European Journal of Human Genetics (2015) 23, 388–394

www.nature.com/ejhg

ARTICLE

Sandwich corrected standard errors in family-based genome-wide association studies

Camelia C Minică*,[1], Conor V Dolan[1], Maarten MD Kampert[2], Dorret I Boomsma[1] and Jacqueline M Vink[1]

Given the availability of genotype and phenotype data collected in family members, the question arises which estimator ensures the most optimal use of such data in genome-wide scans. Using simulations, we compared the Unweighted Least Squares (ULS) and Maximum Likelihood (ML) procedures. The former is implemented in Plink and uses a sandwich correction to correct the standard errors for model misspecification of ignoring the clustering. The latter is implemented by fast linear mixed procedures and models explicitly the familial resemblance. However, as it commits to a background model limited to additive genetic and unshared environmental effects, it employs a misspecified model for traits with a shared environmental component. We considered the performance of the two procedures in terms of type I and type II error rates, with correct and incorrect model specification in ML. For traits characterized by moderate to large familial resemblance, using an ML procedure with a correctly specified model for the conditional familial covariance matrix should be the strategy of choice. The potential loss in power encountered by the sandwich corrected ULS procedure does not outweigh its computational convenience. Furthermore, the ML procedure was quite robust under model misspecification in the simulated settings and appreciably more powerful than the sandwich corrected ULS procedure. However, to correct for the effects of model misspecification in ML in circumstances other than those considered here, we propose to use a sandwich correction. We show that the sandwich correction can be formulated in terms of the fast ML method.
European Journal of Human Genetics (2015) **23,** 388–394; doi:10.1038/ejhg.2014.94; published online 11 June 2014

INTRODUCTION

Given the availability of large datasets of genotyped and phenotyped family members, it is of interest to determine which statistical test is most efficient in genome-wide association studies (GWAS), where computational efficiency and statistical power are important. One option is to use Plink,[1] which employs the standard Unweighted Least Squares (ULS) estimator in combination with the ULS sandwich[2,3] to correct the standard errors for the model misspecification of ignoring the clustering. This approach is non-iterative, and produces unbiased estimates and correct standard errors, without the need to specify a background covariance model. However, given clustered data, ULS is not necessarily the most powerful estimator.[4] Maximum Likelihood (ML) is an important alternative, but is computationally more demanding. Fast algorithms have been developed, but these employ a model for the background covariance, which is limited to additive genetic and unshared environmental effects.[5,6] We note that shared environmental effects are often found in lifestyle and psychiatric phenotypes, such as substance use.[7–10] This raises a practical question: in conducting a family-based analysis, should one use the sandwich corrected ULS, which is fast, robust and requires no model to be specified for the background covariance matrix, or should one use ML, which is efficient and fast, provided one commits to a background model limited to additive genetic and unshared environmental effects? In the latter case, one may ask whether discarding shared environmental effects, affects the results of the ML procedure.[11]

The present aim is to compare the ULS procedure with the ML procedure using simulated data. We consider the performance in terms of type I and type II error rates, with correct and incorrect background specification in ML. To correct for the effects of this misspecification, we propose to use a sandwich correction (as in Plink[1]). We show that the sandwich correction can be formulated in terms of the fast ML method of Lippert *et al.*[5]

MATERIALS AND METHODS

Family-based model for genetic association

Let $\mathbf{y}_{ij}$ be the vector of observed phenotypes, where subscript j stands for individual ($j=1\ldots n_i$) and subscript i stands for family ($i=1\ldots N$). Let $\mathbf{g}_{ij}$ be the vector of observed genetic markers coded as an additive genetic model, as 0 (aa), 1 (Aa) or 2 (AA).[12] We test the statistical association between each observed genetic marker and the phenotype in an appropriate regression model:

$$\boldsymbol{y}_{ij} = \boldsymbol{b}_0 + \boldsymbol{b}_1 * \boldsymbol{g}_{ij} + \boldsymbol{\varepsilon}_{ij} \quad (1)$$

where $\mathbf{b}_0$ represents the intercept, $\mathbf{b}_1$ is the regression coefficient and $\boldsymbol{\varepsilon}_{ij}$ is the residual term. Let k equal $\sum_i^N n_i$, $\mathbf{b}^t$ equal the vector $[\mathbf{b}_0\ \mathbf{b}_1]$ and $\mathbf{X}$ equal the $k \times 2$ matrix with the first column the unit vector, and the second, the k vector $\mathbf{g}$ containing the genetic information. Other covariates may be included, if desired (for example, age, sex). The k vector of residuals $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{Xb}$ is normally distributed with $k \times k$ background covariance matrix $\mathbf{V}$ (positive definite), that is, $\boldsymbol{\varepsilon}|\mathbf{g} \sim N(\mathbf{0}, \mathbf{V})$. We assume that $\mathbf{V}$ is block diagonal (but see Lippert *et al*,[5] Pirinen *et al*[6] and Visscher *et al*[13]), with diagonal blocks, $\mathbf{V}_i$, representing the residual positive definite covariance matrix of each family. An advantage of retaining the full matrix $\mathbf{V}$ (and not reformulating the likelihood given the sparseness) is that the block diagonal structure can be relaxed to accommodate distant genetic relatedness.[5,6,14] This makes the linear mixed approach very flexible. We assume that the elements in the diagonal blocks in $\mathbf{V}$ parameter vector $\boldsymbol{\theta}$ contains the estimated elements of the conditional covariance matrix. Given MZ and DZ

[1]Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; [2]Mathematical Institute, Leiden University, Leiden, The Netherlands
*Correspondence: CC Minică, Department of Biological Psychology, Vrije Universiteit Amsterdam, Van der Boechorststraat 1, Room 2B03, 1081 BT Amsterdam, The Netherlands. Tel: +31 20 59 83035; Fax: +31 20 59 88832; E-mail: c.c.minica@vu.nl
Received 2 December 2013; revised 27 March 2014; accepted 30 April 2014; published online 11 June 2014

families, the covariance matrix $\mathbf{V}_i$ may be calculated conditional on zygosity, but otherwise unstructured and homoskedastic. We denote this the unstructured estimate of $\mathbf{V}(\boldsymbol{\theta})$. Alternatively, $\mathbf{V}$ may be parameterized, that is, $\mathbf{V}(\boldsymbol{\theta})$, where the parameter vector $\boldsymbol{\theta}$ may contain shared (C) and unshared (E) environmental variance components (σ^2_C, σ^2_E), and additive (A) and dominance (D) variance components (σ^2_A, σ^2_D).[15,16] In this case, MZ and DZ relatedness is expressed in terms of these genetic variance components.

Estimation

We compare tests of $\mathbf{b}_1$ based on ML estimation and ULS estimation, with regular and sandwich corrected standard errors. The log-likelihood function is:

$$LogL(\boldsymbol{\theta}, \boldsymbol{b}) = log\left[(2\pi)^{-1/k}|\mathbf{V}(\boldsymbol{\theta})|^{-1/2}exp\{-1/2(\boldsymbol{y}-\boldsymbol{Xb})^t V(\boldsymbol{\theta})^{-1}(\boldsymbol{y}-\boldsymbol{Xb})\}\right] \quad (2)$$

where $\mathbf{b}$ represents the fixed effects, and $\boldsymbol{\theta}$ the random effects.[17] Maximization of the log-likelihood function subject to the correct specification of the background structure, yields the ML estimate of $\mathbf{b}$, $\hat{\mathbf{b}}_{ML}$, which can be tested by means of the Wald test.[4,18] The parameterization of $\mathbf{V}(\boldsymbol{\theta})$ in the linear mixed model, given family data, is well known.[13,19–23]

The ML estimator of $\mathbf{b}$ is based on solving $\mathbf{b}$ in the first order derivative of the ML function with respect to $\mathbf{b}$:

$$\hat{\boldsymbol{b}}_{ML} = \left(\boldsymbol{X}^t\boldsymbol{V}\left(\hat{\boldsymbol{\theta}}\right)^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}\left(\hat{\boldsymbol{\theta}}\right)^{-1}\boldsymbol{y} \quad (3)$$

If $\boldsymbol{\theta}$ is unknown, this requires iteration. Note that the covariance matrix $\mathbf{V}\left(\hat{\boldsymbol{\theta}}\right)$ can also be estimated once and then used as fixed in the generalized least squares estimator (see, for example, Pirinen *et al*[6] and Li *et al*[24]). The Wald test of $\mathbf{b}_{1ML}$ is based on $\mathrm{var}\left(\hat{\mathbf{b}}_{ML}\right) = \left(\mathbf{X}^t\mathbf{V}\left(\hat{\boldsymbol{\theta}}\right)^{-1}\mathbf{X}\right)^{-1}$. ULS is a special case with $\hat{\boldsymbol{\theta}} = \left[\hat{\sigma}^2_E\right]$, that is, $\mathbf{V}\left(\hat{\boldsymbol{\theta}}\right) = \hat{\sigma}^2_E\mathbf{I}$. The ULS estimator can be expressed as:[4,18,25]

$$\hat{\boldsymbol{b}}_{ULS} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}, \quad (4)$$

with

$$\mathrm{var}\left(\hat{\boldsymbol{b}}_{\mathrm{ULS}}\right) = \hat{\sigma}^2_E(\boldsymbol{X}^t\boldsymbol{X})^{-1} \quad (5)$$

The ULS procedure involves misspecification in the case of family data, as $\mathbf{V}\left(\hat{\boldsymbol{\theta}}\right) = \hat{\sigma}^2_E\mathbf{I}$ is almost certainly incorrect. To correct the standard errors, we employ the sandwich correction of $\mathrm{var}\left(\hat{\mathbf{b}}_{\mathrm{ULS}}\right)$,[1]

$$\mathrm{var}\left(\hat{\boldsymbol{b}}_{\mathrm{R-ULS}}\right) = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t(\boldsymbol{y}-\boldsymbol{Xb})(\boldsymbol{y}-\boldsymbol{Xb})^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1} \quad (6)$$

We note that the sandwich correction is equally applicable to ML, given misspecified $\mathbf{V}\left(\hat{\boldsymbol{\theta}}_{\mathrm{m}}\right)$. For instance (eg, Dobson[18]):

$$\mathrm{var}\left(\hat{\boldsymbol{b}}_{\mathrm{R-ML}}\right) = \left(\boldsymbol{X}^t\boldsymbol{V}\left(\hat{\boldsymbol{\theta}}_{\mathrm{m}}\right)^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}\left(\hat{\boldsymbol{\theta}}_{\mathrm{m}}\right)^{-1}(\boldsymbol{y}-\boldsymbol{Xb})(\boldsymbol{y}-\boldsymbol{Xb})^t\boldsymbol{V}\left(\hat{\boldsymbol{\theta}}_{\mathrm{m}}\right)^{-1}\boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{V}\left(\hat{\boldsymbol{\theta}}_{\mathrm{m}}\right)^{-1}\boldsymbol{X}\right)^{-1} \quad (7)$$

where we employ the subscript m to denote misspecification.

Below we consider various tests of $\mathbf{b}_1$ in family data of two full sibs and MZ and DZ twins with and without parents. First, we compare the ULS and ML procedures given correct specification of the background in ML, that is, $\boldsymbol{\theta} = [\sigma^2_A, \sigma^2_E]$. Specifically, we consider the standard ULS and ML procedures (ie, based on the so-called naive variance, which incorporates the assumption that the background model is correctly specified). We also consider the sandwich corrected ULS procedure (as in Plink[1]) and the sandwich corrected ML procedure with the background $\mathbf{V}(\boldsymbol{\theta})$ conditioned on zygosity, but otherwise unconstrained. That is, the family covariance matrix is freely estimated within the MZ and DZ families, which is consistent with the true model. We include the sandwich corrected ML procedure to investigate whether robustification does result in an overcorrection when the underlying model is in fact correct. Second, to assess the effects of misspecification, we consider standard ML estimation, with the (true) background $\boldsymbol{\theta} = [\sigma^2_A, \sigma^2_C, \sigma^2_E]$ misspecified as (a) $\hat{\boldsymbol{\theta}}_{\mathrm{m}} = \left[\hat{\sigma}^2_A, \hat{\sigma}^2_E\right]$, or as (b) $\hat{\boldsymbol{\theta}}_{\mathrm{m}} = \left[\hat{\sigma}^2_C, \hat{\sigma}^2_E\right]$. In addition, we use the misspecified $\mathbf{V}\left(\hat{\boldsymbol{\theta}}_{\mathrm{m}}\right)$ with $\hat{\boldsymbol{\theta}}_{\mathrm{m}} = \left[\hat{\sigma}^2_A, \hat{\sigma}^2_E\right]$ (and the misspecified $\mathbf{V}\left(\hat{\boldsymbol{\theta}}_{\mathrm{m}}\right)$ with $\hat{\boldsymbol{\theta}}_{\mathrm{m}} = \left[\hat{\sigma}^2_C, \hat{\sigma}^2_E\right]$) – estimated with standard ML using the incorrect background model – in the sandwich corrected ML procedure. We also include the standard and the sandwich corrected ULS procedures. Finally, we test $\mathbf{b}_1$ using the standard ML procedure, with the background correctly parameterized (ie, estimating the variance components of the true model). We consider both the type I and type II error rates.

Simulation details

We generated family data for MZ and DZ families consisting of two sibs and MZ and DZ twins, with and without parents. Each simulated sample had a size of 4000 individuals. We simulated a diallelic genetic variant (GV) in Hardy-Weinberg equilibrium, with a minor allele frequency of 0.5, and explaining one percent (1%) of the phenotypic variance. We simulated the background covariance structure according to two models: (1) a model with additive (A) and unshared (E) environmental effects, that is, an AE model, $\boldsymbol{\theta} = [\sigma^2_A, \sigma^2_E]$, with $h^2 = \sigma^2_A/(\sigma^2_A + \sigma^2_E)$ equal to 0.3, 0.5 or 0.7); (2) a model with additive genetic, shared (C) and unshared environmental effects, that is, an ACE model, $\boldsymbol{\theta} = [\sigma^2_A, \sigma^2_C, \sigma^2_E]$, with $h^2 = \sigma^2_A/\sigma^2_{ph} = 0.2$, $\sigma^2_C/\sigma^2_{ph} = 0.6$ and $\sigma^2_E/\sigma^2_{ph} = 0.2$. We also considered an ACE model, with $h^2 = \sigma^2_A/\sigma^2_{ph} = 0.6$, $\sigma^2_C/\sigma^2_{ph} = 0.2$ and $\sigma^2_E/\sigma^2_{ph} = 0.2$ (see Tables 2 and 3, Supplementary Material). These models were chosen to represent a range of complex phenotypes. For example, data generated based on the parameter values in the first cell of Table 1 are illustrative for family-based association studies of highly heritable traits such as height in adults,[26] whereas the data generated based on the parameter values in Table 3 may inform genome-wide analyses of ACE traits, such as initiation of substance use (eg, Vink *et al*[7]). We used the R package MASS[27] for data generation. We implemented the sandwich corrected ULS and the sandwich corrected ML procedures in R. We obtained the standard ML results using linear mixed modeling as implemented in the R-package nlme.[28] Observed power equals the proportion of datasets out of 10 000 replications, in which the *P*-value associated with the Wald test was smaller than our chosen alpha $= 10^{-7}$. Type I error rate was assessed at alpha = 0.05, 0.01, 0.001 and 0.0001, using 1 000 000 datasets, simulated under the null hypothesis of $\mathbf{b}_1 = 0$. Otherwise, given $\mathbf{b}_1 \neq 0$, we used 10 000 replications. Simulations were run on the Lisa Computer Cluster (www.surfsara.nl). The R script used to obtain the results is available at http://cameliaminica.nl/scripts.php.

RESULTS

Correctly specified background model: type I and type II error rates

First we checked the distribution of the four Wald tests given $\mathbf{b}_1 = 0$, and the correct specification of the AE background, that is, $\boldsymbol{\theta} = [\sigma^2_A, \sigma^2_E]$ (except standard ULS which assumes independence). As expected, the null distributions of the ML-based Wald tests (standard and sandwich corrected) and of the sandwich corrected ULS-based Wald test were correct (see Table 1, Supplementary Material). In contrast, the standard ULS procedure (without a sandwich correction) produced an excess of false positives. For instance, in the four sibs condition and with a 70% heritable trait, the observed type I error rate was 0.0024 given an alpha of 0.0001.

Given $\mathbf{b}_1 = -0.141$ ($\mathbf{b}_1$ given the chosen effect size of 1%) and the correct specification of the AE background covariance matrix in ML (with $h^2 = \sigma^2_A/(\sigma^2_A + \sigma^2_E)$ equal to 0.3, 0.5 or 0.7), we obtained the results in Table 1 concerning the power to detect the GV effect.

The mean parameter estimates as produced by ML and ULS are equal, across all conditions. This is expected, as the estimators are all asymptotically unbiased and consistent.[4] The standard errors as produced by the ML standard and by the sandwich corrected ML are identical. This is expected, as both procedures are based on the correct background covariance structure, be it correctly structured (ie, $\boldsymbol{\theta} = [\sigma^2_A, \sigma^2_E]$) or unstructured (the sandwich corrected ML). Therefore, the use of the sandwich does not result in any overcorrection. The ULS procedures are consistent, but differ in terms

Table 1 Power (alpha = 10^{-7}) and parameter estimates for the ML linear mixed (standard and sandwich corrected) and the ULS (standard and sandwich corrected) procedures

Family structure	*ML standard true model*	*Sandwich corrected ULS*	*Sandwich corrected ML (unstructured)*	*ULS standard*
$h^2 = 70\%$				
Two parents and four sibs				
Mean (b_1)	−0.141	−0.141	−0.141	−0.141
Mean (SE)	0.025	0.031	0.025	0.022
Mean (*t*-value)	−5.60	−4.62	−5.67	−6.35
Power	60.3	24.4	62.6	76.8
Four sibs				
Mean (b_1)	−0.141	−0.141	−0.141	−0.141
Mean (SE)	0.025	0.029	0.025	0.022
Mean (*t*-value)	−5.70	−4.95	−5.73	−6.35
Power	63.5	35.1	64.9	78.9
$h^2 = 50\%$				
Two parents and four sibs				
Mean (b_1)	−0.141	−0.141	−0.141	−0.141
Mean (SE)	0.025	0.028	0.025	0.022
Mean (*t*-value)	−5.56	−4.96	−5.62	−6.34
Power	59.1	36.4	61.5	78.4
Four sibs				
Mean (b_1)	−0.141	−0.141	−0.141	−0.141
Mean (SE)	0.025	0.027	0.025	0.022
Mean (*t*-value)	−5.68	−5.25	−5.71	−6.34
Power	63.1	46.6	65.0	80.0
$h^2 = 30\%$				
Two parents and four sibs				
Mean (b_1)	−0.141	−0.141	−0.141	−0.141
Mean (SE)	0.025	0.026	0.025	0.022
Mean (*t*-value)	−5.68	−5.40	−5.74	−6.34
Power	64.0	53.2	66.0	80.8
Four sibs				
Mean (b_1)	−0.142	−0.142	−0.142	−0.142
Mean (SE)	0.024	0.025	0.024	0.022
Mean (*t*-value)	−5.81	−5.63	−5.84	−6.36
Power	67.8	61.3	69.2	81.4

Abbreviations: ML, maximum likelihood; SE, standard error; ULS, unweighted least squares. We simulated a genetic marker having an effect of 1% explained phenotypic variance and a minor allele frequency = 0.5. The sample consisted of $N = 4000$ individuals. The trait was simulated according to an AE background model (the true model) given various heritabilities (h^2) (10 000 simulated samples for each cell). The background model in the ML procedure is correctly specified (true or saturated, ie, unstructured).

of power. The power of the standard ULS procedure appears to be greatest, but this is due to the fact that the standard errors are underestimated, as mentioned above. The sandwich corrected ULS procedure comes at a relative cost in terms of power (compared to ML). The loss in power increases with the family clustering due to the heritability of the trait. For example, in the four sibs condition, with a 70% heritable trait, the power of the sandwich corrected ULS procedure is 35.1%, whereas the power of the ML procedures is about 64%.

Besides the heritability of the trait, the size of the family cluster has a bearing on the power of ULS. For instance, given a 70% heritable trait, the difference in power between the ML and ULS with a sandwich correction is ∼30% and ∼35% when the sample consists of size 4 sibships and when it consists of two parents and four sibs, respectively (see Table 1). Note also the difference in power between the two robust methods as well (the sandwich corrected ULS and ML), with the power of the sandwich corrected ML procedure being higher.

Misspecified background model

We evaluated consequences on type I and II error rates of misspecifying the background model, $\mathbf{V}(\boldsymbol{\theta})$. We employed a background model with additive genetic (σ^2_A) and shared and unshared variance components (σ^2_C and σ^2_E), and discarded the effects of σ^2_A (ML with an incorrect CE structured background) or σ^2_C (ML with an incorrect AE structured background), or discarded both σ^2_A and σ^2_C (ULS with an incorrect E structured background). ML with a correctly specified background is also included. First we considered the type I error rates, given $\mathbf{b}_1 = 0$. Table 2 contains the results.

Based on these results, we conclude that the type I error rates of the ML procedure are not greatly affected by the misspecification. The misspecification $\hat{\boldsymbol{\theta}}_m = [\hat{\sigma}^2_C, \hat{\sigma}^2_E]$ is associated with a slight inflation (eg, 0.0002, given alpha = 0.0001 in the two parents and four sibs cell), but the ML with the CE structured sandwich corrects this (0.00011). The misspecification $\hat{\boldsymbol{\theta}}_m = [\hat{\sigma}^2_A, \hat{\sigma}^2_E]$ hardly affects type I error rates. As expected, the standard ULS procedure $\left(\hat{\boldsymbol{\theta}}_m = [\hat{\sigma}^2_E]\right)$ produced incorrect type I error rates (for example, 0.008, given alpha = 0.0001 in the four sibs cell). However, as above, the ULS sandwich correction yields correct type I rates. The ML with an ACE background is correctly specified and produces correct type I error rates.

Table 3 contains the results relating to the power given $\mathbf{b}_1 \neq 0$ and misspecified background. As expected, all modeling approaches yielded similar mean estimates of $\mathbf{b}_1$, regardless of the specification of the background structure. Given correct background specification ($\boldsymbol{\theta} = [\sigma^2_A, \sigma^2_C, \sigma^2_E]$) and sibships size 4, the power is about 97.4% (standard ML). The power of the standard ML procedure appears to increase to about 98.2%, when σ^2_A is discarded $\left(\hat{\boldsymbol{\theta}}_m = [\hat{\sigma}^2_C, \hat{\sigma}^2_E]\right)$, but this is spurious as it is due to the effect of the misspecification on the type I error (see Table 2). This effect is likely to be more noticeable at more stringent alpha levels (see also Minică *et al*[29]). The ML with a CE structured sandwich, however, preserves the power equal to the power of the (true) ML ACE model, without inflating the type I error rate. Ignoring shared environmental effects, that is, dropping σ^2_C in a $\boldsymbol{\theta} = [\sigma^2_A, \sigma^2_C, \sigma^2_E]$ model results in a loss in power. For instance, in the four sibs condition, the power of the standard ML procedure drops to about 88.1%, when σ^2_C is discarded $\left(\hat{\boldsymbol{\theta}}_m = [\hat{\sigma}^2_A, \hat{\sigma}^2_E]\right)$ (similar results were obtained when dropping σ^2_D in a $\boldsymbol{\theta} = [\sigma^2_A, \sigma^2_D, \sigma^2_E]$ model, where D stands for dominance; see Table 4 Supplementary Material). With an AE structured background, the standard errors as produced by the standard and the sandwich corrected ML are very similar, and so is the power. Given that the latter correctly reflects the parameter variance in the presence of a misspecified model, this result indicates that in the conditions considered here this type of misspecification does not affect estimation (ie, type I error rate is well controlled). However, this is not a general finding. Consider the extreme misspecification of the background employed by the ULS method. This has a clear effect, which is reflected in the notable discrepancy observed between the standard and the robust (correct) ULS standard errors (ie, 0.022 *vs* 0.033). Finally, although both are correct, we note that the sandwich corrected ML procedure is appreciably more powerful than the sandwich corrected ULS procedure (for example, power of 88.1% for the sandwich corrected ML with a misspecified AE structured background *vs* power of 16.4% for the sandwich corrected ULS procedure). Results follow similar trends in the samples consisting of two parents and four sibs.

Table 2 Type I error rates for the ML linear mixed (standard and sandwich corrected) and the ULS (standard and sandwich corrected) procedures

Family structure	*Alpha*	*ML standard ACE model (true)*	*ML standard AE model (false)*	*ML standard CE model (false)*	*Sandwich corrected ML (false: AE structured)*	*Sandwich corrected ML (false: CE structured)*	*ULS standard E model (false)*	*Sandwich corrected ULS E model (false)*
Two parents and four sibs	0.05	0.049	0.049	0.06	0.05	0.049	0.2	0.051
	0.01	0.010	0.010	0.015	0.010	0.010	0.11	0.010
	0.001	0.0010	0.00097	0.0019	0.00097	0.0010	0.045	0.0011
	0.0001	0.0001	0.00009	0.0002	0.0001	0.00011	0.018	0.00012
Four sibs	0.05	0.05	0.05	0.057	0.05	0.05	0.18	0.05
	0.01	0.01	0.01	0.0127	0.01	0.01	0.08	0.01
	0.001	0.001	0.001	0.0014	0.001	0.001	0.025	0.001
	0.0001	0.0001	0.00012	0.00018	0.00012	0.00012	0.008	0.0001

Abbreviations: ML, maximum likelihood; ULS, unweighted least squares.
The background model is (a) correctly specified (true) or (b) misspecified. Background covariance matrix was generated according to an ACE model ($h^2 = 0.2$, $c^2 = 0.6$). The samples comprised 4000 individuals (1 000 000 simulated datasets per cell).

Table 3 Power (given alpha $= 10^{-7}$) and parameter estimates for the ML (standard and sandwich corrected) and the ULS (standard and sandwich corrected) procedures

Family structure	*ML standard ACE model (true)*	*ML standard AE model (false)*	*ML standard CE model (false)*	*Sandwich corrected ML (false: AE structured)*	*Sandwich corrected ML (false: CE structured)*	*ULS standard E model (false)*	*Sandwich corrected ULS E model (false)*
Two parents and four sibs							
Mean (b_1)	−0.141	−0.141	−0.141	−0.141	−0.141	−0.141	−0.141
Mean (SE)	0.019	0.021	0.018	0.021	0.019	0.022	0.037
Mean (*t*-value)	−7.54	−6.59	−7.89	−6.6	−7.44	−6.33	−3.86
Power	98.6	89.4	99.2	89.4	98.1	73.0	7.5
Four sibs							
Mean (b_1)	−0.141	−0.141	−0.142	−0.141	−0.142	−0.141	−0.141
Mean (SE)	0.019	0.022	0.019	0.022	0.020	0.022	0.033
Mean (*t*-value)	−7.27	−6.49	−7.49	−6.50	−7.25	−6.36	−4.33
Power	97.4	88.1	98.2	88.0	97.1	75.9	16.4

Abbreviations: ML, maximum likelihood; SE, standard error; ULS, unweighted least squares.
The background model is (a) correctly specified (true) or (b) misspecified. Background covariance matrix was generated according to an ACE model ($h^2 = 0.2$, $c^2 = 0.6$). The genetic marker explained 1% phenotypic variance and had a minor allele frequency = 0.5. The samples consisted of $N = 4000$ individuals (10 000 simulated datasets per cell).

Given that these results pertain to averages over replications, we also looked at how often the ML *t*-values actually exceed the sandwich corrected ULS *t*-values, considering also the smaller effect sizes to be expected in GWAS. This might be of interest as it will provide an indication on how the two estimators are expected to perform in individual studies involving family data. Dots above the diagonal in Figure 1 show how often the ML-based Wald test is larger than the sandwich corrected ULS-based Wald test, given decline in the size of the genetic effect.

Figure 1 top left shows that the ML (true AE model) almost always produces a larger test statistic, when the effect size is relatively large (effect size of 1% explained phenotypic variance) and the sample is large enough to capture it. In the example, in just about 7.5% of the samples the sandwich corrected ULS test statistic was larger. However, as the effect size decreases, one can observe more and more sandwich corrected ULS-based Wald tests larger than those estimated by the ML procedure (as illustrated in Figure 1 top right). It can be seen that under the null model (Figure 1, bottom) no differences occur between the two estimation methods, which is as expected provided both are correct.

FaST-LMM formulation of the ML sandwich correction

The sandwich correction is computationally relatively simple and quick in the standard formulation of the linear mixed model. We note that the fast full information ML mixed procedures[5,6] are equally amenable to a sandwich correction. The ML sandwich can be presented as follows:

$$\mathrm{var}\left(\hat{\boldsymbol{b}}_{\mathrm{R-ML}}\right) = \left(\boldsymbol{X}^t\boldsymbol{V}\left(\hat{\boldsymbol{\theta}}\right)^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}\left(\hat{\boldsymbol{\theta}}\right)^{-1}(\boldsymbol{y}-\boldsymbol{Xb})(\boldsymbol{y}-\boldsymbol{Xb})^t\boldsymbol{V}\left(\hat{\boldsymbol{\theta}}\right)^{-1}\boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{V}\left(\hat{\boldsymbol{\theta}}\right)^{-1}\boldsymbol{X}\right)^{-1} \quad (8)$$

Given random effects $\hat{\boldsymbol{\theta}} = \left[\hat{\sigma}_a^2, \hat{\sigma}_e^2\right]$, the background covariance matrix is reformulated as $V(\boldsymbol{\theta}) = \left(\sigma_a^2 * \boldsymbol{K} + \sigma_e^2 * \boldsymbol{I}\right) = \left[\sigma_a^2 * (\boldsymbol{K} + \delta * \boldsymbol{I})\right]$, where **K** is the genetic relationship matrix (positive semi-definite), **I** is the identity matrix and $\delta = \sigma_a{}^2/\sigma_e{}^2$. Lippert *et al*[5] (see also Pirinen *et al*[6]) formulate the covariance matrix as follows:

$$\boldsymbol{V}(\boldsymbol{\theta}) = \left[\sigma_a^2 * (\boldsymbol{USU}^t + \delta * \boldsymbol{UIU}^t)\right] = \left[\sigma_a^2 * \boldsymbol{U}(\boldsymbol{S} + \delta * \boldsymbol{I})\boldsymbol{U}^t\right] \quad (9)$$

where $\mathbf{K} = \mathbf{USU}^t$ is the eigen value decomposition of **K**, with **U**, the eigenvectors, orthonormal, and **S** diagonal (eigenvalues). The matrix $\delta^*\mathbf{I}$, being diagonal and constant, can be written as $\delta^*\mathbf{UIU}^t$. The inverse is:

$$\boldsymbol{V}(\boldsymbol{\theta})^{-1} = \left[\sigma_a^{-2} * \boldsymbol{U}(\boldsymbol{S} + \delta * I)^{-1}\boldsymbol{U}^t\right] \quad (10)$$

Note that the addition of off-diagonal terms in $\sigma_e{}^{2*}\mathbf{I}$, that is, terms accommodating shared environmental effects, would render the method invalid, as then the eigenvectors of the environmental

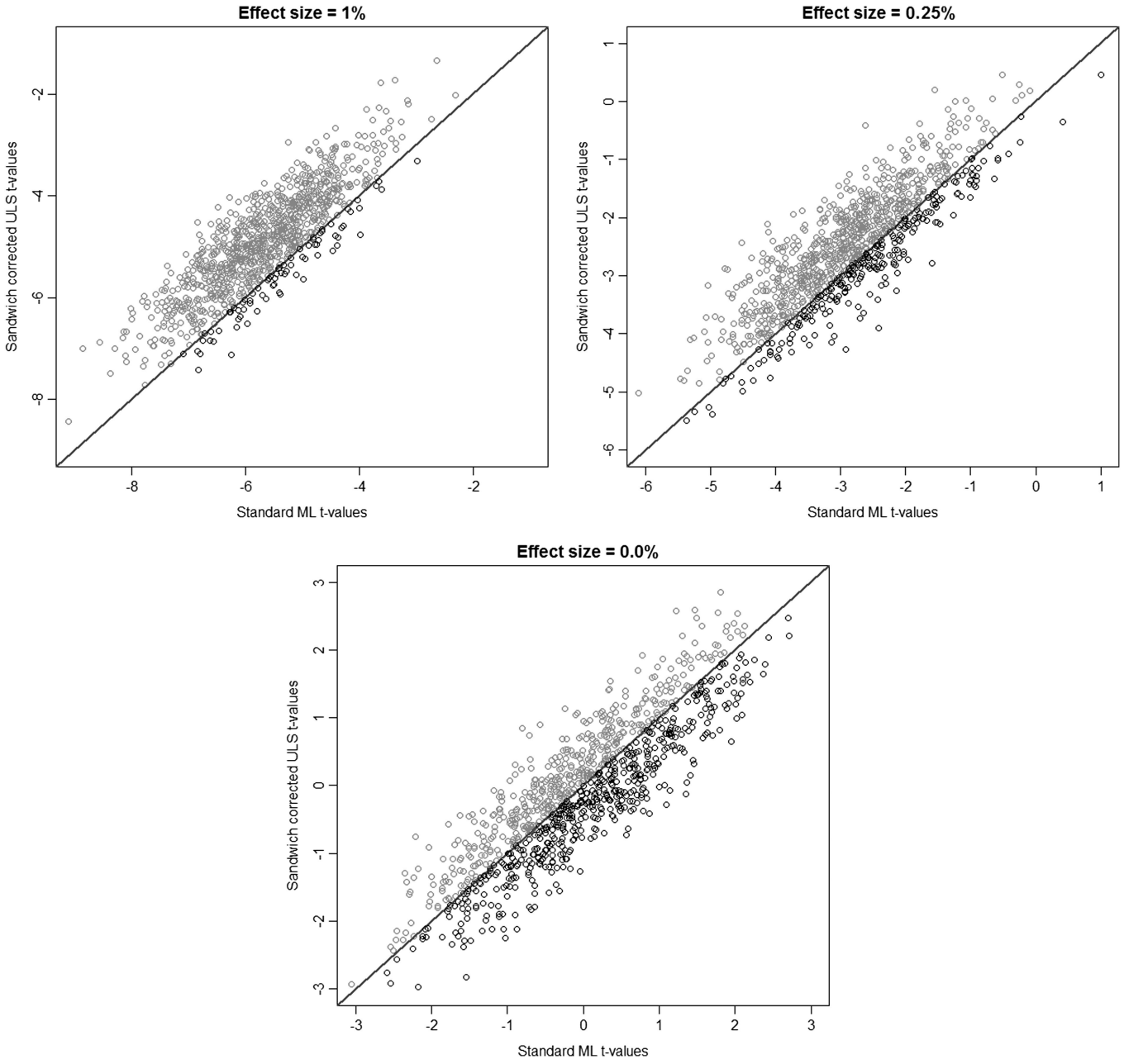

Figure 1 Wald tests produced by the sandwich corrected ULS procedure compared with the test statistic obtained based on full information maximum-likelihood (standard ML) estimation method. We simulated 1000 datasets consisting of 500 MZ and 500 DZ four-sib families, and we varied the effect size of the genetic effect (1%, 0.25% and the null model). The heritability of the trait was $h^2 = 70\%$. The dots above the diagonal show the number of times the standard ML procedure produced a larger test statistic.

covariance matrix cannot be chosen to equal **U.** In terms of this treatment of the matrix $\mathbf{V}(\boldsymbol{\theta})$, the sandwich can be written:

$$var\left(\hat{\boldsymbol{b}}_{\mathrm{R-ML}}\right) = \sigma_a^2 * \left[\boldsymbol{X}^t\boldsymbol{U}(\boldsymbol{S}+\delta * I)^{-1}\boldsymbol{U}^t\boldsymbol{X}\right]^{-1}\sigma_a^{-2} * \boldsymbol{X}^t\boldsymbol{U}(\boldsymbol{S}+\delta * I)^{-1} * (\boldsymbol{U}^t\boldsymbol{y} - \boldsymbol{U}^t\boldsymbol{X}\boldsymbol{b})(\boldsymbol{U}^t\boldsymbol{y} - \boldsymbol{U}^t\boldsymbol{X}\boldsymbol{b})^t * \left[\sigma_a^{-2} * \boldsymbol{X}^t\boldsymbol{U}(\boldsymbol{S}+\delta * \boldsymbol{I})^{-1}\right]^t * \sigma_a^2 * \left[\boldsymbol{X}^t\boldsymbol{U}(\boldsymbol{S}+\delta * \boldsymbol{I})^{-1}\boldsymbol{U}^t\boldsymbol{X}\right]^{-1} \quad (11)$$

In implementing this, the fact that $(\mathbf{S}+\delta^*\mathbf{I})^{-1}$ is diagonal may be exploited to increase computational efficiency.

DISCUSSION

We compared the standard and sandwich corrected ULS and ML procedures, in the context of family-based association analysis of a normally distributed phenotype. Conditional on the correct specification of the background, the standard ML procedure is appreciably more powerful than the sandwich corrected ULS procedure. The actual difference in power depends on the magnitude of the residual correlations, but increases with greater family resemblance.

We also considered the sensitivity of ML to model misspecification. Model misspecification involves the mismatch between the true

background covariance model (say, an ACE or ADE trait) and the background model used in the analyses (a CE or AE model).

This may occur in using fast ML procedures, which employ the background covariance matrix necessarily limited to additive genetic (A) and unshared environmental (E) effects.[5,30] The standard ML procedure was quite robust under model misspecification in the simulated settings, and appreciably more powerful than the sandwich corrected ULS procedure. However, for circumstances other than those considered here, a sandwich correction is equally applicable to ML to correctly capture the parameter variance in the presence of model misspecification. The sandwich corrected standard errors may also be employed as a means to get an indication of the effects of background misspecification on the type I error rate (ie, the larger the discrepancy between the naive and sandwich corrected standard errors, the more likely the type I error rate of the procedure without a sandwich to be affected[31]).

In the present paper, we considered a normally distributed phenotype. Our conclusions apply equally to generalized linear modeling of binary traits, such as disease status. To demonstrate this, we included in the Supplementary Material (Supplementary Tables 5 and 6) results based on continuous and dichotomized (median – split) phenotypes. With respect to binary phenotypes, we note that a general (rather than generalized) linear model is often used in analyzing such variables (eg, Zhou and Stephens[32]). Cogent arguments have been presented that the linear model may suffice in the analysis of binary phenotypes.[5,6]

Although relatively simple to implement and more efficient than the sandwich corrected ULS in correcting for model misspecification, to our knowledge the ML sandwich correction has not yet been implemented by any of the current software for GWAS that can handle family data. With respect to implementation, we note that generalized estimating equations (gee) procedure, as implemented in R[33] has four useful aspects. First, it has a choice of background models, which includes the independence model and exchangeable model (the latter is equivalent to the CE model in linear mixed modeling). Second, it includes sandwich corrected standard errors of the parameters **b**. Third, gee covers generalized linear model. Fourth, as gee is a library it can be accessed from Plink[1] and so provides a computationally feasible strategy for running genome-wide scans in family data. An annotated R script to do this is available at http://cameliaminica.nl/scripts.php.

In conclusion, for traits characterized by moderate to large familial resemblance, using ML with a correctly specified model for the familial covariance matrix should be the strategy of choice. For such traits, the potential loss in power encountered by the sandwich corrected ULS procedure does not outweigh its computational convenience. Using a fast ML algorithm that commits to a background model limited to additive and unshared environmental effects is acceptable even if shared environment has an influence on the phenotype of interest. That is, in the settings considered here, type I error rate of the standard ML was hardly affected by model misspecification. However, a sandwich correction is still of interest when employing ML in genome-wide scans, because (a) it produces correct standard errors regardless of whether the model is correctly parameterized or misspecified; hence it should be useful for situations other than those considered here, (b) it does not result in any overcorrection when the background model is in fact correctly specified, (c) as shown above, it is computationally cheap and can easily be incorporated in the fast ML procedures, and (d) it is a useful diagnostic tool for assessing model misspecification.[31] Currently, Plink often is the preferred software when consortia share GWA results for meta-analyses. When including data from cohorts that include relatives, one should realize that the corrected standard errors while in many circumstances larger than the ML standard errors, are accurate, and so therefore are its type I and II error rates. For ordinary GWAS (ie, not family based), Plink is as good as FastLMM (as then ULS and ML are identical).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

Camelia C Minică and Jacqueline M Vink are supported by the ERC starting grant 284167. Conor V Dolan is supported by the European Research Council (Genetics of Mental Illness; grant number: ERC-230374). Dorret I Boomsma is supported by European Research Council (ERC-230374). The statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org), which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003), the Dutch Brain Foundation and the Department of Psychology and Education of the VU University Amsterdam.

1 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
2 Rogers WH: Regression standard errors in clustered samples. *Stata Tech Bull* 1993; **13**: 19–23.
3 Williams RL: A note on robust variance estimation for cluster-correlated data. *Biometrics* 2000; **56**: 645–646.
4 Greene WH: *Econometric Analysis*. India: Pearson Education, 2003.
5 Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D: FaST linear mixed models for genome-wide association studies. *Nat Meth* 2011; **8**: 833–835.
6 Pirinen M, Donnelly P, Spencer CCA: Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann Appl Stat* 2013; **7**: 369–390.
7 Vink JM, Wolters LMC, Neale MC, Boomsma DI: Heritability of cannabis initiation in Dutch adult twins. *Addict Behav* 2010; **35**: 172–174.
8 van den Bree MBM, Johnson EO, Neale MC, Pickens RW: Genetic and environmental influences on drug use and abuse/dependence in male and female twins. *Drug Alcohol Depend* 1998; **52**: 231–241.
9 Kendler KS, Schmitt E, Aggen SH, Prescott CA: Genetic and environmental influences on alcohol, caffeine, cannabis, and nicotine use from early adolescence to middle adulthood. *Arch Gen Psychiatry* 2008; **65**: 674–682.
10 Vink J, Willemsen G, Boomsma D: Heritability of smoking initiation and nicotine dependence. *Behav Genet* 2005; **35**: 397–406.
11 Litière S, Alonso A, Molenberghs G: Type I and Type II error under random-effects misspecification in generalized linear mixed models. *Biometrics* 2007; **63**: 1038–1044.
12 Falconer DS, Mackay TFC: *Introduction to Quantitative Genetics*, 4th edn. Harlow: Pearson Education Limited, 1996.
13 Visscher PM, Benyamin B, White I: The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. *Twin Res Hum Genet* 2004; **7**: 670–674.
14 Zaitlen N, Kraft P, Patterson N *et al*: Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet* 2013; **9**: e1003520.
15 Martin NG, Eaves LJ: The genetical analysis of covariance structure. *Heredity (Edinb)* 1977; **38**: 79–95.
16 Eaves LJ: Inferring the causes of human variation. *J Roy Stat Soc A* 1977; **140**: 324–365.
17 Pinheiro J, Bates D: *Mixed-Effects Models in S and S-PLUS*. New York: Springer, 2000.
18 Dobson A: *An Introduction to Generalized Linear Models*. London: Chapman & Hall/CRC, 2002.
19 Beem AL, Boomsma DI: Implementation of a combined association-linkage model for quantitative traits in linear mixed model procedures of statistical packages. *Twin Res Hum Genet* 2006; **9**: 325–333.
20 Guo G, Wang J: The mixed or multilevel model for behavior genetic analysis. *Behav Genet* 2002; **32**: 37–49.
21 McArdle JJ, Prescott CA: Mixed-effects variance components models for biometric family analyses. *Behav Genet* 2005; **35**: 631–652.
22 Rabe-Hesketh S, Skrondal A, Gjessing HK: Biometrical modeling of twin and family data using standard mixed model software. *Biometrics* 2008; **64**: 280–288.
23 van den Oord E: Estimating effects of latent and measured genotypes in multilevel models. *Stat Methods Med Res* 2001; **10**: 393–407.

24 Li X, Basu S, Miller MB, Iacono WG, McGue M: A rapid generalized least squares model for a genome-wide quantitative trait association analysis in families. *Hum Hered* 2011; **71**: 67–82.
25 Draper NR, Smith H: *Applied Regression Analysis*. New York: John Wiley and Sons, 1981.
26 Silventoinen K, Sammalisto S, Perola M *et al*: Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res Hum Genet* 2003; **6**: 399–408.
27 Venables WN, Ripley BD: *Modern Applied Statistics with S*, 4th edn. New York: Springer, 2002.
28 Pinheiro J, Bates D, DebRoy S, Sarkar D, Team RC: nlme: Linear and Nonlinear Mixed Effects Models; R package version 3.1-111. 2013.
29 Minică C, Dolan C, Hottenga J-J, Willemsen G, Vink J, Boomsma D: The use of imputed sibling genotypes in sibship-based association analysis: on modeling alternatives, power and model misspecification. *Behav Genet* 2013; **43**: 254–266.
30 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; **30**: 97–101.
31 Chavance M, Escolano S: Misspecification of the covariance structure in generalized linear mixed models. *Stat Methods Med Res* 2012 e-pub ahead of print 14 October 2012; doi:10.1177/0962280212462859.
32 Zhou X, Stephens M: Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 2012; **44**: 821–824.
33 Carey VJ: gee: Generalized Estimation Equation solver. http://CRANR-projectorg/package=gee, R package version 4.13-418, 2012.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)