Combining likelihood and significance functions

Journal:	*Statistica Sinica*
Manuscript ID	SS-2016-0508.R2
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Fraser, Donald; University of Toronto, Statistical Sciences Reid, Nancy; University of Toronto, Statistical Sciences
Keywords:	composite likelihood, Meta-analysis, p-value functions

COMBINING LIKELIHOOD AND SIGNIFICANCE FUNCTIONS

D.A.S. Fraser and N. Reid

Department of Statistical Sciences, University of Toronto

Abstract: The need for combining likelihood information arises widely in the analysis of complex models, and in meta-analysis, where information is to be combined from various studies. We work to first order and show that full accuracy for combining scalar or vector parameter information is available from asymptotic analysis using score variables. We also use this approach to combine p-values for scalar parameters of interest.

Key words and phrases: composite likelihood, meta-analysis, p-value functions

1. Introduction

Statistical models presented in the form of a family of densities $\{f(y;\theta); y \in \mathbb{R}^d, \theta \in \Theta \subset \mathbb{R}^p\}$ are usually analyzed using the likelihood function $L(\theta) \propto f(y;\theta)$, or equivalently the log-likelihood function $\ell(\theta) = \log\{L(\theta)\}$. Evaluated at the observed data this provides all the data-dependent information for a standard Bayesian analysis, and almost all the data-dependent information for frequentist-based analysis; as described in Fraser & Reid (1993) full third-order inference requires in addition sample-space derivatives of the log-likelihood function.

In some modelling situations however the full joint density, and hence the likelihood function, may not be available. In such cases work-arounds have been developed, using for example marginal models for single coordinates, or marginal models for pairs of coordinates, or other variants called pseudo- or composite likelihood functions, studied in Lindsay (1988) and reviewed in Varin et al. (2011). For example, the composite pairwise log-likelihood function is

$$\ell_{\text{pair}}(\theta) = \sum_{r<s} \log\{f_2(y_r, y_s; \theta)\},$$

where $f_2(y_r, y_s; \theta)$ is the marginal model for a pair of components (y_r, y_s), obtained by marginalizing the joint density $f(y;\theta)$. If we were considering what is

called the independence likelihood we would have $\ell_{\rm ind}(\theta) = \Sigma_r \log\{f_1(y_r;\theta)\}$.

Our approach here is to assume that the model is sufficiently smooth that the usual asymptotic theory for composite likelihood inference applies; see, for example, the summary in Varin et al. (2011, §2.3). In particular, we assume that the likelihood components have the property that their score functions are asymptotically normal with finite variances and covariances. This can arise if we have a fixed number of components, each constructed from an underlying sample of size n, or if we have an increasing number of components with appropriate short-range dependence, so that information accumulates at a rate proportional to the number of components, as could arise for example in a time series or spatial setting with correlations decreasing with distance.

We denote an arbitrary composite log-likelihood function by $\sum_{i=1}^m \ell_i(\theta)$; for the pairwise log-likelihood function above we have $m = d(d-1)/2$. We let θ_0 be some trial or reference value of the parameter and examine the model to first derivative about θ_0; we will see that to first order the model has a simple linear regression form that is invariant to the starting point. In practice a starting value could be any consistent estimate, for example that obtained by maximizing the unadjusted composite likelihood function.

We assume that each component log-likelihood function admits an expansion of the form

$$\ell_i(\theta) = a + (\theta-\theta_0)^{\rm T} s_i - \frac{1}{2}(\theta-\theta_0)^T \jmath_i (\theta-\theta_0) + o(||\theta-\theta_0||), \qquad (1.1)$$

where $s_i = s_i(\theta_0) = (\partial/\partial\theta)\ell_i(\theta)|_{\theta_0}$ is the component score variable, and $\jmath_i = \jmath_i(\theta_0)$ is the corresponding negative second derivative. The Bartlett identities hold for component log-likelihood function:

$$E(s_i;\theta_0) = 0, \quad \mathrm{var}(s_i;\theta_0) = E(\jmath_i;\theta_0) = v_{ii}, \qquad (1.2)$$

where v_{ii} is the $p\times p$ expected Fisher information matrix from the ith component. We also have the moment relations (Cox & Hinkley 1974),

$$E(s_i;\theta) = v_{ii}(\theta-\theta_0) + o(||\theta-\theta_0||), \quad \text{and} \quad \mathrm{var}(s_i;\theta) = v_{ii} + o(||\theta-\theta_0||). \qquad (1.3)$$

We stack the m score vectors $s = (s_1^{\rm T}, \ldots, s_m^{\rm T})^{\rm T}$ and write

$$E(s;\theta) \doteq V(\theta-\theta_0), \quad \text{and} \quad \mathrm{var}(s;\theta) \doteq W, \qquad (1.4)$$

where $V = (v_{11}, \ldots, v_{mm})^{\mathrm{T}}$ is the $mp \times p$ matrix of the stacked information matrices v_{ii}, and W is the $mp \times mp$ matrix with v_{ii} on the diagonal, and off-diagonal matrix elements $v_{ij} = \mathrm{cov}(s_i, s_j)$. Expression (1.4) enables the construction of the optimally weighted score vector, using Gauss-Markov theory, as described in the next section. As the error of approximation is $o(||\theta - \theta_0||)$, the mean and variance results (1.2) and (1.3) are valid for any θ_0 within moderate deviations of the true value.

2. First-order combination of component log-likelihood functions

We express (1.4) to first order by the regression model,

$$s = V(\theta - \theta_0) + e, \tag{2.1}$$

with $e \sim N(0, W)$. From this approximation we obtain the log-likelihood function

$$\begin{aligned} \ell^*(\theta) &= a - \frac{1}{2}\{s - V(\theta - \theta_0)\}^{\mathrm{T}} W^{-1} \{s - V(\theta - \theta_0)\}, \\ &= a - \frac{1}{2}(\theta - \theta_0)^{\mathrm{T}} V^{\mathrm{T}} W^{-1} V(\theta - \theta_0) + (\theta - \theta_0)^{\mathrm{T}} V^{\mathrm{T}} W^{-1} s, \end{aligned} \tag{2.2}$$

with score function $s^*(\theta) = V^T W^{-1}\{s - V(\theta - \theta_0)\}$, which has expected value 0 and variance $V^T W^{-1} V$. This log-likelihood function is maximized at

$$\hat{\theta}^* = \theta_0 + (V^{\mathrm{T}} W^{-1} V)^{-1} V^{\mathrm{T}} W^{-1} s,$$

which has expected value θ, variance $(V^{\mathrm{T}} W^{-1} V)^{-1} = \bar{W}$, say, and we can equivalently write

$$\ell^*(\theta) = a - \frac{1}{2}(\theta - \hat{\theta}^*)^{\mathrm{T}} V^{\mathrm{T}} W^{-1} V(\theta - \hat{\theta}^*) = c - \frac{1}{2}(\theta - \hat{\theta}^*)^{\mathrm{T}} \overline{W}^{-1} (\theta - \hat{\theta}^*); \tag{2.3}$$

this makes the location form of the log-likelihood more transparent.

If θ is a scalar parameter then from (2.2)

$$\ell^*(\theta) = a - \frac{1}{2} V^{\mathrm{T}} W^{-1} V(\theta - \theta_0)^2 + V^{\mathrm{T}} W^{-1} s(\theta - \theta_0) \tag{2.4}$$

$$\doteq V^{\mathrm{T}} W^{-1} \underline{\ell}(\theta), \tag{2.5}$$

where $\underline{\ell}(\theta)$ is the vector of components $\ell_i(\theta)$, equivalent to first order to $a - (1/2)\{s_i - v_{ii}(\theta - \theta_0)\}^2 v_{ii}^{-1}$.

In (2.5) we have an optimally weighted combination of component log-likelihood functions, which agrees with (2.4) or (2.3), up to quadratic terms, but will usually be different in finite samples, as the individual component log-likelihood functions are not constrained to be quadratic.

The linear combination (2.5) is not generally available for the vector parameter case as different combinations of components are needed for different coordinates of the parameter, as indicated by different rows in the matrix V^{T} in (2.3).

Lindsay (1988) studied the choice of weights in scalar composite likelihood by seeking an optimally weighted combination of score functions $\partial\ell_i(\theta)/\partial\theta$, in his notation $S_i(\theta)$; the optimal weights depend on θ. Our approach is to work within moderate deviations of a reference parameter value, and use the first-order model for the observed variables s_i, thus leading directly to a first-order log-likelihood function.

This is closely related to indirect inference, widely used in econometrics, where a set of estimating functions $\{g_1(\theta), \ldots, g_K(\theta)\}$ are available, and the goal is to find an estimate of θ based on an optimal combination of these estimating functions. Combining estimating functions into a quadratic form was explored in Jiang & Turnbull (2004), where the result was called an indirect log-likelihood function. Often in indirect inference the model for the data is specified by a series of dynamic equations, so it is feasible to simulate from the true model, but not to write down the true log-likelihood function. Estimation of the model parameters proceeds by matching the simulated data to the indirect log-likelihood function. In the present setting we are instead concerned with optimal combinations of given components, under the assumption that each component is a genuine log-likelihood function satisfying (1.3) and (1.2).

For a scalar or vector parameter of interest ψ of dimension r, with nuisance parameter λ, so that $\theta = (\psi^{\mathrm{T}}, \lambda^{\mathrm{T}})^{\mathrm{T}}$, we have from (2.3) that the first-order log-likelihood function for the component ψ is

$$\ell^*(\psi) = c - \frac{1}{2}(\psi - \hat{\psi}^*)^{\mathrm{T}} \overline{W}^{\psi\psi} (\psi - \hat{\psi}^*), \tag{2.6}$$

where $\overline{W}^{\psi\psi}$ is the $\psi\psi$ submatrix of $\bar{W}^{-1}$ and $\hat{\psi}^* = \psi(\hat{\theta}^*)$ is the relevant component of $\hat{\theta}^*$ above. Pace et al. (2016) consider the use of profile log-likelihood

components for scalar parameters of interest.

3. Illustrations

The first illustrations use latent independent normal variables, as this captures the essential elements and makes clear the role of correlation in the reweighting. The basic underlying densities are assumed to be independent responses x from a $N(\theta, 1)$ distribution, with corresponding log-likelihood function $-\theta^2/2 + \theta x$; we take $\theta_0 = 0$.

Example 1: Independent components. Consider component variables $y_1 = x_1$ and $y_2 = x_2$. Then $m = 2$ and the component log-likelihood functions are $\ell_i(\theta; y_i) = -\theta^2/2 + y_i\theta$ giving $s_i = y_i$, $V = (1, 1)^{\mathrm{T}}$, and $W = \mathrm{diag}(1, 1)$. Thus $\ell^*(\theta) = \ell_1(\theta) + \ell_2(\theta)$ is the independence log-likelihood function, as would be expected.

Example 2: Dependent and exchangeable components. Consider $y_1 = x_1 + x_2$ and $y_2 = x_1 + x_3$. The component log-likelihood functions are $\ell_i(\theta) = -\theta^2 + \theta y_i$ giving $s_i = y_i$,

$$V = (2, 2)^{\mathrm{T}}, \quad W = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad V^{\mathrm{T}} W^{-1} = (2/3, 2/3), \tag{3.1}$$

and combined first-order log-likelihood function

$$\ell^*(\theta) = (2/3, 2/3)^{\mathrm{T}} \underline{\ell}(\theta) = -\frac{4}{3}\theta^2 + \frac{2}{3}\theta(y_1 + y_2). \tag{3.2}$$

In contrast the unadjusted composite log-likelihood function obtained by adding the marginal log-likelihood functions is

$$\ell_{UCL}(\theta) = -2\theta^2 + \theta(y_1 + y_2),$$

with score variable $y_1 + y_2$, which has variance 6, but second derivative 4: the second Bartlett identity does not hold and $\ell_{UCL}(\theta)$ is not a proper log-likelihood. We can recover the Bartlett identity by rescaling: $\ell_{ACL} = a\ell_{UCL} = -2a\theta^2 + \theta a(y_1 + y_2)$, with negative second derivative $4a$ and score variance $6a^2$. These latter become equal with $a = 2/3$, and the adjusted composite log-likelihood is

$$\ell_{ACL}(\theta) = \frac{2}{3}\ell_{UCL}(\theta) = -\frac{4}{3}\theta^2 + \frac{2}{3}\theta(y_1 + y_2),$$

which agrees with $\ell^*(\theta)$. The next example shows that this agreement does not hold generally.

Example 3: Dependent but not exchangeable components. Now let $y_1 = x_1$ and $y_2 = x_1 + x_3$. The individual log-likelihood functions are $\ell_1(\theta) = -\theta^2/2 + \theta y_1$ and $\ell_2(\theta) = -\theta^2 + \theta y_2$ with $s_1 = y_1$ and $s_2 = y_2$. We then have $V^{\mathrm{T}} = (1, 2)$, the off-diagonal elements of W equal to 1, and $V^{\mathrm{T}} W^{-1} = (0, 1)$, leading to

$$\ell^*(\theta) = -\theta^2 + \theta y_2, \tag{3.3}$$

with maximum likelihood estimate $\hat{\theta}^* = y_2/2$, which has variance 1/2. This log-likelihood function reflects the fact that $y_2 = x_1 + x_3$ provides all information about θ.

In contrast, the unadjusted composite log-likelihood function is $\ell_{UCL}(\theta) = -(3/2)\theta^2 + \theta(y_1 + y_2)$, with associated maximum likelihood estimate $\hat{\theta}_{UCL} = (y_1 + y_2)/3$. The rescaling factor to adjust for the Bartlett property is $a = 3/5$ giving the adjusted composite log-likelihood function

$$\ell_{ACL}(\theta) = \frac{3}{5}\ell_{UCL}(\theta) = -\frac{9}{10}\theta^2 + \frac{3}{5}\theta(y_1 + y_2),$$

which is again maximized at $(y_1 + y_2)/3$. Although the second Bartlett identity is satisfied, the adjusted composite log-likelihood function leads to the same inefficient estimate of θ as the unadjusted version. Some discussion related to this point is given in Freedman (2006).

An asymptotic version of these two illustrations is obtained by having n replications of y_1 and y_2, or equivalently assuming x_1, x_2 and x_3 have variances $1/n$ instead of 1.

Example 4: Bivariate Normal. Suppose we have n pairs (y_{i1}, y_{i2}) independently distributed as bivariate normal with mean vector (θ, θ) and a known covariance matrix. The sufficient statistic is the pair of sample means, $(\bar{y}_{\cdot 1}, \bar{y}_{\cdot 2})$, and the component log-likelihood functions are taken as those from the marginal densities of $\bar{y}_{\cdot 1}$ and $\bar{y}_{\cdot 2}$, so that $\ell_1(\theta) = -n(\bar{y}_{\cdot 1} - \theta)^2/(2\sigma_1^2)$ and $\ell_2(\theta) = -n(\bar{y}_{\cdot 2} - \theta)^2/(2\sigma_2^2)$. The score components s_1 and s_2 are, respectively, $n(\bar{y}_{\cdot 1} - \theta)/\sigma_1^2$ and $n(\bar{y}_{\cdot 2} - \theta)/\sigma_2^2$, with variance-covariance matrix

$$W = n \begin{pmatrix} 1/\sigma_1^2 & \rho/(\sigma_1\sigma_2) \\ \rho/(\sigma_1\sigma_2) & 1/\sigma_2^2 \end{pmatrix}, \tag{3.4}$$

giving

$$V^{\mathrm{T}}W^{-1} = (1-\rho^2)^{-1}(1-\rho\sigma_1/\sigma_2, 1-\rho\sigma_2/\sigma_1),$$

leading to

$$\ell^*(\theta) = -\frac{n}{2(1-\rho^2)}\left\{\left(\frac{\bar{y}_{.1}-\theta}{\sigma_1}\right)^2(1-\rho\frac{\sigma_1}{\sigma_2}) + \left(\frac{\bar{y}_{.2}-\theta}{\sigma_2}\right)^2(1-\rho\frac{\sigma_2}{\sigma_1})\right\}. \quad (3.5)$$

As a function of θ this can be shown to be equivalent to the full log-likelihood function based on the bivariate normal distribution of $(\bar{y}_{.1}, \bar{y}_{.2})$, and the maximum likelihood estimate of θ is a weighted combination of $\bar{y}_{.1}$ and $\bar{y}_{.2}$.

If the parameters in the covariance matrix are also unknown then the reduction by sufficiency is more complicated, and the vector version of the combination as described at (2.3) would be needed.

Example 5: Two parameters. Suppose that x_i follow a $N(\theta_1, 1)$ distribution, and independently z_i follow a $N(\theta_2, 1)$ distribution. We base our component log-likelihood functions on the densities of the vectors

$$y_1 = \begin{pmatrix} x_1 \\ z_2 + z_3 \end{pmatrix}, \quad y_2 = \begin{pmatrix} x_1 + x_3 \\ z_2 \end{pmatrix},$$

giving the score variables $s_1 = (y_{11}, y_{12})^T$ and $s_2 = (y_{21}, y_{22})^T$. The needed variances and covariances are:

$$V = \begin{pmatrix} 1 & 0 \\ 0 & 2 \\ 2 & 0 \\ 0 & 1 \end{pmatrix}, \quad W = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \quad W^{-1} = \begin{pmatrix} 2 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 2 \end{pmatrix},$$

and $V^T W^{-1} V = \mathrm{diag}(2, 2)$. This gives

$$\ell^*(\theta_1, \theta_2) = -(\theta_1 - \hat{\theta}_1^*)^2 - (\theta_2 - \hat{\theta}_2^*)^2,$$

where $\hat{\theta}^* = (y_{21}/2, y_{12}/2)^{\mathrm{T}}$. This combines the log-likelihood functions for θ based on s_{12} and s_{21}, as we might reasonably have expected from the presentations in terms of the latent x_i and z_i variables. Meanwhile, the usual composite log-likelihood function derived from the sum of those for s_1 and s_2 has additional terms.

Example 6. Two parameters, without symmetry. Within the structure of the previous example, suppose our component vectors are

$$y_1 = \begin{pmatrix} x_1 \\ z_1 \end{pmatrix}, \quad y_2 = \begin{pmatrix} x_1 + x_2 \\ z_2 \end{pmatrix},$$

The variances and covariances again are directly available:

$$V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 0 \\ 0 & 1 \end{pmatrix}, \quad W = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad W^{-1} = \begin{pmatrix} 2 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

From Example 3 we can see that for inference about θ_1 the weights are $(0, 1)$ and from Example 1 the weights for θ_2 are $(1, 1)$; clearly these are incompatible. For the direct approach using (2.5) we have $V^T W^{-1} V = \mathrm{diag}(2, 2)$, giving

$$\ell^*(\theta_1, \theta_2) = -(\theta_1 - \hat{\theta}_1^*)^2 - (\theta_2 - \hat{\theta}_2^*)^2,$$

where $\hat{\theta}^* = (y_{21}/2, (y_{12} + y_{22})/2)^{\mathrm{T}}$. This simple sum of component likelihood functions for (θ_1, θ_2) is to be expected as the measurements concerning θ_1 are independent of those for θ_2; in addition, all the information concerning θ_1 comes from the first coordinate of y_2, and all the information for θ_2 comes from the second coordinates of both y_1 and y_2.

Example 7: Time series correlation structure. As an illustration a little closer to those that might arise in practice, we consider the underlying model to be a q-dimensional normal with mean zero and with correlations $R_{ss'}(\theta)$ between pairs $(y_s, y_{s'})$. We compare $\ell^*(\theta)$ to the unadjusted composite log-likelihood function created from all possible pairs. In computing the elements of W, each score component s_i corresponds to a pair (s, s') and similarly s_j to a pair (t, t'). Thus W_{ij} depends on up to fourth moments of the original components of the vector y. While the explicit construction is tedious, it can be automated.

In more detail, we have

$$\ell_j(\theta; y_s, y_{s'}) = -\frac{1}{2}\log(1 - R_{ss'}^2) - \frac{y_s^2 + y_{s'}^2 - 2y_s y_{s'} R_{ss'}}{2(1 - R_{ss'}^2)}, \tag{3.6}$$

and

$$s_j = \frac{\partial l_j(\theta)}{\partial \theta}\Big|_{\theta=\theta_0} = \frac{R_{ss'}\dot{R}_{ss'}}{1-R_{ss'}^2} + \frac{y_s y_{s'}\dot{R}_{ss'}}{1-R_{ss'}^2} - \frac{(y_s^2+y_{s'}^2-2y_s y_{s'}R_{ss'})R_{ss'}\dot{R}_{ss'}}{(1-R_{ss'}^2)^2}, \quad (3.7)$$

where $R_{ss'} = R_{ss'}(\theta_0)$ and $\dot{R}_{ss'} = (d/d\theta)R_{ss'}(\theta_0)$. From this we have

$$v_{jj} = \mathrm{var}\{s_j(\theta_0)\} = \frac{\dot{R}_{ss'}^2}{1-R_{ss'}^2} + \frac{2R_{ss'}^2\dot{R}_{ss'}^2}{(1-R_{ss'}^2)^2}, \quad (3.8)$$

and a lengthy formula for the covariance elements, similarly determined, and taking account of pairs (s, s') and (t, t') that have $s' = t'$ and $s' \neq t'$, for example.

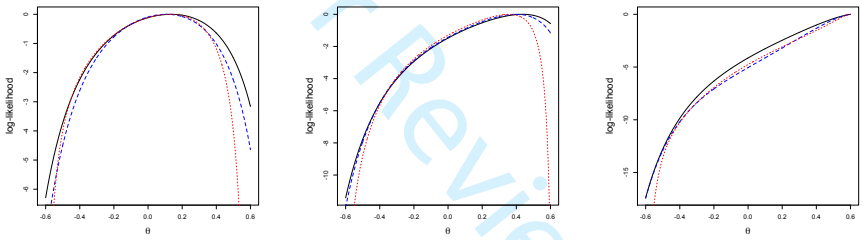

Figure 1: Illustrations of the the proposed combination $\ell^*(\cdot)$ (black, solid), the pairwise composite log-likelihood function (blue, dashed), and the full log-likelihood function (red, dotted) for three simulations of length 11 from the model $N\{0, R(\theta)\}$. In the third plot the likelihood functions do not have the approximate quadratic behaviour needed for first-order theory.

For illustration we chose $R_{ss'}(\theta) = \theta^{|s-s'|}$ if $|s-s'| \leq 2$, and 0 otherwise; only pairs differing by one or two places contribute to $\ell^*(\cdot)$ and to the unadjusted composite likelihood function $\ell_{UCL}(\theta) = 1^{\mathrm{T}}\underline{\ell}(\theta)$. This is a time series version of correlation only between near neighbours; similar structures are often used in spatial modelling. Figure 1 illustrates $\ell_{UCL}(\theta)$ and $\ell^*(\theta)$ for some sample data sets of length 10, and compares them to the true full log-likelihood function. Some simulations from this model are summarized in Table 1. In computing the averages of the point estimates and standard error simulations leading to a maximum on the boundary were removed, as indicated in the table.

Asymptotic theory is being strongly tested in these simulations of a single time series of length 11 or 21. There is some replication as correlations are zero beyond two lags, but we can see from Table 1 that even the full maximum likelihood estimate has appreciable bias. The covariance matrix is positive definite only over a restricted range for θ, although the 2×2 submatrices needed for the weighted and unweighted pairwise likelihood calculations do not require a restricted range: both ℓ^* and ℓ_{UCL} can be computed when there does not exist a full multivariate normal model. Whether or not there may be another multivariate model compatible with these marginal densities is not clear. This is a disadvantage of composite likelihood methods from the point of view of modelling, but possibly an advantage from the point of view of robust estimation. Simulations not shown here suggest that both ℓ^* and the unadjusted composite pairwise likelihood functions give accurate point estimates when the range of θ is expanded to $(-1, 1)$. In Table 2 we explore the show the effect of using incorrect weights in computing ℓ^*. The simulation results seem not very sensitive to changing the point at which the weights are computed.

4. Comparison to composite likelihood

Composite likelihood combines information from different components, often by adding the log-likelihood functions, but care is needed in constructing inference from the resulting function, as the curvature at the maximum does not give an accurate reflection of the precision. Corrections for this in the scalar parameter setting involve either rescaling the composite log-likelihood function, or accommodating the dependence among the components in the estimate of the variance of the composite likelihood estimator. In the vector parameter setting adjustments to the composite log-likelihood function are more complex than a simple rescaling; see Pace et al. (2011).

This rescaling is not enough: the location of the composite log-likelihood function is incorrect to first order, and confidence intervals so obtained are not correctly located to first order. This is corrected by the the use of $\ell^*(\theta)$ from Section 2.

As we are using only first-order log-likelihood functions, it suffices to illustrate this with normal distributions. Suppose $y^{\mathrm{T}} = (y_1, ..., y_m)$, where the marginal models for the individual coordinates y_i are normal with mean θv_{ii} and

Table 1: *Example 7.* Averages over 10000 simulations from the model $N\{0, R(\theta)\}$. We deleted simulation runs in which the estimates were on the boundary of the parameter space; N^* is the number remaining. The weights in ℓ^* use the true value of θ. The theoretical standard error is based on the second derivative at the maximum.

	true $\theta = 0.2$							
	$q = 11$				$q = 21$			
	estimate	simulation st. err.	theoretical st. err.	N^*	estimate	simulation st. err.	theoretical st. err.	N^*
ℓ	0.149	0.309		9056	0.182	0.238		9890
ℓ^*	0.140	0.290	0.288	8586	0.178	0.230	0.219	9673
ℓ_{UCL}	0.135	0.278	0.304	8551	0.172	0.225	0.187	9604
	true $\theta = 0.4$							
	$q = 11$				$q = 21$			
	estimate	simulation st. err.	theoretical st. err.	N^*	estimate	simulation st. err.	theoretical st. err.	N^*
ℓ	0.314	0.269		8437	0.366	0.190		9665
ℓ^*	0.289	0.251	0.270	7479	0.347	0.188	0.194	8684
ℓ_{UCL}	0.279	0.246	0.295	7504	0.340	0.187	0.152	8658

variance v_{ii}, and $\text{cov}(y_i, y_j) = v_{ij}$, all elements of the matrix W. The unadjusted composite log-likelihood function is

$$\ell_{UCL}(\theta) = -\frac{1}{2}\theta^2 \Sigma_{i=1}^m v_{ii} + \Sigma_{i=1}^m y_i \theta,$$

with maximum likelihood estimate $\hat{\theta}_{CL} = \Sigma y_i / \Sigma v_{ii}$ and curvature at the maximum point Σv_{ii}: this curvature is not the inverse variance of $\hat{\theta}_{CL}$ as the second Bartlett identity does not hold.

As indicated in Example 2, the rescaled version that recovers the second Bartlett identity is

$$\ell_{ACL}(\theta) = \frac{H}{J}\ell_{UCL}(\theta) = -\frac{1}{2}\theta^2 \frac{(\Sigma v_{ii})^2}{\Sigma v_{ij}} + \theta \sum y_i \frac{\Sigma v_{ii}}{\Sigma v_{ij}},$$

Table 2: *Example 7. Simulations from the model* $N\{0, R(\theta)\}$*;* ℓ^* *uses weights* $W(\theta_0)$ *computed at a different value of* θ*. Simulation size is 10000.*

Weights W and V computed at $\theta_0 = 0.2$			
q	true θ	estimate	standard error
11	0.4	0.289	0.260
21	0.4	0.347	0.192
11	0.3	0.216	0.273
21	0.3	0.264	0.215
11	0.1	0.075	0.294
21	0.1	0.090	0.240
11	0.0	0.002	0.293
21	0.0	0.001	0.238
11	−0.1	−0.078	0.291
21	−0.1	−0.090	0.240

where $H = E\{-\ell''_{UCL}(\theta)\} = \Sigma_i v_{ii}$ and $J = \text{var}\{\ell'_{UCL}(\theta)\} = \Sigma_{i,j} v_{ij}$; in this context neither H nor J depend on θ. The maximum likelihood estimate from this function is the same, $\hat{\theta}_{UCL}$, but the inverse of the second derivative gives the correct asymptotic variance.

What is less apparent is that the location of the log-likelihood function needs a correction. This is achieved using the weighted version from Section 2:

$$\ell^*(\theta) = -\frac{1}{2}\theta^2(V^{\mathrm{T}}W^{-1}V) + \theta V^{\mathrm{T}}W^{-1}y,$$

which has maximum likelihood estimate $\hat{\theta}^* = (V^{\mathrm{T}}W^{-1}V)^{-1}V^{\mathrm{T}}W^{-1}y$ with variance $(V^{\mathrm{T}}W^{-1}V)^{-1}$. Note that $\ell^*(\theta)$ has the same linear and quadratic coefficients for θ as the full log-likelihood function for the model $N(\theta V, W)$. Computation of both $\ell_{ACL}(\theta)$ and $\ell^*(\theta)$ require variances and covariances of the score variables.

Writing the uncorrected composite log-likelihood function as $1^{\mathrm{T}}\underline{\ell}(\theta)$, where $\underline{\ell}(\theta)$ is the vector $\{\ell_1(\theta), \ldots, \ell_m(\theta)\}$, with $\ell_i(\theta) = -(1/2)v_{ii}\theta^2 + y_i\theta$, we have $\text{var}(\hat{\theta}_{UCL}) = (1^{\mathrm{T}}W)/(1^{\mathrm{T}}V^2)$, $\text{var}(\hat{\theta}^*) = (V^{\mathrm{T}}W^{-1}V)^{-1}$, and $\text{cov}(\hat{\theta}_{UCL}, \hat{\theta}^*) =$

$(V^{\mathrm{T}}W^{-1}V)^{-1}$, giving

$$\mathrm{var}(\hat{\theta}_{UCL} - \hat{\theta}^*) = \frac{1^{\mathrm{T}}W1}{(1^{\mathrm{T}}V)^2} - \frac{1}{V^{\mathrm{T}}W^{-1}V}$$

and

$$\frac{\mathrm{var}(\hat{\theta}_{UCL})}{\mathrm{var}(\hat{\theta}^*)} = \frac{(1^{\mathrm{T}}V)^2}{(1^{\mathrm{T}}W1)(V^{\mathrm{T}}W^{-1}V)}.$$

5. Combining significance or p-value functions

In the case of a scalar parameter, we can directly link the score variable for each component to a standard normal variable, and hence to a p-value. Using the regression formulation (2.1), and taking $\theta_0 = 0$ as in §3, we have

$$s_i - v_{ii}\theta \longrightarrow z_i = v_{ii}^{-1/2}(s_i - v_{ii}\theta) \longrightarrow p_i = \Phi(z_i),$$

where z_i is standard normal, and p_i is the first order p-value for assessing $\theta = \theta_0$. Similarly we can make the inverse sequence of transformations

$$p_i \longrightarrow z_i = \Phi^{-1}(p_i) \longrightarrow (s_i - v_{ii}\theta) = (v_{ii})^{1/2} z_i.$$

As we have seen, to first order the optimal combination is linear in $s_i - v_{ii}\theta$, and this directly gives a route to combine the associated p-values:

$$V^{\mathrm{T}}W^{-1}(s - V\theta) = V^{\mathrm{T}}W^{-1}V^{1/2}\Phi^{-1}\{p(\theta; s)\}, \tag{5.1}$$

where $V^{1/2}\Phi^{-1}\{p(\theta; s)\}$ is the vector with coordinates $v_{ii}^{1/2}\Phi^{-1}\{p(\theta; s_i)\}$. We can then convert this to a combined first-order p-value:

$$\tilde{p}(\theta; s) = \Phi[(V^T W^{-1}V)^{-1/2}V^{\mathrm{T}}W^{-1}V^{1/2}\Phi^{-1}\{p(\theta; s)\}]. \tag{5.2}$$

Example 2 continued: Dependent and exchangeable components. The composite score variable relative to the nominal parameter value $\theta_0 = 0$ is

$$V^{\mathrm{T}}W^{-1}s = \frac{2}{3}(y_1 + y_2),$$

which is the score variable from the proposed log-likelihood function $\ell^*(\theta)$, and the relevant quantile is

$$z = \left(\frac{8}{3}\right)^{-1/2}\left\{\frac{2}{3}(y_1 + y_2) - \frac{8}{3}\theta\right\},$$

which has a standard normal distribution, exact in this case. The corresponding composite p-value function is then

$$\tilde{p}(\theta; s) = \Phi(z).$$

Example 3 continued: Dependent but not exchangeable components. The combined score variable for $\theta_0 = 0$ is $V^{\mathrm{T}}W^{-1}s = y_2$, which is the score variable from $\ell^*(\theta)$. The corresponding quantile is $z = 2^{-1/2}(y_2 - 2\theta)$ and the corresponding composite p-value function is

$$\tilde{p}(\theta; s) = \Phi(z) = \Phi\{2^{-1/2}(y_2 - 2\theta)\},$$

which is in accord with the general observation that y_2 here provides full information on the parameter θ.

Example 8: Combining three p-values. Suppose three investigations of a common scalar parameter θ have yielded the following p-values for assessing a null value θ_0 : $1.15\%, 3.01\%, 2.31\%$. For combining these we need the measures of precision as provided by the information, or variance of the score, for each component, say, $v_{11} = 3.0, v_{22} = 6.0, v_{33} = 9.0$. The corresponding z-values and score values s are

$$\begin{aligned} z_1 &= \Phi^{-1}(0.0115) = -2.273 \quad & s_1 - 3\theta_0 &= 3^{1/2}(-2.273) = -3.938, \\ z_2 &= \Phi^{-1}(0.0301) = -1.879 \quad & s_2 - 6\theta_0 &= 6^{1/2}(-1.879) = -4.603, \\ z_3 &= \Phi^{-1}(0.0231) = -1.994 \quad & s_3 - 9\theta_0 &= 9^{1/2}(-1.994) = -5.981. \end{aligned} \tag{5.3}$$

First suppose for simplicity that the investigations are independent, so that $W = \mathrm{diag}(V)$ and $V^T W^{-1} = (1, 1, 1)$and we add the scores, as one would expect; this gives the combined score -14.522. Then standardizing by the root of the combined information $18^{1/2} = 4.243$ we obtain $\tilde{p} = \Phi(-3.423) = 0.00031$.

To examine the effect of dependence among the scores we consider a cross-correlation matrix of the form

$$R = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

with corresponding covariance matrix W having entries $3, 6, 9$ on the diagonal and the appropriate covariances otherwise. To illustrate a low-level of correlation we take $\rho = 0.2$. The coefficients for combining the scores s_i in the array (17) are given in the array

$$\begin{aligned} V^T W^{-1} &= (3,6,9) \begin{pmatrix} 3.000 & 0.848 & 1.039 \\ 0.848 & 6.000 & 1.470 \\ 1.039 & 1.470 & 9.000 \end{pmatrix}^{-1} \\ &= (0.510, 0.726, 0.822). \end{aligned}$$

The resulting z and p-value are -2.81 and $\tilde{p} = 0.0025$, an order of magnitude larger than that obtained assuming independence. The combined p-value increases with ρ; for example if $\rho = 0.8$ the combined p-value is 0.075.

Fisher's combined p-value, obtained by referring $-2\Sigma \log p_i$ to a χ^2_6 distribution is 0.0006, independently of the value of ρ, as Fisher's method assumes the p-values are independent. The Bonferroni p-value is $3 \min(p_i) = 0.0345$, and while valid under dependence, is known to be conservative.

Many modern treatments of meta-analysis concentrate instead on combining the effect estimates, typically weighted by inverse variances, and the approach here is similar, although we work in the space of score functions. More specifically, we could combine estimates $\hat{\theta}_i = -v_{ii}^{-1/2} z_i$ with weights v_{ii}; under independence the combined estimate of θ is 0.807 with standard error 0.236 leading to the same p-value 0.0003 under independence. Similarly weighted linear combinations of $\hat{\theta}_i$ incorporating correlation give the same p-values as above. As was pointed out by a reviewer, the combination of point estimates here is analogous to a random effects meta-analysis, except that we assume the within-study and between-study variances are known; ρ here plays the role of the between study correlation. Of course in more practical applications of meta-analysis both the within-study and between-study variances must be estimated.

5. Conclusion

In this paper we use likelihood asymptotics to construct a fully first-order accurate log-likelihood function for the composite likelihood context. It requires the covariance matrix of score variables, which is also needed for inference based on composite likelihood.

The advantage of the first-order approach is that it expresses each component log-likelihood function as equivalent to that from a normal model with unknown mean and known variance. This in turn provides a straightforward way to describe the optimal combination. This is achieved through a linear combination of score variables, which can be converted both from p-values, for components, and to p-values, for the combination, and thus gives a procedure for meta-analysis.

Acknowledgements

We are grateful to the reviewers for helpful comments which improved the paper substantially. We express special thanks to Ruoyong Xu, U Toronto, for editorial and computational assistance with the revision.

Funding in partial support of this research was provided by the National Science and Engineering Research Council of Canada and the Senior Scholars Fund of York University.

References

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

Fraser, D. A. S. & Reid, N. (1993), 'Third order asymptotic models: likelihood functions leading to accurate approximation to distribution functions.', *Statistica Sinica* **3**, 67 – 82.

Freedman, D. A. (2006), 'On the so-called "Huber sandwich estimator" and "Robust standard errors"', *The American Statistician* **60**, 299 – 302.

Jiang, W. & Turnbull, B. (2004), 'The indirect method: inference based on intermediate statistics – a synthesis and examples', *Statistical Science* **19**, 239 – 263.

Lindsay, B. G. (1988), Composite likelihod methods., *in* N. Prabhu, ed., 'Statistical Inference from Stochastic Processes', Vol. 80, American Mathematical Society, Providence, Rhode Island, pp. 221 – 239.

Pace, L., Salvan, A. & Sartori, N. (2011), 'Adjusting composite likelihood ratio statistics', *Statistica Sinica* **21**, 129 – 148.

Pace, L., Salvan, A. & Sartori, N. (2016), Combining dependent log likelihoods for a scalar parameter of interest. preprint.

Varin, C., Reid, N. & Firth, D. (2011), 'An overview of composite likelihood methods', *Statist. Sinica* **21**, 5 – 42.

Department of Statistical Sciences, University of Toronto, Toronto, Canada

E-mail: dfraser@utstat.toronto.edu

E-mail: reid@utstat.utoronto.ca

Dear Dr. Cheng,
Editor,
Statistica Sinica, SS-2016-0508.R1

Thank you for your comments and those from the reviewer on the revision. We hope the paper is now suitable for publication.

With best wishes,

Don Fraser
Nancy Reid

Response to Reviewer

Thank you for your comments; we have responded to them as follows.

1. In the simulation of Example 7 I would add a comparison with: (i) the maximum likelihood estimator of the full model; (ii) the estimator from the combined log-likelihood when θ_0 used in calculating the weights is not the true value. The first estimator would yield a gold standard for comparison, while (ii) would give an idea of the sensitivity of the results to the choice of θ_0. Moreover, I would also add in the setting of the simulation at least another value of θ and one or two additional values of n.

We discovered an error in our programming, and the new plots do not show as much difference between the likelihood functions. It is possible that the choice of example for illustration does not entail enough dependence to make a substantial difference. At this point it seemed unwise to embark on further numerical work with different examples.

We have added a comparison to the full maximum likelihood estimator, and expanded the simulations. Thanks to this suggestion, we note that the covariance matrix is only positive definite over a restricted range of values for θ. Some comments on this point have been added.

We have included some simulations at a different value of θ, although a thorough exploration of sensitivity would require quite a bit more numerical work, which would be a longer project. Some caveats have been added to the text.

2. In Example 8 it would be nice to have a numerical comparison also with a standard meta-analysis model that assumes that the estimates from the ith study are $Y_i \sim N(\beta, \hat{\sigma}_i^2+\tau^2)(i = 1, \ldots, K)$, where β is the common effect, $\hat{\sigma}_i^2$ is the (known) variance of Y_i, τ^2 is a variance component, and K is the number of studies. I suspect that the approach proposed in the paper would lead to similar results, with τ^2 playing somehow the role of ρ. The only difference is that τ^2 is estimated jointly with β, while ρ is assumed as known by the Authors.

We have added some text about the calculation of the point estimate of θ, which is a weighted linear combination of the individual estimates. Because we are relating everything back to standard normal observations through the z_i, the p-values associated with the estimates are the same as those associated with the scores, and ρ is identical to $\tau^2/(\sigma_i\sigma_j)$, if we write, for example: $\hat{\beta}_i = \beta + epsilon_i + U$, where $\epsilon_i \sim N(0, \sigma_i^2)$ and $U \sim N(0, \tau^2)$ in the meta-analysis notation. (In our notation $\sigma_i^2 = v_{ii}^{-1}$.)

3. Minor comments

All the minor comments have been addressed; thank you for pointing out these errors.