Integrated likelihood inference in small sample meta-analysis for continuous outcomes

RUGGERO BELLIO
University of Udine, Italy
and
ANNAMARIA GUOLO
University of Padova, Italy

Abstract

This paper proposes the use of the integrated likelihood for inference on the mean effect in small sample meta-analysis for continuous outcomes. The method eliminates the nuisance parameters given by variance components through integration with respect to a suitable weight function, with no need to estimate them. The integrated likelihood approach takes into proper account the estimation uncertainty of within-study variances, thus providing confidence intervals with empirical coverage closer to nominal levels than standard likelihood methods. The improvement is remarkable when either *i*) the number of studies is small to moderate or *ii*) the small sample size of the studies does not allow to consider the within-study variances as known, as common in applications. Moreover, the use of the integrated likelihood avoids numerical pitfalls related to the estimation of variance components which can affect alternative likelihood approaches. The proposed methodology is illustrated via simulation and applied to a meta-analysis study in nutritional science.

Key words: Frequentist inference; Integrated likelihood; Meta-analysis; Nuisance parameters; Small sample inference.

Running title: Integrated likelihood in meta-analysis

1 Introduction

Meta-analysis is a diffuse approach to combine evidence from different studies about the same issue of interest. The usage of meta-analysis pervades almost any area of research, such as, for example, biological sciences, medicine, epidemiology and, more recently, economics and behavioral investigations (Roberts, 2005; Sutton & Higgins, 2008).

Meta-analysis is typically performed by specifying an appropriate random effects model, with the random component associated to the different studies providing summary information about the common issue of interest. Inference is then carried out by relying on the procedure by DerSirmonian & Laird (1986), traditionally, or on more recent likelihood approaches developed either from a frequentist or a Bayesian perspective (van Houwelingen et al., 2002). The reliability of the inferential conclusions is strictly related to the amount of information available from the meta-analysis studies. This paper investigates the problem of *small sample inference* as a consequence of small sample size for the studies included in the meta-analysis or as a consequence of a small number of studies recruited in the meta-analysis.

A common strategy in meta-analysis assumes that the within-study variances provided by each study are known and equal to the variances associated to the estimates of the mean effect (van Houwelingen et al., 2002, Section 3). The justification is that the sample size of each study is large enough to guarantee a good estimate of the true within-study variance, with little or no impact on the results. Actually, such an assumption is justifiable in case of large studies, as, for example, many medical or epidemiological investigations. Conversely, standard inference performed on studies of small sample size can provide misleading results, if the uncertainty related to variance estimation is not properly taken into account. Several authors pointed out the relevance of the problem, e.g., Hardy & Thompson (1996), Brockwell & Gordon (2001), Sidik & Jonkman (2007), Sánchez-Meca & Marín-Martínez (2008), with the suspicion that consequences could affect the variance estimator of the mean effect and related inferential procedures. Simulations by Böhning et al. (2002) illustrate that the DerSimonian and Laird estimator of the between-study variance can be prone to consider-

able bias when estimates of the within-study variances are employed and Jackson & Bowden (2009) show that changes in the distribution of the within-study variances can notably affect the performance of the quantile approximation method by Brockwell & Gordon (2007). Several solutions have been proposed in the literature to face the problem. Böhning et al. (2002) rely on population-averaged study specific variances, although this is not generally applicable. For meta-analysis of standardized differences, Malzahn et al. (2000) take into account within-study variance estimates when proposing a nonparametric estimation of the between-study variance, while Di Gessa (2008) investigates the use of shrinkage approaches for variance estimation. Johnson & Huedo-Medina (2013) show the advantages of using the standardized mean difference in place of the unstandardized version as a tool to incorporate within-study variances directly in the effect measure. The problem of the estimation of the within-study variances has been recently faced by Sharma & Mathew (2011) with reference to the consensus mean in inter-laboratory studies. Although Sharma & Mathew (2011) never directly refer to meta-analysis and related terminology, the framework they focus on is analogous. For the purpose of investigation on the consensus mean in inter-laboratory studies, Sharma & Mathew (2011) propose to improve on likelihood results by applying higher-order asymptotics via second-order likelihood ratio statistic (Skovgaard, 1996). Nevertheless, the approach can suffer from some computational problems, as illustrated in this paper, requiring a lot of care for its application.

Small sample inference in meta-analysis can also arise as consequence of the limited number of studies recruited, a concern which has been raised by several authors in the literature (e.g., Hardy & Thompson, 1996; Normand, 1999; van Houwelingen et al., 2002). Within a likelihood-based approach, for example, the small number of studies can give rise to inaccurate inferential conclusions when relying on first-order approximations, such as the χ^2 distribution for the likelihood ratio statistic. Guolo (2012) exploits the theory of higher-order asymptotics (Severini, 2000) to refine first-order likelihood solutions in meta-analysis, when the within-study variances are assumed to be known. The attention is paid to the

Skovgaard's second-order statistic, which is implemented within the `R` (R Core Team, 2015) package `metaLik` (Guolo & Varin, 2012).

In this paper we consider the problem of small study inference in meta-analysis for continuous outcomes when information is available as summary data. We suggest to perform the meta-analysis by using the integrated likelihood (Severini, 2000, Section 8.4). The approach replaces the elimination of the nuisance parameters given by variance components through maximization with their elimination by integration. We show that this method provides a good accuracy of inferential results and it is free of numerical pitfalls. The proposed approach is evaluated through a simulation study covering scenarios of practical interest and it is applied to a meta-analysis study in nutritional science.

2 Likelihood inference

Consider a meta-analysis of n independent studies about a common effect β. Let Y_i be the summary measure of β obtained from study i, $i = 1, \ldots, n$, such as, for example, the mean difference. The classical model for meta-analysis is the random effects model (DerSirmonian & Laird, 1986)

$$Y_i = \beta_i + \varepsilon_i,\ \varepsilon_i \sim \text{Normal}(0, \sigma_i^2),$$

where β_i is the random effects component associated to each study,

$$\beta_i = \beta + u_i,\ u_i \sim \text{Normal}(0, \tau^2).$$

Variance components are the within-study variances σ_i^2, $i = 1, \ldots, n$, and the between-study variance τ^2. Thus, marginally, $Y_i \sim \text{Normal}(\beta, \sigma_i^2 + \tau^2)$. The traditional approach to meta-analysis is based on the assumption that each within-study variance σ_i^2 is known and equal to the variance estimate reported in the i-th study. This assumption is justifiable when the sample size of each study included in the meta-analysis is large. Otherwise, inference can provide misleading results, if the uncertainty related to the variance estimation is not properly taken into account. Let S_i^2 denote the measure of the within-study variance σ_i^2

obtained from study i having f_i degrees of freedom, with S_i^2 following a scaled chi-square distribution, $S_i^2 f_i/\sigma_i^2 \sim \chi^2_{f_i}$. For example, f_i is equal to $n_i - 1$ where n_i is the sample size of each study, in case of a single group or a paired t test, or f_i is equal to $n_{i1} + n_{i2} - 2$ in case of a two-group comparison, with n_{i1} and n_{i2} denoting the sample size of each group in study i. When the outcome is derived from the analysis of covariance, then $f_i = n_i - p_i - 1$, if the number of observations n_i and the number of covariates p_i for study i are available.

According to the specifications above, the log likelihood function for the $(n + 2)$-dimensional parameter vector $\boldsymbol{\theta} = (\beta, \tau, \sigma_1, \ldots, \sigma_n)^T$ is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ -\frac{1}{2}\log(\sigma_i^2 + \tau^2) - \frac{1}{2}\frac{(y_i - \beta)^2}{\sigma_i^2 + \tau^2} - \frac{f_i}{2}\log\sigma_i^2 - \frac{f_i s_i^2}{2\sigma_i^2} \right\}. \tag{1}$$

Inferential interest is usually focused on the mean effect β, while variance components are considered as nuisance parameters. Accordingly, we can partition $\boldsymbol{\theta}$ into $\boldsymbol{\theta} = (\beta, \boldsymbol{\lambda})^T$, where $\boldsymbol{\lambda} = (\tau, \sigma_1, \ldots, \sigma_n)^T$. Let $\hat{\boldsymbol{\theta}} = (\hat{\beta}, \hat{\boldsymbol{\lambda}})^T$ denote the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ and let $\hat{\boldsymbol{\lambda}}_\beta$ denote the constrained MLE of $\boldsymbol{\lambda}$ for a given value of β. Let $\ell_{\mathrm{P}}(\beta)$ indicate the corresponding profile log likelihood for β, $\ell_{\mathrm{P}}(\beta) = \ell(\beta, \hat{\boldsymbol{\lambda}}_\beta)$. Inference on β can be based on the signed profile log likelihood ratio statistic

$$r_{\mathrm{P}}(\beta) = \mathrm{sgn}(\hat{\beta} - \beta)\sqrt{2\left\{\ell_{\mathrm{P}}(\hat{\beta}) - \ell_{\mathrm{P}}(\beta)\right\}},$$

which is asymptotically distributed as a standard normal up to first-order error, under mild regularity conditions (Severini, 2000, Section 4.4).

Despite the feasibility, a serious drawback of first-order asymptotic results is that they can be inaccurate in case of large dimension of the nuisance parameter $\boldsymbol{\lambda}$ compared to the available information. In meta-analysis, this corresponds to either small study sizes, which leads to imprecise estimation of the within-study variances, or to a small number of studies, which leads to imprecise estimation of the between-study variance. To face the problem, it is preferable to resort to the theory of likelihood asymptotics (Severini, 2000), which makes several solutions available for the task.

Skovgaard (1996) proposes to base inference on a scalar component of interest on statistic

$$r_{\mathrm{P}}^{*}(\beta) = r_{\mathrm{P}}(\beta) + \frac{1}{r_{\mathrm{P}}(\beta)}\log\frac{u(\beta)}{r_{\mathrm{P}}(\beta)}, \tag{2}$$

which is asymptotically standard normally distributed up to second-order error. The component $u(\beta)$ included in (2) is a function of the observed and the expected information matrices and of covariances of likelihood quantities, evaluated at the MLE and the constrained MLE. Skovgaard's statistic is well defined for a wide class of sufficiently regular parametric models and it is invariant with respect to interest-respecting re-parameterizations. Guolo (2012) investigates the applicability of Skovgaard's statistic in meta-analysis and meta-regression problems, following the convention of assuming known within-study variances. The approach is satisfactory in improving on the accuracy of standard first-order likelihood analysis when the sample size n is small to moderate. Sharma & Mathew (2011) examine the performance of Skovgaard's statistic in inter-laboratory studies where interest relies on the consensus mean, assuming unknown different within-laboratory variances. The simulation studies performed highlight a better accuracy of results based on r_{P}^{*} with respect to its first-order counterpart. The computational difficulties and numerical instabilities encountered by Skovgaard's approach in case of small sample sizes (see Section 5) are mainly due to the fact that the nuisance parameters are eliminated via maximization. The same difficulty is also shared by alternative solutions like the modified profile likelihood (Severini, 2000, Chapter 9). For an illustration, see Vangel & Rukhin (1999), where an example of the profile likelihood for $(\beta, \tau)^T$ exhibiting some local maxima is provided.

A different route is provided by the integrated likelihood (Severini, 2007), which eliminates the nuisance parameters by integration of the likelihood with respect to a weight function. For the model of interest here, the integrated log likelihood function for β is

$$\ell_{\mathrm{Int}}(\beta) = \log \int_{\Lambda} L(\boldsymbol{\theta})\pi(\boldsymbol{\lambda}|\beta)d\boldsymbol{\lambda},$$

where $L(\boldsymbol{\theta}) = \exp\{\ell(\boldsymbol{\theta})\}$, $\pi(\boldsymbol{\lambda}|\beta)$ denotes a weight function for $\boldsymbol{\lambda}$ for fixed β and $\boldsymbol{\lambda} \in \Lambda$. Once the integrated log likelihood is obtained, it can be used as a standard log likelihood function

for inference. For example, let $\bar{\beta}$ be the estimate of β obtained from the maximization of $\ell_{\text{Int}}(\beta)$. Then, inference on β can be performed via the signed integrated log likelihood ratio statistic (Severini, 2010)

$$r_{\text{Int}}(\beta) = \text{sgn}(\bar{\beta} - \beta)\sqrt{2\left\{\ell_{\text{Int}}(\bar{\beta}) - \ell_{\text{Int}}(\beta)\right\}}. \tag{3}$$

Advantages of the integrated likelihood approach include better accuracy of the inferential results if compared with those from r_{P} as well as reduced numerical instabilities in case of large dimension of $\boldsymbol{\lambda}$ (Severini, 2010). The main drawback is the specification of the weight function $\pi(\boldsymbol{\lambda}|\beta)$. Severini (2007) provides several suggestions about how to choose the weight function in order to make the integrated likelihood share the frequentist properties of a genuine likelihood function and be suitable for non-Bayesian inference. Possible choices are discussed in Section 3.

3 Integrated likelihood in meta-analysis

In our context, the parameter space for $\boldsymbol{\lambda} = (\tau, \sigma_1, \ldots, \sigma_n)^T$ is $\Lambda = [0, \infty) \times (0, \infty)^n$, and the integrated log likelihood has the following form

$$\ell_{\text{Int}}(\beta) = \log \int_0^\infty \underbrace{\int_0^\infty \cdots \int_0^\infty}_{\text{n times}} L(\beta, \tau, \sigma_1, \ldots, \sigma_n)\pi(\tau, \sigma_1, \ldots, \sigma_n|\beta)d\sigma_1 \cdots d\sigma_n d\tau. \tag{4}$$

The usage of (4) requires to overcome two main obstacles. The first one is the choice of the weight function for the nuisance parameter vector $\boldsymbol{\lambda}$, for fixed β. The second obstacle pertains to the computation of $\ell_{\text{Int}}(\beta)$.

For the choice of the weight function for $\boldsymbol{\lambda}$, we can follow the recommendations by Severini (2007, 2010). He advocates the use of an orthogonal parameterization of the nuisance parameters and the consequent choice of the weight function for $\boldsymbol{\lambda}$ free of β. From a frequentist perspective, he shows that the best inferential results are achieved when the model parameterization is expressed so that the nuisance parameter is *strongly unrelated* to β. A

nuisance parameter $\boldsymbol{\phi}$ is strongly unrelated to β if

$$E\{\ell_{\boldsymbol{\lambda}}(\beta, \boldsymbol{\lambda}); \beta_0, \boldsymbol{\lambda}_0\}\big|_{(\beta_0, \boldsymbol{\lambda}_0)=(\hat{\beta}, \boldsymbol{\phi})} = 0,$$

where $\ell_{\boldsymbol{\lambda}}$ is the score vector for $\boldsymbol{\lambda}$ and the expected value is computed before the evaluation at $(\beta_0, \boldsymbol{\lambda}_0)^T = (\hat{\beta}, \boldsymbol{\phi})^T$. The function $\boldsymbol{\phi} = \boldsymbol{\phi}(\beta, \boldsymbol{\lambda}; \hat{\beta})$ defines a data-dependent parameterization and $\boldsymbol{\phi}$ is called the *zero-score-expectation parameter*. When such a parameterization is employed, the resulting integrated likelihood is a high-order approximation to the modified profile likelihood (Severini, 2000, Section 9.3), which achieves optimal elimination of the nuisance parameters (Severini, 2007). With the zero-score-expectation parameterization, the choice of the weight function for the nuisance parameter becomes largely inconsequential. With reference to model (1), the nuisance parameter vector $\boldsymbol{\lambda}$ is orthogonal to β, i.e., the corresponding $\beta\boldsymbol{\lambda}$-block of the expected Fisher information is nil. Moreover, parameters β and σ_i are also strongly unrelated; indeed, $\hat{\sigma}_i^2 \doteq s_i^2$. Let $\boldsymbol{\phi} = (\zeta, \delta_1, \ldots, \delta_n)^T$, then

$$\begin{aligned} \tau^2 &= \zeta^2 + (\hat{\beta} - \beta)^2, \\ \delta_i &= \sigma_i, \qquad i = 1, \ldots, n. \end{aligned} \tag{5}$$

Details about the derivation of the zero-score-expectation parameterization are reported in the Supporting Information.

Once a strongly unrelated parameterization for the nuisance parameters is obtained, the weight function for ζ and $\sigma_1, \ldots, \sigma_n$ can be chosen with some liberty. A simple choice is given by separate weights for all the components of ϕ, with $\pi(\zeta) \propto 1$ and $\pi(\sigma_i) \propto 1/\sigma_i^k$, for fixed k. In the following, we set $k = 1$, after checking that different choices of k would lead to similar results. The choice of the weight function, coupled with the algebraic form of the score function for β, $\ell_\beta(\beta, \boldsymbol{\lambda})$, implies that the signed integrated log likelihood ratio statistic $r_{\text{Int}}(\beta)$ in (3) is asymptotically standard normally distributed with high accuracy (Severini, 2010, Section 5). The latter property is shared also by the integrated likelihood computed using the original parameterization, provided that similar weights, free of β, are adopted.

Computation of $\ell_{\text{Int}}(\beta)$ is less demanding than it might seem at first sight. Indeed, under the assumption of independent meta-analysis information, $L(\beta, \tau, \sigma_1, \ldots, \sigma_n)$ in (4) is the product of n similar terms, which can be readily recovered from formula (1). The aforementioned choice of the weight function for $\boldsymbol{\phi}$ with separate components implies that $\ell_{\text{Int}}(\beta)$ can be written as

$$\ell_{\text{Int}}(\beta) = \log \int_0^\infty \left\{ \prod_{i=1}^{n} g_i(\beta, \zeta) \right\} \pi(\zeta) d\zeta, \tag{6}$$

where $g_i(\beta, \zeta) = \int_0^\infty L(\beta, \zeta, \sigma_i)\pi(\sigma_i)d\sigma_i$ and $L(\beta, \zeta, \sigma_i)$ is the likelihood term for study i. In other words, each of the n integrals $g_i(\beta, \zeta)$ as well as the main integral in (6) amount to one-dimensional integrals, that can be approximated via standard numerical methods. In our study, the inner integrals for $g_i(\beta, \zeta)$ in (6) is computed by adaptive Gauss-Kronrod quadrature, using the `C` function `Rdqags`, which is the port to the `R` library of `C` functions of the `QUADPACK` routine `dqags` (Piessens et al., 1983). The outer integral is computed by a standard Gaussian quadrature. The resulting integrated log likelihood is quite a smooth function of β in all the experiments performed and its maximization by means of a derivative-free optimizer is usually not an issue.

4 Cocoa intake and blood pressure reduction

Increasing consumption of sources of polyphenols is recommended by physicians as coadjutant therapy to face hypertension and prevent cardiovascular risks. Taubert et al. (2007) perform a meta-analysis of randomized controlled studies to evaluate blood pressure-lowering effects of cocoa and tea intake, which represent a high proportion of total polyphenol intake in Western countries. We focus on a portion of the data about the effectiveness on lowering diastolic blood pressure after two-weeks of cocoa consumption. Data refer to five studies, with sample size ranging from 21 to 41. The estimate of the effect provided by each study is the mean difference in diastolic blood pressure before and after the cocoa consumption. Figure 1, left panel, reports the forest plot of the data, that is, a graphical display of the

information provided by each study in the meta-analysis. Information includes the estimated mean difference from each study together with the associated 95% confidence interval. The summary estimate obtained from the likelihood analysis based on r_{P} is added.

Figure 1 here

Likelihood approach provides an estimate of the treatment effect equal to -2.799 (s.e. 1.009), which is found to be significant, given the P-value equal to 0.030 associated to r_{P}. The associated 95% confidence interval for the parameter is $(-5.262, -0.397)$. A comparable result is obtained by the standard likelihood approach assuming known within-study variances. The integrated likelihood approach based on the zero-score-expectation parameterization suggests a non-significant effect of cocoa consumption on lowering diastolic blood pressure, with the estimate of the treatment effect equal to -2.805 (s.e. 1.270) and the P-value for the effectiveness of the treatment equal to 0.071. The integrated likelihood accounts for the variability of the estimated within-study variances and the associated 95% confidence interval for the parameter is wider, equal to $(-6.027, 0.349)$. The profile log likelihood function and the integrated log likelihood function are compared in Figure 1, right panel.

5 Simulation studies

The performance of the integrated likelihood has been investigated via simulation.

5.1 Experiment based on the design of cocoa data

As a first experiment, we consider the same setting of the cocoa data and generate 10,000 data sets from the model of Section 2, with parameters equal to the maximum likelihood estimates, namely, $\beta = -2.8, \tau^2 = 4.27, (\sigma_1^2, \ldots, \sigma_5^2)^T = (0.34, 0.86, 1.37, 1.38, 0.29)^T$. Inference on β is based on the signed profile log likelihood ratio statistic r_{P}, Skovgaard's statistic r_{P}^* and their counterparts assuming known within-study variances, $r_{\mathrm{P,k}}$ and $r_{\mathrm{P,k}}^*$, respectively. The solutions are compared to the following specifications of the signed integrated

log likelihood ratio statistic:

- r_{Int}, based on (4) expressed in the original parameterization, with $\pi(\tau) \propto 1$ and $\pi(\sigma_i) \propto 1/\sigma_i$;
- $\tilde{r}_{\text{Int}}$, based on the re-parameterized model using the zero-score-expectation parameter $\boldsymbol{\phi}$, with $\pi(\zeta) \propto 1$ and $\pi(\sigma_i) \propto 1/\sigma_i$;
- $\bar{r}_{\text{Int}}$, based on the re-parameterized model using the zero-score-expectation parameter $\boldsymbol{\phi}$, with $\pi(\zeta) \propto 1$ and $\pi(\sigma_i) \propto 1/\sigma_i$. Here $\hat{\beta}$ in ζ is replaced by the maximizer of (4) expressed in the original parameterization.

The latter choice has the virtue of not requiring the MLE of β, thus bypassing all the numerical problems related to likelihood maximization.

The simulation studies evaluate the empirical one-sided rejection rates for the competing approaches according to different nominal levels. Results are reported in Table 1.

Table 1

The standard first-order statistic r_{P} provides empirical one-sided rejection rates which are far from the target levels, with coverages of confidence intervals substantially below the nominal level. An improvement over first-order results is provided by Skovgaard's statistic, although such an amelioration is not free of pitfalls. From a practical point of view, the evaluation of r_{P}^* suffers from numerical instabilities when estimating the between-study variance in about 10% of the simulation trials. In most of these cases, either the MLE or the constrained MLE of τ is close to zero and the $\boldsymbol{\lambda\lambda}$-part of the observed Fisher information matrices entering the definition of the adjustment $u(\beta)$ in (2) fails to be definite positive. In such cases r_{P}^* is not computable. The same problems are also experienced by Sharma & Mathew (2011) when applying Skovgaard's statistic in inter-laboratory studies with unknown within-laboratory variances. They suggest some practical measures to face the computational difficulties related to the evaluation of r_{P}^* in case of limited data. For example, the observed information matrix,

when not positive definite, can be substituted with a positive quantity, e.g., the expected information matrix. After such a careful computation, the resulting statistic is always well defined, though the theoretical consequences of the various modifications are not clear. As a final note, the $r^*_{\rm P}$ statistic can be unstable when the value under testing is close to $\hat{\beta}$, thus requiring some further adjustments. See, for example, the discussion in Fraser et al. (2003).

The results provided by $r^*_{\rm P,k}$ are much more satisfactory than those from $r_{\rm P}$ and $r_{\rm P,k}$, showing that for study size between 21 and 41 the effect of taking the within-study variances as fixed is minor.

The use of the integrated likelihood approach provides a substantial improvement of the results accuracy with respect to $r_{\rm P}$, and overall it is the most accurate solution. Moreover, the application of the approach is not affected by any computational inconvenience, especially when $\bar{r}_{\rm Int}$ is employed. Empirical rejection rates are close to the target levels, with no appreciable difference among the integrated likelihood specifications, see Table 1.

5.2 Experiments based on a planned design

For a more systematic investigation, we design a study with three experimental factors, namely, *i*) the number of studies $n \in \{5, 20\}$; *ii*) the study degrees of freedom $f_i \in \{9, 24\}$; *iii*) the between-study variance $\tau^2 \in \{0.1, 0.5, 2\}$. For each combination of the experimental factors, 10,000 data sets are generated following the model specification given in Section 2 with parameters $\beta = 1$ and σ_i^2 generated from a Uniform variable on $[0.1, 2.0]$. Inference on β is performed using statistics $r_{\rm P}$, $r^*_{\rm P,k}$ and the three statistics based on the integrated likelihood introduced in §5.1. Skovgaard's statistic $r^*_{\rm P}$ is not used for comparison because of the overwhelming percentage of datasets with numerical problems encountered in the simulation (up to over 70% for some of the experiments), which makes the results unreliable. Statistic $r_{\rm P,k}$ is not considered as well, as in the previous simulation study the adjusted version $r^*_{\rm P,k}$ turned out to be uniformly preferable. The simulation study evaluates the empirical coverages of the confidence intervals at nominal level 0.95 for the competing approaches, see

Figure 2.

Figure 2

A notable result from the simulation study is the crucial role of the number of studies and the true value of the between-study variance τ^2 in determining the performance of the various methods. The liberal behaviour of r_{P} is apparent across different settings. Skovgaard's statistic $r^*_{\mathrm{P,k}}$ assuming known within-study variances provides a remarkable improvement, although it underestimates the nominal level for small number of studies and large values of τ^2. The integrated likelihood is the most satisfactory solution overall. A conservative performance is experienced for small values of τ^2 and small number of studies. Results from $\tilde{r}_{\mathrm{Int}}$ and $\bar{r}_{\mathrm{Int}}$ essentially overlap, so that only the latter are displayed. The two solutions are both preferable to r_{Int}, with a coverage of confidence intervals which tends to be close to 0.95 or slightly higher.

6 Concluding remarks

This paper considers small sample inference in meta-analysis of normally distributed measures of the effect of interest. Instead of assuming the within-study variances as known, we considered an integrated likelihood approach to account for the additional uncertainty related to the estimation of the within-study variances. The methodology is shown to provide accurate inferential results when the sample size of the studies or the number of studies included in the meta-analysis is limited. In the meanwhile, the method avoids the computational difficulties related to the large number of nuisance parameters. Typically, the usage of the integrated likelihood proposed here would lead to more cautious inferences, a commendable result in several settings with limited sample information.

The focus of the paper is on meta-analysis of summary data. When individual patient data are available, the small sample size of meta-analysis studies is typically not a drawback and appropriate hierarchical models are commonly adopted for inference. The small number

of studies, instead, can still represent a source of inaccurate inference to deal with.

The estimation of the within-study variances for binary data differs from the case of normally distributed outcomes examined in this paper since the available information is analogous to that from individual patient data. For this situation, several approaches in the literature have been proposed, which mainly focus on hierarchical models, e.g., Smith et al. (1995), Turner et al. (2000), Hamza et al. (2008). The integrated likelihood for binary data maintains an interesting analogy with that examined in this paper under the normal case, despite obtaining a weight function for the nuisance components is less immediate. Details about the binary data case are reported in the Supporting Information.

From the methodological side, the problem studied in this paper is an instance of two-index asymptotics (Sartori, 2003), meaning that the available sample information grows with both the observations within each study and the number of studies. Although we did not formally cast the study of the available methodology within the two-index setting, it seems worth mentioning that recent results presented in De Bin et al. (2014) substantiate the good properties of the integrated likelihood using the zero-score-expectation parameterization for general statistical models within two-index asymptotics.

Albeit the methodology discussed here is embedded in a frequentist approach, an extension to a full Bayesian formulation is possible. In fact, the integrated likelihood (4) can be used along with an a priori distribution for β to obtain a marginal posterior distribution.

Although the paper considers the integrated likelihood approach within the meta-analysis context, the extension to the meta-regression case is straightforward. Deriving the explicit form of the zero-score-expectation parameterization, however, requires to assume equal within-study variances. Details are reported in the Supporting information.

Supporting information

Additional information for this article is available online. The additional information includes details about the derivation of the zero-score-expectation parameterization, the

illustration of the integrated likelihood approach for binary data and a description of the integrated likelihood within the meta-regression context.

Acknowledgements

The authors are grateful to the Associate Editor and the Referees for their detailed comments and suggestions that led to an improved version of the paper.

References

Böhning, D., Malzahn, U., Dietz, E., Schlattmann, P., Viwatwongkasem, C. & Biggeri, A. (2002). Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. *Biostatistics* **3**, 445–457.

Brockwell, S. E. & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Stat. Med.* **20**, 825–840.

Brockwell, S. E. & Gordon, I. R. (2007). A simple method for inference on an overall effect in meta-analysis. *Stat. Med.* **26**, 4531–4543.

De Bin, R., Sartori, N. & Severini, T. A. (2014). Integrated likelihoods in models with stratum nuisance parameters. Technical Report Number 157, Department of Statistics, University of Munich.

DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.

Di Gessa, G. (2008). Simple strategies for variance uncertainty in meta-analysis. MSc(R) thesis, University of Glasgow, UK. `http://theses.gla.ac.uk/128`.

Fraser, D. A. S., Reid, N., Li, R. & Wong, A. (2003). p-value formulas from likelihood asymptotics: Bridging the singularities. *J. Statist. Res.* **37**, 1–15.

Guolo, A. (2012). Higher-order likelihood inference in meta-analysis and meta-regression. *Stat. Med.* **31**, 313–327.

Guolo, A. & Varin, C. (2012). The `R` package `metaLik` for likelihood inference in meta-analysis. *Journal of Statistical Software* **50** (7), 1–14.

Hamza, T. H., van Houwelingen, H. C. & Stijnen, T. (2008). The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of Clinical Epidemiology* **61**, 41–51.

Hardy, R. J. & Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Stat. Med.* **15**, 619–629.

van Houwelingen, H. C., Arends, L. R. & Stijnen, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat. Med.* **21**, 589–624.

Jackson, D. & Bowden, J. (2009). A re-evaluation of the 'quantile approximation method' for random-effects meta-analysis. *Stat. Med.* **28**, 338–348.

Johnson, B. T. & Huedo-Medina, T. B. (2013). Meta-analytic statistical inferences for continuous measure outcomes as a function of effect size metric and other assumptions. Rockville (MD): Agency for Healthcare Research and Quality (US); Report No.: 13-EHC075-EF. `http://www.ncbi.nlm.nih.gov/books/NBK140575`.

Malzahn, U., Bohning, D. & Holling, H. (2000). Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika* **87**, 619–632.

Normand, S.-L. T. (1999). Meta-analysis: Formulating, evaluating, combining, and reporting. *Stat. Med.* **18**, 321–359.

Piessens, R., de Doncker-Kapenger, E., Ueberhuber, C. & Kahaner, D. (1983). *QUADPACK, A subroutine package for automatic integration.* Springer-Verlag.

R Core Team (2015). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, `http://www.R-project.org`.

Roberts, C. J. (2005). Issues in meta-regression analysis: an overview. *Journal of Economic Surveys* **19**, 295–298.

Sánchez-Meca, J. & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods* **13**, 31–48.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90**, 533–549.

Severini, T. A. (2000). *Likelihood methods in Statistics*. Oxford University Press, Oxford.

Severini, T. A. (2007). Integrated likelihood functions for non-Bayesian inference. *Biometrika* **94**, 529–542.

Severini, T. A. (2010). Likelihood ratio statistics based on an integrated likelihood. *Biometrika* **97**, 481–496.

Sharma, G. & Mathew, T. (2011). Higher order inference for the consensus mean in inter-laboratory studies. *Biom. J.* **53**, 128–136.

Sidik, K. & Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Stat. Med.* **26**, 1964–1981.

Skovgaard, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2**, 145–165.

Smith, T. C., Spiegelhalter, D. J. & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Stat. Med.* **14**, 2685–2699.

Sutton, A. J. & Higgins, J. P. T. (2008). Recent developments in meta-analysis. *Stat. Med.* **27**, 625–650.

Taubert, D., Roesen, R. & Schömig, E. (2007). Effect of cocoa and tea intake on blood pressure: A meta-analysis. *Archives of Internal Medicine* **167**, 626–634.

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H. & Thompson, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Stat. Med.* **19**, 3417–3432.

Vangel, M. G. & Rukhin, A. L. (1999). Maximum likelihood analysis for heteroscedastic one-way random effects ANOVA in interlaboratory studies. *Biometrics* **55**, 129–136.

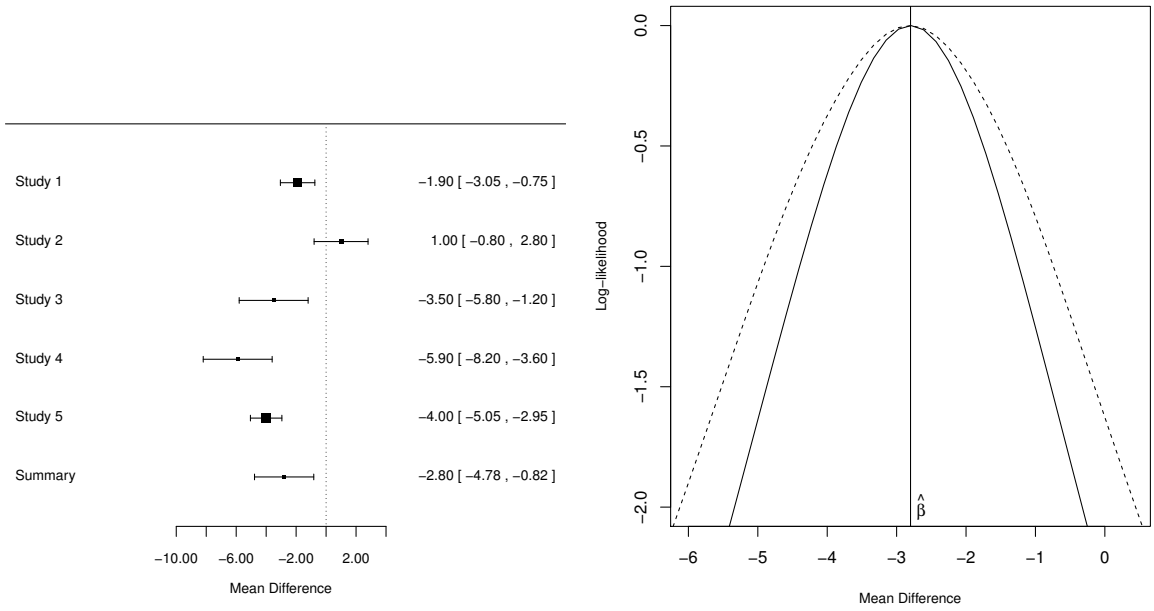

Figure 1: Cocoa data. Left panel: forest plot reporting the estimated mean difference from each study and the associated 95% confidence interval. Right panel: profile log likelihood function (solid line) and integrated log likelihood function (dashed line).

Table 1: Empirical one-sided rejection rates for the signed profile log likelihood ratio statistic r_{P}, Skovgaard's statistic r_{P}^*, their counterparts $r_{\mathrm{P,k}}$ and $r_{\mathrm{P,k}}^*$ assuming known within-study variances, and different specifications of the signed integrated log likelihood ratio statistic, r_{Int}, $\tilde{r}_{\mathrm{Int}}$, and $\overline{r}_{\mathrm{Int}}$, based on 10,000 replicates. Maximum simulation standard error smaller than 0.001.

	Rejection rates	r_{P}	r_{P}^*	$r_{\mathrm{P,k}}$	$r_{\mathrm{P,k}}^*$	r_{Int}	$\tilde{r}_{\mathrm{Int}}$	$\overline{r}_{\mathrm{Int}}$
Lower	0.010	0.030	0.018	0.023	0.016	0.013	0.010	0.010
	0.025	0.055	0.032	0.049	0.032	0.034	0.026	0.026
	0.050	0.086	0.056	0.082	0.058	0.066	0.050	0.050
	0.100	0.139	0.102	0.136	0.102	0.115	0.092	0.091
Upper	0.900	0.850	0.892	0.852	0.890	0.875	0.904	0.904
	0.950	0.910	0.941	0.914	0.939	0.932	0.946	0.946
	0.975	0.942	0.965	0.948	0.965	0.964	0.971	0.972
	0.990	0.968	0.980	0.974	0.982	0.984	0.989	0.988

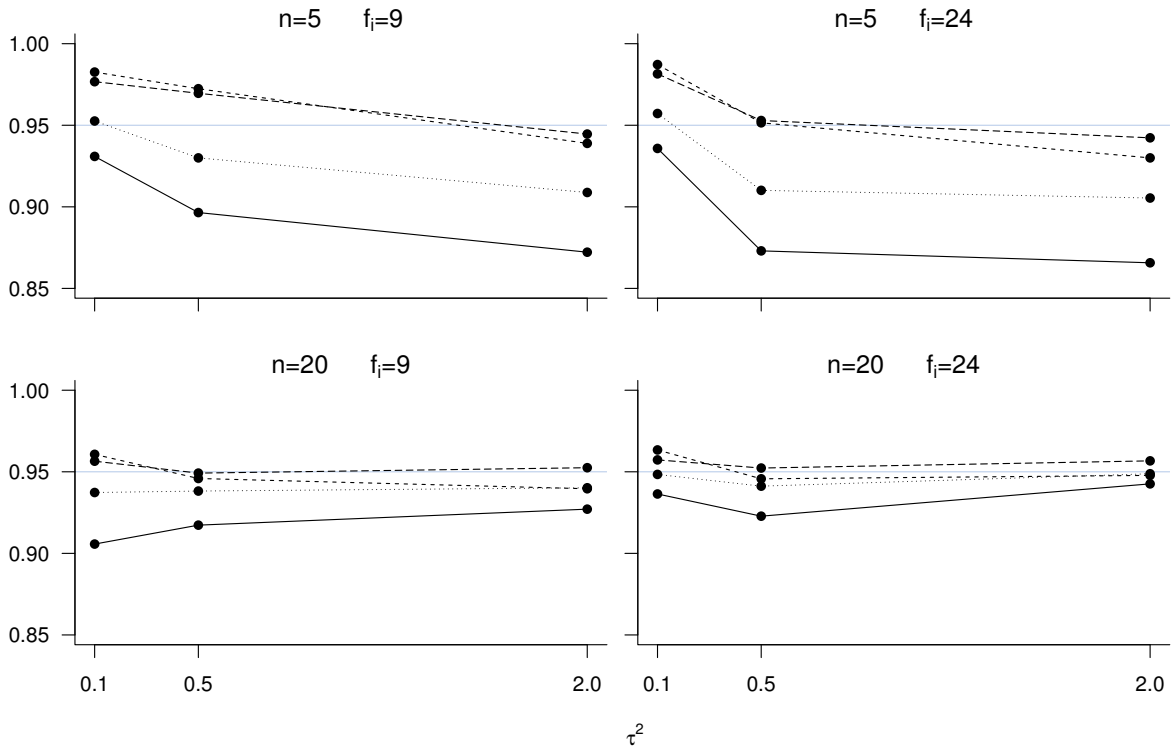

Figure 2: Empirical coverages of confidence intervals at nominal level 0.95 for increasing values of τ^2 and different combinations of number of studies n and degrees of freedom f_i. Lines correspond to r_{P} (solid line), $r^*_{\mathrm{P,k}}$ (dotted line), r_{Int} (dashed line), $\bar{r}_{\mathrm{Int}}$ (long dashed line). The solid grey horizontal line corresponding to the 0.95 confidence level is superimposed.

Annamaria Guolo

Department of Statistical Sciences, University of Padova

Via C. Battisti, 241, I-35121 Padova, Italy

annamaria.guolo@unipd.it