TECHNICAL
REPORT

0651

THE EFFECTIVE SAMPLE SIZE AND A NOVEL SMALL SAMPLE DEGREES OF FREEDOM METHOD

FAES, C., MOLENBERGHS, H., AERTS, M., VERBEKE, G. and M.G. KENWARD

IAP STATISTICS
NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

The Effective Sample Size and a Novel Small Sample Degrees of Freedom Method

Christel Faes, Geert Molenberghs, Marc Aerts

Center for Statistics, Hasselt University, Agoralaan, Diepenbeek, Belgium

Email: christel.faes@uhasselt.be

Geert Verbeke

Biostatistical Centre, Katholieke Universiteit Leuven, Belgium

Michael G. Kenward

Medical Statistics Unit, London School of Hygiene and Tropical Medicine, UK

Summary

In statistics, one is often confronted with the analysis of correlated data. The amount of information in such data depends on the correlation among the observations. A general concept is derived, the *effective sample size*, as a way to quantify the amount of information in such data. It is defined as the sample size one would need in an independent sample to equal the amount of information in the actual correlated sample. This concept has the advantage of general applicability and provides important insight into the setting of correlated data. For example, using the concept of the effective sample size, it is seen that the amount of information is not always infinite, but rather that there exists a limit of information in some situations. Also, the effective sample size can be used as a building block in the construction of a novel degrees-of-freedom determination method for a scaled Wald statistic. The performance of the proposed method is investigated through a simulation study.

Keywords: Correlated Data, Effective Sample Size, Degrees of Freedom, Information Limit

1 Introduction

The size of a sample is a very important measure for the amount of information available in the data. In the simplest situation of independent continuous data, the sample size is defined as the number of individuals in the study and the information is proportional to the sample size. For binary data, the information is proportional to the sample size, and for, possibly censored, survival data, the information is proportional to the number of events. Here, information is used in the precise mathematical sense as defined in, for example, Cox (1974). However, as soon as one ventures away from independence, there is considerable uncertainty as to how to define the sample size except in a number of well understood cases, mostly in a multivariate normal setting (Johnson and Wichern, 1992). At first sight, if several measurements are taken for each independent unit, one could take the number of individuals or the number of measurements as the sample size. While the first approach underestimates the amount of information, the second approach overestimates it. The reason is that the amount of information depends on the correlation among the observations. Such considerations are important to determine the degrees of freedom, an essential component when selecting an appropriate null distribution for a variety of hypothesis tests. The best known examples include the t, F, and U tests, and their corresponding distributions.

Several methods to estimate the appropriate number of degrees of freedom needed for these tests are available. The best known methods for continuous data are the Satterthwaite-type approximations (Satterthwaite, 1941) and the Kenward-Roger method (Kenward and Roger , 1997). Satterthwaite's degrees of freedom are obtained by matching moments with those of a χ^2-distribution, and using an approximation of the denominator based on a Taylor expansion (Giesbrecht and Burns, 1985). Kenward and Roger proposed a scaled Wald statistic, based on an adjusted covariance estimate, which accounts for small-sample bias and incorporates the extra variability arising from the estimation of the variance-covariance matrix. They show that the small-sample distribution can be approximated well by an F-distribution with denominator degrees of freedom, obtained by a multivariate moment-matching argument, to properly capture the internal stochastic structure of the estimated covariance matrix, of which Satterthwaite's approximation is a univariate special case. In case of binary data, one typically uses the residual number of degrees of freedom, defined as the number of measurements minus the number of parameters to be estimated, or switches to an asymptotic method. While acceptable in large samples of independent data, such an approach generally overstates the number of degrees of freedom, with associated p-values that then are too small.

Here, we will derive a generic concept, the *effective sample size*, loosely defined as the sample size one would need if repeated measures were independent, to equal the information in the actual sample of correlated data. This concept will be derived and formulated in Section 3. It has the advantage of general applicability and provides important insight in the setting of correlated data. Also, the effective sample size can be used, of course with due adaptation, as a component of a novel degrees of freedom determination method. In subsequent sections, we illustrate how the effective sample size can be used in a simple testing situation. The specific cases of normally distributed repeated measures with either compound-symmetric or auto-regressive covariance structures will be considered in detail. Simulations will assess the method's performance, primarily in terms of test size, and compare it to Satterthwaite's method (Satterthwaite, 1941) and to the technique of Kenward and Roger (Kenward and Roger , 1997).

We organize this paper as follows. In Section 2 we introduce the case studies which we use to illustrate the proposed ideas. In Section 3 we explain the concept of the effective sample size, which yields nice insight in the analysis of correlated data, such as the information limit, which is discussed in Section 4. Also, we discuss how the idea of the effective sample size can be used as a method to approximate the number of degrees of freedom in a scaled Wald test. In Section 6, a simulation study exploring the behavior of the proposed method is shown. The case study is analyzed in Section 7. A discussion follows in Section 8.

2 Applications

We motivate and illustrate the ideas in three different, generic settings.

2.1 Cancer of the Ovaries

Our methods will first be illustrated using data from a meta-analysis of two large multicenter trials in advanced ovarian cancer (Ovarian Cancer Meta-Analysis Project, 1991). The trials contain 411 and 382 patients, respectively. The survival time (in years) of individual patients are available in these trials. The endpoint of interest is the logarithm of survival, defined as time (in years) from randomization to death from any cause. The full results of this meta-analysis were published with a minimum follow-up of 5 years in all trials (Ovarian Cancer Meta-Analysis Project, 1991). The dataset was subsequently updated to include a minimum follow-up of 10 years in all trials (Ovarian Cancer Meta-Analysis Project, 1998).

We consider the random intercepts model:

$$T_{ij} = \beta_0 + b_i + \varepsilon_{ij},$$

where T_{ij} is the log-survival time of individual j in trial i. The random intercepts b_i are used to account for the correlation within a trial, and are assumed to follow a normal distribution with mean zero and variance τ^2. It is further assumed that the residual error ε_{ij} is independently normally distributed with mean zero and variance σ^2. We are interested in testing the null hypothesis $H_0 : \beta_0 = 0$. It should be noted that this example is different from a typical longitudinal study, since here only two trials contribute independent information. It will be shown that different estimation methods for the degrees of freedom may lead to severe differences in the resulting p-values.

2.2 The National Toxicology Program Data

The National Toxicology Program (NTP) develops scientific information about potentially toxic chemicals that can be used for protection of public health and prevention of chemically induced diseases. The study considered in this example was performed to investigate the effects of Ethylene glycol (EG) on the developing fetus. Price *et al* (1985) describe a study in which timed-pregnant CD-1 mice were dosed by gavage with EG in distilled water. Dosing occurred during the period of major organogenesis and structural development of the foetuses (gestational days 6 through 15). The doses selected for the study were 0, 750, 1500, or 3000 mg/kg/day, with 25, 24, 23, and 23 timed-pregnant mice randomly assigned to each of these dose groups, respectively. For each viable fetus, the birth weight is recorded, since this is an important indicator of toxicity.

We consider a generalized linear mixed model:

$$Y_{ij} \quad = \quad \beta_0 + b_{0i} + \beta_1 d_i + \varepsilon_{ij}, \tag{1}$$

where Y_{ij} is the birth weight of fetus j in litter i and d_i is the treatment dosage for dam i. Further, litter-specific intercepts b_i are used to account for the clustering of foetuses within litters. They are assumed normally distributed with mean zero and variance τ^2. Further, the residual error ε_{ij} is independently normally distributed with mean zero and variance σ^2. In this example, it is of interest to test for a dose effect:

$$H_0 : \beta_1 = 0.$$

2.3 The Rat Data

The data from this example resulted from a randomized longitudinal experiment (Verdonck *et al*, 1998), in which 50 male Wistar rats were randomized to either a control group or one of the two treatment groups, where treatment consisted of a low or high dose of the testosterone inhibitor Decapeptyl. The treatment started at the age of 45 days, and measurements were taken every 10 days, starting at the age of 50 days. Of interest was skull height, measured as the distance (in pixels) between two well-defined points on X-ray pictures taken under anesthesia. Some rats have incomplete follow-up because they did not survive anesthesia.

Let Y_{ij} denote the response taken at time t_j, for rat i. Similar as in Verbeke and Molenberghs (2000), we model subject-specific profiles as linear functions of $t =$

$\ln(1 + (\mathsf{Age} - 45)/10)$:

$$Y_{ij} \quad = \quad \beta_0 + b_i + \beta_1 t_{ij} + \varepsilon_{ij}. \tag{2}$$

Here, β_0 is the average response at the time of randomization, while β_1 is the average slope in the three different treatment groups. Further, the b_i are rat-specific intercepts, representing natural heterogeneity between rats, relative to baseline values. They are assumed to be zero-mean normally distributed with variance τ^2. The residual error terms ε_{ij} are independently normally distributed with zero mean and variance σ^2. In this example, we are interested in testing the linear time trend

$$H_0 : \beta_1 = 0.$$

In contrast with the previous example, where a cluster-specific effect is tested, here we test for a subject-specific effect.

3 The Effective Sample Size

In this section, the general concept of the effective sample size is explained. Let Y_{ij} be the jth measurement for the ith individual, $i = 1, \dots, N, j = 1, \dots, n_i$. Consider the general Gaussian linear model of the form

$$\boldsymbol{Y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i),$$

with $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{in_i})'$ the n_i dimensional vector with all measurements for subject i, $\boldsymbol{X}_i$ a $n_i \times p$ design matrix, $\boldsymbol{\beta}$ a $p \times 1$ vector of unknown parameters, and $\boldsymbol{V}_i$ a general $n_i \times n_i$ covariance matrix. The design matrix may contain both an intercept as well as subject-specific covariates. $\boldsymbol{V}_i$ can be left unstructured or assumed to be of a specific parametric form.

The amount of information in the data to estimate a fixed-effects parameter can now be represented conveniently by the number of independent measurements one would need to reach the same amount of information. We define this as the effective sample size $\widetilde{N}$. We estimate the effective sample size by comparing the variance of a specific parameter $\boldsymbol{\beta}$, with the variance of this parameter under independence. The fixed-effects parameter $\boldsymbol{\beta}$ can be estimated as (Laird and Ware, 1982):

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{N} X_i' V_i^{-1} \boldsymbol{Y}_i.$$

This is an unbiased estimate for $\boldsymbol{\beta}$ if the mean of the response is correctly specified, even if the variance V_i is misspecified. The variance of $\widehat{\boldsymbol{\beta}}$, provided V_i is properly specified, is equal to

$$\widehat{\mathsf{Var}}(\widehat{\beta}) = \Big(\sum_{i=1}^{N} X_i' V_i^{-1} X_i \Big)^{-1}.$$

Under the assumption of an independent sample, this variance would be estimated as

$$\widetilde{\mathsf{Var}}(\widehat{\beta}) = \Big(\sum_{i=1}^{N} X_i' W_i^{-1} X_i \Big)^{-1},$$

with W_i a diagonal matrix. By assuming that the variance under independence is equal to the true variance, we have that the effective sample size is equal to

$$\widetilde{N} = \sum_{i=1}^{N} \Big[J_{n_i}' (W_i^{-1/2} V_i W_i^{-1/2})^{-1} J_{n_i} \Big]^{-1} = \sum_{i=1}^{N} \Big(J_{n_i}' C_i^{-1} J_{n_i} \Big)^{-1},$$

with C_i the correlation matrix and J_{n_i} an $n_i \times 1$ vector consisting of ones. Derivation of this expression is given in Appendix A. Note that it is valid for both an intercept and a subject-specific covariate, since, by equating the variances, the subject-specific covariate x_i drops from the equation. These considerations also indicate the definition is specific for the parameter being considered. It can be more general, for individual- and/or measurement-specific covariates, to

$$\widetilde{N} = \sum_{i=1}^{N} \Big(X_i' W_i^{-1} X_i \Big)^{-1/2} \Big(X_i' V_i^{-1} X_i \Big) \Big(X_i' W_i^{-1} X_i \Big)^{-1/2}.$$

Note that this expression is valid for the intercept and depends on the design, through the design matrix X_i.

3.1 Compound-symmetry Structure

We specialize the idea of the effective sample size to the simple but important context of a continuous response Y on a set of measurements j which are grouped in a cluster

i of size n. Assume the random-intercepts model

$$Y_{ij} = \beta + b_i + \varepsilon_{ij}, \tag{3}$$

where b_i and ε_{ij} are normally distributed with mean zero and variance τ^2 and σ^2, respectively. In this case, we have that $V_i = \tau^2 J_n + \sigma^2 I_n$. The regression parameter β can be estimated as (Laird and Ware, 1982):

$$\widehat{\beta} = \frac{1}{Nn} \sum_{i=1}^{N} \sum_{j=1}^{n} Y_{ij}$$

and the variance of $\widehat{\beta}$ equals

$$\widehat{\mathsf{Var}}(\widehat{\beta}) = \frac{\sigma^2 + n\tau^2}{Nn}. \tag{4}$$

Assuming that measurements are independent, we would have that $W_i = (\sigma^2 + \tau^2) I_n$ and the variance of $\widehat{\beta}$ would equal

$$\widetilde{\mathsf{Var}}(\widehat{\beta}) = \frac{\sigma^2 + \tau^2}{Nn}. \tag{5}$$

The effective sample size $\widetilde{n}$ can be calculated by equating (4) and (5):

$$\frac{\sigma^2 + n\tau^2}{Nn} = \frac{\sigma^2 + \tau^2}{N\widetilde{n}},$$

yielding

$$\widetilde{n} = \frac{n}{1 + \rho(n-1)}, \tag{6}$$

with $\rho = \tau^2/(\tau^2 + \sigma^2)$.

In general, when cluster sizes are not equal, the correction required for the effective sample size is different for different cluster sizes. Thus, the effective sample size $\widetilde{N}$ for the entire sample equals $\sum_i \widetilde{n}_i$, yielding

$$\widetilde{N} = \sum_i \frac{n_i}{1 + \rho(n_i - 1)}. \tag{7}$$

In Table 1, the effective sample size for clusters of size $n = 5$ is presented for different correlations ρ. For example, if $\rho = 0.2$ then the information obtained from

Table 1: Effective sample size for a cluster of size n with correlation ρ, calculated under the CS-model and under the AR(1)-model.

ρ	n	$\widetilde{n}$(CS)	$\widetilde{n}$(AR(1))	ρ	n	$\widetilde{n}$(CS)	$\widetilde{n}$(AR(1))
0	5	5	5	0.5	1	1	1
0.2	5	2.8	3.7	0.5	2	1.33	1.33
0.4	5	1.9	2.7	0.5	5	1.67	2.33
0.6	5	1.5	2.0	0.5	10	1.82	4
0.8	5	1.2	1.4	0.5	100	1.98	34
1	5	1	1	0.5	∞	2	∞

$n = 5$ measurements on the same individual is similar to what would be obtained from 2.8 independent measurements. There are some interesting special cases. When measurements are independent within a cluster ($\rho = 0$), the effective sample size equals $\widetilde{N} = \sum_i n_i$, the total number of measurements. In case the measurements within a cluster are perfectly correlated ($\rho = 1$), the effective sample size equals the number of clusters, since $\widetilde{N} = \sum_i \frac{n_i}{n_i} = N$. Further, Table 1 shows the effective sample size for different cluster sizes (n) and within-correlation $\rho = 0.5$. The effective sample size increases very slowly with growing cluster size. This will be discussed further in Section 4.

Note that above derivations are valid for non-negative correlations. The effective sample size is positive only, and hence well-defined, for correlations $\rho > -1/(n-1)$. Thus, our argument can be used for mildly negative correlation, down to this bound. Negative correlations are fully acceptable in case one merely seeks a marginal interpretation of model 3. Values below this bound do not correspond to valid distributions any longer. Notwithstanding this, when a fully hierarchical interpretation is adopted, then negative correlation is not allowable (Verbeke and Molenberghs, 2000).

3.2 Some Other Covariance Structures

Let us consider the effective sample size for some other correlation structures as well. When the independence correlation structure applies, the effective sample size reduces to

$$\widetilde{N} = \sum_{i=1}^{N} n_i,$$

as would be expected.

A first-order autoregressive model, assuming that the covariance between two measurements Y_{ij} and Y_{ik} is of the form $\sigma^2\rho^{|k-j|}$, has an effective sample size

$$\widetilde{N} = \frac{n-(n-2)\rho}{1+\rho}N.$$

In Table 1, the effective sample size for clusters of size $n = 5$ is presented for various correlations ρ, as well as for a cluster of different size n but fixed correlation $\rho = 0.5$. For example, if $\rho = 0.2$, then the information obtained from $n = 5$ measurements on the same individual is similar to what would be obtained from 3.7 independent measurements, which is larger than the effective sample size when a compound-symmetry structure would apply. When $\rho = 0$ the effective sample size reduces to the number of measurements. When $\rho = 1$ the effective sample size is equal to the number of clusters. In the special cases that $n = 1$, $n = 2$, $\rho = 0$ and $\rho = 1$, the CS and AR(1) cannot be distinguished, and hence also the effective sample sizes for both settings are the same. Finally, note that the effective sample size increases faster with growing cluster sizes, as compared with the compound symmetry structure.

Next, assume the following linear three-level model:

$$Y_{ijk} = \beta_0 + u_i + v_{ij} + \varepsilon_{ijk},$$

where β_0 is a fixed-effects parameter, u_i is a random effect at the third level ($i = 1,\ldots,N$), u_{ij} is a random effect at the second level ($j = 1,\ldots,J$), and ε_{ijk} is an error term ($k = 1,\ldots,K$). All random terms in the model are assumed to be mutually independent and normally distributed: $u_i \sim N(0,\sigma_u^2)$, $v_{ij} \sim N(0,\sigma_v^2)$, and $\varepsilon_{ijk} \sim$

$N(0, \sigma_\varepsilon^2)$. In this case, the effective sample size is equal to

$$\widetilde{N} = \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{1}{(n_j n_i - 1)\rho_1 + (n_i - 1)\rho_2 + 1},$$

with n_i the number of individuals in group i, n_j the number of measurements on subject j, ρ_1 and ρ_2 the intra-class correlations within a group and within a subject, respectively.

3.3 Contrast Parameter

Thus far, the effective sample size for the overall mean parameter was calculated for different covariance structures. Here, we will focus rather on a contrast. Consider the following random-intercepts model:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \varepsilon_{ij},$$

where b_i and ε_{ij} are normally distributed with mean zero and variances τ^2 and σ^2, respectively. Consider the simple setting where there are 2 measurements for each individual i, i.e., $j = 1, 2$. The covariate x_{ij} can be either measurement- or individual-specific. When assuming independence among the fixed effect parameters, the effective sample size for β_1 is equal to

$$\widetilde{N} = \left(1 + \frac{(x_{i0} - x_{i1})^2}{(x_{i0}^2 + x_{i1}^2)} \frac{\tau^2}{\sigma^2}\right) . \frac{N}{1 + (2-1)\rho}. \tag{8}$$

In case the covariate x_{ij} is individual-specific, (8) reduces to

$$\widetilde{N} = \frac{N}{1 + (2-1)\rho},$$

which is equal to the effective sample size for the overall mean parameter, which is in line with intuition. Then, once again, the higher the correlation among the measurements, the smaller the effective sample size will be. However, when x_{ij} is measurement-specific with, for example, $x_{i0} = 0, x_{x1} = 1$, such as in a pre-test post-test design, the parameter β_1 becomes a contrast parameter, and the effective sample size reduces becomes

$$\widetilde{N} = \frac{N}{(1 + (2-1)\rho)(1 - \rho)}.$$

When the correlation is $\rho = 0$, the effective sample size for the contrast parameter is equal to 1, $\rho = 0.5$ yields an effective sample size of 3, and when measurements are perfectly correlated, i.e., $\rho = 1$, the effective sample size reaches infinity, meaning that one pair of measurements corresponds to the asymptotic situation of perfect knowledge about the contrast. Thus clearly, and again in line with intuition, for a contrast parameter, the larger the correlation in the data is, the larger the amount of information about this contrast will be.

4 Information Limit

Let us now focus on a mean parameter. When data are independent, more measurements yield more information, and such information grows unboundedly with sample size, an intuitive and well-known result. This is important since this implies that a better accuracy can be obtained by gathering more data. It is worth considering whether this is also the case for correlated measurements. We see that in terms of the effective sample size a larger sample size is needed with correlated data to achieve the same accuracy as compared with independent data. The larger the sample size, the more information and the better the accuracy. In some situations, however, there is a limit to this information. For example, when measurements within a cluster are exchangeable, the information limit equals

$$\lim_{n_i \to \infty} \frac{n_i}{1 + \rho(n_i - 1)} = \frac{1}{\rho}. \tag{9}$$

This implies that, when a compound-symmetry structure applies, there is a maximum amount of information. Only when observations are independent ($\rho = 0$), is this limit reached. For example, when $\rho = 0.2$, the limit is equal to 5; hence a cluster can never contribute more information for the fixed-effects parameters than would be obtained from 5 independent measurements. Similarly, when $\rho = 0.5$, a cluster cannot contribute more information than from 2 independent measurements. This implies that there are no conventional asymptotic arguments possible for $n \to \infty$ in such cases.

In contrast to the compound symmetry structure, the information limit is infinite when observations follow an AR(1) covariance structure, since

$$\lim_{n_i \to \infty} \frac{n_i - (n_i - 2)\rho}{1 + \rho} = \infty, \tag{10}$$

Thus, depending on the covariance structure, the amount of information is different. This is an important feature of CS and AR(1) models, which needs to be considered when designing experiments. The contrast between (9) and (10) is dramatic in this respect.

5 Degrees of Freedom in Wald Tests

When comparing a Wald test statistic with a normal distribution, the variance of the parameter of interest is known. However, typically, the variance of $\widehat{\beta}$ is derived from an estimate of the variance-covariance matrix $\boldsymbol{V}$. In this way, approximate Wald-type tests for parameters β can easily be constructed. Since the standard errors are obtained by replacing the variance components by their ML or REML estimates and therefore underestimate the true variability in $\widehat{\beta}$, one often uses t- or F-distributions, where the denominator degrees of freedom need to be estimated from the data. Several methods are available for estimating the appropriate number of degrees of freedom needed for the specific t- or F-test: Satterthwaite's approximation (Satterthwaite, 1941), and the Kenward and Roger (Kenward and Roger , 1997) approximation. Note that these methods are only fully developed for the case of linear mixed models and related multivariate normally based models. In the analysis of longitudinal data, subjects contribute independent information, which results in numbers of degrees of freedom which are typically large enough, whatever estimation method is used, to lead to very similar p-values. However, for parameters about which there is very little information, the different estimation methods for degrees of freedom may lead to severe differences in the resulting p-values. An important example is when a small number of clinical trials, all encompassing a large number of patients, are combined into a meta-analysis, i.e., N small and n_i large.

Suppose that inference is to be made about a single parameter β from the fixed-effects structure. The Wald statistic, taking the form

$$T = \frac{\widehat{\beta}}{\sqrt{\widehat{\mathsf{Var}}(\widehat{\beta})}}, \tag{11}$$

can be used to test the null hypothesis $H_0 : \beta = 0$. Here, a novel method to test whether a significant effect exists, is proposed, with the effective sample size an important building block in the degrees of freedom approximation of a scaled Wald test. It is assumed that a scaled form $T^* = \lambda T$ of the T-statistic follows a t-distribution with ν degrees of freedom, where λ and ν are unknown quantities. This is similar to the method proposed by Kenward and Roger, where a scaled form of the F-statistic is used.

Derivation of the scale factor λ follows from matching the first two moments of T^* with the moments of a t-distribution, leading to

$$\lambda^2 \;\; = \;\; \frac{\nu}{(\nu-2)V(T)}, \tag{12}$$

with $V(T)$ the variance of the t-statistic T. The variance $V(T)$ can be approximated by use of the multivariate delta method. Derivation of $V(T)$ is given in Appendix B. The degrees of freedom ν are calculated from the data, by assuming that it is equal to the degrees of freedom for a similar but independent set of data. Note that the degrees of freedom in an independent data set are given by the sample size minus the number of parameters to be estimated in the fixed-effects structure. If we denote the effective sample size, as derived in Section 3, by $\widetilde{N}$ and the numbers of parameters to be estimated in the fixed effects structure as ℓ, then we have

$$\nu \;\; = \;\; \widetilde{N} - \ell. \tag{13}$$

For the random-intercepts model with a compound-symmetry structure, this leads to

$$\lambda^2 = \frac{\nu}{(\nu-2)V(T)} \qquad \text{where} \qquad \nu \;\; = \;\; \sum_{i=1}^{N} \frac{n_i}{1+(n_i-1)\rho} - \ell, \tag{14}$$

which are straightforward to compute.

The major difference between the proposed method, and Satterthwaite's or Kenward-Roger's method, is that in the latter methods the degrees-of-freedom are calculated directly from approximating the distribution for the Wald tests of the individual parameter estimates. The proposed method is more general, in the sense that the concept of the effective sample size is not restricted to a normally distributed response.

Note that the scaled Wald test is defined only when the degrees of freedom are larger than 2, since the variance of the t-distribution is infinite otherwise. Therefore, in case the calculated degrees of freedom are less than 2, no scaling is applied to the test statistic. This is similar to the Kenward-Roger methodology as implemented in the SAS procedure MIXED. Furthermore, a lower bound of 1 on the degrees of freedom is assumed in the Kenward-Roger methodology.

6 A Simulation Study

A simulation study was conducted to explore the behavior of the method as proposed in previous sections, and to compare the proposed methodology with (i) the unadjusted test, which uses the t-distribution with the number of measurements minus the number of estimated parameters as the degrees of freedom, (ii) the Satterthwaite method and (iii) the Kenward-Roger method. Three different settings were used. We present the results from each of these in turn.

6.1 Interest in Mean of Exchangeable Correlated Data

In a first simulation study, we generate data from compound symmetry model (3) with an unbalanced design and study the $t-$test for the overall mean. We vary the mean cluster size, number of clusters, and the intra-class correlation $\rho = \tau^2/(\sigma^2+\tau^2)$ to investigate the performance of the proposed method in different situations.

For each setting, $10,000$ sets of data are simulated from the compound symmetry model, with zero mean, and for each set the fixed effects are estimated together with the REML variance estimates of the variance components. Tables 2 and 3, display the

Table 2: Simulation Study (part 1). Mean of Exchangeable Correlated Data: mean estimated effective sample size ($\widetilde{N}$) and mean of scale parameter (λ), and observed size of nominal 5% Wald t-test from the simulation study. (Unadj: unadjusted; Satterth: Satterthwaite; KR: Kenward-Roger; effSS: effective sample size.)

Nr.	Mean				Observed Size			
clusters	cluster size	ρ	$\widetilde{N}$	λ	Unadj	Satterth	KR	effSS
10	4	0.2	19.77	1.02	11.0	5.5	5.5	5.0
		0.5	10.69	1.09	13.0	5.4	5.4	4.3
		0.8	6.07	1.26	11.9	4.7	4.7	3.7
10	3	0.2	16.07	1.05	12.9	7.5	7.5	7.3
		0.5	9.66	1.20	16.2	6.0	6.0	4.9
		0.8	5.66	1.73	17.0	5.1	5.0	2.8
10	2	0.0	18.16	1.05	3.7	3.4	3.4	3.4
		0.2	16.72	1.05	5.3	4.6	4.7	4.6
		0.5	14.02	1.06	6.6	5.0	5.0	5.2
100	4	0.2	230.13	1.00	4.9	4.6	4.6	4.7
		0.5	153.24	1.00	5.0	4.9	4.9	5.0
		0.8	115.92	1.01	5.1	4.9	4.9	4.9
100	3	0.2	191.53	1.00	5.6	5.4	5.4	5.5
		0.5	140.31	1.00	5.5	5.3	5.3	5.3
		0.8	112.53	1.01	5.4	5.2	5.2	5.2
100	2	0.2	147.29	1.00	5.5	5.3	5.3	5.3
		0.5	122.75	1.01	5.8	5.6	5.6	5.6
		0.8	107.39	1.01	5.6	5.4	5.4	5.4

observed size of a nominal 5% t-test, using an unadjusted t-test, the Satterthwaite test, and the Kenward-Roger test. Also, we show the observed size of the proposed effective sample size t-test, together with the average effective sample size and the average scale factor which was used in the proposed scaled Wald test.

In a typical longitudinal setting, the number of clusters is larger than the number of observations. Simulation results where data are generated under such a setting are presented in Table 2. The behavior of the proposed method is generally quite good, with an observed size close to the nominal level. The proposed method is comparable to the Satterthwaite and Kenward-Roger method. Note also that the effective sample size decreases with increasing intra-class correlation. The scale parameter is always close to 1. When the number of clusters is large, all methods perform equally well since all approximate t-statistics are close to each other.

In a typical meta-analytic setting, as in the first example, one encounters a small number of clusters (trials) combined with a large sample size within clusters (number of patients per trial). Table 3 presents the results of such a simulation study. In most settings, the proposed method works well. However, when both the correlation is large and the number of clusters is very small, the proposed method tends to deteriorate. Note that this is the situation where there is very little information in the data. Also, when we have only 2 clusters and each clusters contains about 10 to 100 observations, it is observed that the scale-parameter can become infinite, especially when the correlation is large. This is due to the infinite variance of a t-distribution when the number of degrees of freedom is smaller then 2. This might occur in situations when there is only a very small amount of information in the data, with an effective sample size smaller then 3.

6.2 Interest in the Mean of AR(1)-Correlated Data

Next, we consider a longitudinal study with a balanced design and an AR(1) correlation structure. Again, interest is in a test for the overall mean of the response. Various settings for cluster size, number of clusters and correlation are considered and results are summarized in Table 4

Also in this setting, the proposed method works very well. Note that the effective sample size under the AR(1) model is much higher when compared to its counterpart under compound-symmetry, in line with our theoretical developments. Thus, the same number of measurements leads to a different amount of information, owing to a differ-

Table 3: Simulation Study (part 2). Mean of Exchangeable Correlated Data: mean estimated effective sample size ($\widetilde{N}$) and mean of scale parameter (λ), and observed size of nominal 5% Wald t-test from the simulation study. (Unadj: unadjusted; Satterth: Satterthwaite; KR: Kenward-Roger; effSS: effective sample size.)

Nr.	Mean				Observed Size			
clusters	cluster size	ρ	$\widetilde{N}$	λ	Unadj	Satterth	KR	effSS
4	400	0.2	49.56	0.95	14.7	5.4	5.4	5.1
		0.5	15.96	1.05	14.8	5.3	5.3	4.8
		0.8	6.89	1.23	15.1	5.4	5.4	4.6
2	400	0.2	118.70	0.92	31.0	6.8	6.8	6.4
		0.5	57.42	1.12	32.0	4.2	4.2	4.0
		0.8	29.38	1.10	31.9	4.6	5.2	4.0
4	100	0.2	42.18	0.96	14.6	5.3	5.3	4.9
		0.5	14.83	1.06	14.8	5.1	5.1	4.4
		0.8	6.97	1.23	14.8	5.0	5.0	4.2
3	100	0.0	255.93	0.98	4.0	3.4	3.4	3.4
		0.2	45.31	0.94	18.7	5.5	5.5	2.9
		0.5	17.75	1.11	18.8	4.9	4.9	1.4
		0.8	7.73	1.61	18.3	4.9	4.9	1.0
2	100	0.0	169.43	0.98	3.5	3.2	3.2	3.2
		0.2	56.53	0.99	27.9	11.0	11.0	11.0
		0.5	31.45	1.18	29.5	7.5	7.5	7.5
		0.8	16.93	1.09	29.9	5.7	5.7	5.7
4	10	0.2	19.89	1.02	10.7	5.6	5.6	5.3
		0.5	10.88	1.10	12.9	5.0	4.9	4.2
		0.8	6.21	1.26	12.9	4.7	4.7	3.9
3	10	0.0	27.00	1.03	4.32	3.7	3.7	3.8
		0.2	16.04	1.05	12.1	6.8	6.8	6.6
		0.5	9.67	1.21	16.0	5.8	5.8	4.8
		0.8	5.64	1.74	16.9	4.8	4.8	2.6
2	10	0.0	17.23	1.05	4.2	3.7	3.7	3.7
		0.2	12.09	1.20	14.6	10.5	10.5	10.7
		0.5	8.65	1.30	22.2	12.2	13.2	12.2
		0.8	5.80	1.12	26.2	9.3	9.3	11.1

Table 4: Simulation Study. Mean of AR(1)-Correlated Data: estimated effective sample size ($\widetilde{N}$) and scale parameter (λ), and observed size of nominal 5% Wald t-test from the simulation study. (Unadj: unadjusted; Satterth: Satterthwaite; KR: Kenward-Roger; effSS: effective sample size.)

Nr.	Mean				Observed Size			
clusters	cluster size	ρ	$\widetilde{N}$	λ	Unadj	Satterth	KR	effSS
3	10	0.2	22.98	1.02	7.0	4.7	5.2	4.8
		0.5	13.53	1.05	8.4	4.7	5.9	4.8
		0.8	7.12	1.13	11.7	4.9	6.4	3.9
10	3	0.2	25.20	1.02	6.7	5.2	5.4	5.5
		0.5	17.76	1.04	6.9	5.1	5.5	5.0
		0.8	12.68	1.06	7.1	4.9	5.3	4.1
3	100	0.2	203.02	1.00	6.0	5.6	5.7	5.7
		0.5	103.35	1.00	6.2	5.4	5.8	5.6
		0.8	37.06	1.01	6.7	5.2	5.7	5.2
100	3	0.2	235.07	1.00	6.0	5.8	5.9	5.9
		0.5	167.67	1.00	6.0	5.8	5.9	5.9
		0.8	122.63	1.00	5.9	5.7	5.7	5.6

ence in correlation structure. Recall that, while the information limit for the compound symmetry covariance structure is finite, it is infinite for the AR(1) correlation structure. We also see that the scale parameter λ is virtually 1 when the cluster size is large or when the number of clusters is large.

6.3 Interest in a Dose Parameter

In this simulation, we generate data from the model

$$Y_{ij} = \beta_0 + \beta_1 d_i + b_i + \varepsilon_{ij},$$

for $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, $b_i \sim N(0, \tau^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, and $d_i = 0, 1$ depending on whether the unit belongs to the control or active dose group. For each setting, 5000

Table 5: Simulation Study. CS-Correlated Data: estimated effective sample size ($\widetilde{N}$) and scale parameter (λ), and observed size of nominal 5% Wald t-test from the simulation study for intercept and dose parameter. (Unadj: unadjusted; Satterth: Satterthwaite; KR: Kenward-Roger; effSS: effective sample size.)

	Nr.	Mean				Observed Size			
Effect	clust	clust size	ρ	$\widetilde{N}$	λ	Unadj	Satterth	KR	effSS
Dose	20	10	0.2	75.42	1.00	6.1	4.9	4.9	5.1
			0.33	52.61	1.01	6.5	5.1	5.1	5.3
			0.5	37.90	1.02	6.4	5.0	5.0	5.1
	10	5	0.2	29.95	1.02	7.0	4.5	4.5	4.7
			0.33	23.22	1.02	7.6	5.0	4.8	5.3
			0.5	18.06	1.04	7.9	5.0	4.8	5.1
Int.	20	10	0.2	75.42	1.00	5.6	4.2	4.2	4.5
			0.33	52.61	1.01	5.7	4.4	4.4	4.6
			0.5	37.90	1.02	5.8	4.4	4.4	4.7
	10	5	0.2	28.95	1.01	6.9	4.5	4.5	4.7
			0.33	23.22	1.02	7.4	4.7	4.6	4.9
			0.5	18.06	1.04	7.7	4.7	4.6	4.8

sets of data are simulated from this compound symmetry model, and for each set the fixed effects are estimated together with the REML variance estimates of the variance components. The results are displayed in Table 5.

We observe a very good behavior of the effective sample size method, in line with Satterthwaite and Kenward-Roger, and oftentimes slightly outperforming these. All of these three methods are definitely superior to the unadjusted approach.

7 Analysis of Applications

7.1 Cancer of the Ovaries

The mean log-survival time is estimated to be 0.7906 (s.e. 0.1726). The residual error degrees of freedom, equal to the number of individuals minus the number of parameters to be estimated, equals 792, resulting in $p < 0.0001$ for the t-test corresponding to the null hypotheses of one-year survival. The correlation of individuals within a trial is estimated to be 0.038. This correlation should be accounted for in the analysis, and has a huge effect on the degrees of freedom. Both Satterthwaite and Kenward-Roger estimate the degrees of freedom equal to 1, resulting in a $p = 0.1368$. Thus, our previous conclusions on the one-year survival do no longer hold. Finally, we estimate the effective sample size based on the estimated correlation and the sample sizes of the two trials, as 49. As a result, the scaled t-test, with scale parameter 0.30, has 48 degrees of freedom, resulting in a p-value of 0.173.

Further, it should be noted that, due to the correlation ($\rho = 0.038$) in the trials, a trial cannot obtain more information to estimate the mean parameter as corresponds to about 26 independent measurement.

7.2 The National Toxicology Program Data

The EG data contain 111 dams with a total of 1368 foetuses. The dose-effect on fetal birth weight is estimated as -0.2099 (s.e. 0.017). The intra-class correlation equals 0.61. In this setting, all methods for the approximation of the degrees of freedom yield $p < 0.0001$. However, the estimated degrees of freedom are very different. The residual degrees of freedom equal 1366, Satterthwaite degrees of freedom are estimated as 105, Kenward-Roger degrees of freedom are 105, and the effective sample size degrees of freedom is amount to 171. Based on the effective sample size, we can show that the asymptotic information limit for this setting equals 182.35.

7.3 The Rats Data

In the rats data, the effect of time is estimated as 0.1934 (s.e. 0.0059). The residual degrees of freedom are 250. The Satterthwaite and Kenward-Roger method estimate the degrees of freedom as 214, while the effective sample size equals 98.48. All methods result in a significant t-test statistic, with $p < 0.0001$.

8 Concluding Remarks

We have introduced the concept of effective sample size. It is, broadly speaking, the sample size needed in independent data, to retrieve the amount of information obtained from a dependent sample. In a number of cases, such as the linear mixed model or clustered binary data, the expressions are simple and insightful.

The concept of effective sample size has led us to two important ramifications. First, an information limit can be constructed: the amount of information retrieved from a subject or cluster, when the cluster size goes to infinity. For example, it is finite for the compound-symmetry case, while it is infinitely large under AR(1) assumptions. Second, the effective sample size concept can be employed to construct an alternative degrees-of-freedom calculation method. For the Gaussian case, it is similar in behavior to the well-established, time-honored Satterthwaite and Kenward-Roger methods. This, rather than replacing these methods, our approach adds intuition and insight. Importantly, unlike Satterthwaite and Kenward-Roger, our method can also be applied in important non-Gaussian settings, such as clustered binary data.

Other promising application of the effective sample size are in sample size calculations and in the use of information criteria (Faes *et al*, 2004). Use of the effective sample size in non-Gaussian settings is topic of current research.

References

Cox, D. R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Faes, C., and Aerts, M., and Geys, H., and Molenberghs, G. and Declerck, L. (2004). Bayesian testing for trend in a power model for clustered binary data. *Environmental and Ecological Statistics* **11**, 305–322.

Giesbrecht, G.F. and Burns, J.C. (1985) Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results. *Biometrics* **41**, 853–862.

Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall.

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

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–974.

Ovarian Cancer Meta-Analysis Project (1998). Cyclophosphamide plus cisplatin versus cyclophosphamide, doxorubicin, and cisplatin chemotherapy of ovarian carcinoma: a meta-analysis. *Classic Papers and Current Comments* **3**, 237–243.

Ovarian Cancer Meta-Analysis Project (1991). Cyclophosphamide plus cisplatin versus cyclophosphamide, doxorubicin, and cisplatin chemotherapy of ovarian carcinoma: a meta-analysis. *Journal of Clinical Oncology* **9**, 1668–1674.

Price, C.J. and Kimmel, C.A. and Tyl, R.W. and Marr, M.C. (1985). The developmental toxicity of ethylene glycol in rats and mice. *Toxicology and Applied Pharmacology* **81**, 113–127.

Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika* **6**, 309–316.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Verdonck, A., De Ridder, L., Verbeke, G., Bourguignon, J.P., Carels, C., Kuhn, E.R., Darras, V., and de Zegher, F. (1998). Comparative effects of neonatal and prepubertal castration on craniofacial growth in rats. *Archives of Oral Biology*, **43**, 861–871.

Appendix A: Derivation of Effective Sample Size

For simplicity, we focus on a Gaussian response. The information contributed by a subject i, regarding the estimation of a fixed effects parameter β, is contained in the variance $\mathsf{Var}(\widehat{\beta})$:

$$\widehat{\mathsf{Var}}(\widehat{\beta}) = \Big(\sum_{i=1}^{N} X_i' V_i^{-1} X_i\Big)^{-1}.$$

In case the experiment was conducted with independent measurements on each individual, the corresponding quantity would be:

$$\widetilde{\mathsf{Var}}(\widehat{\beta}) = \Big(\sum_{i=1}^{N} X_i^{*'} W_i^{-1} X_i^{*}\Big)^{-1},$$

with W_i a diagonal matrix and X_i^* a $\widetilde{n} \times p$ design matrix. Equating both yields

$$\Big(\sum_{i=1}^{N} X_i' V_i^{-1} X_i\Big)^{-1} = \Big(\sum_{i=1}^{N} X^{*'} W_i^{-1} X_i^{*}\Big)^{-1},$$

satisfied by

$$X_i' V_i^{-1} X_i = X^{*'} W_i^{-1} X_i^{*}, \tag{15}$$

for all $i = 1, \dots, N$. Under the assumption of a homogeneous covariance structure, the diagonal matrix W_i is equal to $\mathsf{Var}(Y_i) I_{\widetilde{n}_i}$. Thus, the right hand side of (15) can be written as

$$X^{*'} W_i^{-1} X_i^{*} = \frac{1}{\mathsf{Var}(Y_i)} X^{*'} I_{\widetilde{n}_i} X_i^{*}. \tag{16}$$

If x_i is the value in the design matrix X_i, corresponding to the parameter β, then (16) can be written as $x_i J_{\widetilde{n}_i}$, with $J_{\widetilde{n}_i}$ an $\widetilde{n}_i \times 1$ vector consisting of ones only. Thus, we have that

$$x_i^2 J_{n_i}' V_i^{-1} J_{n_i} = \frac{x_i^2}{\mathsf{Var}(Y_i)} J_{\widetilde{n}_i}' I_{\widetilde{n}_i} J_{\widetilde{n}_i},$$

or

$$\widetilde{n}_i = J_{n_i}' \Big(\frac{V_i}{\mathsf{Var}(Y_i)}\Big)^{-1} J_{n_i}. \tag{17}$$

If the variance-covariance matrix is not homogeneous, then (17) can be extended to

$$\widetilde{n}_i = J_{n_i}' \Big(W_i^{-1/2} V_i W_i^{-1/2}\Big)^{-1} J_{n_i}.$$

As a result

$$\widetilde{N} = \sum_{i=1}^{N} J'_{n_i} \left(W_i{}^{-1/2} V_i W_i{}^{-1/2} \right)^{-1} J_{n_i}.$$

Appendix B: Derivation of V(T)

We are interested in the variance of the test-statistic T:

$$\widehat{\mathsf{Var}}(T) = \widehat{\mathsf{Var}} \left(\frac{\widehat{\beta}}{\sqrt{\widehat{\mathsf{Var}}(\widehat{\beta})}} \right).$$

Let us derive this in the context of a compound-symmetry model. The parameters in this model are $(\beta, \sigma^2, \tau^2)$. Use the delta method, to calculate the variance of T:

$$\widehat{\mathsf{Var}}\,(T) \quad = \quad \begin{pmatrix} \frac{\partial T}{\partial \beta} & \frac{\partial T}{\partial \sigma^2} & \frac{\partial T}{\partial \beta} \end{pmatrix} \widehat{\mathsf{Var}}(\beta, \sigma^2, \tau^2) \begin{pmatrix} \frac{\partial T}{\partial \beta} \\ \frac{\partial T}{\partial \sigma^2} \\ \frac{\partial T}{\partial \beta} \end{pmatrix},$$

with derivatives equal to:

$$\begin{aligned} \frac{\partial T}{\partial \beta} &= \frac{1}{\sqrt{\widehat{\mathsf{Var}}(\beta)}}, \\ \frac{\partial T}{\partial \sigma^2} &= -\frac{\beta}{2\left(\widehat{\mathsf{Var}}(\beta)\right)^{3/2}} \frac{\partial \widehat{\mathsf{Var}}(\beta)}{\partial \sigma^2}, \\ \frac{\partial T}{\partial \tau^2} &= -\frac{\beta}{2\left(\widehat{\mathsf{Var}}(\beta)\right)^{3/2}} \frac{\partial \widehat{\mathsf{Var}}(\beta)}{\partial \tau^2}. \end{aligned}$$

Since the variance of $\widehat{\beta}$ in the compound-symmetry model is equal to

$$\widehat{\mathsf{Var}}(\widehat{\beta}) = \left(\sum_{i=1}^{N} \frac{n_i}{\sigma^2 + n_i \tau^2} \right)^{-1},$$

the derivatives of $\widehat{\mathsf{Var}}(\widehat{\beta})$ equal:

$$\frac{\partial \widehat{\mathsf{Var}}(\widehat{\beta})}{\partial \sigma^2} = \left(\widehat{\mathsf{Var}}(\widehat{\beta}) \right)^2 \left(\sum_{i=1}^{N} \frac{n_i}{(\sigma^2 + n_i \tau^2)^2} \right), \tag{18}$$

$$\frac{\partial \widehat{\mathsf{Var}}(\widehat{\beta})}{\partial \tau^2} = \left(\widehat{\mathsf{Var}}(\widehat{\beta}) \right)^2 \left(\sum_{i=1}^{N} \frac{n_i^2}{(\sigma^2 + n_i \tau^2)^2} \right). \tag{19}$$

Note that, if all samples sizes are equal, i.e., $n_i \equiv n$, (18) and (19) reduce to

$$\begin{aligned}\frac{\partial \widehat{\mathsf{Var}}(\widehat{\beta})}{\partial \sigma^2} &= \frac{1}{Nn},\\ \frac{\partial \widehat{\mathsf{Var}}(\widehat{\beta})}{\partial \tau^2} &= \frac{1}{N}.\end{aligned}$$

As a result, we obtain

$$\begin{aligned}\frac{\partial T}{\partial \beta} &= \frac{1}{\sqrt{\widehat{\mathsf{Var}}(\widehat{\beta})}},\\ \frac{\partial T}{\partial \sigma^2} &= -\frac{\widehat{\beta}\sqrt{\widehat{\mathsf{Var}}(\widehat{\beta})}}{2}\left(\sum_{i=1}^{N}\frac{n_i}{(\sigma^2+n_i\tau^2)^2}\right),\\ \frac{\partial T}{\partial \tau^2} &= -\frac{\widehat{\beta}\sqrt{\widehat{\mathsf{Var}}(\widehat{\beta})}}{2}\left(\sum_{i=1}^{N}\frac{n_i^2}{(\sigma^2+n_i\tau^2)^2}\right).\end{aligned}$$

Finally, we assume that β and (σ^2, τ^2) are uncorrelated, such that the variance of the test statistic is equal to:

$$\begin{aligned}\widehat{\mathsf{Var}}(T) &= 1+\left(\frac{\partial T}{\partial \sigma^2}\right)^2\widehat{\mathsf{Var}}(\sigma^2)+\left(\frac{\partial T}{\partial \tau^2}\right)^2\widehat{\mathsf{Var}}(\tau^2)+2\left(\frac{\partial T}{\partial \sigma^2}\right)\left(\frac{\partial T}{\partial \tau^2}\right)\widehat{\mathsf{Cov}}(\sigma^2,\tau^2)\\ &= 1+\left(\frac{\widehat{\beta^2\mathsf{Var}(\beta)}}{4}\right)\left\{\left(\sum_{i=1}^{N}\frac{n_i}{(\sigma^2+n_i\tau^2)^2}\right)^2\widehat{\mathsf{Var}}(\sigma^2)\right.\\ &\quad+\left(\sum_{i=1}^{N}\frac{n_i^2}{(\sigma^2+n_i\tau^2)^2}\right)^2\widehat{\mathsf{Var}}(\tau^2)\\ &\quad\left.+2\left(\sum_{i=1}^{N}\frac{n_i}{(\sigma^2+n_i\tau^2)^2}\right)\left(\sum_{i=1}^{N}\frac{n_i^2}{(\sigma^2+n_i\tau^2)^2}\right)\widehat{\mathsf{Cov}}(\sigma^2,\tau^2)\right\}.\end{aligned}$$

In general, we have that

$$\widehat{\mathsf{Var}}(T) = 1+\left(\frac{\widehat{\beta}^2}{4\widehat{\mathsf{Var}}(\widehat{\beta})^3}\right)\widehat{\mathsf{Var}}\left[\widehat{\mathsf{Var}}(\widehat{\beta})\right],$$

with

$$\widehat{\mathsf{Var}}\left[\widehat{\mathsf{Var}}(\widehat{\beta})\right] = \sum_{l}\left(\frac{\partial \widehat{\mathsf{Var}}(\widehat{\beta})}{\partial \sigma_l}\right)^2\mathsf{Var}(\sigma_l)+\sum_{l}\sum_{k\neq l}\left(\frac{\partial \widehat{\mathsf{Var}}(\widehat{\beta})}{\partial \sigma_l}\right)\left(\frac{\partial \widehat{\mathsf{Var}}(\widehat{\beta})}{\partial \sigma_k}\right)\widehat{\mathsf{Cov}}(\sigma_l,\sigma_k)$$

and

$$\left(\frac{\partial \widehat{\mathsf{Var}}(\widehat{\beta})}{\partial \sigma_l}\right) = \widehat{\mathsf{Var}}(\widehat{\beta})^2\left(\sum_{i=1}^{N}X_i'V_i^{-1}\frac{\partial V_i}{\partial \sigma_l}V_i^{-1}X_i\right).$$