

Using Non-Normal SEM to Resolve the ACDE Model in the Classical Twin Design

Koken Ozaki · Hideki Toyoda · Norikazu Iwama · Saori Kubo · Juko Ando

Received: 8 June 2009 / Accepted: 22 July 2010 / Published online: 12 August 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract One of the biggest problems in classical twin studies is that it cannot estimate additive genetic (A), non-additive genetic (D), shared environmental (C), and non-shared environmental (E) effects, simultaneously, because the model, referred to as the ACDE model, has negative degrees of freedom when using Structural Equation Modeling (SEM). Therefore, instead of the ACDE model, the ACE model or the ADE model is actually used. However, using the ACE or ADE models almost always leads to biased estimates. In the present paper, the univariate ACDE model is developed using non-normal Structural Equation Modeling (nnSEM). In SEM, (1st- and) 2nd-order moments, namely, (means and) covariances are used as information. However, nnSEM uses higher-order moments as well as (1st- and) 2nd-order moments. nnSEM has a number of advantages over SEM. One of which is that nnSEM can specify models that cannot be specified using SEM because of the negative degrees of freedom. Simulation studies have shown that the proposed method can decrease the biases. There are other factors that have possible

effects on phenotypes, such as higher-order epistasis. Since the proposed method cannot estimate these effects, further research on developing a more exhaustive model is needed.

Keywords Biases in estimators · Higher-order moments · Model identification · nnSEM · Non-normality · Univariate ACDE model

Introduction

In behavior genetics, two types of genetic effects on a trait are considered, namely, additive genetic (A) effects and non-additive genetic (D or NA) effects. In addition, two types of environmental effects on a trait are considered, namely, shared environmental (C) effects and non-shared environmental (E) effects. Therefore, in behavior genetics, the above four effects on phenotypes are assumed. To estimate these effects, structural equation modeling (Bollen 1989; Eaves et al. 1978a, b) is used extensively.

Figure 1 illustrates the path diagram in which all four factors are included. This is called the univariate ACDE model. In Fig. 1, p_1 and p_2 are phenotypes of the same trait for twin 1 and twin 2. In addition, let a , c , d , and e be the additive genetic effect, the shared environmental effect, the sum of the dominance effect and the interaction effect between genes (epistatic effect), and the nonshared environmental effect, respectively.

To estimate a , c , d , and e , SEM using 2nd-order (and sometimes 1st-order) moments are used in current behavior genetic models (Eaves et al. 1978a, b; Neale and Cardon 1992). A significant problem in current behavior genetics models is that the ACDE model in Fig. 1 cannot be identified. Therefore, reduced ACE or ADE models are actually used in most research. However, the use of these reduced

Edited by Pak Sham.

K. Ozaki (✉)
Research Organization of Information and Systems,
The Institute of Statistical Mathematics, 10-3 Midori-cho,
Tachikawa, Tokyo 190-8562, Japan
e-mail: koken@ism.ac.jp

H. Toyoda · N. Iwama · S. Kubo
Waseda University, Tokyo, Japan

N. Iwama
Japan Society for the Promotion of Science,
Tokyo, Japan

J. Ando
Keio University, Tokyo, Japan

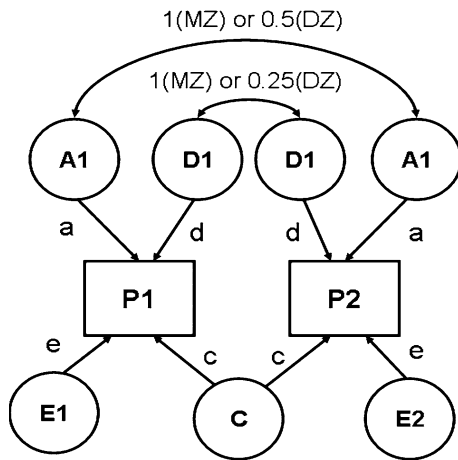


Fig. 1 The ACDE model

models may lead to biased estimators of these effects (Eaves et al. 1978a; Grayson 1989). The purpose of the present paper is to identify the ACDE model using non-normal structural equation modeling (nnSEM; Shimizu and Kano 2008; Ozaki and Ando 2009) to overcome this problem.

The reason why the ACDE model cannot be identified is as follows. The expected covariance matrices of the ACDE model for MZ and DZ twins are as follows:

$$\Sigma_{MZ} = \begin{bmatrix} E[p_1^2(mz)] & E[p_1p_2(mz)] \\ E[p_1p_2(mz)] & E[p_2^2(mz)] \end{bmatrix} = \begin{bmatrix} a^2 + c^2 + d^2 + e^2 & a^2 + c^2 + d^2 \\ a^2 + c^2 + d^2 & a^2 + c^2 + d^2 + e^2 \end{bmatrix} \tag{1}$$

$$\Sigma_{DZ} = \begin{bmatrix} E[p_1^2(dz)] & E[p_1p_2(dz)] \\ E[p_1p_2(dz)] & E[p_2^2(dz)] \end{bmatrix} = \begin{bmatrix} a^2 + c^2 + d^2 + e^2 & 0.5a^2 + c^2 + 0.25d^2 \\ 0.5a^2 + c^2 + 0.25d^2 & a^2 + c^2 + d^2 + e^2 \end{bmatrix} \tag{2}$$

Here, $E[p_1p_2(mz)]$ is the expectation of p_1p_2 of MZ twins (here, let both p_1 and p_2 be mean centered phenotypes). The variances of A, C, D, and E are fixed to 1s. This model cannot be identified because there are three unique elements (one variance and two covariances). However, there are four parameters in the model (number of degrees of freedom: -1). Therefore, instead of the ACDE model, the ACE or ADE models are actually used.

In general, when twice the correlation between the DZ pair (r_{dz}) exceeds the correlation between the MZ pair (r_{mz}), the ACE model, in which D is removed, is used. Otherwise, the ADE model, in which C is removed, is used. In the latter case, MZ pairs are much similar than DZ pairs. Therefore, genetic effects are expected to be more

Table 1 Biases in estimators when the ACE model or the ADE model is fit

	ACE	ADE
a_E^2	$1.5d_T^2$	$3c_T^2$
c_E^2	$-0.5d_T^2$	$-c_T^2$
d_E^2	$-d_T^2$	$-2c_T^2$
e_E^2	0	0

Subscript E denotes the estimator, and subscript T denotes the true value

important than common environmental effects, and d is included in the model. However, this model fit strategy is used because of the statistical limitation described above.

Actually, when twice the correlation between the DZ pair does not exceed the correlation between the MZ pair, this indicates only that the d effect is smaller than twice the c effect, because

$$r_{mz} < 2r_{dz} \iff a^2 + c^2 + d^2 < a^2 + 2c^2 + 0.5d^2 \iff d^2 < 2c^2.$$

Similarly when twice the correlation between the DZ pair exceeds the correlation between the MZ pair, the d effect is larger than twice the c effect. Keller et al. (2005) noted that $r_{mz} > 2r_{dz}$ suggests only that c is less powerful than d , and not that c is non-existent.

Including only three factors (ACE or ADE) often leads to incorrect results. Table 1 shows the biases in estimators (explained variance) for each factor when the ACE or ADE models are fit, even though all four factors actually affect a phenotype (which is realistic in most cases). The table shows that when the ACE model is fit, the explained variance for A is overestimated by $1.5d^2$, C is underestimated by $0.5d^2$, and D is underestimated by d^2 , and when ADE model is fit, A is overestimated by $3c^2$, C is underestimated by c^2 , and D is underestimated by $2c^2$ (as shown in the Appendix; see also Keller and Coventry 2005). In both the cases of the ACE and ADE models, the explained variance for E is estimated without bias. Therefore, both the ACE and ADE models yield biased estimators, except for E. In particular, C is underestimated whether the ACE model or the ADE model is fit. The constant results of small C effects in behavior genetics are to a degree due to the use of the ACE or ADE model (Grayson 1989).

Two numerical examples are given. First, in reality, if the explained percentage of variance for A, C, D, and E of a phenotype are all 0.25, then the MZ correlation is 0.75 ($= 0.25 + 0.25 + 0.25$) and the DZ correlation is 0.4375 ($= 0.25 \times 0.5 + 0.25 + 0.25 \times 0.25$). In this case, the ACE model is fit because $r_{mz} < 2r_{dz}$. Therefore, D is removed. However, in reality, D contributes as much as A, C, and E.

Second, when the explained variance for A, C, D, and E of a phenotype are 0.1, 0.2, 0.6, and 0.1, respectively, the MZ correlation is 0.9 ($= 0.1 + 0.2 + 0.6$), and the DZ correlation is 0.4 ($= 0.1 \times 0.5 + 0.2 + 0.6 \times 0.25$).

In this case, the ADE model is fit because $r_{mz} > 2r_{dz}$. Therefore, C is removed. However, in reality, C has the second largest effect on a phenotype.

Table 1 and these two examples demonstrate that misspecification of the analysis model leads to incorrect results. In the present paper, in order to overcome this problem, the ACDE model is identified using nnSEM. The developed method provides more accurate estimators and is therefore the more accurate method for examining the etiology of variation in human behaviors. Eaves et al. (1978a) and Grayson (1989) noted that there are two primary problems in classical twin design, (which refers to behavior genetics research that uses only twin data and the ACE model or the ADE model is used), namely, (1) the inability to simultaneously estimate both C and D effects and (2) the inability to estimate higher-order epistasis effects. In addition, Keller and Coventry (2005) showed that problem (1) leads to larger bias in estimators than problem (2) (see also Coventry and Keller 2005). Therefore, although it does not perfectly explain variations in human behaviors, the ACDE model covers the primary behavioral factors.

nnSEM

The idea of using higher-order moments in SEM to accommodate non-normal variables was first suggested by Bentler (1983). Moreover, in 1985, a factor analysis model using non-normal variables and higher-order moments to eliminate rotational indeterminacy was developed by Mooijaart (1985). Therefore, the idea of using non-normal variables and higher-order moments is not a recent concept. However, Shimizu and Kano (2008) were the first to develop a simple regression model that uses non-normal variables and higher-order moments in the framework of SEM, where SEM using higher-order moments as well as (1st- and) 2nd-order moments is referred to as nnSEM. Also, Kano and Shimizu (2003) showed that using 2nd, 3rd, and 4th order moments a simple regression model considering confounding a latent variable can be identified.

And in the framework of Independent component analysis (ICA; Comon 1994), there are some researches using non-normally distributed disturbance variables. For example, Shimizu et al. (2006) showed that under the assumptions that the data generating process is linear, no confounding latent variable exists, and disturbance variables have non-normal distributions of non-zero variances, the best fit acyclic model can be identified. They called the model LiNGAM (a linear, non-gaussian, acyclic model). Hoyer et al. (2008) then showed that even if some confounding latent variables exist, parameter estimation of an acyclic model can be performed.

nnSEM has some advantages over SEM. For example, nnSEM can identify models that cannot be identified using SEM, because of a negative degree of freedom. Another advantage is that nnSEM can detect the direction of causation between two cross-sectional variables. Ozaki and Ando (2009) used this advantage to develop a method to detect the direction of causations between C factors and between E factors in bivariate behavior genetics models.

As an illustration of the former advantage, Toyoda (2007) showed that the spurious correlation model and the reciprocal causation model can be identified within the framework of nnSEM. The numbers of parameters of the spurious correlation model and the reciprocal causation model are both four (two path coefficients and two residual variances) when specifying models using only second-order moments. However, the number of unique elements of the covariance matrix between two variables is three. Therefore, neither the spurious correlation model nor the reciprocal causation model can be identified using SEM (number of degrees of freedom: -1). However, as shown in the Appendix of Ozaki and Ando (2009), nnSEM using second- and third-order moments can identify both models.

The two models can be identified using nnSEM because the use of third-order moments increases the number of moments of observed variables and therefore the number of degrees of freedom. Actually, the number of degrees of freedom of the spurious correlation model using nnSEM is 0 and that of the reciprocal causation model is 1. Therefore, these two models can be identified using nnSEM. In the present paper, using this advantage of nnSEM, the univariate ACDE model will be developed.

Model specifications

In addition to SEM using up to second-order moments, third-order moments are used in the present study. Sample third-order moments ($s_{p_1^3}$, $s_{p_1^2 p_2}$, $s_{p_1 p_2^2}$, and $s_{p_2^3}$) are calculated as follows when there are two phenotypes p_1 and p_2 (for twin 1 and twin 2, respectively):

$$s_{p_1^3} = \frac{1}{n} \sum_{i=1}^n (p_{i1} - \bar{p}_1)^3$$

$$s_{p_1^2 p_2} = \frac{1}{n} \sum_{i=1}^n (p_{i1} - \bar{p}_1)^2 (p_{i2} - \bar{p}_2)$$

$$s_{p_1 p_2^2} = \frac{1}{n} \sum_{i=1}^n (p_{i1} - \bar{p}_1) (p_{i2} - \bar{p}_2)^2$$

$$s_{p_2^3} = \frac{1}{n} \sum_{i=1}^n (p_{i2} - \bar{p}_2)^3$$

where i indicates an observation (in this case, an observation pair). These sample statistics become information in the estimation of a , c , d , and e effects.

When three latent variables C, D, and E are assumed to follow non-normal distributions, the expected third-order moments for the MZ twins are as follows:

$$E[p_1^3(mz)] = c^3\sigma_{C^3} + d^3\sigma_{D^3} + e^3\sigma_{E^3} \quad (3)$$

$$E[p_1^2p_2(mz)] = c^3\sigma_{C^3} + d^3\sigma_{D^3} \quad (4)$$

$$E[p_1p_2^2(mz)] = c^3\sigma_{C^3} + d^3\sigma_{D^3} \quad (5)$$

$$E[p_2^3(mz)] = c^3\sigma_{C^3} + d^3\sigma_{D^3} + e^3\sigma_{E^3} \quad (6)$$

In addition, the expected third-order moments for the DZ twins are as follows:

$$E[p_1^3(dz)] = c^3\sigma_{C^3} + d^3\sigma_{D^3} + e^3\sigma_{E^3} \quad (7)$$

$$E[p_1^2p_2(dz)] = c^3\sigma_{C^3} + d^3\sigma_{D_1^2D_2} \quad (8)$$

$$E[p_1p_2^2(dz)] = c^3\sigma_{C^3} + d^3\sigma_{D_1D_2^2} \quad (9)$$

$$E[p_2^3(dz)] = c^3\sigma_{C^3} + d^3\sigma_{D^3} + e^3\sigma_{E^3} \quad (10)$$

Note that A is assumed to follow a normal distribution when many loci affect the phenotype of interest (Fisher 1918). Here, σ_{C^3} , σ_{D^3} , and σ_{E^3} are the skewnesses of C, D, and E, respectively. (In this case, the variances of these independent factors are fixed to 1s. Therefore, these parameters are the skewnesses of the factors.) $\sigma_{D_1^2D_2}$ expresses the expected value of the second power of D of twin 1 times D of twin 2, and $\sigma_{D_1D_2^2}$ expresses the expected value of D of twin 1 times the second power of D of twin 2, respectively. $\sigma_{D_1^2D_2}$ and $\sigma_{D_1D_2^2}$ are assumed to be equal, because the order of twins is arbitrary.

The above equations are calculated as follows. For example,

$$\begin{aligned} E[p_1^2p_2(dz)] &= E[(aA_1 + cC + dD_1 + eE_1)^2 \\ &\quad (aA_2 + cC + dD_2 + eE_2)] \\ &= a^3 \times 0 + c^3\sigma_{C^3} + d^3\sigma_{D_1^2D_2}. \end{aligned}$$

Assumptions

However, the above model has not yet been identified. There are eight parameters (a , c , d , e , σ_{C^3} , σ_{D^3} , σ_{E^3} , and $\sigma_{D_1^2D_2} = \sigma_{D_1D_2^2}$) in the model. However, the number of unique equations is only six (three for second-order moments and three for third-order moments), resulting in the number of degrees of freedom being negative. Therefore, constraints are needed in order to identify the model. The following are models that can be identified.

1. C and E are non-normal, and D is normal. (Therefore, $\sigma_{D^3} = 0$, and $\sigma_{D_1^2D_2} = 0$.) Furthermore, the skewnesses

of C and E are the same (σ_3 .) In this case, there are five parameters (and a , c , d , e , and σ_3), and the number of unique equations is five ($E[p_1^2(mz)]$ (Eq. 1) = $E[p_2^2(mz)]$ (Eq. 1) = $E[p_1^2(dz)]$ (Eq. 2) = $E[p_2^2(dz)]$ (Eq. 2), $E[p_1p_2(mz)]$ (Eq. 1), $E[p_1p_2(dz)]$ (Eq. 2), $E[p_1^3(mz)]$ (Eq. 3) = $E[p_2^3(mz)]$ (Eq. 6) = $E[p_1^3(dz)]$ (Eq. 7) = $E[p_2^3(dz)]$ (Eq. 10), and $E[p_1^2p_2(mz)]$ (Eq. 4) = $E[p_1p_2^2(mz)]$ (Eq. 5) = $E[p_1^2p_2(dz)]$ (Eq. 8) = $E[p_1p_2^2(dz)]$ (Eq. 9)), resulting in the number of degrees of freedom being 0. However, since there are 14 (six for second-order moments and eight for third-order moments) sample statistics and five parameters, the number of degrees of freedom used to calculate model fit indices is nine (=14 – 5). If the phenotypes are standardized, since the four variances are all 1s, the number of degrees of freedom is six (=11 – 5).

2. D and E are non-normal, and C is normal. (Therefore, $\sigma_{C^3} = 0$.) Furthermore, the skewnesses of D and E are the same (σ_3) and $\sigma_{D_1^2D_2} = \sigma_{D_1D_2^2}$. In this case, there are six parameters (a , c , d , e , σ_3 , and $\sigma_{D_1^2D_2}$), and the number of unique equations is six, resulting in the number of degrees of freedom being 0. The number of degrees of freedom used to calculate model fit indices is eight (=14 – 6). If the phenotypes are standardized, then the number of degrees of freedom is five (=11 – 6).
3. D, C, and E are non-normal. Furthermore, the skewnesses of D, C, and E are the same (σ_3). In this case, there are six parameters (a , c , d , e , σ_3 , and $\sigma_{D_1^2D_2}$), and the number of unique equations is six, resulting in the number of degrees of freedom being 0. The number of degrees of freedom used to calculate model fit indices is eight (=14 – 6). If the phenotypes are standardized, then the number of degrees of freedom is five (=11 – 6).

Note that if all four latent factors are distributed normally, nnSEM cannot estimate these effects. The non-normal distributional assumption on D is likely. For example, if genotypes AA and Aa have the same effect on a trait and the gene frequencies of A and a are both 0.5, then D has a skewed distribution because the genotypic frequency of aa is 0.25 and the genotypic frequency of AA + Aa is 0.75. When researchers use the present method, they should compare the fits of the three identified models, and the results of the best fit model should be interpreted. An example R script can be downloaded from <http://www.010.upp.so-net.ne.jp/koken/bg.html>.

Note also that the developed ACDE model can be used only with continuous data and is not suitable for binary or ordinal data because, in order to analyze the latter two types of data, the liability model that assumes latent variables that follow some normal distributions is usually used in Behavior Genetics. However, if all four latent variables (A, C, D, and E) are normally distributed, then nnSEM

cannot be used, because it is impossible to calculate the 3rd order moments of the assumed normally distributed latent variables.

Estimator

The asymptotically distribution-free (ADF) method (Browne 1982, 1984) is used to estimate parameters. In the two-group case (MZ and DZ), the ADF estimator can be expressed as follows:

$$F = n_m(\sigma_m(\theta) - s_m)'W_m^{-1}(\sigma_m(\theta) - s_m) + n_d(\sigma_d(\theta) - s_d)'W_d^{-1}(\sigma_d(\theta) - s_d) \tag{11}$$

Here, F is the function to be minimized, n is the sample size, $\sigma(\theta)$ is the expected second- and third-order moments vector, s is the observed second- and third-order moments vector, W is the weight matrix, and subscripts m and d denote MZ and DZ, respectively. Each element of W consists of n times the asymptotic covariance of estimators of moments of the observed phenotypes.

Simulation study

Settings

This simulation study examines the effects of two factors on the average biases and standard deviations of estimates of explained variance for A, C, D, and E. The two factors are (1) the true explained variance for A, C, D, and E (21 patterns, see Table 2), and (2) the skewness of C, D, and E (seven patterns, see Table 3). For all of the 147 (= 21 × 7) conditions, 300 replications were performed. Second- and third-order moments were used in the analyses. In each replication, the BICs of the three identified models are calculated, and the estimates of the best fit model were considered as the estimates in the replication. For the same generated data, when $r_{mz} < 2r_{dz}$, the ACE model was fit, otherwise the ADE model was fit. This enables comparisons between the results of the current method (SEM) and the developed method (nnSEM). The average biases and standard deviations of the estimates were calculated for each condition. In all cases, the numbers of MZs and DZs were both 600. The ADF estimator was used for both nnSEM and SEM analyses. For nnSEM, second- and third-order moments were used, on the other hand, for SEM only second-order moments were used. The analysis program was written using the R programming language (Ihaka and Gentleman 1996; version 2.4.0). We wrote the estimator, which was not a pre-existing R package. However, the ‘BFGS’ method, which is a quasi-Newton method, in the ‘optim’ function was used for optimization calculation.

Table 2 True explained variances for A, C, D, and E

VC (nnSEM is true)				VC (nnSEM and SEM are true)			
25	25	25	25	100/3	100/3	0	100/3
45	35	10	10	60	20	0	20
45	10	35	10	20	60	0	20
45	10	10	35	20	20	0	60
35	45	10	10	100/3	0	100/3	100/3
10	45	35	10	60	0	20	20
10	45	10	35	20	0	60	20
35	10	45	10	20	0	20	60
10	35	45	10				
10	10	45	35				
35	10	10	45				
10	35	10	45				
10	10	35	45				

VC (nnSEM is true) denotes the conditions in which all four factors affect a trait. Therefore, nnSEM will be the appropriate method. On the other hand, VC (nnSEM and SEM are true) denotes the conditions in which ACE or ADE are the true models. Therefore, both nnSEM and SEM will be appropriate methods

Table 3 Skewness of C–E

C	D	E	T/ F
0	0	1	F
0	1	0	F
1	0	0	F
0	1	1	T
1	0	1	T
1	1	0	F
1	1	1	T

When the skewness is 0, the factor score was generated from the standard normal distribution. On the other hand, when the skewness is 1, the factor score was generated from the $\chi^2(8)$ distribution. When one of the three models identified is the true model (patterns 4, 5, and 7), T/F is T and when none of the three models identified is the true model (patterns 1, 2, 3, 6, and 7), T/F is F

Twenty one patterns of the true explained variance were divided into two cases. In one case, all four factors affect a trait. Therefore, nnSEM is the true model. In the other case, three of the four factors (ACE or ADE) affect a trait. Therefore, both nnSEM and SEM are the true model. Figures 2 and 3 show the results for the former cases, and Figs. 4 and 5 show the results for the latter cases.

The seven patterns of skewness are shown in Table 3. These seven patterns can be divided into two cases, in which either (1) one of the three models identified is the true model (patterns 4, 5, and 7; denoted as T in Figs. 2, 3, 4, 5) or (2) none of the three models identified is the true model (patterns 1, 2, 3, and 6; denoted as F in Figs. 2, 3, 4, 5). In order to simplify the results, rather than seven patterns, the

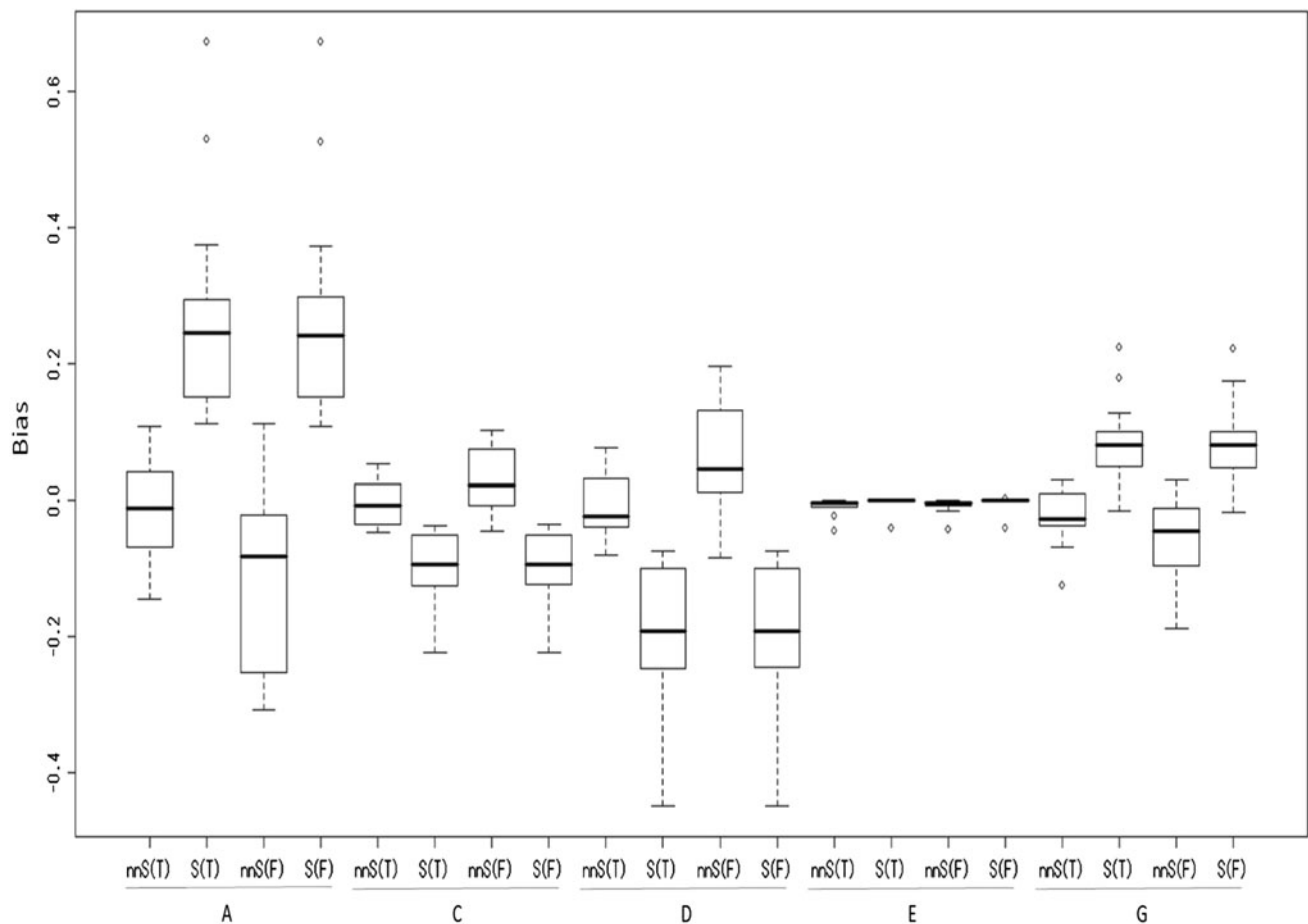


Fig. 2 Biases (estimated value–true value) when all four factors affect a trait

average of the results of the three T patterns and the average of the results of the four F patterns are presented in Figs. 2, 3, 4, 5. In all cases, the factor scores for A were generated from the standard normal distribution.

Results and discussion

Figures 2, 3, 4, 5 show the box plots for A, C, D, E, and G. Here, G indicates the results of A + D, which can be considered as the broad-sense heritability (Keller and Coventry 2005). In each figure, nnS indicates the case in which nnSEM was used for the analyses, and S indicates the case in which SEM was used. In addition, T and F in parentheses indicate whether one of the three models identified is the true model (T) or not (F). A, C, D, E, and G at the bottom of each figure indicate that the plots are the results for the corresponding factors. Figures 2 and 4 show the results for the average biases (estimated value–true value), and Figs. 3 and 5 show the results for the standard deviation of the estimates. The box plots are shown for the

averages of the three T cases or the four F cases. For example, the box plot for A of the nnS(T) case in Fig. 2 was shown for the following values: 0.041, –0.060, –0.089, –0.145, –0.012, 0.031, 0.107, –0.080, 0.055, –0.016, –0.069, 0.098, and –0.004. The first value, 0.041, is the average of the 300 biases for A when using nnS(T) and the variance components for the four factors are the same (first case in Table 2). The results will be discussed based on (1) an examination of the estimation accuracy of the proposed method and (2) comparison of the estimation accuracies of the conventional method (SEM) and the proposed method (nnSEM).

Examination of estimation accuracy of the proposed method

The simulation study revealed the following characteristics of the proposed method. Note that the following interpretations are for T + F cases (T and F are pooled), except for point 11.

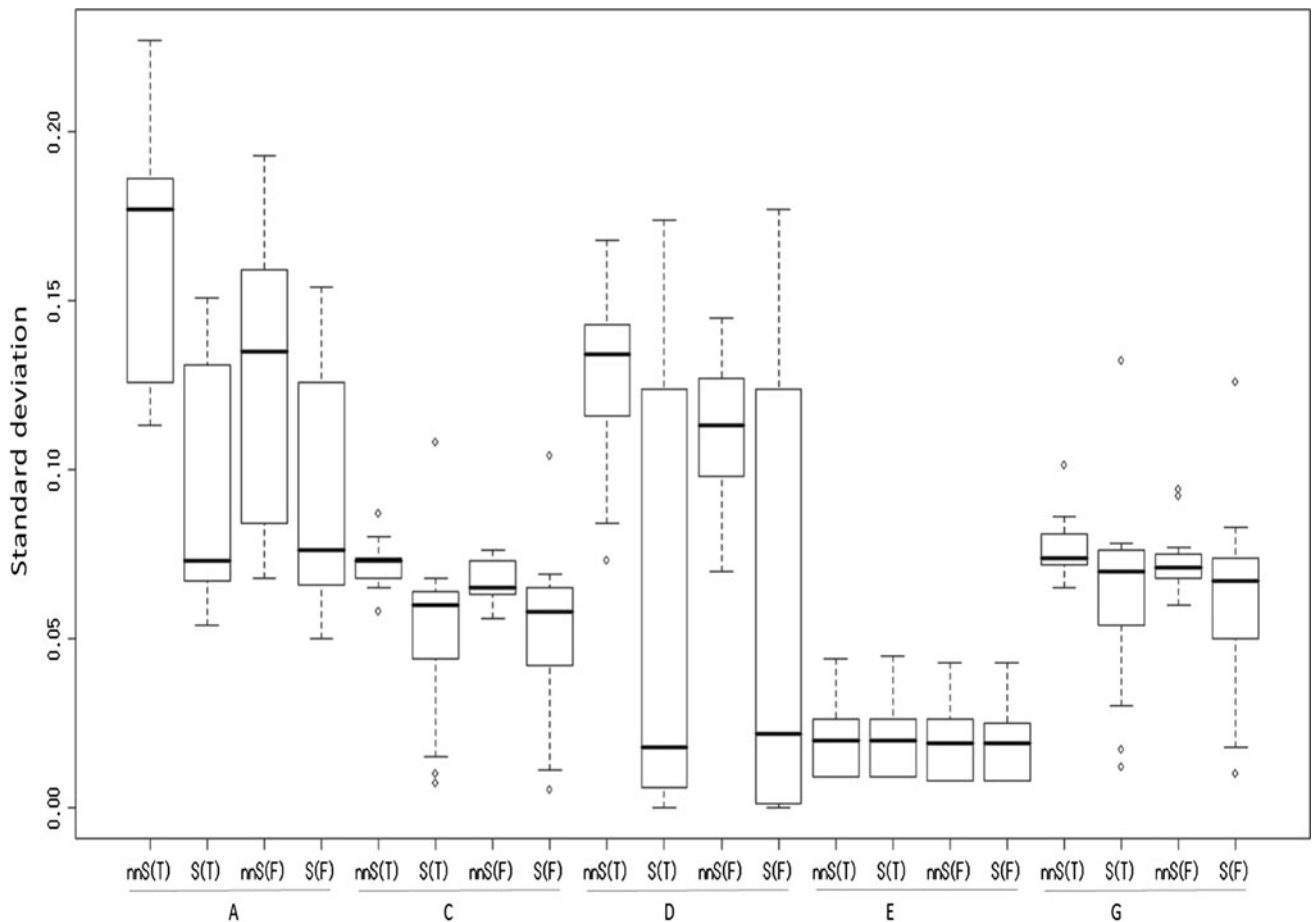


Fig. 3 Standard deviations of estimates when all four factors affect a trait

1. Figure 2 shows that when all four factors affect a trait, the explained variance of A tends to be slightly underestimated (-7.9% on average), and Fig. 3 shows that the average standard deviation of the estimates of the variance of A is 14.8% and occasionally exceeds 20% , which is a large value. Consequently, the variance of A is not accurately estimated under these conditions.
2. Figure 3 shows that the average standard deviation is larger in the T case (16.8%) than in the F case (12.8%). However, this does not mean that the F case yields better estimates than the T case, because the average bias is much smaller in the T case (-1.1%) than in the F case (-11.4%).
3. Figure 4 shows that when three of the four factors affect a trait, the explained variance of A tends to be somewhat underestimated (-19.5% on average), and Fig. 5 shows that the average standard deviation of the estimates of the variance of A is 10.8% , which is not a small value. Consequently, the variance of A is not accurately estimated under these conditions.
4. Figure 2 shows that when all of the four factors affect a trait, the biases of the explained variance of C are $<5\%$ in most cases and are at most 10.1% . Furthermore, Fig. 3 shows that the standard deviations of the estimates are $<9\%$, which indicates that the variance of C is accurately estimated under these conditions.
5. Figure 4 shows that when three of the four factors affect a trait, the explained variance of C tends to be slightly overestimated (6.1% on average), and Fig. 5 shows that the average standard deviation of the estimates of the variance of C is 5.7% . Consequently, the variance of C is accurately estimated under these conditions.
6. Figure 2 shows that when all four factors affect a trait, the biases of the explained variance of D are $<10\%$ in most cases, and the explained variance of D tends to be overestimated. Figure 3 shows that standard deviation of the estimates is not larger than in the case of A. Consequently, the variance of D is not accurately estimated, but is estimated more accurately than the variance of A.
7. Figure 4 shows that when three of the four factors affect a trait, the biases of the explained variance of

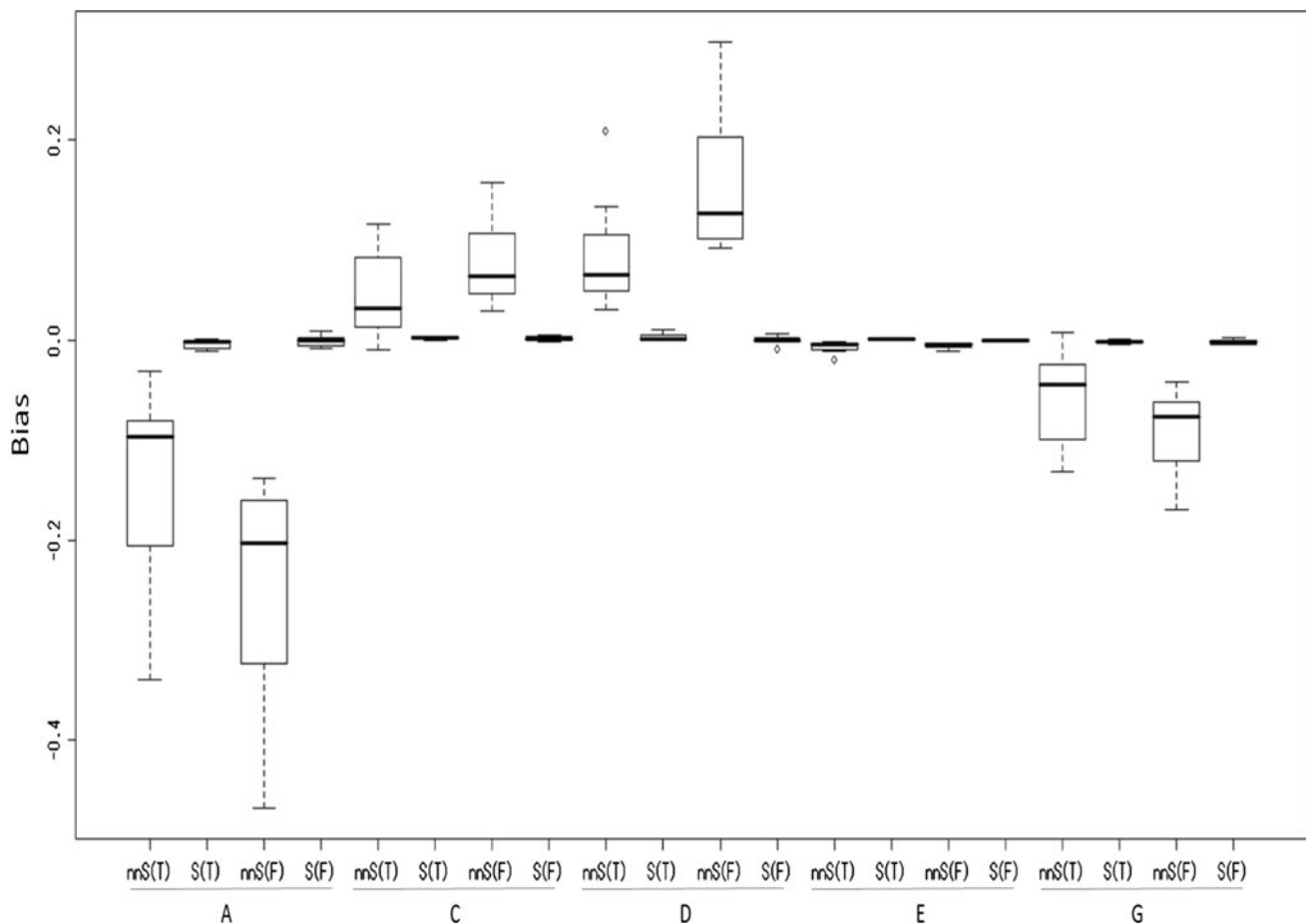


Fig. 4 Biases (estimated value–true value) when three factors affect a trait

D are overestimated (12.0% on average). Figure 5 shows that the standard deviation of the estimates is not larger than in the case of A. Consequently, the variance of D is not accurately estimated, but is estimated more accurately than the variance of A. These results are the same as in the case in which all four factors affect a trait.

8. Figures 2, 3, 4, 5 show that whether three or four factors affect a trait, the biases of the explained variance of E are less than 5% in all cases. Furthermore, the standard deviations of the estimates are <5%, which indicates that the variance of E is very accurately estimated.
9. Figure 2 shows that when all four factors affect a trait, the biases of the average explained variance of G is –4%, and Fig. 3 shows that the standard deviations of the estimates are <10%.
10. Figure 4 shows that when all four factors affect a trait, the biases of the average explained variance of G is –7.5%, and Fig. 5 shows that the standard deviations of the estimates are <9%.

11. Figures 2, 3, 4, 5 show that whether three or four factors affect a trait, even if data is generated from a model that is not one of the three identified models (F cases), the results of biases and standard deviations of the estimates are not much larger than in the T cases, except for the biases for the A factor. This indicates that the proposed method can provide a similar level of accuracy for these two cases.
12. Figures 2, 3, 4, 5 show that if $G = A + D$ is considered to be global effects of genes, then the explained variances of G, C, and E can be estimated accurately.

Comparison of the results obtained using SEM and nnSEM

The simulation study revealed the following characteristics. Note that the following interpretations are for T + F cases (T and F are pooled).

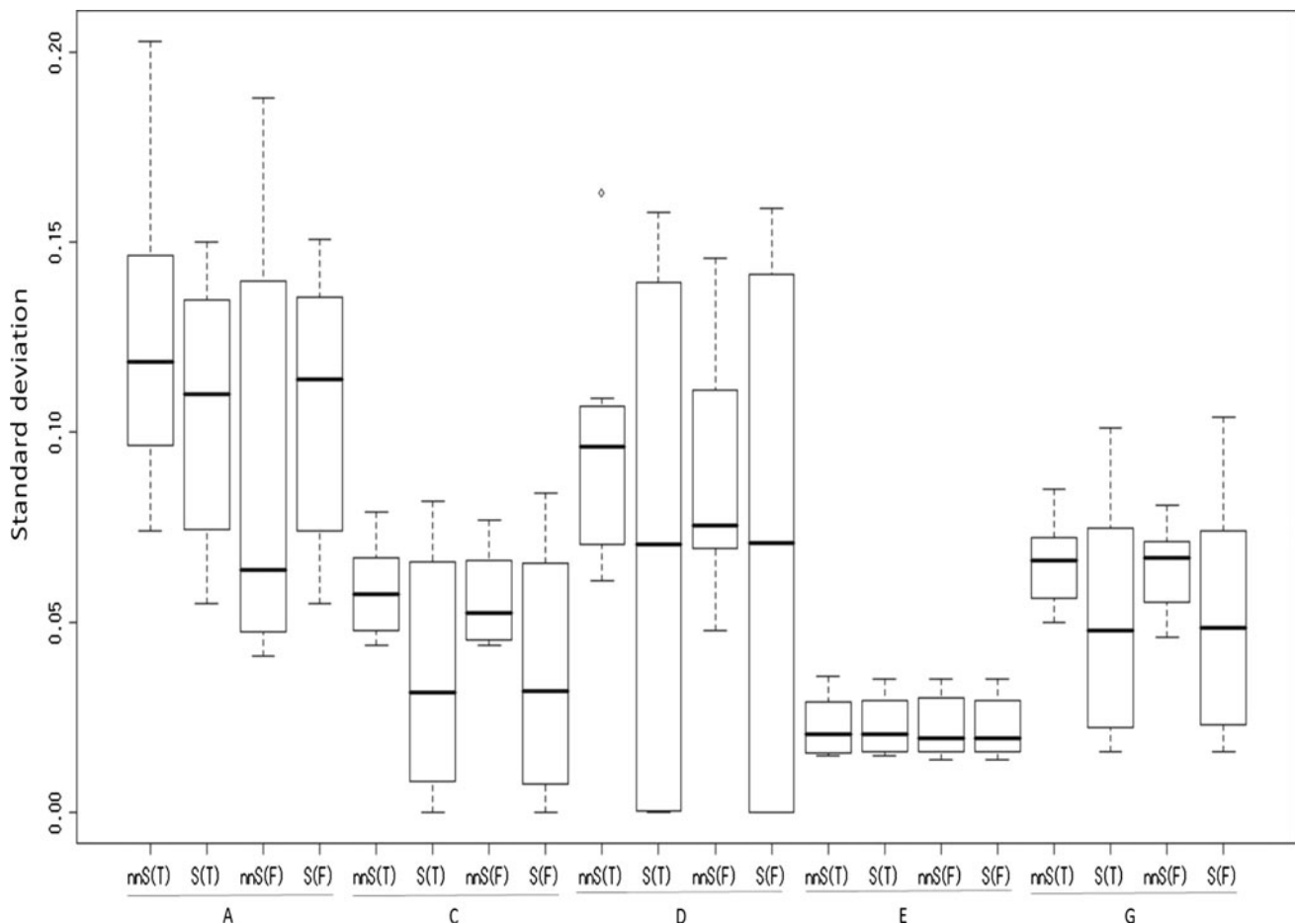


Fig. 5 Standard deviations of estimates when three factors affect a trait

1. When all four factors affect a trait, for A, C, D, and G, there is a high probability of providing a smaller bias by nnSEM than by SEM under these conditions ($(22/26) \times 100 = 85\%$ for A, $(23/26) \times 100 = 89\%$ for C, $(24/26) \times 100 = 92\%$ for D, and $(21/26) \times 100 = 81\%$ for G).
2. In contrast, when three of the four factors affect a trait, for A, C, D, and G, SEM provided smaller biases than nnSEM in every case.
3. For E, both SEM and nnSEM provide very small bias whether three or four factors affect a trait.
4. The explained variance of C is always underestimated using SEM when all four factors affect a trait. However, this does not occur when using nnSEM.
5. SEM provides a smaller standard deviation of the estimates. Therefore, although nnSEM provides a smaller bias when all four factors affect a trait, a greater number of twins are needed in order to obtain the same level of standard deviation of the estimates using SEM.

Consequently, nnSEM was shown to provide better estimators than SEM in the cases in which four latent

factors affect a trait. In particular, when G is considered as broad-sense heritability, G, C, and E can be estimated with biases of $<5\%$ in most cases for this simulation setting. When three latent factors affect a trait, SEM provides better estimators than nnSEM. However, even in these cases, C can be estimated with biases of $<6.1\%$ on average, E can be estimated with almost no bias, and G can be estimated with biases of $<8\%$ in most cases for this simulation setting using nnSEM.

Discussion

The ACE and ADE models yield biased estimates if the ACDE model is the true model. However, structural equation modeling cannot identify the ACDE model because of a negative degree of freedom. In the present paper, the ACDE model was developed using nnSEM. Simulation studies have suggested that the proposed method can decrease the biases and can be used even in cases in which three latent factors affect a trait if G, C, and

E are used as outcomes. The effects of the MZ/DZ ratio on the results were not examined in the simulation study described above. However, a small simulation study was conducted, in which the cases of MZ:DZ = 800:400 and 400:800 were considered. The results showed that the MZ/DZ ratio has almost no effect on the estimates in this simulation setting.

In the present paper, although power calculation was not discussed, the power would be calculated by setting the true second- and third-order moments and analyzing the moments using a null model. This is the same method used to calculate the power in the SEM using Mx (Neale et al. 2006, p. 96).

There is another method of identifying the ACDE model. Kathryn et al. (2004) identified the ACDE model using sibling data, where one of the siblings is an adopted person, as well as twin data. Such sibling data also provides information of the etiology of human variation, because the siblings (one of whom is an adopted person) are thought to share the same environment, however, there is no genetic relationship between the siblings. Therefore, their similarity is thought to be caused by their shared environment. Unfortunately, gathering data for siblings for the case in which one sibling is adopted is very difficult and costly. In the present paper, however, nnSEM was used to identify the ACDE model in which higher-order moments are used as additional information. Since the method of the present paper can estimate a , c , d , and e effects using only twin data, the proposed method requires a much lower cost and is much easier to perform. Therefore, the proposed method contributes to behavior genetic research.

There are other factors that have possible effects on phenotypes, such as higher-order epistasis (Eaves et al. 1978a), assortative mating (people choose to marry persons similar to themselves; Eaves et al. 1984; Loelin and DeFries 1987), gene–environment correlation (Plomin et al. 1977; Eaves et al. 1977), and gene–environment interaction (Plomin et al. 1977; Eaves et al. 1977; Purcell 2002; Eaves and Erkanli 2003). The presence of higher-order epistasis makes the correlation between D factors of DZ twins lower than 0.25. The presence of assortative mating leads to overestimation of the effect of C and underestimation of the effect of A if these are not included in models. An example of gene–environment correlation is that genetically intellectual parents tend to provide an intellectual environment for their children, which in turn increases the intelligence of the children. The presence of positive gene–environment correlation also leads to overestimation of the effect of C and underestimation of the effect of A if these are not included in models. An extended twin-family design (ETFD, Truett et al. 1994) can estimate the effects of assortative mating and gene–environment correlation, but cannot estimate the higher-order epistatic

effect or gene–environment interaction. Although Purcell (2002) proposed a relatively easy method of estimating gene–environment interaction, at present, there is no method that incorporates the higher-order epistatic effect in the model. Therefore, further research on developing a more exhaustive model is needed.

Higher-order epistasis is expressed as the correlation between D factors of DZ twins, and assortative mating is expressed as the correlation between A factors of DZ twins and D factors of DZ twins. gene–environment correlation is expressed as the correlation between A and C factors and between D and C factors, and gene–environment interaction is the interaction between one of the genetic factors and one of the environmental factors. Therefore, these four effects can be expressed as parameters associated with A, C, D, or E. Consequently, proper estimation of a , c , d , and e is an important basis for including these phenomena (such as assortative mating). Although, in the present paper, only the ACDE model was proposed, this is an important first step in the development of a more exhaustive model.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Appendix

When the ACE model is fit, the expected variance of a phenotype is $a^2 + c^2 + e^2$, the expected correlation between MZs (r_{mz}) is $a^2 + c^2$, and the correlation between DZs (r_{dz}) is $0.5a^2 + c^2$. Therefore, the simplest method of estimating a^2 , c^2 , and e^2 when the ACE model is fit is as follows (when the phenotypes are standardized):

$$\hat{a}^2 = 2(r_{mz} - r_{dz})$$

$$\hat{c}^2 = 2r_{dz} - r_{mz}$$

$$\hat{e}^2 = 1 - r_{mz}$$

When the ADE model is fit, the expected variance of a phenotype is $a^2 + c^2 + d^2$, the expected correlation between MZs (r_{mz}) is $a^2 + d^2$, and the correlation between DZs (r_{dz}) is $0.5a^2 + 0.25d^2$. Therefore, the simplest method of estimating a^2 , d^2 , and e^2 when the ADE model is fit is as follows:

$$\hat{a}^2 = 4r_{dz} - r_{mz}$$

$$\hat{d}^2 = 2(r_{mz} - 2r_{dz})$$

$$\hat{e}^2 = 1 - r_{mz}$$

If the ACDE model is the true model and the true explained variances are a_T^2 , c_T^2 , d_T^2 , and e_T^2 , respectively,

then the expected correlation between MZs (r_{mz}) is $a_T^2 + c_T^2 + d_T^2$ and the expected correlation between DZs (r_{dz}) is $0.5a_T^2 + c_T^2 + 0.25d_T^2$. Therefore, when the ACE model is fit, the estimators of the explained variances are as follows:

$$\hat{a}^2 = 2(r_{mz} - r_{dz}) = a_T^2 + 1.5d_T^2$$

$$\hat{c}^2 = 2r_{dz} - r_{mz} = c_T^2 - 0.5d_T^2$$

$$\hat{e}^2 = 1 - r_{mz} = e_T^2$$

Therefore, \hat{a}^2 and \hat{d}^2 are biased ($1.5d_T^2$ and $-0.5d_T^2$).

When the ADE model is fit, the estimators of the explained variance are as follows:

$$\hat{a}^2 = 4r_{dz} - r_{mz} = a_T^2 + 3c_T^2$$

$$\hat{d}^2 = 2(r_{mz} - 2r_{dz}) = d_T^2 - 2c_T^2$$

$$\hat{e}^2 = 1 - r_{mz} = e_T^2$$

Therefore, \hat{a}^2 and \hat{d}^2 are biased ($3c_T^2$ and $-2c_T^2$).

References

- Bentler PM (1983) Some contributions to efficient statistics in structural models: specification and estimation of moment structures. *Psychometrika* 48:493–517
- Bollen KA (1989) *Structural equations with latent variables*. Wiley, New York
- Browne MW (1982) Covariance structures. In: Hawkins DM (ed) *Topics in applied multivariate analysis*. Cambridge University Press, Cambridge, pp. 72–141
- Browne MW (1984) Asymptotically distribution-free methods for the analysis of covariance structures. *Br J Math Stat Psychol* 9:665–672
- Comon P (1994) Independent component analysis—a new concept? *Signal Process* 36:287–314
- 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* 8:214–223
- Eaves LJ, Erkanli A (2003) Markov chain Monte Carlo approaches to analysis of genetic and environment components of human developmental change and $G \times E$ interaction. *Behav Genet* 33:279–299
- Eaves LJ, Last KA, Martin NG, Jinks JL (1977) A progressive approach to non-additivity and genotype–environment covariance in the analysis of human differences. *Br J Math Stat Psychol* 30:1–42
- Eaves LJ, Last KA, Young PA, Martin NG (1978a) Model-fitting approaches to the analysis of human behavior. *Heredity* 41: 249–320
- Eaves LJ, Martin NG, Eysenck SBG (1978b) An application of the analysis of covariance structures to the psychogenetical study of impulsiveness. *Br J Math Stat Psychol* 30:185–197
- Eaves LJ, Heath AC, Martin NG (1984) A note on the generalized effects of assortative mating. *Behav Genet* 14:371–376
- Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* 52:399–433
- Grayson DA (1989) Twins reared together: minimizing shared environmental effects. *Behav Genet* 19:593–604
- Hoyer PO, Shimizu S, Kerminen A, Palviainen M (2008) Estimation of causal effects using linear non-gaussian causal models with hidden variables. *Int J Approx Reason* 49:362–378
- Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. *J Comput Graph Stat* 5:299–314. <http://www.R-project.org>
- Kano Y, Shimizu S (2003) Causal inference using nonnormality. In: *Proceedings of international symposium on science of modeling—the 30th Anniversary of the Information Criterion (AIC)*, pp 261–270
- Kathryn AB-B, Kirby D-D, Thalia E, Jennifer JF, Jim S, Plomin R (2004) A genetic analysis of individual differences in dissociative behaviors in childhood. *J Child Psychol Psychiatry* 45:522–532
- Keller MC, Coventry WL (2005) Quantifying and addressing parameter indeterminacy in the classical twin design. *Twin Res Hum Genet* 8:201–213
- Keller MC, Coventry WL, Heath AC, Martin NG (2005) Widespread evidence for non-additive genetic variation in Cloninger's and Eysenck's personality dimensions using a twin plus sibling design. *Behav Genet* 35:707–721
- Loelin JC, DeFries JC (1987) Genotype–environment correlation and IQ. *Behav Genet* 7:263–277
- Mooijaart A (1985) Factor analysis for non-normal variables. *Psychometrika* 50:323–342
- Neale MC, Cardon LR (1992) *Methodology for genetic studies of twins and families*. Kluwer, Dordrecht
- Neale MC, Boker SM, Xie G, Maes HH (2006) *Mx Statistical modeling*, 7th edn. Virginia Commonwealth University, Department of Psychiatry, Richmond, VA
- Ozaki K, Ando J (2009) Direction of causation between shared and non-shared environmental factors. *Behav Genet* 39:321–336
- Plomin R, DeFries JC, Loehlin JC (1977) Genotype–environment interaction and correlation in the analysis of human behavior. *Psychol Bull* 84:310–322
- Purcell S (2002) Variance components models for gene–environment interaction in twin analysis. *Twin Res* 5:554–571
- Shimizu S, Kano Y (2008) Use of non-normality in structural equation modeling: application to direction of causation. *J Stat Plan Inference* 138:3483–3491
- Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A (2006) A linear non-gaussian acyclic model for causal discovery. *J Mach Learn Res* 7:2003–2030
- Toyoda H (2007) *Theory of structural equation modeling*. Asakura-shoten, Tokyo
- Truett KR, Eaves LJ, Walters EE, Heath AC, Hewitt JK, Meyer JM, Siberg J, Neale MC, Martin NG, Kendler KS (1994) A model system for analysis of family resemblance in extended kinships of twins. *Behav Genet* 24:35–49