is the Correlated Factors model (Neale et al. 2006). Here, the predicted *A, C* and *E* covariance matrices are calculated by pre- and post-multiplying an estimated correlation matrix by a diagonal matrix of standard deviations (e.g., $\mathbf{A} = \mathbf{a}(\mathbf{R_A})\mathbf{a}^{\mathbf{T}}$ where $\mathbf{a}$ is diagonal and $\mathbf{R_A}$ is the genetic correlation matrix). As with the Cholesky decomposition, the Correlated Factors model ensures that the variances are positive, algebraically imposing a lower bound on the standard deviation parameters. Furthermore, it is possible to bound the correlations to be less than or equal to 1 in absolute value, which aids interpretability, although it is not sufficient to keep the matrices non-negative definite. Imposing boundaries to prevent the nonsensical correlations enforces another set of explicit boundaries that can influence model fit and affect the Type I error rate. While this is preferable in practice, we do not impose this boundary condition in the following simulations and allow the estimated correlations to take nonsensical values.

For both the Cholesky decomposition and the Correlated Factors models, the divergence between numerical and theoretical in Type I error rates is exacerbated in multivariate models. As the number of variables in a twin model increases, the number of implicit boundaries in the model increase, and subsequently, the divergence between the theoretical and numerical Type I error rates increases. In both cases, the numerical Type I error rate is lower than the theoretical threshold. Therefore, the null hypotheses that either $a^2 = 0$ or $c^2 = 0$ are rejected less frequently than would be expected due to chance. This, in turn, causes an increase in Type II errors, where the researcher falsely concludes the variance component is not significant.

The alternative approach to parameterizing twin models proposed here is to directly estimate the symmetric *A, C* and *E* matrices with no restrictions. We call this the Direct Symmetric approach as it directly estimates a set of symmetrical variance components matrices. While this approach may return nonsensical values in some situations (e.g. heritability estimates larger than 1, or non-positive definite covariance matrices), the absence of boundaries on the estimates yields asymptotically unbiased parameter estimates and correct Type I and Type II error rates.

Adjusting Type I error rates in the multivariate ACE model

While adjusting the p-values for the univariate case is fairly straightforward, doing so for multivariate twin models is more complicated (Wu and Neale 2012). The sampling distribution of the parameters under the null follows a mixture of zero and chi-squared distributions from 1 to the number of implicitly bounded parameters being set to zero. Worse, the mixture proportions are unknown because they depend

upon: (i) the parameters' estimates; (ii) the covariance between the estimates; and relatedly (iii) the study design, particularly ratio of MZ to DZ twins in the sample and the amounts of missing data. For example, in a bivariate twin model, the probability that all three shared environmental parameters are zero follows a mixture of chi-squared distributions with 0, 1, 2 and 3 df, with mixing proportions that are not known a priori. This issue also affects likelihood-based confidence intervals, which can be adjusted (and are automatically with OpenMx under certain circumstances; see Pritikin et al. 2017).

Hypothesis-driven multivariate behavioral genetic models

The common factor model for data from unrelated individuals has been generalized in two main ways for variance components analyses of twin data. The first, known as the Common Pathway (CPM) or Psychometric Factor model, partitions both the common factor and residual components into biometrical *A, C* and *E* variance components, as shown in Fig. 2. As with non-twin factor analysis, the number of factors, and the pattern of factor loadings relating the latent factors to the observed measures, may be set to represent specific hypotheses. The second generalization of the common factor model, known as the Independent Pathway model (IPM) or Biometric Factor model, may be viewed as a modification of the CPM model, where each latent factor is specified to consist of only one source of variation, *A, C* or *E*, as shown in Fig. 3. Depending on the number of measured phenotypes, more than one of each of these variance component factors may be specified.

Twin models that estimate *A, C* and *E* covariance matrices, such as the Cholesky, are often used as a comparison against which these more restricted, hypothesis-driven models are compared. In particular, it is common practice to compare the fit of the IPM and CPM models to that of the Cholesky. Unfortunately, if the Type I error rate of the saturated model does not follow theoretical expectations, likelihood ratio tests and other measures of relative model fit of may diverge from expectations. In practice, a researcher who finds no significant loss of fit when, e.g., the *C* matrix is fixed to zero, may decide to exclude it from further consideration when fitting IPM or CPM models. However, if that initial decision was erroneous, due to the use of an incorrect Type I error rate, the error may be perpetuated throughout the hypothesis-testing framework.

An alternative approach to addressing biases in the multivariate models

We address the previously identified biases in multivariate genetic models by focusing on three goals. First, using simulation, we compare the Type I error rates for three models: the Cholesky decomposition, correlated factors, and the direct symmetric models. Second, we explore how the Type I error rates of these three models vary as a function of the number of phenotypes being analyzed. Third, we examine how the choice of saturated model affects the Type I error rates for the hypothesis-driven independent and common pathway models. Lastly, we discuss the implications of the results for prior research and consider potential limitations of the bias-free Direct Symmetric model.

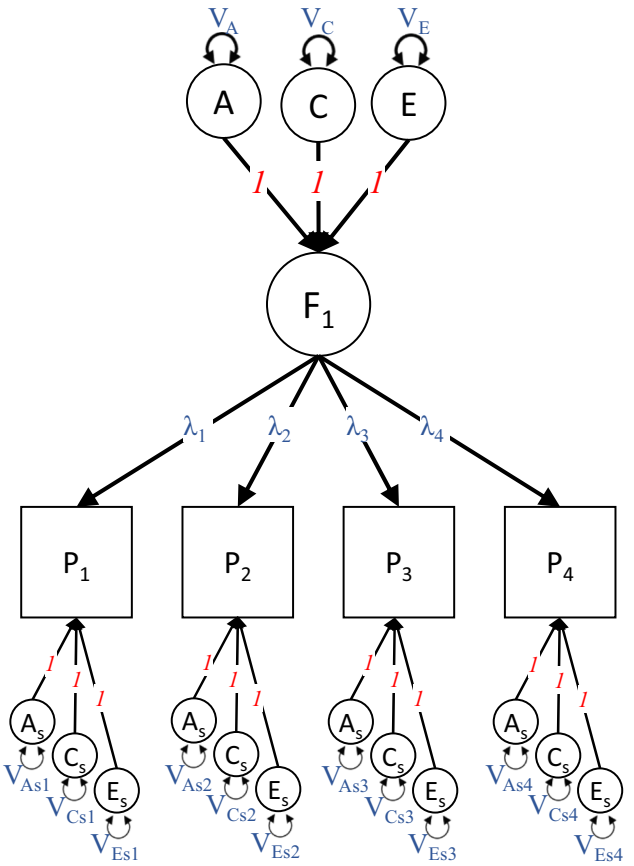

Fig. 2 Common Pathway model or Biometric Factor model for one twin. *Note* the CPM twin model is an extension of the common factor model. The latent factor, F_1, is caused by additive genetic (A), common (C) and specific (E) environmental factors. The four individual phenotypes P_1–P_4 are each caused by these latent factors and by residual variance components, which are also partitioned into additive genetic, common and specific environment components (A_{s1}–A_{s4}, C_{s1}–C_{s4} and E_{s1}–E_{s4}). Path labels in blue are estimated parameters and paths labels in red italics are fixed at the specified values. The variance components of the latent variable sum to 1 (the variance of the latent variable), while the residual variance components sum to the residual phenotypic variation. Only one twin is presented to simplify the schematic diagram. (Color figure online)

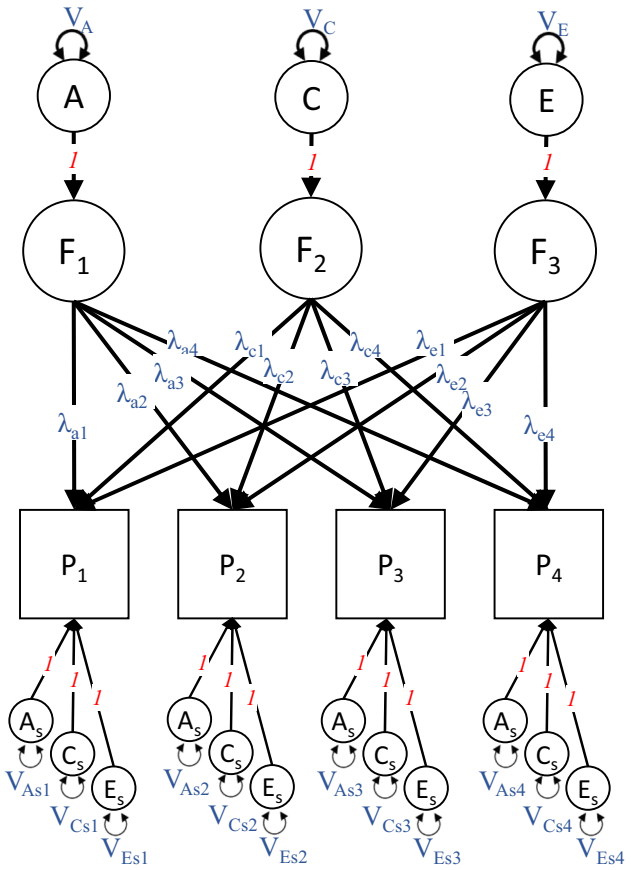

Fig. 3 Independent Pathway model or Psychometric Factor model for one twin. *Note* the IPM twin model is an extension of a three-factor confirmatory factor analysis. The latent factors F_1 is exclusively caused by additive genetic (A) factors; the latent factors F_2 is exclusively caused by common (C) environmental factors; and the latent factors F_3 is exclusively caused by specific (E) environmental factors. The association between F_1, F_2, and F_3, and the phenotypes P_1–P_4 are a function of the additive genetic, common or specific environmental factors that contribute to F_1, F_2, and F_3 and the respective factor loadings (e.g. $V_A \times \lambda_{a1}$) as well as the sum of the residual variance components. Path labels in blue are estimated parameters and paths labels in red italics are fixed at the specified values. Only one twin is presented to simplify the schematic diagram. (Color figure online)

Methods

Two simulation studies were conducted to examine the research questions. All simulations were conducted in R version 3.4.2 (R Core Team 2017), using OpenMx 2.7.12 (Neale et al. 2016; Boker et al. 2017) with the NPSOL 5.0 optimization algorithm, and MASS 7.3 (Venables and Ripley 2002) for generating continuous data from a multivariate normal distribution.

Study 1

Type I error rates are estimated under the three model specifications: (i) direct symmetric; (ii) Cholesky; and (iii) correlated factors. The number of variables per twin is varied from 1 to 4. To generate data under the null hypothesis of no common environment variation, data were generated using an AE model, in which half of the variance in each phenotype was due to additive genetic factors and half to unique environmental factors. In the multivariate simulations, all genetic correlations were set to $r_G = 0.6$. All unique environmental correlations were set to zero ($r_E = 0$). All three models were fitted to the same simulated data, to control stochastic variation between simulated datasets. The simulation was repeated 100,000 times to improve the empirical resolution in the tails of the sampling distributions that are particularly informative for the numerical Type I error rate.

For simulations with more than one phenotype, we also examined the Type I error rate for the common environmental *covariances* (i.e., *C* variation, but no *C* covariation). In summary, three models were fitted to each dataset: (i) the full ACE model, (ii) the AE model, and (iii) the no *C* covariance model.

Study 2

In study 2 we examined the Type I error rates when the data were generated according to a CPM. For each simulation, the latent phenotype was specified to have unit variance, with equal parts due to additive genetic and specific environment components. Factor loadings for the four items were set to [λ = 1, 0.9, 0.8, 0.7]. Residual variance for each item was set to have an AE structure, such that the A and E residual variances were set to $\sqrt{0.5}$. Each condition was repeated 100,000 times.

We fitted the following models to each dataset: (1) IPM, (2) CPM, (3) direct symmetric model, (4) the Cholesky decomposition, and (5) correlated factors model. For the IPM and CPM, three submodels were fit: (i) an *AE* model, (ii) the no latent *C* model, and (iii) a no residual *C* model. For the direct symmetric, Cholesky decomposition, and correlated factors models, the same models were fit as in the previous simulation study. The comparison between the saturated models with the IPM and CPM models provides information about the rejection rate of the hypothesis-driven models.

Scripts used to conduct the simulations studies are available at http://psychology.psy.msu.edu/QuantGen/T1E/T1E.html.

Table 1 Numerical estimates of the Type I error rate for different parameterizations of the saturated multivariate biometrical twin model

Number of variables	Estimation method	Comparison model	df	p-value threshold		
				0.1	0.05	0.01
1 Variable	Estimating the variance	AE	1	0.099 (0.097, 0.101)	0.049 (0.047, 0.050)	0.0097 (0.0091, 0.0103)
	Estimating the SD	AE	1	0.049 (0.048, 0.050)	0.024 (0.023, 0.025)	0.0049 (0.0044, 0.0053)
2 Variables	Direct symmetric	AE	3	0.100 (0.098, 0.102)	0.050 (0.049, 0.051)	0.0106 (0.0100, 0.0113)
		No Cov	1	0.097 (0.097, 0.100)	0.049 (0.048, 0.051)	0.0103 (0.0097, 0.0109)
	Cholesky decomposition	AE	3	0.021 (0.020, 0.022)	0.009 (0.009, 0.010)	0.0017 (0.0014, 0.0119)
		No Cov	1	0.018 (0.018, 0.019)	0.007 (0.006, 0.007)	0.0008 (0.0007, 0.0010)
	Correlated factors	AE	3	0.048 (0.047, 0.050)	0.023 (0.022, 0.024)	0.0045 (0.0041, 0.0049)
		No Cov	1	0.105 (0.103, 0.107)	0.053 (0.051, 0.054)	0.0111 (0.0104, 0.0117)
3 Variables	Direct symmetric	AE	6	0.101 (0.099, 0.103)	0.051 (0.050, 0.053)	0.0105 (0.0099, 0.0111)
		No Cov	3	0.101 (0.099, 0.102)	0.050 (0.049, 0.051)	0.0099 (0.0093, 0.0105)
	Cholesky decomposition	AE	6	0.021 (0.020, 0.022)	0.009 (0.009, 0.010)	0.0015 (0.0013, 0.0018)
		No Cov	3	0.018 (0.018, 0.019)	0.007 (0.006, 0.007)	0.0008 (0.0006, 0.0010)
	Correlated factors	AE	6	0.049 (0.048, 0.051)	0.023 (0.022, 0.024)	0.0041 (0.0037, 0.0045)
		No Cov	3	0.106 (0.104, 0.108)	0.053 (0.052, 0.054)	0.0113 (0.0106, 0.0119)
4 Variables	Direct symmetric	AE	10	0.100 (0.098, 0.102)	0.050 (0.048, 0.051)	0.0103 (0.0097, 0.0109)
		No Cov	6	0.100 (0.098, 0.102)	0.050 (0.049, 0.051)	0.0099 (0.0093, 0.0105)
	Cholesky decomposition	AE	10	0.012 (0.012, 0.013)	0.006 (0.005, 0.006)	0.0009 (0.0007, 0.0011)
		No Cov	6	0.010(0.009, 0.010)	0.004 (0.003, 0.004)	0.0005 (0.0003, 0.0006)
	Correlated factors	AE	10	0.046 (0.045, 0.047)	0.022 (0.021, 0.023)	0.0039 (0.0035, 0.0043)
		No Cov	6	0.105 (0.103, 0.107)	0.053 (0.051, 0.054)	0.0111 (0.0104, 0.0118)

95% confidence intervals in parentheses

The p-value threshold indicates the theoretical p-value and the cell entries indicate the percentage of the observed test statistics that exceeded the critical value of the χ^2 distribution with the correct number of degrees of freedom (or the numerical Type I error rate). Data were simulated 100,000 times under the assumption that the common environmental variance–covariance matrix was null, and then evaluated nine times using each parameterization method for the saturated model (ACE), the no common environmental model (AE), and the no common environmental covariance model (No Cov). The 95% confidence intervals are presented for each numerical estimate of the Type I error rate

Results

Type I error rates

Table 1 shows the results of Study 1. The top panel shows that the univariate twin model Type I error rate is consistent with previous research (Carey 2005; Dominicus et al. 2006; Visscher 2006). When the standard deviation of the variance component is estimated (as is the case in both the univariate Cholesky decomposition and correlated factors model) the Type I error rate is half that of the theoretical rate. This is consistent with the notion that with this parameterization, the test statistic is distributed as a 50:50 mixture of a χ^2 distribution with 1 df, and χ^2 distribution with 0 df. However, when the *A, C* and *E* variance components are estimated directly (i.e., not bounded to be positive), the empirical Type I error rate approximates χ^2 with 1 df.

A similar picture emerges in the multivariate case. With the direct symmetric model, the empirical and theoretical Type I error rates are consistent for both the AE model and the no C covariance conditions. That is, the likelihood ratio test for the existence of C components is distributed as χ^2 with df equal to the number of parameters fixed to zero. However, under the Cholesky, the empirical Type I error rates for both the AE and the no C covariance models are much lower than theoretically expected for a χ^2 distribution for the appropriate df. For the correlated factors model, the empirical Type I error rate for the AE model is consistent with the explanation that the test statistics are mixture of multiple χ^2 distributions with increasing dfs. By contrast, the Type I error rate of the test for no *C* contribution to covariance, is lower than would be expected if it were distributed as χ^2 with df equal to the number of parameters fixed to zero.

Based upon these results, we conclude that when boundaries are included, be they implicit or explicit, the Type I error rate is substantially lower than would be expected. Furthermore, as the number of phenotypes in the model increases, the greater the divergence from the nominal Type I error rates.

Table 2 shows the results of the second simulation study. The general pattern of results for the Cholesky, correlated

Table 2 Numerical estimates of the Type I error rate for different parameterizations of common multivariate biometrical twin models

		df	p-value threshold		
			0.1	0.05	0.01
Direct symmetric	AE	10	0.101 (0.099, 0.103)	0.050 (0.049, 0.052)	0.0103 (0.0098, 0.0110)
	No C Cov	6	0.100 (0.098, 0.102)	0.049 (0.048, 0.051)	0.0099 (0.0093, 0.0106)
Cholesky decomposition	AE	10	0.012 (0.011, 0.013)	0.005 (0.005, 0.006)	0.0008 (0.0006, 0.0009)
	No C Cov	6	0.011 (0.010, 0.011)	0.004 (0.004, 0.004)	0.0005 (0.0003, 0.0006)
Correlated factors	AE	10	0.060 (0.058, 0.061)	0.034 (0.033, 0.035)	0.0155 (0.0148, 0.0163)
	No C Cov	6	0.122 (0.120, 0.124)	0.068 (0.067, 0.070)	0.0231 (0.0222, 0.0231)
Independent pathway	AE	8	0.097 (0.095, 0.099)	0.050 (0.048, 0.051)	0.0099 (0.0093, 0.0105)
	No latent C	4	0.096 (0.094, 0.098)	0.048 (0.047, 0.049)	0.0100 (0.0094, 0.0107)
	No residual C	4	0.091 (0.089, 0.093)	0.051 (0.050, 0.053)	0.0210 (0.0201, 0.0219)
Common pathway	AE	5	0.101 (0.099, 0.103)	0.051 (0.049, 0.052)	0.0103 (0.0097, 0.0110)
	No latent C	1	0.101 (0.099, 0.103)	0.051 (0.049, 0.052)	0.0093 (0.0088, 0.0100)
	No residual C	4	0.101 (0.099, 0.102)	0.051 (0.050, 0.052)	0.0100 (0.0094, 0.0106)
Direct symmetric vs IPM		6	0.109 (0.107, 0.110)	0.055 (0.054, 0.056)	0.0114 (0.0108, 0.0121)
Direct symmetric vs CPM		16	0.099 (0.097, 0.101)	0.050 (0.048, 0.051)	0.0100 (0.0094, 0.0107)
Cholesky vs IPM		6	0.017 (0.016, 0.018)	0.007 (0.007, 0.008)	0.0011 (0.0009, 0.0013)
Cholesky vs CPM		16	0.025 (0.024, 0.026)	0.010 (0.010, 0.011)	0.0016 (0.0014, 0.0019)
Correlated factors vs CPM		6	0.064 (0.063, 0.066)	0.032 (0.031, 0.033)	0.0060 (0.0055, 0.0065)
Correlated factors vs IPM		16	0.065 (0.064, 0.067)	0.031 (0.030, 0.032)	0.0060 (0.0055, 0.0065)

95% confidence intervals in parentheses

The p-value threshold indicates the theoretical p-value and the cell entries indicate the percentage of the observed test statistics that exceeded the critical value of the χ^2 distribution with the correct number of degrees of freedom. Data for a 4-variable common pathway model were simulated 100,000 times under the assumption that the common environmental did not contribute to variation in the items. The data were then evaluated 17 times using each saturated model an each of the various reduced models listed. The 95% confidence intervals are presented for each numerical estimate of the Type I error rate

factors and direct symmetric models follows those of Study 1, replicating the results under different simulation conditions. The results from the IPM and CPM models, presented in the middle panels of Table 2, show that when the reduced IPM or CPM is compared to the full IPM or CPM, respectively, the numerical and theoretical Type I error rates are consistent with nominal rates. Thus, the deviation in the Type I error rate is not an inevitable feature of the IPM or CPM. Furthermore, when the IPM and CPM are compared with the direct symmetric model, the numerical and theoretical Type I error rates converge.[1]

Bias in the variance estimates

Another consequence of the implicit lower bound of zero on variance components is that the under the null hypothesis, the parameters will be biased for the Cholesky decomposition and Correlated Factors models. Figure 4 presents histograms for Cholesky decomposition, correlated factors, and direct symmetric methods for the common environmental and additive genetic variance components of the first variable in simulation study 2. This specific variable was chosen because the algebraic calculation of this common environmental and additive genetic variance is exactly the same regardless of the number of variables in the model. In the figure, the solid red line indicates the mean of the numerically observed estimate and the dotted blue line indicates the simulated value.

As can be seen in the top row of the figure, because the variance components of the Cholesky decomposition and the correlated factors models cannot be negative, a large portion of the sampling distribution of the common environmental variance component is truncated. Further, because the true value of this parameter is zero, which is the implicit boundary, any deviations from the simulated value must be positive, inducing a small positive bias in the estimated common environmental parameters. This is not the case for the Direct Symmetric sampling distribution, where the distribution is approximately normal around zero.

[1] We repeated the second simulation study to examine whether the observed results were specific to the likelihood ratio test or whether they could generalize to other hypotheses testing techniques such as the Wald Test. The Wald test results were effectively equivalent with the LRT results above, making their presentation here unnecessary.

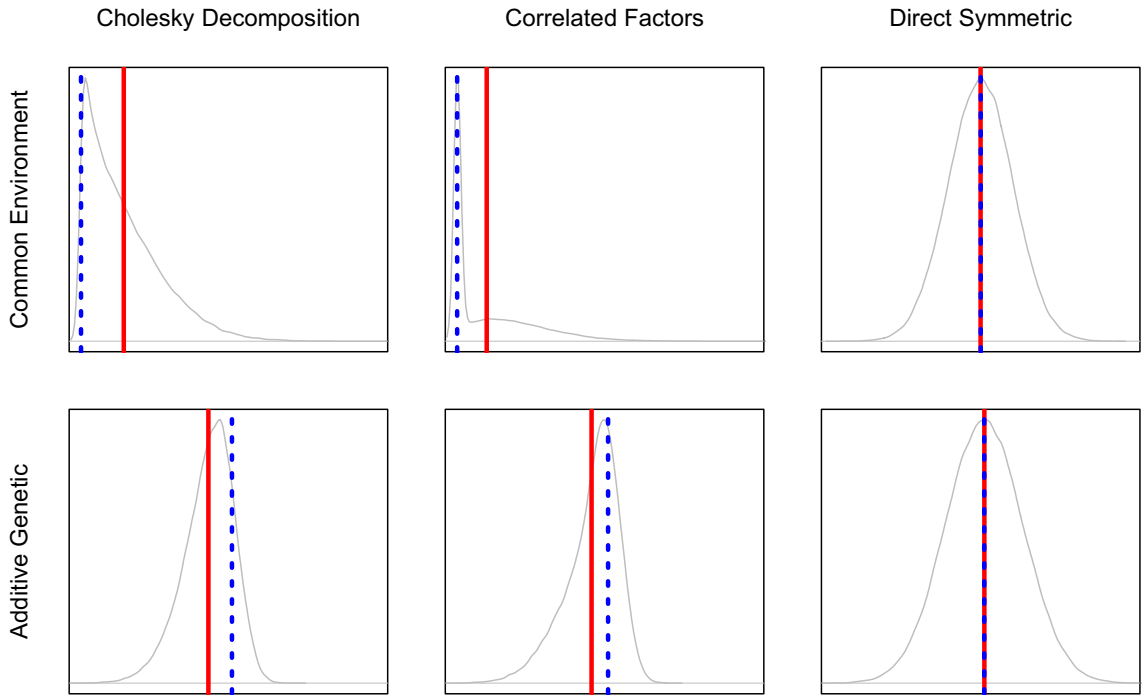

Fig. 4 Density plots of the estimated common environmental and additive genetic variance components for the Cholesky, Correlated Factors, and Direct Symmetric estimation methods. *Note* the density plots depict the common environmental and additive genetic variance components for the first variable in simulation study 2. The solid red lines indicate the observed mean of the distribution while the dotted blue lines indicated the simulated value for the parameter. If the solid red line is on the right of the dotted blue line, then the parameter is overestimated and if the solid red line is on the left of the dotted blue line the parameter is underestimated. (Color figure online)

The bottom row of the figure presents the distribution of the additive genetic variance component from the same analyses. Because there is a dependency between the common environmental and additive genetic variance components, the slight upward bias in the common environmental variance corresponds with a slight downward bias in the additive genetic variance for the Cholesky decomposition and the Correlated Factors methods. Therefore, in repeated sampling, if the Cholesky decomposition or Correlated Factors models were used, we would expect a slight inflation of any variance components that are truly zero, and a corresponding deflation of the alternative variance components.

Model convergence rates

The NPSOL optimizer used to fit the models here can result in one of five status codes, of which 0, 1, 4, 5, and 6 are of the most interest for the current project (Neale et al. 2016: see the note for Table 3 or the OpenMx manual for more detail on the status codes). Typically, codes 0 or 1 indicate good convergence, whereas codes 5 and 6 require additional scrutiny and caution because the optimizer may not have found a global minimum and report incorrect parameter estimates. Code 4 indicates that insufficient major iterations were undertaken (i.e., the iteration limit has been reached) and the results should not be trusted.

Table 3 presents the optimization status codes for both simulation studies. The Direct Symmetric method invariably returned optimization status code 0 or 1 (in 400,000 repetitions). By contrast, the Cholesky method usually returned an optimization status of 0 or 1, but occasionally returned a status of 5. The probability of status 5 codes seems to increase with the number of phenotypes being analyzed. Finally, the Correlated Factors model usually returned optimization status 5 or 6, sometimes returns a code 4 (all of which demand additional scrutiny) and only occasionally returned code 0 or 1. Moreover, the probability of observing a code 0 or 1 decreased as the number of phenotypes increased.

Therefore, contrary to expectation, the direct symmetric method does not appear to increase optimization problems relative to the two other methods. Follow-up analyses reveal that optimization time is, on average, shortest for the direct symmetric method, slightly longer for the Cholesky, and longest for the correlated factors method, consistent with the expectation that it will take longer to estimate models that have issues with optimization (where the NPSOL status is greater than 1).

Comparison of the −2 log-likelihood (−2lnL) across the parameterization methods provides further insight into their

Table 3 NPSOL status codes for all of the models that were estimated separated by the estimation algorithm and the data generation process

Number of variables	Estimation method	NPSOL status codes				
		0	1	4	5	6
1 Variable	Variance	100,000				
	Std. dev	100,000				
2 Variables	Direct symmetric	85,721	14,279			
	Cholesky	60,768	36,276		2956	
	Correlated factors	15,894	13,649	1487	67,603	1367
3 Variables	Direct symmetric	85,498	14,502			
	Cholesky	60,844	36,038		3118	
	Correlated factors	15,872	13,334	1445	68,020	1329
4 Variables	Direct symmetric	96,792	3208			
	Cholesky	45,414	49,921		4665	
	Correlated factors	3776	4440	459	89,359	1966
5 Variables	Direct symmetric	99,924	76			
	Cholesky	43,888	48,073		8039	
	Correlated factors	2507	4104	212	91,829	1348
	IPM	3477	20,735	2	5335	70,451
	CPM	77,598	22,402			

The five NPSOL status codes indicate: successful optimization (0), successful optimization but optimization did not converge to a single likelihood (1: Mx status GREEN), the iteration limit was reached before a solution could be found (4: Mx status BLUE), the Hessian matrix is not convex at the solution (5), and the gradient is not close enough to zero to satisfy the optimization requirements, but that no improvement in the solution could be found (6: Mx status RED). The top four panels use the results from the first simulation study and the bottom panel uses the results from the second simulation study

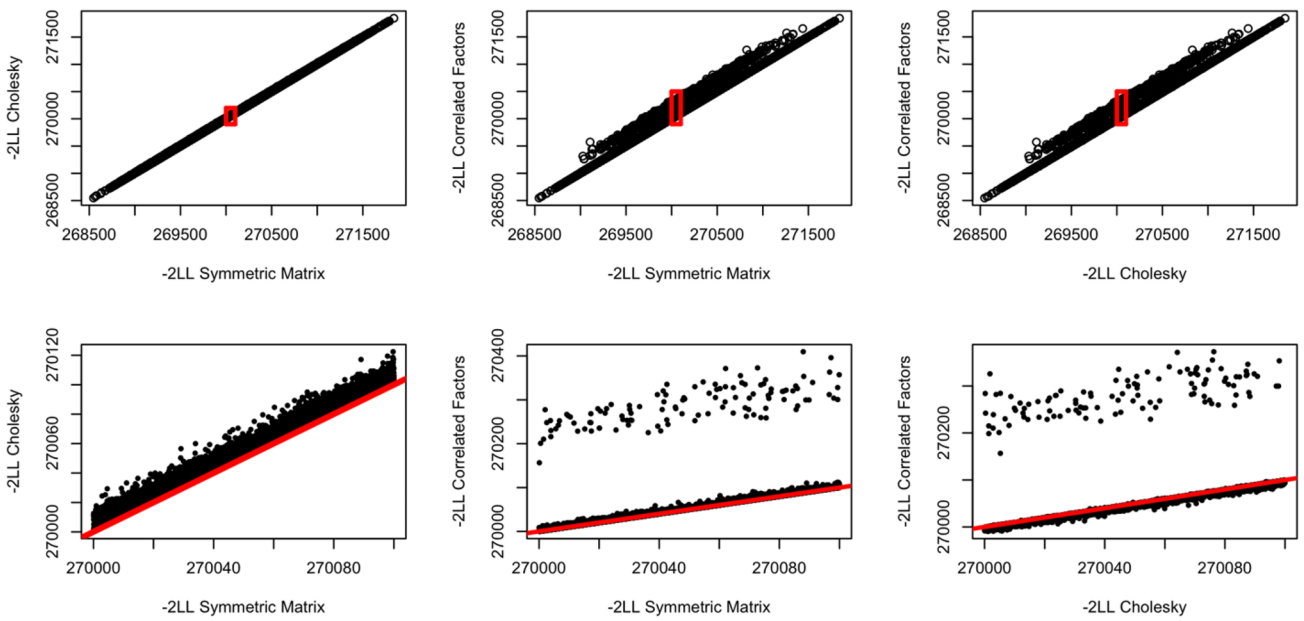

Fig. 5 Pairwise comparisons between the negative 2 log-likelihoods of the Direct Symmetric, Cholesky and Correlated Factors methods of estimating variance components. *Note* the top panels show the full range of the −2 log-likelihoods for each pairwise estimation method. The bottom panels depict a zoomed in view of the scatterplot consistent with the red box in the panel above. The red line in the bottom panels indicates the equality of the −2 log-likelihood for each parameterization method. The data are taken from the 4-variable model from the first simulation study. (Color figure online)

differences, as presented in Fig. 5. The three parameterization methods fit models that are theoretically equivalent across the region of the parameter space when all three variance components are greater than zero. The −2lnL should be equal across methods in these cases. However, the Cholesky and correlated factors models have a restricted parameter space, being unable to fit the data as well when one or more of the variance components is estimated to be less than zero. Figure 5 compares the −2lnL between the three parameterizations. The top panels show the entire range of the −2lnL values for each model, and the bottom panels focus on a narrow range of the values to explore the differences in model fit in more detail. The red line depicts equality of the −2lnL.

The top panels show high correspondence between the −2lnL of the three parameterization methods. The correlations between the models' −2lnL are all > 0.99. These high correlations in the upper panels overstate the agreement in fit across the models. The bottom panels, with the narrower range of log-likelihoods, emphasize their differences. All points in the Figure are either on or above the diagonal, showing that the Direct Symmetric method always fits as well as, or better than, the other two methods. There is, however, considerable variation in the difference in model fit, with the alternative parameterizations occasionally having a difference in −2lnL of up to approximately 50 (for the Cholesky) or 450 (for the Correlated Factors). With the Correlated Factors model, there is a cluster of −2lnL that appear to be much worse than expected, likely due to the optimizer failing to find the global minimum. Such results highlight the robustness of the Direct Symmetric method.

Discussion

These simulation studies demonstrate that implicit and explicit boundaries lead to a deviation from the expected Type I error rate and can induce bias in the parameter estimates under the null hypothesis. Specifically, the more boundaries that exist within an estimation procedure, the more the numerical Type I error rate departs from the nominal rate. Each implied boundary contributes to this divergence, but not necessarily at the same rate. On a practical level this means that Type I error rates from analyses that directly estimate the *A, C* and *E* covariance matrices without boundaries will follow the theoretical Type I error rate, while analyses that place explicit or implicit boundaries on parameters, such as a Cholesky decomposition, will deviate from the theoretical expectations, often by an order of magnitude or more. These incorrect Type I error rates can cause errors of inference, where researchers conclude that a statistically significant parameter is not significant because they are testing at a much more stringent alpha level than they realize. Furthermore, the implicit boundary of zero on variance components will also upwardly bias estimates of any variance components that are at, or near, zero. If the sampling distribution of a parameter includes an implicit or explicit boundary, the sampling distribution for the parameter will be truncated with all estimates outside the bounded space returning the bounded value. This will alter the observed mean of the sampling distribution, resulting in biased parameter estimates. Because the common environment and additive genetic variance components in twin models are compensatory, an upward bias for one variance component will result in a downward bias in the estimate for the compensatory variance component.

While the current simulations focus on the common environment parameters, the results directly translate to the additive genetic variance components. Under the null hypothesis of no genetic variation, the expected value of the additive genetic variance component would be upwardly biased if implicit boundaries were included in the model, and the common environmental variance component would be downwardly biased.

The method of estimating saturated models can influence subsequent statistical decision making for hypothesis-driven IPM and CPM models. Note that comparing the reduced IPM or CPM to the full IPM or CPM is not the comparison that is generally made when assessing the fit of multivariate twin models. Typically, the IPM and CPM model fits are compared to that of the Cholesky. For both the Cholesky and correlated factors models, the Type I error rate is lower than nominal. This divergence results in concluding that the fit of the more parsimonious IPM or CPM is worse than the saturated model less often than would be expected by chance. Historically, the restricted IPM or CPM may have been accepted too frequently, by comparing their fit relative to that of the Cholesky. Comparison to the direct symmetric model is to be preferred. Accordingly, the choice of baseline model can strongly affect the statistical properties of the test statistics as well as the inferences that can be drawn from the results.

When discussing Type I error rate violations, it is typically in the context of inflation of the family-wise error rate (FEWR). In such cases, the nominal error rate is much higher than the specified error rate, due, for example, to multiple testing. In the current context, however, the error rate is downwardly biased, making it less likely that the null hypothesis will be rejected. For multivariate twin studies, this means that while researchers may think they are testing their hypotheses with a significance level of $\alpha = 0.05$, in many cases they are actually testing with a significance level of $\alpha = 0.01$ or less (depending on the number of variables in the model). When the test statistic is significant, however, the results are even more robust than the researcher implied. In cases where the statistic

did not reach the theoretical level of significance, however, Type II errors are likely quite common.

The fact that the Type I error rate is conservative implies that the Type II error rate is inflated. This problem is compounded by the routine practice of dropping non-significant parameters to present more parsimonious models with narrower parameter confidence intervals. Moreover, in twin models, the A and C parameters have a high dependency, such that by excluding one parameter the other parameter is inflated in a compensatory fashion. Because the evidence for common environmental variation is often much weaker than that for additive genetic variation, *C* is routinely dropped from analyses. It is possible that this practice has upwardly biased estimates of additive genetic variance. There has been a push in the literature to avoid dropping non-significant parameters, minimizing the overestimation of the A variance components (Sullivan and Eaves 2002). Alternatively, utilizing the Cholesky decomposition or Correlated Factors approach and retaining null parameters will have the opposite effect by downwardly biasing the compensatory variance component. The recent meta-analyses of Polderman et al. (2015) seem consistent with upward bias of variance components. They found smaller meta-analytic estimates of C when Holzinger's formulas (commonly but incorrectly referred to as Falconer's formulas) were used than from the implicitly bounded maximum likelihood (Newman et al. 1937). In adequately powered studies, however, this bias is likely relatively small.

Before twin researchers started using structural equation modeling to fit twin models (Martin and Eaves 1977), heritability estimates were often calculated using Holzinger's formulas (Newman et al. 1937), and these formulas also did not place boundaries on the heritability estimates (also see Falconer 1960). As such, heritability estimates were routinely calculated that were not bounded by 0 and 1 (when heritability is estimated from LD score regression methods, the estimates are also not bounded by 0 and 1; Bulik-Sullivan 2015; Zheng et al. 2016). This "inconvenience" played a role in the development of SEM methods to estimate twin models.

It is useful to consider the conditions where negative estimates of variance components are likely to be observed. We discuss three scenarios where we expect they are most likely to be observed but other situations may arise with additional research.

First, it is possible that the negative variance components accurately reflect the underlying genetic or environmental mechanism under investigation (Steinsaltz et al. 2017). For example, genes that have the opposite effects in subsequent generations, as has been observed for neonatal jaundice (Haldane 1996). Alternatively, while most heritability estimates for epigenetic factors are positive, a non-trivial proportion appear to have negative heritability estimates (Steinsaltz et al. 2017). While we expect this to be rare in practice, it is a theoretically plausible expectation in some situations.

Second, due to sampling error, the observed MZ correlations could be underestimated and the DZ correlations could be overestimated by chance alone. This is particularly likely in smaller samples, where there is more variability in the sampling distribution of the respective correlations. If this is the case, the estimate of the variance component will often be negative but not statistically significant, implying the parameter is not statistically distinguishable from zero. The substantially larger sample sizes in modern twin studies may ameliorate this problem, but it is unclear how frequently this will occur in practice.

Finally, an interesting situation arises when thinking about the possibility of using negative variance components as a method of evaluating model misspecification. For example, in the analysis of MZ and DZ twin data, researchers are forced to choose between estimating a shared environmental variance component (*C*) or a non-additive genetic variance component (*D*), as one variance component must be fixed to zero to allow for model identification. If the shared environmental variance component is negative, however, it may be due to stochastic variation in the estimate, or a genuinely different source of variation, such as genetic dominance. In fact, it is possible to calculate the expected value of *C* or *D* from the obtained negative parameter estimates. Specifically, $D = -2C$ and $C = -D/2$. While the prevalence of these illogical estimates is currently unknown, their presence will provide an opportunity to explore the biometrical model in a new light and hopefully lead to a better understanding of the phenotypes under investigation.

The above scenario assumes that either *C* or *D* is truly zero. Previous research has demonstrated that if both *C* and *D* are non-zero (i.e. positive), their estimates in a twins only model are confounded, and the observed value of *A* is inflated (Coventry and Keller 2005). It is necessary to note that the direct symmetric approach does not resolve any confounding between *C* and *D*. If both *C* and *D* are present for a phenotype, simply allowing the variance components to go negative will not negate the bias in *A*.

When analyses are expanded to multiple phenotypes, which exacerbates the Type I error rate issues, potential problems with negative definite matrices can arise. For example, in situations where one phenotype has a negative variance component, or the genetic or environmental correlations are greater than 1, the entire matrix will be negative definite and difficult to interpret. If this should occur, analysts should delve deeper into the potential factors that give rise to the aberrant result, such as a peculiar pattern of covariation between variables, or marked differences between groups. These scenarios may provide the opportunity to learn something that would otherwise have been

masked by forcing matrices to be positive definite. It is worth noting that if the genetic and environmental covariance between phenotypes are in the opposite direction (e.g. the genetic correlation is positive and the common environmental correlation is negative), the proportion of genetic or environmental covariation between phenotypes can be greater than one or less than zero with any multivariate biometrical variance decomposition method. Researchers should also examine whether the parameters driving the results are statistically significant, or whether the observed parameter is potentially due to sampling variability. If the anomalous parameter is not significant, it should be noted, but drastic measures may not be justified. If, however, the results are completely uninterpretable, it may be necessary to impose constraints.

It is important to highlight that the Type I error rate and parameter bias issues do not affect all twin models equally. For example, models that test the basic assumptions of twin data, such as equal means, thresholds and variances across twin order and zygosity, are unaffected by the current issues because the statistical analyses do not include implicit or explicit boundaries. Furthermore, in some circumstances, all three methods are equivalent. If all of the estimated variance components are positive, the same model fit will be obtained for each approach, as can be seen in Fig. 5, and the parameters from one specification will be transformations of another. In this situation, the numerical Type I error rates from each model will follow the theoretical distribution and there will be no bias in any of the parameters. If any of the variance components from the direct symmetric approach are negative, the model fit for each algorithm will diverge and the any boundaries that are encountered will induce potential bias.

Despite the Type I error issue, the Cholesky decomposition approach to fitting models to twin data has several very useful properties. First, it generalizes well to fitting models to multigenerational data, e.g., twins and their parents or children (Rice et al. 1979; Neale and Fulker 1984). It is also of practical value in the specification of models of sex-limitation and genotype by environment interaction (Purcell 2002; Neale et al. 2006). Analogous direct symmetric implementations of these more elaborate models to handle more complex research designs and questions have yet to be developed. There are some issues with an expected opposite sex DZ twin covariance of $0.5\sqrt{(VAm*VAf)}$ if either the male or female additive genetic components, VAm and VAf, is less than zero. Research is currently being conducted to account for this situation.

As researchers consider modeling twin data in the future, it is important to consider the goals of such an exercise. First, does the model fit the data? And second, does the model make biological sense? In this paper we have demonstrated that the divergence of the Type I error rates for the Cholesky decomposition and correlated factors models from their theoretical expectations is a result of implicit boundary conditions. Furthermore, these implicit boundaries can induce bias in the estimated parameters. The primary trade-off between the Cholesky, the correlated factors, and the direct symmetric model pits the statistical properties of the estimates and the inferences that can be drawn from the analyses against the interpretability of the estimates. While we do not wish to downplay the importance of interpretability, we strongly believe that the statistical properties and subsequent inferences are of paramount concern and therefore urge future twin researchers to use the direct symmetric approach to fit twin models. In summary, the direct symmetric approach has several advantages over other multivariate twin models as it corrects the Type I error rate and parameter bias issues, is easy to implement in current software, and has fewer optimization problems.

Acknowledgements An earlier draft of this paper was circulated to the faculty of the 2018 International Workshop on Statistical Genetic Methods for Human Complex Traits in Boulder, Colorado and was presented at the 48th meeting of the Behavioral Genetics Association in Boston Mass., June 20 to June 23, 2018. We would like to thank the workshop faculty and students, conference attendees for their suggestions to improve the paper.

Funding This study was supported by NIDA Grants R01DA-018673 and R25DA-26119.

Compliance with ethical standards

Conflict of interest Brad Verhulst, Elizabeth Prom-Wormley, Matthew C Keller, Sarah Medland, and Michael C. Neale declare that they have no conflict of interest.

Informed consent For this type of study formal consent is not required.

Statement of human and animal rights This article does not contain any studies with human participants or animals performed by any of the authors.

References

Boker SM, Neale MC, Maes HH, Wilde MJ, Spiegel M, Brick TR, Estabrook R, Bates TC, Mehta P, von Oertzen T, Gore RJ, Hunter MD, Hackett DC, Karch J, Brandmaier A, Pritikin JM, Zahery M, Kirkpatrick RM, Wang Y, Driver C, Johnson SG, Kraft D, Wilhelm S, Manjunath BG (2017) OpenMx 2.7.17-23 User Guide.

Bulik-Sullivan BK (2015) Relationship between LD score and haseman-Elston, bioRxiv. https://doi.org/10.1101/018283

Carey G (2005) Cholesky problems. Behav Genet 35(5):653–665

Coventry WL, Keller MC (2005) Estimating the extent of parameter bias in the classical twin design: a comparison of parameter estimates from extended twin-family and classical twin designs. Twin Res Hum Genet 8(3):214–223

Dominicus A, Skrondal A, Gjessing HK, Pedersen NL, Palmgren J (2006) Likelihood ratio tests in behavioral genetics: problems and solutions. Behav Genet 36(2):331–340

Falconer DS (1960) Introduction to quantitative genetics. Oliver and Boyd, London

Haldane JBS (1996) The negative heritability of neonatal jaundice. Ann Hum Genet 60:3–5

Martin NG, Eaves LJ (1977) The genetical analysis of covariance structure. Heredity 38(1):79–95

Neale MC, Cardon LR (1992) Methodology for genetic studies of twins and families. Kluwer Academic Publishers BV, Dordrecht

Neale MC, Fulker DW (1984) Heritability of item responses on the eysenck personality questionnaire. Personal Individ Differ 7:771–779

Neale MC, Røysamb E, Jacobson K (2006) Multivariate genetic analysis of sex limitation and G E interaction. Twin Res Hum Genet 9(4):481–489

Neale MC, Hunter MD, Pritikin JN, Zahery M, Brick TR, Kickpatrick RM, Estabrook R, Bates TC, Maes HH, Boker SM (2016) OpenMx 2.0: Extended structural equation and statistical modeling. Psychometrika 80(2):535–549. https://doi.org/10.1007/s11336-014-9435-8

Newman HH, Freeman FN, Holzinger KJ (1937) Twins: a study of heredity and environment. The University of Chicago Press, Chicago, IL

Pritikin JN, Rappaport LM, Neale MC (2017) Likelihood-based confidence intervals for a parameter with an upper or lower bound. Struct Equ Modeling 24(3):395–401

Purcell S (2002) Variance components models for gene-environment interaction in twin analysis. Twin Res 5(6):554–571

R Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/

Steiger JH (1980) Tests for comparing elements of a correlation matrix. Psychol Bull 87(2):245–251

Steinsaltz D, Dahl A, Wachter KW (2017) On negative heritability and negative estimates of heritability. bioRxiv. https://doi.org/10.1101/232843

Sullivan PF, Eaves LJ (2002) Evaluation of analyses of univariate discrete twin data. Behav Genet 32(3):221–227

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer: New York

Visscher PM (2006) A note on the asymptotic distribution of likelihood ratio tests to test variance components. Twin Res Hum Genet 9(4):490–495

Wu H, Neale MC (2012) Adjusted confidence intervals for a bounded parameter. Behav Genet 42(6):886–898

Zheng J, Erzurumluoglu AM, Elsworth BB, Kemp JP, Howe L, Haycock PC, Hemani G, Tansey K, Laurin C, Early Genetics and Lifecourse Epidemiology (EAGLE), Eczema Consortium, Warrington NM, Finucane HK, Price AK, Bulik-Sullivan BK, Anttila V, Paternoster L, Gaunt TR, Evans DM, Neale BM (2016) LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics 33(2):272–279