A CAVEAT ON THE ROBUSTNESS OF COMPOSITE LIKELIHOOD ESTIMATORS: THE CASE OF A MIS-SPECIFIED RANDOM EFFECT DISTRIBUTION

Helen E. Ogden

University of Warwick

Abstract: Composite likelihoods are a class of alternatives to the full likelihood which may be used for inference in many situations where the likelihood itself is intractable. A composite likelihood estimator will be robust to certain types of model misspecification, since it may be computed without the need to specify the full distribution of the response. This potential for increased robustness has been widely discussed in recent years, and is considered a secondary motivation for the use of composite likelihood. The purpose of this paper is to show that there are some situations in which a composite likelihood estimator may actually suffer a loss of robustness compared to the maximum likelihood estimator. We demonstrate this in the case of a generalized linear mixed model under misspecification of the random-effect distribution. As the amount of information available on each random effect increases, we show that the maximum likelihood estimator remains consistent under such misspecification, but various marginal composite likelihood estimators are inconsistent. We conclude that composite likelihood estimators cannot in general be claimed to be more robust than the maximum likelihood estimator.

Key words and phrases: Consistency, generalized linear mixed model, laplace approximation, pairwise interactions.

1. Introduction

Suppose we observe independent samples $y^{(1)}, \ldots, y^{(r)}$, where each $y^{(i)} = (y_1^{(i)}, \ldots, y_m^{(i)})$ is assumed to be an independent sample from a model depending on an unknown parameter θ. The likelihood

$$L(\theta) = \prod_{i=1}^{r} L(\theta; y^{(i)})$$

is sometimes difficult to compute, and composite likelihoods (Lindsay (1988)) provide a class of alternatives for conducting inference about θ in such circumstances. A marginal composite likelihood

$$L^C(\theta) = \prod_{i=1}^{r} \prod_{s \in S} L(\theta; y_s^{(i)})^{w_s}$$

is formed by taking a product of component likelihoods, each of which is the likelihood given some subset of the data y_s, where w_s is a weight assigned to component s. A review of composite likelihoods and their many uses is given by Varin, Reid, and Firth (2011).

If the model is correctly specified, the composite likelihood estimator is consistent as $r \to \infty$, provided that the parameter of interest remains identifiable, although the estimator typically has a higher asymptotic variance than the maximum likelihood estimator. There is a hope that some compensation for this loss of efficiency can be provided by an increased robustness of the composite likelihood estimator to misspecification of the model. Lindsay , Yi, and Sun (2011) motivate the use of composite likelihoods in this way, stating that 'Compared to the traditional likelihood method, the composite likelihood method may be less efficient, but it could be a lot computationally faster to implement and be more robust to model misspecification.'

This notion is motivated by the fact that it is not necessary to specify the full distribution of the response in order to be able to compute a composite likelihood. If the marginal distributions of Y_s for each subset $s \in S$ are correctly specified, then the corresponding estimator of θ is consistent as $r \to \infty$, even if the full model is misspecified. The maximum likelihood estimator need not be consistent in such a setting, since the likelihood relies on the full, misspecified, distribution of Y. Xu and Reid (2011) discuss this type of robustness in some detail, and provide a formal proof of the consistency of the composite likelihood estimator in this setting.

In some situations the relevant marginal distributions themselves may be misspecified, in which case the marginal composite likelihood estimator no longer retains this robustness property. However, in that case the full distribution of Y must also be incorrect, so the maximum likelihood estimator need not be robust to this misspecification either. From this, it is tempting to conclude that a marginal composite likelihood estimator must always be at least as robust to model misspecification as the full likelihood estimator. We show that this is not the case.

To see this, we consider the impact of incorrectly specifying the random-effects distribution in a generalized linear mixed model. Generalized linear mixed models are a widely-used class of models, but one in which the likelihood is often difficult to compute. Because of this intractability, composite likelihood methods have been proposed as alternatives to full-likelihood inference.

In Section 3, we consider a class of generalized linear mixed models with simple nested structure, in which the likelihood is tractable. In this setting, we obtain some asymptotic results to show that under misspecification of the random-effect distribution the maximum likelihood estimator is consistent, while various composite likelihood estimators are inconsistent.

In Section 4, we consider a class of generalized linear mixed models with intractable likelihoods, in which each observation involves a pair of random effects. We compare the pairwise likelihood estimator proposed by Bellio and Varin (2005) with an estimator found by maximizing the Laplace approximation to the likelihood, in terms of the limiting value of each estimator under misspecification of the random-effect distribution. When only a small amount of information is available on each random effect, we find that the asymptotic bias in the pairwise likelihood estimator is smaller than that of Laplace estimator. However, as the amount of information available on each random effect increases, the magnitude of the asymptotic bias in the Laplace estimator decreases towards zero, but the asymptotic bias in the pairwise likelihood estimator remains fixed away from zero, in agreement with the asymptotic results of Section 3.

2. Generalized Linear Mixed Models

In a generalized linear mixed model, the distribution of $Y^{(i)}$ takes exponential family form, with distribution determined by the linear predictor

$$\eta^{(i)} = X^{(i)}\beta + Z(\sigma)u^{(i)},$$

where X is the design matrix for the fixed effects, $u^{(i)} = (u_1^{(i)}, \ldots u_n^{(i)})$ are random effects, where each $u_j^{(i)}$ is assumed to have independent standard normal distribution, and $Z(\sigma)$ is a design matrix for the random effects, whose entries may depend on a parameter σ.

The addition of random effects often makes the model more realistic, but the likelihood of $\theta = (\beta, \sigma)$ given each $y^{(i)}$ is

$$L(\theta; y^{(i)}) = \int_{\mathbb{R}^n} f(y^{(i)}|\eta^{(i)} = X^{(i)}\beta + Z(\sigma)u^{(i)}) \prod_{j=1}^{n} \phi(u_j^{(i)}) du^{(i)},$$

where $\phi(.)$ is the standard normal density, and it might be difficult to approximate this n-dimensional integral well. Moreover, there is often no good reason to suppose that the random effects are normally distributed, so inferences made should be checked for sensitivity to this assumption.

There is a large literature on the impact of random-effects misspecification on the maximum likelihood estimator. A detailed review is provided by McCulloch and Neuhaus (2011), who conclude that the asymptotic bias due to such misspecification may be expected to be small in many cases, and highlight some situations where it may be of more concern.

In situations where the likelihood is intractable, it is common to replace the true likelihood with a Laplace approximation, and to use the approximated likelihood for inference (Pinheiro and Bates (1995)). A disadvantage of this approach is that the resulting estimator need not be consistent as the amount of data increases, even if the model is correctly specified.

Composite likelihood estimators are consistent under correct model specification, albeit with some loss of efficiency compared to the maximum likelihood estimator. Bellio and Varin (2005) construct a pairwise likelihood for inference in these models, with a contribution from each pair $(y_j^{(i)}, y_k^{(i)})$ which are dependent under the model. Thus, defining t_j as the set of non-zero elements in the jth row of $Z(\sigma)$, a pairwise likelihood may be defined by

$$L^{\mathrm{pair}}(\theta; y) = \prod_{i=1}^{r} \prod_{(j,k):t_j \cap t_k \neq \emptyset} L(\theta; y_j^{(i)}, y_k^{(i)}).$$

Bellio and Varin (2005) focus on those cases in which each t_j contains two elements, although they note that essentially the same method could be applied more generally. Each term $L(\theta; y_j^{(i)}, y_k^{(i)})$ can be written as an integral over the $|t_j \cup t_k|$ random effects involved in the pair of observations. In the case where there are two random effects involved in each observation, this is a three-dimensional integral that is relatively easy to approximate well.

3. Robustness in a Two-level Model

3.1. A two-level model

Taking $n = 1$ in the generalized linear mixed model, and writing $x_j^{(i)}$ for the jth column of $X^{(i)}$, we obtain the two-level model $\eta_j^{(i)} = \beta^T x_j^{(i)} + \sigma u^{(i)}$. Here the $u^{(i)}$ are independent (scalar) random effects, which are assumed to be drawn from a $N(0, 1)$ distribution. Since $n = 1$, the likelihood is just a product of one-dimensional integrals. For ease of notation, we further assume that $x_j^{(i)} = x^{(i)}$ is constant across j, so that we can write $\eta_j^{(i)} = \eta^{(i)} = \beta^T x^{(i)} + \sigma u^{(i)}$; our results also hold for the more general case.

We are interested in the robustness of estimators of β to deviations from the assumed random-effects distribution. We suppose that each $u^{(i)}$ is drawn independently from a non-normal distribution with mean zero and variance one and that, for some parameter values $(\beta, \sigma) = (\beta_0, \sigma_0)$, the rest of the model is correctly specified.

Under this misspecification, we are interested in the limit β_m^∞ of the maximum likelihood estimator of β as $r \to \infty$, and how this limit varies with m. In particular, we show that as $m \to \infty$, $\beta_m^\infty \to \beta_0$, so that if r and m simultaneously tend to infinity, the maximum likelihood estimator is consistent.

3.2. Consistency of the maximum likelihood estimator

The likelihood for $\theta = (\beta, \sigma)$ is given by $L(\theta; y) = \prod_{i=1}^{r} L(\theta; y^{(i)})$, where

$$L(\theta; y^{(i)}) = \int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{m} f(y_j^{(i)} \mid \eta^{(i)} = \beta^T x^{(i)} + \sigma u^{(i)}) \right\} \phi(u^{(i)}) du^{(i)}.$$

We write $\ell_i(\theta; y^{(i)}) = \log L(\theta; y^{(i)})$ for the contribution to the log-likelihood from $y^{(i)}$, and $s_i(\theta; y) = \nabla_\theta \ell_i(\theta; y)$ for the corresponding score function. In the case that m is fixed and $r \to \infty$, the results of White (1982) show that the maximum likelihood estimator $\hat{\theta}_m^r = (\hat{\beta}_m^r, \hat{\sigma}_m^r)$ of θ converges to the value θ_m^∞ which solves $\bar{s}(\theta) = E\left\{ s_i(\theta, Y^{(i)}) \right\} = 0$, where the expectation is taken over the true distribution of $Y^{(i)}$ and the covariates $x^{(i)}$.

Intuitively, for large m, it should be possible to obtain an estimate of the value of each linear predictor $\eta^{(i)}$ from the data $y^{(i)}$ that is close to the true value $\eta_0^{(i)}$. This means that for sufficiently large m, inference given the data y should be similar to the inference we would obtain if we were given the true value $\eta_0^{(i)}$ of each linear predictor $\eta^{(i)}$, and assume a linear model $\eta^{(i)} = \beta^T x^{(i)} + \sigma u^{(i)}$, with the u_i independent $N(0,1)$ error terms. Thus, for large m, the problem is reduced to studying the impact of incorrectly assuming that the errors in a linear model are normally distributed.

To formalize this argument, write

$$\ell_i(\theta; \eta_0^{(i)}) = -\frac{1}{2} \log(2\pi\sigma) - \frac{1}{2\sigma^2}(\eta_0^{(i)} - \beta^T x^{(i)})^2$$

for the log-likelihood for θ in the linear model given $\eta_0^{(i)}$, and $s_i(\theta; \eta_0^{(i)}) = \nabla_\theta \ell_i(\theta; \eta_0^{(i)})$ for the corresponding score function. We obtain the following result, whose proof is given in the appendix.

Lemma 1. *As $m \to \infty$, $s_i(\theta; y^{(i)}) = s_i(\theta; \eta_0^{(i)}) + o(1)$.*

In the linear model setting, the impact of the distribution of the error term has been well studied. We may use the results from this setting to show that the asymptotic bias in $\hat{\beta}$ diminishes with m.

Lemma 2. *As $m \to \infty$, $\beta_m^\infty \to \beta_0$.*

Proof. If we observe $\eta^{(i)}$ from the linear model $\eta^{(i)} = \beta^T x^{(i)} + \epsilon^{(i)}$, the ordinary least squares estimator of β, the maximum likelihood estimator if we assume that $\epsilon^{(i)} \sim N(0, \sigma^2)$, is consistent irrespective of the true distribution of the error $\epsilon^{(i)}$. Writing $\bar{s}(\theta; \eta) = E\left\{ s_i(\theta, \eta^{(i)}) \right\}$, $\bar{s}(\theta^*; \eta) = 0$ has unique solution $\theta^* = (\beta_0, \sigma^*)$, for some σ^*. But for any θ as $m \to \infty$, $\bar{s}(\theta; y) = \bar{s}(\theta; \eta) + o(1)$, so $\bar{s}(\theta^*; y) \to 0$ but $\bar{s}(\tilde{\theta}; y) \not\to 0$ for any fixed $\tilde{\theta} \neq \theta^*$. So θ_m^∞, which satisfies $\bar{s}(\theta_m^\infty; y) = 0$, must tend to θ^* as $m \to \infty$, and $\beta_m^\infty \to \beta_0$ as claimed.

Theorem 1. *As $r, m \to \infty$, $\hat{\beta}_m^r \to \beta_0$ in probability.*

The maximum likelihood estimator is consistent in this setting because the increasing number of observations made per random effect allow us to obtain a good estimate of each $\eta^{(i)}$. Intuitively, if we define a composite likelihood which uses small subsets of the data in each component likelihood, then we will be unable make use of this increasing amount of information available on each random effect. The resulting composite likelihood estimator cannot be expected to be consistent under the same misspecification.

3.3. Inconsistency of fixed-order composite likelihood estimators

We now show that a marginal composite likelihood estimator may be inconsistent under random-effects misspecification, as r and m tend to infinity. We define the composite likelihood by specifying a set S_m that contains the set of subsets of $\{1, \ldots, m\}$ used to construct the component likelihoods. For ease of notation, we assume that $w_t = 1$ for each subset $t \in S_m$, and restrict attention to composite likelihoods where S_m consists of $N_m \geq 1$ sets of fixed size k. We define a k-wise log-likelihood as $\ell_i^{(k)}(\theta, y^{(i)}) = \sum_{t \in S_m} \log L(\theta, y_t^{(i)})$. For the all-pairwise likelihood, $k = 2$ and $N_m = m(m-1)/2$. Other pairwise likelihoods could also be constructed: for instance S_m could contain the $N_m = m - 1$ pairs of the form $\{j, j+1\}$, for $j = 1, \ldots, m-1$.

We write $\tilde{\theta}_k^{m,r} = (\tilde{\beta}_k^{m,r}, \tilde{\sigma}_k^{m,r})$ for a k-wise composite likelihood estimator. Suppose that $m \geq k$ is fixed. We now show that as $r \to \infty$, any k-wise composite likelihood estimator (with $N_m \geq 1$) tends to θ_k^∞, the limit of the maximum likelihood estimator in an alternative two-level model in which m is fixed at k, as $r \to \infty$. The alternative model uses exactly the same modeling assumptions as the original, but only the observations $y_{1:k}^{(i)}$ are used, rather than the full set $y_{1:m}^{(i)}$. The limit θ_k^∞ is found by taking the root of the expected score $E\{s_i(\theta, Y_{1:k}^{(i)})\}$, where the expectation is taken under the true distribution of $Y_{1:k}^{(i)}$ and $x^{(i)}$.

Lemma 3. *Suppose m is fixed such that $N_m \geq 1$. As $r \to \infty$, $\tilde{\theta}_k^{m,r} \to \theta_k^\infty$ in probability.*

Proof. We define $s_i^{(k)}(\theta; y^{(i)}) = \nabla_\theta \ell_i^{(k)}(\theta; y^{(i)})$ to be the composite score function given $y^{(i)}$. The k-wise likelihood estimator solves the estimating equation $\frac{1}{r} \sum_{i=1}^r s_i^{(k)}(\theta; y^{(i)}) = 0$, so by results of Yi and Reid (2010), it converges to the root of $\bar{s}^{(k)}(\theta) = E\{s_i^{(k)}(\theta; Y^{(i)})\}$, where the expectation is taken over the true distribution of $Y^{(i)}$ and the covariates $x^{(i)}$.

But

$$\bar{s}^{(k)}(\theta) = E\left\{ \sum_{t \in S_m} s_i(\theta, Y_t^{(i)}) \right\} = \sum_{t \in S_m} E\{s_i(\theta, Y_t^{(i)})\} = N_m E\{s_i(\theta, Y_{1:k}^{(i)})\}$$

since $Y_t^{(i)}$ are identically distributed for each $t \in S_m$. The limit of the k-wise likelihood estimator is given by the root of $\bar{s}^{(k)}(\theta)$, or equivalently the root of $E\{s_i(\theta, Y_{1:k}^{(i)})\}$, which is θ_k^∞.

Since the same limit holds for any fixed m, it also holds in the limit as $m \to \infty$, which leads directly to our main result.

Theorem 2. *As $r, m \to \infty$, $\tilde{\beta}_k^{m,r} \to \beta_k^\infty$ in probability.*

For example, the limit of the pairwise likelihood estimator of β as r and m tend to infinity is β_2^∞, the limit of the maximum likelihood estimator when m is fixed at two. In most cases, the pairwise likelihood estimator will be inconsistent as r and m tend to infinity, unless the random-effect distribution is correctly specified. The same conclusion holds for any k-wise likelihood estimator, provided k is fixed and does not grow with m, although we expect the asymptotic bias to diminish with k, as a result of Lemma 2.

3.4. A numerical example

Suppose that each $x^{(i)} \sim \text{Bernoulli}(1/2)$, and that each observation $y_j^{(i)}$ is binary, with $pr(Y_j^{(i)} = 1 \mid \alpha, \beta, \sigma, u^{(i)}, x^{(i)}) = \text{logit}^{-1}(\alpha + \beta x^{(i)} + \sigma u^{(i)})$. In the model, we assume that $u^{(i)} \sim N(0,1)$ but, in truth, suppose that $u^{(i)}$ has a skew-normal distribution (Azzalini (1985)); here we change the level of skewness of the distribution by varying the shape parameter a, altering the other parameters of the distribution as we do so to ensure that the mean remains fixed at 0 and the variance at 1. The case $a = 0$ corresponds to the $N(0,1)$ distribution, in which case the model is correctly specified, and moving a further from 0 increases the degree of misspecification.

Figure 1 shows a contour plot of the limit α_m^∞ for various values of m and a, when $\alpha_0 = -0.5$, $\beta_0 = 1$, and $\sigma_0 = 1$. As expected, for each fixed $a \neq 0$, the asymptotic bias in $\hat{\alpha}$ diminishes with m. Figure 2 shows a cut across this contour plot at $m = 2$, which gives the limit of the pairwise likelihood estimator of α. If $a \neq 0$, the pairwise likelihood estimator of α is not consistent. The limiting value of a similarly defined k-wise likelihood estimator may be obtained by a cut across the contour plot at $m = k$. In the limit as r and m tend to infinity, the maximum likelihood estimator is consistent, but the k-wise likelihood estimator is not, for any fixed k.

The asymptotic bias in $\hat{\beta}$ is far smaller than that for $\hat{\alpha}$, and β_m^∞ is close to 1 even for small m. We compute the integrals involved in the likelihood numerically, leading to small errors in our computation of the limit of the estimators. Since the asymptotic bias in $\hat{\beta}$ is small, it is hard to distinguish the true asymptotic bias from the error in the computed limit. However, for each a and m we consider,
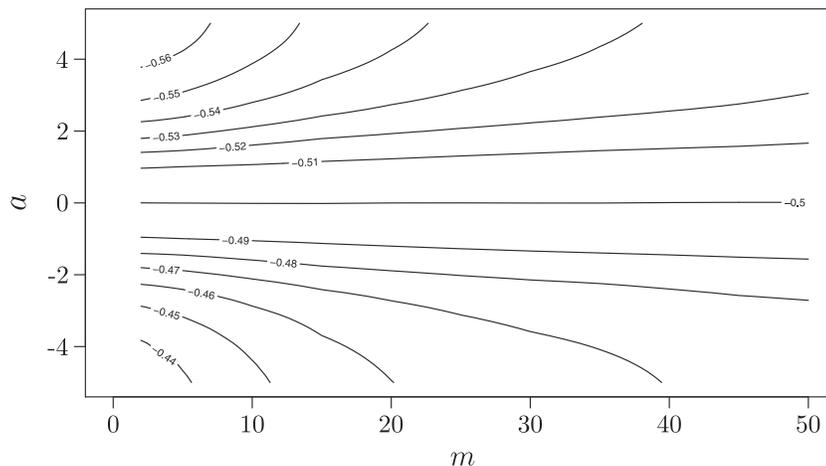
Figure 1. The limit of the maximum likelihood estimator of α as $r \to \infty$ in the two-level model, where $\alpha_0 = -1$, for various skew-normal random-effect distributions. The model is correctly specified in the case $a = 0$.
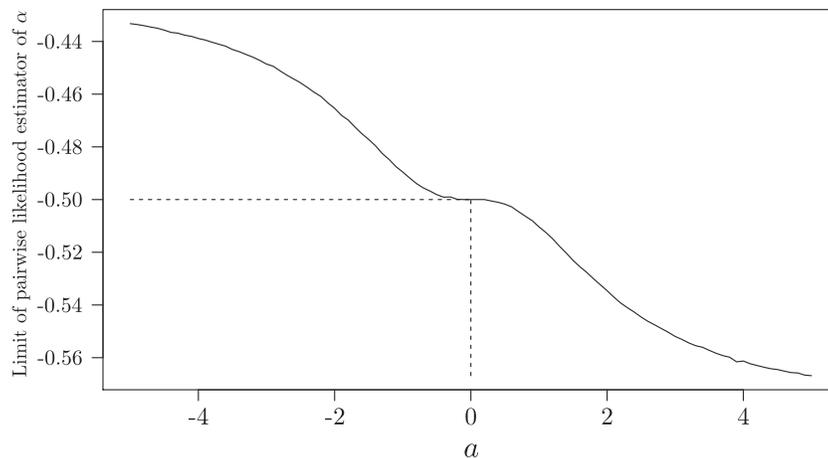

Figure 2. The limit of the pairwise likelihood estimator of α in the two-level model, for various skew-normal random-effect distributions. The estimator is only consistent when the model is correctly specified, at $a = 0$.

we compute $\beta_m^\infty \in (0.99, 1.01)$, so the asymptotic bias in each estimator of β is negligible.

This agrees with the conclusions of McCulloch and Neuhaus (2011), who argue that the extent of asymptotic bias due to random effect misspecification is often larger when estimating an intercept than for other parameters. In practice, we are typically more interested in the effect of a covariate than in the intercept. In that case, the impact of asymptotic bias due to random-effect misspecification may be relatively small.

4. Robustness in Pairwise Interaction Models

We now consider models of the type considered by Bellio and Varin (2005). Here, each observation involves two items, one from each of two groups. We call these models pairwise interaction models. The linear predictor can be written as

$$\eta_j^{(i)} = \beta^T x_j^{(i)} + \sigma_1 u_{p1(j)}^{(i)} + \sigma_2 v_{p2(j)}^{(i)},$$

where $p1(j)$ and $p2(j)$ indicate the type-1 and type-2 items involved in observation j. We suppose that there are q type-1 and q type-2 items, and consider the complete matching, where there is one observation involving each possible combination of type-1 and type-2 items. There are $n = 2q$ random effects and $m = q^2$ observations in each replication. We consider the limiting value of various estimators, under misspecification of the random-effect distribution, as the number of independent replications $r \to \infty$, for different values of q. As q increases, both the number of random effects $(rn = 2rq)$ and the number of observations per random effect $(m/n = q/2)$ increase. Because the data provide an increasing amount of information on each random effect, we expect the asymptotic bias in the maximum likelihood estimator to decrease toward zero as q increases. The limit of each estimator as q tends to infinity, for any fixed r, is equal to the limit as r and q tend simultaneously to infinity, since in both cases the limit is a root of the expected score function for a single replication $Y^{(i)}$. Our findings therefore also apply in the more realistic situation where r is small.

For simplicity, we consider the case where there is no covariate, and we fit only an intercept term. We also suppose that the two types of random effect have a common variance, so that $\sigma_1 = \sigma_2 = \sigma$, and we may write

$$\eta_j^{(i)} = \alpha + \sigma u_{p1(j)}^{(i)} + \sigma v_{p2(j)}^{(i)}.$$

Suppose that we have binary responses, and we model $pr(Y_j^{(i)} = 1|\eta_j^{(i)}) = \text{logit}^{-1}(\eta_j^{(i)})$. As before, suppose that in fact u_i and v_i are independent samples from a skew-normal distribution with shape parameter a and other parameters chosen so that the random effects have mean 0 and variance 1.

Since the true maximum likelihood estimator is unavailable, we instead compare the limit of the pairwise likelihood estimator with that of the estimator maximizing the Laplace approximation to the likelihood. Figure 3 shows the limit of the Laplace estimator of α in the case $\alpha_0 = -1$, for various values of q and a. For each fixed q and a, these limiting values were found by taking 5,000 samples from the true distribution of $Y^{(i)}$, computing the Laplace approximation to the log-likelihood surface for each, and averaging to give an approximation to the expected Laplace log-likelihood surface. The limiting value of the Laplace estimator of θ was found by maximizing this surface over θ. Notice that even when
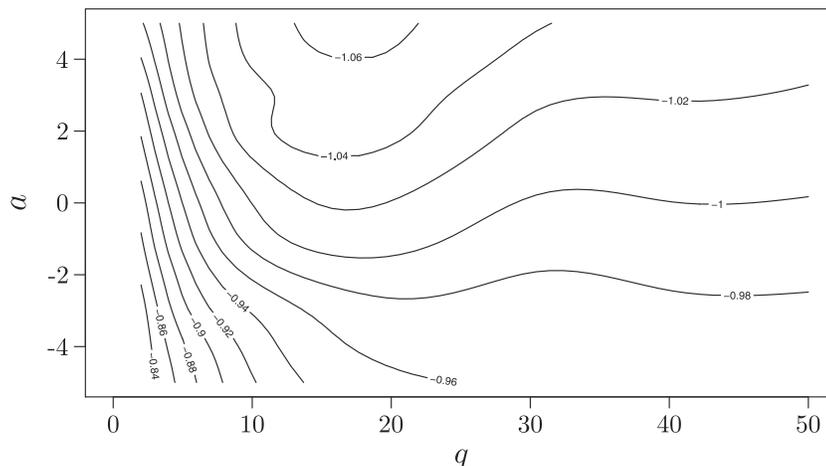
Figure 3. The limit as $r \to \infty$ of the Laplace estimator of α in the pairwise interaction model, where $\alpha_0 = -1$, for different values of q and various skew-normal random-effect distributions. The model is correctly specified in the case $a = 0$.
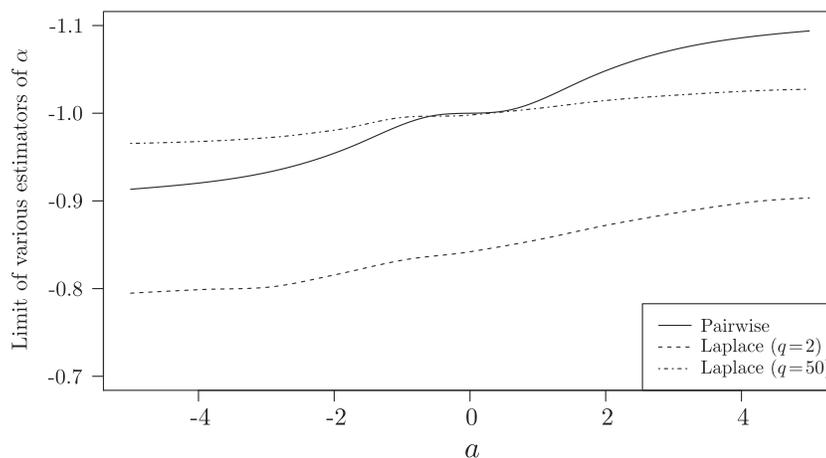

Figure 4. The limit of various estimators of α in the pairwise interaction model, where $\alpha_0 = -1$. The model is correctly specified in the case $a = 0$.

there is no misspecification (at $a = 0$), the Laplace estimator is not consistent and, for small q, the asymptotic bias is quite large.

The limit of the Laplace estimator changes with q, while the limit of the pairwise likelihood estimator remains fixed. The solid line in Figure 4 shows the limit of the pairwise likelihood estimator (for any q), with the Laplace estimator for $q = 2$ (dashed line) and $q = 50$ (dot-dash) overlaid. In the case $q = 2$, the pairwise likelihood performs better than the Laplace approximation, since the

asymptotic bias induced by model misspecification is much smaller than that from error in the Laplace approximation. However, once q is as large as 50, the asymptotic bias of the Laplace estimator in the correctly specified case is negligible, and the Laplace estimator has much lower asymptotic bias than the pairwise likelihood estimator when the model is misspecified.

5. Discussion

We have provided some counterexamples to the the notion that a marginal composite likelihood estimator is at least as robust to model misspecification as the maximum likelihood estimator. Researchers using composite likelihood methods should be aware of these findings, and take care to ensure that claims of robustness made in any particular case are fully justified.

A composite likelihood estimator cannot be claimed, in general, to be either more or less robust to model misspecification than the maximum likelihood estimator. It is natural to ask what characterizes the situations in which either one of these claims do hold. Much more work is required to fully answer this question, and here we provide only some limited intuition. A composite likelihood is constructed from a set of components of the data, and only uses those parts of the model that affect the distributions of these components. The maximum likelihood estimator also uses extra between-component information. If the full distribution is incorrectly specified, but the component distributions are correct, then the composite likelihood estimator achieves increased robustness by discarding the unreliable between-component information.

However, the maximum likelihood estimator is robust to some other types of model misspecification precisely because it makes use of all of the available information. For example, the impact of the random-effects distribution diminishes as the amount of information per random effect grows, so the maximum likelihood estimator is robust to random-effect misspecification in this setting. A fixed-order composite likelihood estimator only uses a fixed amount of information per random effect, leading to a loss of robustness. The choice of random-effect distribution belongs to a wider class of model assumptions whose impact diminishes as the amount of information provided by the data increases. Under deviations from other assumptions of this type, we might also expect fixed-order composite likelihood estimators to suffer a loss of robustness relative to the maximum likelihood estimator.

Acknowledgements

Appendix: Proof of Lemma 1

Proof. We first reparameterize the integral, to write the likelihood as an integral over $\eta^{(i)}$. Let

$$g_i(\eta^{(i)} \mid y^{(i)}, \theta) = \Big[\prod_{j=1}^{m} f(y_j^{(i)} \mid \eta^{(i)}) \Big] \frac{1}{\sigma} \phi\Big(\frac{\eta^{(i)} - \beta^T x^{(i)}}{\sigma}\Big),$$

so that $L(\theta; y^{(i)}) = \int_{-\infty}^{\infty} g_i(\eta^{(i)} \mid y^{(i)}, \theta) d\eta^{(i)}$. We may think of $g_i(. \mid y^{(i)}, \theta)$ as a non-normalized posterior density for $\eta^{(i)}$, given a prior

$$\frac{1}{\sigma} \phi\Big(\frac{\eta^{(i)} - \beta^T x^{(i)}}{\sigma}\Big)$$

which shrinks $\eta^{(i)}$ towards $\beta^T x^{(i)}$. Provided $\sigma > 0$, as m increases, the $\prod_{j=1}^{m} f(y_j^{(i)} \mid \eta^{(i)})$ term, which does not depend on θ, dominates the prior, so that $\hat{\eta}^{(i)}(\theta)$, the maximizer of $g_i(. \mid y^{(i)}, \theta)$ over $\eta^{(i)}$, loses its dependence on θ, and tends towards its true value $\eta_0^{(i)} = \beta_0^T x^{(i)} + b_i^0$.

As m increases, $g_i(. \mid y^{(i)}, \theta)$ becomes well approximated by a normal density about $\hat{\eta}^{(i)}(\theta)$, and the relative error in a Laplace approximation to $L(\theta; y^{(i)})$ tends to zero. Writing $\ell_i(\theta; y^{(i)}) = \log L(\theta; y^{(i)})$, as $m \to \infty$,

$$\ell_i(\theta; y^{(i)}) = \log g_i(\hat{\eta}^{(i)}(\theta) \mid y^{(i)}, \theta) + \frac{1}{2} \log H_\theta(\hat{\eta}^{(i)}(\theta)) - \frac{1}{2} \log 2\pi + o(1),$$

where

$$H_\theta(\eta^{(i)}) = \frac{\partial^2}{\partial \eta^{(i)2}} \log g_i(\eta^{(i)} \mid y^{(i)}, \theta).$$

So, for any two distinct θ_1, θ_2, the difference in log-likelihoods $\ell_i(\theta_1; y^{(i)}) - \ell_i(\theta_2; y^{(i)})$ is

$$\log g_i(\hat{\eta}^{(i)}(\theta_1) \mid y^{(i)}, \theta_1) - \log g_i(\hat{\eta}^{(i)}(\theta_2) | y^{(i)}, \theta_2)$$
$$+ \frac{1}{2} \log H_{\theta_1}(\hat{\eta}^{(i)}(\theta_1)) - \frac{1}{2} \log H_{\theta_2}(\hat{\eta}^{(i)}(\theta_2)) + o(1)$$
$$= \log g_i(\eta_0^{(i)} \mid y^{(i)}, \theta_1) - \log g_i(\eta_0^{(i)} \mid y^{(i)}, \theta_2)$$
$$+ \frac{1}{2} \log H_{\theta_1}(\eta_0^{(i)}) - \frac{1}{2} \log H_{\theta_2}(\eta_0^{(i)}) + o(1),$$

since for any θ, $\hat{\eta}^{(i)}(\theta) \to \eta_0^{(i)}$ in probability as $m \to \infty$. But

$$\frac{H_{\theta_2}(\eta^{(i)})}{H_{\theta_1}(\eta^{(i)})} = \frac{\frac{1}{m}\Big[\sum_{j=1}^{m} \frac{\partial^2}{\partial \eta^{(i)2}} \log f(y_j^{(i)} \mid \eta^{(i)}) + \frac{\partial^2}{\partial \eta^{(i)2}} \log \Big\{\frac{1}{\sigma_2} \phi\Big(\frac{\eta^{(i)} - \beta_2^T x^{(i)}}{\sigma_2}\Big)\Big\}\Big]}{\frac{1}{m}\Big[\sum_{j=1}^{m} \frac{\partial^2}{\partial \eta^{(i)2}} \log f(y_j^{(i)} \mid \eta^{(i)}) + \frac{\partial^2}{\partial \eta^{(i)2}} \log \Big\{\frac{1}{\sigma_1} \phi\Big(\frac{\eta^{(i)} - \beta_1^T x^{(i)}}{\sigma_1}\Big)\Big\}\Big]} \to 1$$

in probability as $m \to \infty$, so

$$
\begin{aligned}
\ell_i(\theta_1; & y^{(i)}) - \ell_i(\theta_2; y^{(i)}) \\
&= \log g_i(\eta_0^{(i)} \mid y^{(i)}, \theta_1) - \log g_i(\eta_0^{(i)} \mid y^{(i)}, \theta_2) + o(1) \\
&= \log \left\{ \frac{1}{\sigma_1} \phi\Big(\frac{\eta_0^{(i)} - \beta_1^T x^{(i)}}{\sigma_1} \Big) \right\} - \log \left\{ \frac{1}{\sigma_2} \phi\Big(\frac{\eta_0^{(i)} - \beta_2^T x^{(i)}}{\sigma_2} \Big) \right\} + o(1) \\
&= \ell_i(\theta_1; \eta_0^{(i)}) - \ell_i(\theta_2; \eta_0^{(i)}) + o(1).
\end{aligned}
$$

Letting $\theta_1 = \theta$, $\theta_2 = \theta + h$ and considering the limit as $h \to 0$, we therefore have $s_i(\theta; y^{(i)}) = s_i(\theta; \eta^{(i)}) + o(1)$, as claimed.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* **12**, 171-178.

Bellio, R. and Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Stat. Model.* **5**, 217-227.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemp. Math.* **80**, 221-239

Lindsay, B. G., Yi, G. Y. and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statist. Sinica* **21**, 71-105.

McCulloch, C. E. and Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statist. Sci.* **26**, 388-402.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J. Comput. Graph. Stat.* **4**, 12-35.

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

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-25.

Xu, X. and Reid, N. (2011). On the robustness of maximum composite likelihood estimate. *J. Statist. Plann. Inference* **141**, 3047-3054.

Yi, G. Y. and Reid, N. (2010). A note on mis-specified estimating functions. *Statist. Sinica* **20**, 1749-1769.

Department of Statistics, University of Warwick, Coventry CV4 7AL, UK.

E-mail: h.ogden@warwick.ac.uk