HIERARCHICAL COVARIANCE MODELLING FOR NONLINEAR REGRESSION WITH RANDOM PARAMETERS

A.C. Singh and S. Wu, Statistics Canada
A.C. Singh, Methodology Research Advisory Group, Statistics Canada, 16-A, R.H. Coats, Ottawa, Ont. K1A 0T6

ABSTRACT

Under a semiparametric framework, estimation of parameters for mixed nonlinear models poses, in general, serious problems in ensuring consistency (not to mention asymptotic optimality) of fixed parameter estimates (both first and second order), and unbiasedness of random parameter estimates. The reason for this seems to be the practice of making the random parameters part of the nonlinear predictor function in order to satisfy certain range restrictions on the conditional means. Consequently, computation of the marginal mean and the covariance becomes intractable in general which renders usual estimation methods for fixed nonlinear models inapplicable. Note that for the usual nonlinear models the random parameters are first specified through the nonlinear predictor function, and then attempt is made to obtain the (marginal) covariance. However, such specification of random parameters is not necessary in view of the following three observations. (i) The target random parameters can be alternatively defined as the hierarchical differences of conditional means, and then they can be made additive to the fixed nonlinear predictor function. Note that this may give rise to variance-mean relationships that should be accounted for in modelling the covariance structure. However, the additivity feature should help to overcome the estimation problem mentioned above. (ii) For BLUP-type optimal estimation of additive random parameters, it is sufficient to specify only the covariance structure, and later the random parameters can be specified (in a wide sense) to match the covariance structure. (iii) Although the usual BLUP is not designed to meet range restrictions, it being a Stein-type shrinkage estimator may often work well provided the fixed predictor function (i.e., the marginal mean) does meet the restrictions. However, if necessary a suboptimal BLUP via ridge-modification of the shrinkage coefficient can be constructed to meet range restrictions while preserving unbiasedness. We, therefore, propose a wide sense specification of random parameters by first modelling the conditional covariances in a hierarchical manner while accounting for variance-mean relationships, instead of the customary reverse route in which the covariance structure is obtained after the nonlinear functional form of random parameters is specified. Estimating functions with suitable properties can be constructed for both fixed and random parameters. An illustrative example for estimation for the widely analysed salamander mating experiment data is presented.

KEY WORDS: Estimating and predicting functions; Logistic regression; Random effects; Ridge BLUP; Shrinkage estimation.

1. INTRODUCTION

Mixed nonlinear models (i.e., models with fixed and random parameters; the random parameters may or may not be part of the nonlinear predictor function) arise naturally in analysing the effects of covariates on discrete outcomes from clustered data. Random parameters are introduced to represent clustering effect, and if ignored, over-dispersion is generally manifested. If sample cluster sizes were large, then cluster effects can be reasonably estimated without treating them as random. However, in practice, often the cluster sizes are small but the number of clusters are large which results in a large number of parameters, number being proportional to the sample size. Parsimony in parameters is achieved by treating similar clusters as connected through a random parameter. This approach is known to have worked well in different applications, see e.g., Breslow and Clayton (1993). Note that estimation of random parameters may or may not be of direct interest in any particular application.

To help focus on the problem discussed in this paper, consider a semiparametric model (with only up to second moment assumptions) for a binary outcome variable y_{ik} for unit k in cluster i given by

$$y_{ik} = \mu_{ik} + e_{ik}, \quad e_{ik} \mid \mu_{ik} \sim (0, \mu_{ik}(1 - \mu_{ik})), \qquad (1.1a)$$

where

$$\text{logit}\, \mu_{ik} = x_{ik}' \beta + z_{ik}' u_i, \quad u_i \sim (0, \sigma_u^2), \qquad (1.1b)$$

where β is a p-vector of fixed parameters, u_i $(i = 1, ..., q)$ are random parameters. Above is an example of a nested design and only one dimension of the random parameter. In general, however, the design need not be nested and/or there may be more than one dimension of the random parameter which may interact with each other. We illustrate this point by the following example.

Example

In the widely analysed Salamander Experiment data, the objective is to compare the mating habits of two types (Roughbutt and Whiteside) of salamander – a lizard-like animal. Forty salamanders (10 of each type and gender) were paired in different ways and observed on three occasions in a given season. The experiment was repeated three times, once in Spring '86, and twice in Fall '86. The design is crossed, details of which can be found in McCullagh and Nelder (1989, ch. 14). The outcome variable is binary taking the value of 1 if the mating is successful, and 0 otherwise. The covariate for fixed parameters is indicator for the pair type with four levels, and the covariate for the two dimensional random parameter (dimension corresponding to gender) is simply the gender indicator. In all there are forty random parameters, 10 for each of the two levels (or the salamander type) of the female and male factors, i.e., one for each animal. Each female salamander was allowed to mate with several males and vice-versa, thus creating a clustering effect for each animal. In terms of model (1.1), this problem can be expressed as follows: Denoting by y_{ij} the observation for i-th female and j-th male, we can write for $i, j = 1,..., 20$,

$$\text{logit}\, \mu_{ij} = \beta_0 + x_{1ij} \beta_1 + x_{2ij} \beta_2 + x_{1ij} x_{2ij} \beta_3 + u_{1i} + u_{2j}, \qquad (1.2)$$

where $u_{1i} \sim (0, \sigma^2_{u1})$, $u_{2j} \sim (0, \sigma^2_{u2})$ are random parameters corresponding to the i-th female and j-th male respectively, x_{1ij} is the indicator of the female whiteside type, and x_{2ij} for the male whiteside type. Here estimation of fixed parameters (first and second order, i.e. βs *and* $\sigma^2 s$) and not of random parameters (i.e. $u's$) is of direct interest.

. Under a completely parametric setup, likelihood-based methods can be used. For example, under a frequentist approach, MLE for fixed parameters (first and second order) can be obtained from the marginal distribution although this would require in general a high dimensional integration. Some important contributions are briefly described. Geyer and Thompson (1992) suggested use of simulated likelihood to compute MLE while use of Metropolis algorithm was suggested by McCulloch (1997). Jiang (1998) proposed a method based on simulated moments to get consistent estimates (although their efficiencies may not be high.) For the special case of Poisson mixed models, under small-sigma asymptotics Sutradhar and Qu (1998) propose a simplified method by approximating the likelihood using a log-transformation of the Gamma distribution for random parameters with first two moments matched with respect to the normal distribution. Under a Bayesian approach, some important contributions are due to Zeger and Karim (1991) who used Gibbs sampling to compute posterior means assuming a flat prior on fixed parameters of generalized linear mixed models, and Ghosh et al. (1998) noting that the posterior need not be proper, proposed a hierarchical Bayes approach with diffuse priors and identified general conditions which ensure proper posteriors. The random parameters under both frequentist and Bayesian approaches are estimated by conditional (or posterior) means to minimize the MSE, although analytic expressions for these estimates are generally intractable due to the problem of high dimensional integration needed for marginal moments. The main virtue of parametric approaches (especially Bayesian) seems to be that the MSE of estimated parameters can take account of estimated second order parameters. However, all these results are subject to the validity of the model. It may be remarked that the model for the conditional means given the random parameters is especially hard to validate because of the unavailability of consistent estimates of random parameters under usual scenarios.

Our objective in this paper is to estimate under a semiparametric setup (i.e., up to second moment assumptions) both fixed and random parameters in nonlinear regression for a discrete response variable. The fixed parameter estimates should be asymptotically consistent and at least for first order parameters asymptotically optimal. The random parameter estimates should have BLUP-type optimality. (BLUP being a shrinkage-type estimator may often meet the range restrictions on the conditional mean. However, in practice, a ridge-type modification may be used if necessary without sacrificing unbiasedness.) Inference on second order parameters is not considered in this paper as it will require higher (third and fourth) moment assumptions; see e.g., Lin (1997). Existing methods using a Bayesian-type argument are the maximum quasi-posterior likelihood method of Schall (1991), penalized quasi-likelihood method of Breslow and Clayton (1993), the BLUP-likelihood method of McGilchrist (1994),and the Henderson-likelihood method of Lee and Nelder (1996), while using a frequentist argument the estimating function method of Waclawiw and Liang (1993), and the method of predicting functions of Singh (1995). However, for first order fixed parameters, none of the methods (except that of Waclawiw-Liang which works with the marginal quasi-likelihood after integrating out random parameters at least approximately under the assumption of normality) provides consistent estimates, for first order random parameters, all produce biased estimates, and for second order parameters, all have problems in providing consistent estimates under the usual asymptotic framework of a large number of clusters and bounded sample cluster sizes. These problems arise because of the difficulty in specifying marginal means and covariances, even more so now for the semiparametric case than for the parametric case.

As shown in the next section, the main reason for the inconsistency problem mentioned above is the practice of making the random parameters part of the nonlinear predictor function presumably in the interest of meeting range restrictions on the conditional means without putting any restrictions on the distribution of random parameters. It may be reasonable to do so for the parametric case because it seems natural to specify various conditional means given random parameters hierarchically from the lowest to the highest level of aggregation. However, for the semiparametric case, it is more natural to go from the highest to the lowest level of aggregation in specifying conditional means because of the unavailability of various conditional distributions. This way random parameters (now defined as differences of hierarchical conditional means) remain outside the nonlinear predictor function. Note that this definition of random parameters does meet the practical need, although it differs from the usual definition involving u-parameters. Now range restrictions on various conditional means impose restrictions on distribution of random parameters, but model for the conditional mean is easier to validate. Thus we are led to essentially two different approaches in specification of nonlinear predictor functions.

There do exist methods which keep random parameters outside the nonlinear predictor function and thus avoid the problem of inconsistency. For example, McCullagh and Nelder (1989, Ch. 14) use a first order Taylor expansion under small-sigma asymptotics to linearize the nonlinear predictor function and then use the standard generalized linear model method to estimate fixed parameters taking account of variance-mean relationships. However, the linear representation may be inappropriate because the interaction term (defined in terms of differences of conditional means) is forced to be zero. Thus, estimation of random parameters may not be well-founded although this was not of direct interest in the example considered by the authors. Sutradhar and Rao (1996) consider second order Taylor expansion which would allow for the interaction term to be present; however, an offset term is added to the mean (i.e., the fixed predictor function) which may no longer satisfy the range restrictions. Drum and McCullagh (1993 p. 680) state that the assumption of small-sigma for random parameters may not be reasonable in general and the positive definiteness of the Taylor linearized covariance matrix is not ensured. Vonesh and Carter (1992) propose

a generalized mixed effects nonlinear regression model where random effects are additive to the fixed nonlinear predictor function and variances of random effects are assumed to be constant as in the case of mixed linear models. However, again this may not yield positive definiteness of the covariance matrix because it does not respect relationships between variance and mean in view of the range restrictions on conditional means. Moreover, importance of the interaction term was not considered although their model can accommodate it. In view of the above concerns, estimates of random effects may become questionable, although such estimation was not considered by the authors.

In this paper, we first show that meeting range restrictions on the conditional means should not be the sole motivation for including random parameters inside the nonlinear predictor function. The reason for this is that there are many practical applications where estimation of random parameters is not of direct interest. Moreover, when random parameters are additive to the fixed nonlinear predictor function, the BLUP estimator being a shrinkage estimator would often meet the restrictions, and if necessary can easily be modified via ridge. We, therefore, propose keeping random parameters outside the nonlinear predictor function, and use a hierarchical breakdown to propose a (proper) decomposition of the conditional mean given random parameters. Here, as mentioned earlier, random parameters are defined simply as hierarchical differences of conditional means. Second order moment assumptions are based on variance-mean relationships to ensure nonnegative definiteness of the covariances and are motivated from realistic considerations. Estimation of random parameters is well-founded because of a proper decomposition of the conditional mean and a proper specification of covariances. Estimates of random parameters have BLUP-type optimality, and that of fixed parameters are consistent and have appropriate optimality.

Section 2 presents a motivation for the proposed method of hierarchical covariance modelling while Section 3 contains its description. Estimation of parameters is discussed in Section 4 and an illustrative example in Section 5. Finally, Section 6 contains concluding remarks.

2. THE INCONSISTENCY PROBLEM

We motivate the proposed method described in the next section by analysing the reasons of the problem of inconsistency in estimation of fixed parameters for mixed nonlinear models. First, let us examine why there is no such problem in the case of general mixed linear models defined as

$$y = X\beta + Zv + e, \tag{2.1}$$

where $e \sim (0, \Sigma_e)$, $v \sim (0, \Sigma_v)$, and are uncorrelated with each other.

Using the method of predicting functions (Singh, 1995), the estimating equations for β and v as originally derived by Henderson (1975) can be easily obtained as

$$\begin{pmatrix} X' & 0 \\ Z' & I \end{pmatrix} \begin{pmatrix} \Sigma_e & 0 \\ 0 & \Sigma_v \end{pmatrix}^{-1} \begin{pmatrix} y - X\beta - Zv \\ 0 - v \end{pmatrix} = 0 \tag{2.2}$$

Consider the asymptotic behaviour of the estimating equation for β as number of random parameters (i.e., the clusters) increase to ∞ but number of sample observations (i.e., the sample cluster size) corresponding to any given random parameter remains bounded. Notice that because of additive random parameters, only a fixed number of linear combinations of $\hat{v}(\beta)$ (an unbiased estimate obtained from (2.2) for given β) appears in the estimating equation for β, and therefore the law of large numbers can be applied under regularity conditions. This will imply that the estimating function for β goes to zero in probability which, in turn, implies consistency of $\hat{\beta}$. An alternative way to see this is as follows. Since the random parameters are additive to the fixed predictor function, they are separate and can be combined with the model error to get a suitable covariance matrix $Z'\Sigma_v Z + \Sigma_e$, which can be used in the optimal estimation of β, i.e., by solving

$$X'(Z'\Sigma_v Z + \Sigma_e)^{-1}(y - X\beta) = 0, \tag{2.3}$$

from which the consistency of $\hat{\beta}$ follows under mild regularity conditions. For estimating the second order fixed parameters appearing in Σ_v and Σ_e (such as when $\Sigma_v = \sigma_v^2 I$, $\Sigma_e = \sigma_e^2 I$), one can use ML or REML estimating equations under normality (cf. Jiang, 1996).

With mixed nonlinear models, however, the random parameters are part of the nonlinear predictor function containing fixed parameters, and therefore, they can't be separated out and combined with the model error in order to estimate β and variance components under a semi-parametric framework. (In the parametric case, the random parameters can, in principle, be integrated out). The estimating equations for β and v in the case of mixed nonlinear models are analogous to (2.2), but clearly there is a problem in applying the law of large numbers to the estimate $\hat{\beta}$ because the corresponding estimating function is not in general a function of a fixed number of linear combinations of $\hat{v}(\beta)$ (there is also the additional problem of bias in this estiamtor). Inconsistency of $\hat{\beta}$ in turn affects consistency of the variance component estimates. It may be noted that for other methods (such as Schall, Breslow-Clayton, McGilchrist, and Lee-Nelder), estimating equations for the first order parameters are (essentially) identical to the ones obtained by the method of predicting functions of Singh (1995).

What is the recourse out of this inconsistency problem? Is it essential to put the random parameters inside the nonlinear predictor function? To answer this, let us make some basic observations. The main reason for the practice of keeping v inside the nonlinear predictor function is probably the need to meet the range restrictions on the conditional mean without putting any restrictions on v. However, the random parameters can be alternatively defined as differences of conditional means (in a hierarchical sense), and then they can stay outside the predictor function containing only fixed parameters, and become additive to the fixed predictor function. This, in fact, is often done in practice. For example, in the model $y = \mu + e$, e can be viewed as a random parameter (although hardly of direct interest) defined as the difference of the conditional mean y and the unconditional mean μ. Now if y is a continuous variable taking values in the real line, then there is no restriction on the random component e. However, if y is discrete (e.g., binary) then this would impose range restriction on e and will result in variance-mean relationships. There

exist many distributions which satisfy different types of variance-mean relationships as exemplified in the literature on generalized linear models. A more pertinent example is the model $y = \eta + v + e$, when $\mu = \eta + v$ is the conditional mean of y given the random parameter v. Again for discrete response variables, there would be range restrictions on μ and η and so both the random components v and e will have to satisfy some restrictions; in particular, their variances may depend on means. This is not really a problem in model specification because one can motivate the functional form of variance-mean relationships by appropriate distributions. Moreover, given that the covariance matrices are properly specified, i.e., they are nonnegative definite, the range restrictions can be met by BLUP (or via a ridge-modification of its shrinkage factor). Another point worth noting is that in defining random parameters as hierarchical differences of conditional means, it may be difficult in general to specify the functional dependence on some unobserved random effects. Fortunately, it turns out that this functional specification is not needed. For BLUP-type optimal estimation of random parameters, only specification of a suitable covariance structure is needed which, in turn, specifies random parameters in a wide sense in order to match the covariance structure.

As mentioned in the introduction, there exist models where random parameters are additive to the fixed nonlinear predictor function. However, care should be taken in accounting for variance-mean relationship, and interactions between random parameters as illustrated below in terms of the small sigma Taylor linearization model of McCullagh and Nelder (1989 Ch.14). For the model (1.2), using first order small sigma asymptotics, (i.e., ignoring terms of order $o_p(\sigma)$), we have

$$y_{ij} \simeq \eta_{ij} + h(\eta_{ij})(u_{1i} + u_{2j}) + e_{ij}, \qquad (2.4)$$

where $h(\eta_{ij}) = \eta_{ij}(1 - \eta_{ij})$, $\eta_{ij} = g^{-1}(x'_{ij}\beta)$,

and g is the logit link function. Here the conditional means can be approximately defined hierarchically as

$$E(y_{ij} \mid x_{ij}) = \eta_{ij}, E(y_{ij} \mid x_{ij}, u_{1i}) := \zeta_{1ij} = \eta_{ij} + h(\eta_{ij})u_{1i}$$

$$E(y_{ij} \mid x_{ij}, u_{1i}, u_{2j}) := \mu_{ij} = \eta_{ij} + h(\eta_{ij})(u_{1i} + u_{2j}) \qquad (2.5)$$

Similarly, $E(y_{ij} \mid x_{ij}, u_{2j}) := \zeta_{2ij} = \eta_{ij} + h(\eta_{ij})u_{2j}$, and the random interaction term ($= \mu_{ij} - \zeta_{1ij} - \zeta_{2ij} + \eta_{ij}$) is set to 0 under (2.4).

We make the following observations about variance-mean relationships for (2.4). Variance of e_{ij} given μ_{ij} is $\mu_{ij}(1 - \mu_{ij})$ which depends on μ_{ij} as expected. Variance of μ_{ij} (given η_{ij}) is $h^2(\eta_{ij})\sigma^2_{u12}$ and covariance of μ_{ij} and $\mu_{ij'}$ $(j \neq j')$ is $h(\eta_{ij})h(\eta_{ij'})\sigma^2_{u1}$, and so on. All these variance-covariances depend on the unconditional mean η_{ij}. However, there are no range restrictions on the variance components $\sigma^2_{u1}, \sigma^2_{u2}$. This may be reasonable before linearization but not afterwards. To see this, note that the conditional mean ζ_{1ij} must lie between 0 and 1. This implies that for each i, the standardized variable u_{1i}/σ_{u1} must satisfy

$$-\min_j \eta_{ij}/h(\eta_{ij})\sigma_{u1} < u_{1i}/\sigma_{u1} < \min_j (1 - \eta_{ij})/h(\eta_{ij})\sigma_{u1} \qquad (2.6)$$

This clearly suggests restrictions on the range of u_{1i}/σ_{u1} and hence on its variance. Without any restrictions on $\sigma^2_{u1}, \sigma^2_{u2}$, the positive definiteness of the covariance matrix of y is not ensured. Although, this may not be a problem for sufficiently small $\sigma^2_{u1}, \sigma^2_{u2}$, in some applications such as that of salamander data, it was observed by Drum and McCullagh (1993) that they can take values greater than one (when the marginal covariance is computed exactly), and so the small sigma assumption may not be reasonable. Note that in the logit scale (i.e., before linearization), $\sigma^2_{u1}, \sigma^2_{u2}$ could be large.

The model (2.4) sets the interaction term to zero. For discrete response variables, it may not be appropriate to ignore interaction terms in the mean scale because of range restrictions. In particular, restrictions on the conditional mean given u_{ij} is not likely to be the same as on the conditional mean when u_{2j} is also given. Thus the two random variables u_{1i} and u_{2j} have an interaction effect on μ_{ij}. The interaction term can be approximately recovered by using a second order small sigma asymptotics (i.e., where only terms of order $o_p(\sigma^2)$ are ignored), as follows:

$$\begin{aligned}
&[\eta_{ij} + h(\eta_{ij})(1 - 2\eta_{ij})(\sigma^2_{u1} + \sigma^2_{u2})/2] \\
&+ h(\eta_{ij})[u_{1i} + (1 - 2\eta_{ij})(u^2_{1i} - \sigma^2_{u1})/2] \\
&+ h(\eta_{ij})[u_{2j} + (1 - 2\eta_{ij})(u^2_{2j} - \sigma^2_{u2})/2] \\
&+ h(\eta_{ij})(1 - 2\eta_{ij})u_{1i}u_{2j} + e_{ij} \\
&:= \eta^*_{ij} + h(\eta_{ij})(u^*_{1i} + u^*_{2j}) + h(\eta_{ij}) \mid 1 - 2\eta_{ij} \mid u^*_{12ij}
\end{aligned} \qquad (2.7)$$

The presence of the interaction term u^*_{12ij} signifies that the conditional mean μ_{ij} given u_{1i}, u_{2j} is not simply the sum of the separate conditional means ζ_{1ij} and ζ_{2ij} minus η_{ij}. In other words, it represents the adjustment to the effect of conditioning on u_{1i} when u_{2j} is also conditioned. Note that the marginal mean is changed to η^*_{ij} which may no longer satisfy range restrictions. Also to specify the covariance structure of e_{ij} in (2.7), one would need third and fourth moment assumptions about u_{1i}, u_{2j} which can be motivated from normality. Although (2.7) is an improved approximation over (2.4), the problem of variance components not being sufficiently small still remains. A way out is to directly define random parameters as additive to the fixed predictor function η_{ij} without relying on small sigma asymptotics. This was in fact proposed by Vonesh and Carter (1992) but the issues regarding variance-mean relationships in specifying moments of random parameters, the need for interaction parameters, and positive definiteness of the covariance matrix of y were not considered. Like mixed linear models, Vonesh-Carter suggested an additive mixed nonlinear model given by

$$y = \eta + Zv + e \qquad (2.8)$$

where η as before is nonlinear in β, and $v \sim (0, \Sigma_v)$. The matrix Σ_v was left unspecified resulting in $q(q + 1)/2$ second order parameters arising from Σ_v where q is the number of v-parameters. A method of moments based on residuals was suggested for estimating Σ_v. However,

under our asymptotic framework where q tends to ∞, the resulting estimate of Σ_v will not be consistent.

The existing methods with additive random parameters do not, however, consider estimation of random parameters. In the hierarchical covariance model proposed in the next section, we take a different approach in defining random parameters via a general hierarchical partitioning of the conditional mean. In the process, interaction terms arise naturally, and the issues of suitable variance-mean relationships along with a possible modification of BLUP to meet range restrictions. The model is motivated from very general considerations about the definition of target random parameters, the role of covariance in motivating the wide sense specification of random parameters, and the Stein-type shrinkage property of random parameter estimates.

3. HIERARCHICAL COVARIANCE MODELLING

We will describe the proposed method in terms of the salamander example. Assuming that the design has replicate observations, y_{ijk} will denote the k-th replicate observation for the factors f_{1i}, f_{2j} applied to the pair (i, j) where f_1, f_2 signify female and male factors. Using a hierarchical partition of the conditional mean $E(y_{ijk}|f_{ij}, f_{2j}, x_{ij})$ (the hierarchy defines the order in which the random factors f_{1i}, f_{2j} appear in conditioning), we have

$$y_{ijk} = E(y_{ijk}|x_{ijk}) + \left[E(y_{ijk}|x_{ijk}, f_{1i}) - E(y_{ijk}|x_{ijk})\right] + \left[E(y_{ijk}|x_{ijk}, f_{1i}, f_{2j}) - E(y_{ijk}|x_{ijk}, f_{1i})\right] + \left[y_{ijk} - E(y_{ij_k}|x_{ijk}, f_{1i}, f_{2j})\right] \quad (3.1a)$$

$$:= \eta_{ijk} + (\zeta_{1ijk} - \eta_{ijk}) + (\mu_{ijk} - \zeta_{1ijk}) + e_{ijk} \quad (3.1b)$$

The interpretation of various conditional means is as follows. μ_{ijk} is the average of y_{ijk} over replications having common characteristics x_{ijk}, and common factors f_{1i}, f_{2j}; ζ_{1ijk} is the average of y_{ijk} over observations which have the common factor f_{1i} and characteristics x_{ijk}; and η_{ijk} is the average of y_{ijk} over observations having common x_{ijk}. Note that the above breakdown holds in general and the choice of hierarchy in the order f_1, f_2 is arbitrary. The random parameters are defined simply as hierarchical differences of conditional means, i.e., $\zeta_{1ijk} - \eta_{ijk}$, $\mu_{ijk} - \zeta_{1ijk}$, and e_{ijk} (estimation of e_{ijk} is usually not of direct interest) which have zero means and suitable covariances. To complete the above semiparametric set-up with second moment assumptions, it remains to specify the covariance structure. It can be done as follows.

Denoting by E_1, E_2, E_3 the hierarchical conditional expectations given x_{ijk}, $\{x_{ijk}, f_{1i}\}$, $\{x_{ijk}, f_{1i}, f_{2j}\}$, respectively, we have the conditional variance of y_{ijk} about μ_{ijk} given $\{x_{ijk}, f_{1i}, f_{2j}\}$ as $\mu_{ijk}(1 - \mu_{ijk})$, and therefore the unconditional variance about μ_{ijk} is

$$\begin{aligned} E_{12} V_3(y_{ijk}) &= E_{12}(\mu_{ijk}) - E_{12}(\mu_{ijk}^2) \\ &= \eta_{ijk} - [E_{12}(\mu_{ijk})]^2 - V_{12}(\mu_{ijk}) \\ &:= \eta_{ijk}(1 - \eta_{ijk}) - \eta_{ijk}(1 - \eta_{ijk})\,\xi_{ijk} \\ &= h(\eta_{ijk})\,(1 - \xi_{ijk}), \end{aligned} \quad (3.2)$$

where $h(\eta_{ijk}) = \eta_{ijk}(1 - \eta_{ijk})$, and ξ_{ijk} is necessarily between 0 and 1 because $E_{12} V_3(y_{ijk}) > 0$. Next, $E_2(\mu_{ijk}) = \zeta_{1ijk}$ and

$$\begin{aligned} E_1 V_2(\mu_{ijk}) &= V_{12}(\mu_{ijk}) - V_1 E_2(\mu_{ijk}) = V_{12}(\mu_{ijk}) - V_1(\zeta_{1ijk}) \\ &:= h(\eta_{ijk})\,\xi_{ijk} - h(\eta_{ijk})\,\xi_{ijk}\,\tau_{ijk} \\ &= h(\eta_{ijk})\,\xi_{ijk}(1 - \tau_{ijk}), \end{aligned} \quad (3.3)$$

where τ_{1ijk} is between 0 and 1, and finally $V_1(\zeta_{1ijk}) = h(\eta_{ijk})\,\xi_{1jk}\,\tau_{1ijk}$. Thus the total variance $h(\eta_{ijk})$ is partitioned into three components as

$$h(\eta_{ijk}) = h(\eta_{ijk})\left[\xi_{ijk}\,\tau_{1ijk} + \xi_{ijk}(1 - \tau_{1ijk}) + (1 - \xi_{ijk})\right] \quad (3.4)$$

If in the hierarchy the f_2 factor was placed before f_1, then the analogous partition would be

$$h(\eta_{ijk}) = h(\eta_{ijk})\left[\xi_{ijk}\,\tau_{2ijk} + \xi_{ijk}(1 - \tau_{2ijk}) + (1 - \xi_{ijk})\right] \quad (3.5)$$

It follows that variance of the main effects $(\zeta_{1ijk} - \eta_{ijk})$ and $(\zeta_{2ijk} - \eta_{ijk})$ are respectively $h(\eta_{ijk})\,\xi_{ijk}\,\tau_{1ijk}$ and $h(\eta_{ijk})\,\xi_{ijk}\,\tau_{2ijk}$, and, that of the interaction effect $(\mu_{ijk} - \zeta_{1ijk} - \zeta_{2ijk} + \eta_{ijk})$ is $h(\eta_{ijk})\,\xi_{ijk}(1 - \tau_{1ijk} - \tau_{2ijk})$ which implies that $\tau_{1ijk} + \tau_{2ijk}$ must be between 0 and 1. So the total variance $h(\eta_{ijk})$ can be partitioned into four components corresponding to two main effects, one interaction, and the model error as

$$h(\eta_{ijk}) = h(\eta_{ijk})\left[\xi_{ijk}\,\tau_{1ijk} + \xi_{ijk}\,\tau_{2ijk} + \xi_{ijk}(1 - \tau_{1ijk} - \tau_{2ijk}) + (1 - \xi_{ijk})\right] \quad (3.6)$$

So far the variances of components are specified. To specify covariances, we first assume that the different main effects are uncorrelated with each other and are uncorrelated with the interaction effect. In addition it seems reasonable to make the following assumptions about correlations.

$$\text{corr}\,(\zeta_{1ijk} - \eta_{ijk}, \zeta_{1i'j'k'} - \eta_{i'j'k'}) = \begin{cases} 1 & \text{if } i = i' \\ 0 & \text{otherwise} \end{cases} \quad (3.7a)$$

$$\text{corr}\,(\zeta_{2ijk} - \eta_{ijk}, \zeta_{2i'j'k'} - \eta_{i'j'k'}) = \begin{cases} 1 & \text{if } j = j' \\ 0 & \text{otherwise} \end{cases} \quad (3.7b)$$

$$\text{corr}\,(\mu_{ijk} - \zeta_{1ijk} - \zeta_{2ijk} + \eta_{ijk},\ \mu_{i'j'k'} - \zeta_{1i'j'k'} - \zeta_{2i'j'k'} + \eta_{i'j'k'}) = \begin{cases} 1 & \text{if } (i,j) = (i',j') \\ 0 & \text{otherwise} \end{cases} \quad (3.7c)$$

The above assumption implies existence of random variables v_{1i}, v_{2j}, v_{12ij} with mean 0 and variance 1, such that

$$\zeta_{1ijk} - \eta_{ijk} = \{h(\eta_{ijk})\,\xi_{ijk}\,\tau_{1ijk}\}^{1/2}\,\nu_{1i},$$

$$\zeta_{2ijk} - \eta_{ijk} = \{h(\eta_{ijk})\,\xi_{ijk}\,\tau_{2ijk}\}^{1/2}\,\nu_{2j},$$

$$\mu_{ijk} - \zeta_{1ijk} - \zeta_{2ijk} + \eta_{ijk} =$$

$$\{h(\eta_{ijk})\,\xi_{ijk}\,(1 - \tau_{1ijk} - \tau_{2ijk})\}^{1/2}\,\nu_{12ij}. \qquad (3.8)$$

That the above assumption is plausible can be demonstrated by the following argument. In the case of ζ_{1ijk}, the condition $-\eta_{ijk} < \zeta_{1ijk} - \eta_{ijk} < 1 - \eta_{ijk}$ for all (j, k) for every i, implies that

$$-\min_{j,k} \eta_{ijk} / \{h(\eta_{ijk})\,\xi_{ijk}\,\tau_{1ijk}\}^{1/2} < \nu_{1i} < \min_{j,k} (1 - \eta_{ijk}) / \{h(\eta_{ijk})\,\xi_{ijk}\,\tau_{ijk}\}^{1/2}, \qquad (3.9)$$

which can be expressed as $-L_i < \nu_{1i} < U_i$.

It can be shown that if $L_i\,U_i > 1$, then there exists a beta variable with parameters $L_i\,(L_i\,U_i - 1)/(L_i + U_i)$, and $U_i\,(L_i\,U_i - 1)/(L_i + U_i)$ taking values in $(-L_i, U_i)$ such that it has mean 0 and variance 1. Similarly, the existence of ν_{2j} follows. Existence of ν_{12ij} follows from the existence of a random variable $\nu_{2(1)ij}$ with mean 0 and variance 1 such that

$$\mu_{ijk} - \zeta_{1ijk} = \{h(\eta_{ijk})\;\xi_{ijk}\,(1 - \tau_{1ijk})\}^{1/2}\,\nu_{2(1)ij}, \qquad (3.10)$$

by using a beta-motivated distribution as above and noting that $\mu_{ijk} - \zeta_{1ijk} - \zeta_{2ijk} + \eta_{ijk}$ is simply $(\mu_{ijk} - \zeta_{1ijk}) - (\zeta_{2ijk} - \eta_{ijk})$.

Thus the covariance matrix Σ of the observation vector y can be obtained as

$$\Sigma = \Sigma_{v1} + \Sigma_{v2} + \Sigma_{v12} + \Sigma_e = Z\,\Sigma_v\,Z' + \Sigma_e \qquad (3.11)$$

where Σ_{v1} is covariance of the vector $\{\zeta_{1ijk} - \eta_{ijk}\}$ and others are simmilarly defined. Here the $(i, j, k)^{th}$ row of the Z-matrix corresponds to the coefficients of ν_{1i}, ν_{2j}, and ν_{12ij}, as given in (3.8), ν is the stacked column vector of random parameters ν_{1i}, ν_{2j}, and ν_{12ij}, and Σ_v is blockdiag $\{J_{v1}, J_{v2}, J_{v12}\}$ where J denotes the matrix of ones and the subscripts denote the dimension corresponding to vectors ν_1, ν_2, and ν_{12}. The matrix Σ is positive definite because Σ_e (the unconditional covariance of e) being diagonal is positive definite, and the other matrices $\Sigma_{v1}, \Sigma_{v2}, \Sigma_{v12}$ are nonnegative definite by construction. The diagonal elements of Σ are $h(\eta_{ijk})$ while the off-diagonal ones are:

$$\{h(\eta_{ijk})\,\xi_{ijk}\,\tau_{1ijk}\,h(\eta_{i'j'k'})\,\xi_{i'j'k'}\,\tau_{1i'j'k'}\}^{1/2} \text{ if } i = i', j \neq j', \qquad (3.12a)$$

$$\{h(\eta_{ijk})\,\xi_{ijk}\,\tau_{2ijk}\,h(\eta_{i'j'k'})\,\xi_{i'j'k'}\,\tau_{2i'j'k'}\}^{1/2} \text{ if } i \neq i', j = j'$$

and

$$\{h(\eta_{ijk})\,\xi_{ijk}\,\tau_{1ijk}\,h(\eta_{i'j'k'})\,\xi_{i'j'k'}\,\tau_{1i'j'k'}\}^{1/2} +$$

$$\{h(\eta_{ijk})\,\xi_{ijk}\,\tau_{2ijk}\,h(\eta_{i'j'k'})\,\xi_{i'j'k'}\,\tau_{2i'j'k'}\}^{1/2} +$$

$$\{h(\eta_{ijk})\,\xi_{ijk}\,(1 - \tau_{1ijk} - \tau_{2ijk})$$

$$h(\eta_{i'j'k'})\,\xi_{i'j'k'}\,(1 - \tau_{1i'j'k'} - \tau_{2i'j'k'})\}^{1/2}$$

$$\text{if } (i, j) = (i', j'), k \neq k'. \qquad (3.12b)$$

So far the parameters $\xi_{ijk}, \tau_{1ijk}, \tau_{2ijk}$ lying in the interval $(0, 1)$ such that $0 < \tau_{1ijk} + \tau_{2ijk} < 1$, were assumed to be quite general. However, their functional form need to be specified to avoid overparametrization as well as in the interest of parismony. Choice of a suitable function can be motivated from the form of $\xi_{ijk}\,\tau_{1ijk}$ in the small sigma Taylor linearized model (2.4) where it is equal to $\sigma^2_{u1}\,h(\eta_{ijk})$. This suggests $\xi_{ijk}\,\tau_{1ijk}$ can be a function of $h(\eta_{ijk})$; however, unlike $\sigma^2_{u1}\,h(\eta_{ijk})$ it is restricted to be between 0 and 1. We, therefore, propose to specify ξ_{ijk}, τ_{1ijk}, and τ_{2ijk} as

$$\log\,(\tau_{1ijk} / \tau_{3ijk}) = p_1\,\log\,h(\eta_{ijk}) + q_1 \qquad (3.13)$$

where p_1, q_1 are two unknown parameters and minus $\log\,h(\eta_{ijk})$ is treated like a known covariate taking values in $(0, \infty)$ and τ_{3ijk} denotes $1 - \tau_{1ijk} - \tau_{2ijk}$. Similary, we write

$$\log\,(\tau_{2ijk} / \tau_{3ijk}) = p_2\,\log\,h(\eta_{ijk}) + q_2 \qquad (3.14)$$

and

$$\log\,(\xi_{ijk} / (1 - \xi_{ijk})) = p_3\,\log\,h(\eta_{ijk}) + q_3 \qquad (3.15)$$

It may be remarked that the above specification of ξ_{ijk}, τ_{1ijk}, and τ_{2ijk} is just one of several possibilities. In the beta correlated binary model of Prentice (1988), specification of second order parameters can be seen as a special case of the above formulation with $p's$ set to zero.

4. ESTIMATION OF PARAMETERS

4.1 First Order Parameters

For our model (3.1), the first order fixed parameters are β, and the first order random parameters are ν_{1i}, ν_{2j}, and ν_{12ij} as defined in (3.8) via hierarchical differences of conditional means. Suppose, the second order (fixed) parameters appearing in $\Sigma_{v1}, \Sigma_{v2}, \Sigma_{v12}$ and Σ_e, namely, p's and q's of equations (3.13), (3.14), (3.15) are known. Denote these parameters by λ. It follows from standard maximum quasi-likelihood theory (or the method of estimating functions) that the optimal estimating function for estimating β is given by

$$(\partial \eta / \partial \beta')\,\Sigma^{-1}\,(y - \eta) = 0 \qquad (4.1)$$

where logit $\eta_{ijk} = x_{ijk}'\,\beta$, and $\partial \eta_{ijk} / \partial \beta' = h(\eta_{ijk})\,x_{ijk}'$, and Σ as in (3.11). The usual Newton-Raphson method can be used to get $\hat{\beta}$ where negative of the Hessian matrix is simply $G'\,\Sigma^{-1}\,G$ where $G = \partial \eta / \partial \beta'$; this is also known as the Godambe Information matrix (Jorgensen and Labouriau, 1995). The estimator $\hat{\beta}$ is asymptotically consistent and optimal with covariance given by inverse of the Godambe Information matrix.

Now, for estimating the random v-parameters (assuming β are known in addition to λ), BLUP (best linear unbiased prediction) equations (see 2.2) are

$$(Z'\ I)\begin{pmatrix}\Sigma_e & 0\\ 0 & \Sigma_v\end{pmatrix}^{-1}\begin{pmatrix}y-\eta\ -\ Zv\\ 0-v\end{pmatrix}=0. \qquad (4.2)$$

More specifically, BLUP of v is given by

$$\hat{v}=\Sigma_v Z'\ (\Sigma_e+Z\Sigma_v Z')^{-1}\ (y-\eta), \qquad (4.3)$$

and its MSE matrix is obtained as $\Sigma_v Z'\ (\Sigma_e+Z\ \Sigma_v\ Z')^{-1}\ Z\Sigma_v$. When a consistent estimate of β is substituted in η, the estimator $\hat{v}$ has the property of being empirical BLUP which is also the case when estimates of second order parametric λ are substituted. EBLUPs are asymptotically BLUP because of the consistency of fixed parameter estimates.

Next, for estimating parameters $\mu\ (=\eta+Zv)$, it follows from (4.1) and (4.3) that the EBLUP estimator is given by

$$\begin{aligned}\hat{\mu}&=\hat{\eta}+Z\Sigma_v Z'(\Sigma_e+Z\Sigma_v Z')^{-1}(y-\hat{\eta})\\ &=\Sigma_e(\Sigma_e+Z\Sigma_v Z')^{-1}\hat{\eta}+Z\Sigma_v Z'\ (\Sigma_e+Z\Sigma_v Z')^{-1}y \qquad (4.4)\\ &:=(I-\Lambda)\hat{\eta}\ +\ \Lambda y\end{aligned}$$

and the MSE matrix about $\eta+Zv$ is approximately (because η is nonlinear in β)

$$\begin{aligned}&\left[(I-\Lambda)Z\Sigma_v Z'(I-\Lambda)'+\Lambda\Sigma_e\Lambda'\right]+(I-\Lambda)\,\mathrm{Cov}(\hat{\eta})(I-\Lambda)' \qquad (4.5)\\ &=(Z\Sigma_v Z')\Sigma^{-1}\Sigma_e+(I-\Lambda)G(G'\Sigma^{-1}G)^{-1}G'(I-\Lambda)'\end{aligned}$$

where $\Sigma=\Sigma_e+Z\ \Sigma_v\ Z'$ as defined earlier. Note that the eigen-values of the shrinkage matrix Λ of (4.4) are between 0 and 1 (because $Z\Sigma_v Z'$ and Σ_e are nonnegative definite by construction), the estimates $\hat{\mu}$ have the Stein-type shrinkage property in the transformed scale. In other words, when the orthogonal matrix of eigen-vectors of Λ is used to transform y and η, the transformed $\hat{\mu}$ is elementwise between the transformed y and η. However, this need not be so in the original scale. In practice it may often happen that the individual components of $\hat{\mu}$ may not be between 0 and 1, but averages over domains of interest may be in the range. More specifically, interest may center on estimating a weighted average of μ_{ijk}'s, i.e., $\Sigma_D a_{ijk}\mu_{ijk}$, over a population domain D, $\Sigma_D a_{ijk}=1$, where only sampled y's are observed. Thus, there is no information on y for the non-sampled part of D, although the corresponding x_{ijk} and z_{ijk} are known and hence η_{ijk} and μ_{ijk} can be estimated. The BLUP estimate of μ_{ijk} for the nonsampled part takes advantage of the correlation with $y-\eta$ for the sampled units and is given by

$$\hat{\mu}_{N-n}=\hat{\eta}_{N-n}+Z_{N-n}\Sigma_v Z_n'\ (\Sigma_{en}+Z_n\Sigma_v Z_n')^{-1}(y_n-\hat{\eta}_n) \qquad (4.6)$$

where N denotes the size of the domain D, and y_n is the n-vector of sampled observations, and so on. In the above equation, subscripts $n, N-n, N$ are inserted whenever appropriate to emphasize reference to sample, nonsample, and population units.

If the weighted average of $\hat{\mu}_N$ does not satisfy restrictions, it suggests need of a restricted BLUP. Note that in our semiparametric framework with only second moment assumptions, it seems difficult to define directly a restricted BLUP (i.e., with a built-in feature to satisfy restrictions) without making further distributional assumptions. We, therefore, start with the usual BLUP and suggest a ridge-type adjustment to the shrinkage matrix Λ to define a suboptimal BLUP as follows.

$$\hat{\mu}(\gamma)=\hat{\eta}+\Lambda(\gamma)(y-\hat{\eta}) \qquad (4.7a)$$

where

$$\Lambda(\gamma)=Z\Sigma_v Z'(\Sigma\ +\gamma I)^{-1} \qquad (4.7b)$$

and γ is a positive constant specified a priori. This also implies that the estimate of v from (4.3) is modfied accordingly. Clearly for a given γ, $\hat{\mu}(\gamma)$ is unbiased but is less efficient than BLUP. However, its MSE will be close to that of BLUP for small γ. The MSE of $\hat{\mu}(\gamma)$ for known η is given by

$$\left[(I-\Lambda(\gamma))\,Z\Sigma_v Z'\,(I-\Lambda(\gamma))'+\Lambda(\gamma)\,\Sigma_e\Lambda(\gamma)'\right] \qquad (4.8)$$

4.2 Second Order Parameters

It follows from Jiang (1996) that as in the case of mixed linear models, we can get under regularity conditions consistent estimates of the second order parameters λ appearing in Σ using Normal-based MLE for the observation vector $y\sim(\eta,\Sigma)$ asuming that η (alternatively, the first order fixed parameters β) are given. In practice, it may be advisable to transform y to Ay so that $A\,\Sigma\,A'$ is diagonal. This can be achieved, for example, by the Gram-Schmidt procedure. Then, the working loglikelihood under normality can be computed without matrix inversion. Specifically, if the n-vector $y^*=A(y-\eta)\sim(0,\mathrm{diag}\ (\psi_1,\ldots\psi_n))$, then $\hat{\lambda}$ is obtained by

$$\min_{\lambda}\ \Sigma_{i=1}^{n}\left[y_i^{*2}/\psi_i\ +\ \log\ \psi_i\right] \qquad (4.9)$$

The Newton-Raphson method with numerical derivatives can be used to minimize the above objective function. To find REML in the case of nonlinear η, one can first linearize it using $\hat{\beta}$ from (4.1), and then use REML (similar to Schall, 1991) for estimating λ. However, for obtaining consistent estimates, MLE would be adequate.

If the data do not have replications, i.e., there is only one observation per pair (i,j), then interaction parameters cannot be estimated because Σ_e is not estimable. However, Σ_{v12} and Σ_e can be collapsed together to define a new model error, and then the remaining parameters can be estimated using MLE.

5. APPLICATION

The example mentioned in the introduction was analysed by many authors; however, estimation of random parameters does not seem to have been considered before. Here we consider estimation of random parameters for illustration purposes only as they may not be of direct interest. Two cases are considered. In the first case, the three repetitions (corresponding to Spring '86 and Fall I and II '86) are treated as separate observations from different animal pairs. In other words, it is assumed that there are no replications, and therefore, random parameters corresponding to the interaction cannot be estimated in general. In the second case, the

three repetitions are treated as replications, and therefore interaction parameters can be estimated.

Case I - Without Replication

We consider three models. (i) Small-sigma Taylor linearization (first order) model and the corresponding covariance as defined by McCullagh and Nelder (1989), (ii) Additive Mixed Nonlinear model of Vonesh and Carter (1992) with some modifications and (iii) the proposed hierarchical covariance model.

The covariance matrix Σ was compared for the three models. In the case of additive mixed nonlinear model, the offdiagonal elements of Σ were modelled in a manner somewhat similar to that for the hierarchical covariance model (see equation 3.12) except that $\xi_{ijk}\tau_{1ijk}$, $\xi_{1ijk}\tau_{2ijk}$, $\xi_{ijk}(1-\tau_{1ijk}-\tau_{2ijk})$ were replaced by arbitrary positive parameters σ^2_{u1i}, σ^2_{u2j} and σ^2_{u12ij} respectively. Moreover σ^2_{u1i} was allowed to depend on x_{ij} through i only, σ^2_{u2j} on x_{ij} through j only, and σ^2_{u12ij} on x_{ij} through both i and j. (In fact, for the present case of no replication, interaction is combined with the model error). For all the three models, the diagonal elements of the covariance matrix are common (the four distinct values are 0.222, 0.247, 0.167, and 0.222 corresponding to pairs WW, WR, RW, and RR respectively) because the four distinct $\hat{\eta}_{ij}$ are same due to data being balanced (cf. Drum and McCullagh, 1993, p 678). This is true when the data from the three repetitions are analysed separately as well as when pooled. The estimates of the first order fixed parameters β (which are used in computing $\hat{\eta}_{ij}$) are 0.693, -2.011, -0.470, and 2.481 for the pooled data and are identical for the three models. Also for the pooled data, estimates for the second order fixed parameters λ are : (i) for the Taylor method, (0.699, 0.631) for $(\sigma^2_{u1}, \sigma^2_{u2})$, (ii) for the additive method, (0.130, 0.000) for the two values of σ^2_{u1i} corresponding to W (whiteside) and R (roughbutt) and (0.194, 0.095) for the two values of σ^2_{u2j}, and (iii) for the hierarchical method, (0.143, 0.121, 0.218, 0.143) for the four values of $\xi_{ij}\tau_{1ij}$ corresponding to pairs WW, WR, RR, and RR; and (0.131, 0.115, 0.176, 0.130) for $\xi_{ij}\tau_{2ij}$. For the nonpooled data, estimates look similar although not shown here. It can be seen that there are only twelve distinct covariance elements (i.e., offdiagonals) as shown in Table 1 for the pooled data based on estimates of λ parameters for each model. The eigenvalues of the shrinkage matrix Λ were also computed to check if they lie between 0 and 1. For the hierachical method , this is of course expected but it turns out that for this data even for other methods, there is no problem with the range of eigenvalues.

EBLUP estimates of μ_{ij} for the sampled pairs were also computed. For the pooled data, Figure 1 shows box plots of estimates (90 of them) for each of the four pair types and for each of the three models. The subscripts T, A, and H denote respectively the methods Taylor, Additive, and Hierarchical. Note that some estimates lie outside the interval (0,1). This, however, can be easily modified using ridge although the results are not shown here. Similar results were obtained for the nonpooled case.

Case II - With Replication

In this case, we allow interaction terms of the type u_{12ij} in all the models and the corresponding variance components. For model (i), this is done by the second order small-sigma Taylor expansion. For this model, variance of the interaction term is a function of σ^2_{u1} and σ^2_{u2} and as such it does not introduce any new parameters. However, their estimates would be different from case I because the model is different now. Also, the present model (i) introduces an offset term in the mean η_{ij} and therefore the estimates of the first order fixed parameters β differ from those for the case of without replication (pooled data). They are obtained as 0.764, -2.200, 0.517, and 2.717. For models (ii) and (iii) estimates for these parameters β remain the same as in the without replication case. Estimates for the second order (fixed) parameters λ are: (i) for the Taylor method, (0.091, 0.305) for $(\sigma^2_{u1}, \sigma^2_{u2})$, (ii) for the additive method, (0.016, and 0.025) for the two values of σ^2_{u1i} corresponding to W (whiteside) and R (roughbutt), (0.043, 0.082) for the two values of σ^2_{u2j}, and nearly zeroes for the four values of σ^2_{u12ij}, and (iii) for the hierarchical method, (0.200, 0.136, 0.469, 0.200) for the four values of τ_{1ij} corresponding to pairs WW, WR, RR, and RR, (0.800, 0.864, 0.531, 0.800) for τ_{2ij}, and (0.082, 0.077, 0.098, 0.082) for ξ_{ij}. In the covariance matrix Σ, there will now be 16 distinct offdiagonal terms. Table 2 shows these values under the three models. The eigenvalues of the shrinkage matrix as in the previous case do lie between 0 and 1 for models (i) and (ii). Also, as in the previous case, Figure 2 compares EBLUPs of μ_{ij} for four types of sampled pairs (i, j) for the three models. Note that the estimates are now different from case I because of the presence of interactions. In fact they all lie in the interval (0, 1) as desired and so ridge-modification is not required. The difference in estimates from the three methods is also affected by the modified $\hat{\eta}^*_{ij}$ for model (i).

6. CONCLUDING REMARKS

In the usual mixed nonlinear models, the random parameters are specified as part of the nonlinear predictor function in much the same way as the specification of fixed parameters. Although this approach seems attractive because of the simplicity in the specification of random parameters (in the nonlinear scale, it is not necessary to introduce restrictions for random parameters because the conditional means would automatically satisfy the appropriate restrictions), this simplicity does not carry over to the estimation stage. In practice, the prediction of random parameters (main effects and interactions) is generally done in the mean scale, and it may be analytically intractable to get these parameter estimates (in the mean scale) and the corresponding MSE. More importantly, complete parametric assumptions are necessary for estimation purposes. However, if only semiparametric assumptions up to second moments are made, then with the available estimation methods problems of inconsistency in the estimation of fixed parameters, and bias in the estimation of random parameters arise. As an alternative, hierarchical covariance models were proposed in this paper, which shift the burden of computing marginal and conditional means and their covariances (when the random parameters are specified as part of the nonlinear predictor function) to that of first specifying the conditional covariances in a hierarchical manner and then specifying the random parameters (additive to the fixed nonlinear predictor function) in a wide sense to match the covariance structure. It was shown that this alternative is not really burdensome because a reasonable guidance can be obtained from small-sigma asymptotics applied to the usual mixed nonlinear models. Most of all, only a semiparametric

framework is needed which may be quite desirable in many applications.

The existing method of small-sigma Taylor linearization (McCullagh and Nelder, 1989, Ch. 14) shares some properties of the proposed method and, in fact, is useful in motivating the form of the covariance structure. Its main limitation seems to be the lack of a proper specification of the covariance matrix in that it need not be positive definite. Similarly, the additive mixed nonlinear model of Vonesh and Carter (1992), like the proposed model, makes random parameters additive to the fixed nonlinear predictor function (as in the case of mixed linear models), but the question of a proper specification of the covariance structure is not considered. In the proposed model, we take a reverse route in that a proper covariance is first specified using hierarchical considerations, and then the form of the random parameters is specified although in a wide sense only.

It would be very useful to compare the customary mixed nonlinear model and the proposed hierarchical covariance model, and in principle, this can be carried out term by term by computing hierarchical differences of conditional means for the mixed nonlinear model and the corresponding covariances. However, for actual comparison, more work is needed which we plan to undertake in the future. Another area requiring further work is that of adjusting MSE approximations to take account of the estimation of second order parameters. Note that for the special case of mixed linear models, various approximations are available, see Singh et al. (1998).

Finally, it may be remarked that although the proposed model was described in the context of binary data, the method is applicable in general to other discrete response variables

ACKNOWLEDGEMENTS

The authors would like to thank Colin McLeod for assistance in computing. The first author's research was supported in part by a grant from NSERC of Canada held at Carleton University under adjunct research professorship.

REFERENCES

Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear models. **JASA, 88**, 9-25.

Drum, M.L., and McCullagh, P. (1993). REML estimation with exact covariance in the logistic mixed model. **Biometrics**, 677-689.

Geyer, C.J., and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. **JRSS (B), 54**, 657-699.

Ghosh, M., Natarajan, K., Stroud, T.W.F., and Efron, B.P. (1998). Generalized linear models for small-area estimation. **JASA, 93,** 273-282

Henderson, C.R. (1975). Best linear unbiased estimation and pediction under a selection model. **Biometrics, 31**, 423-447.

Jiang, J. (1996). REML estimation: asymptotic behaviour and related topics. **Ann. Statist., 24**, 255-286.

Figure 1. Predicted conditional means for sampled pairs (pooled, no replication)

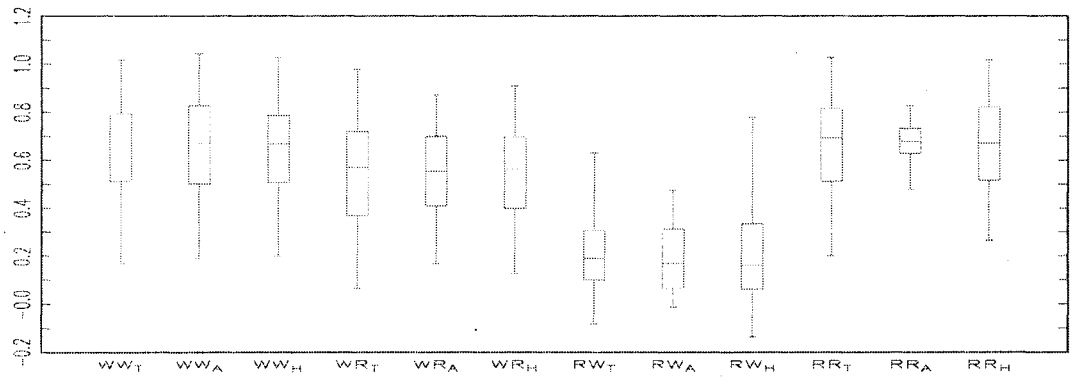
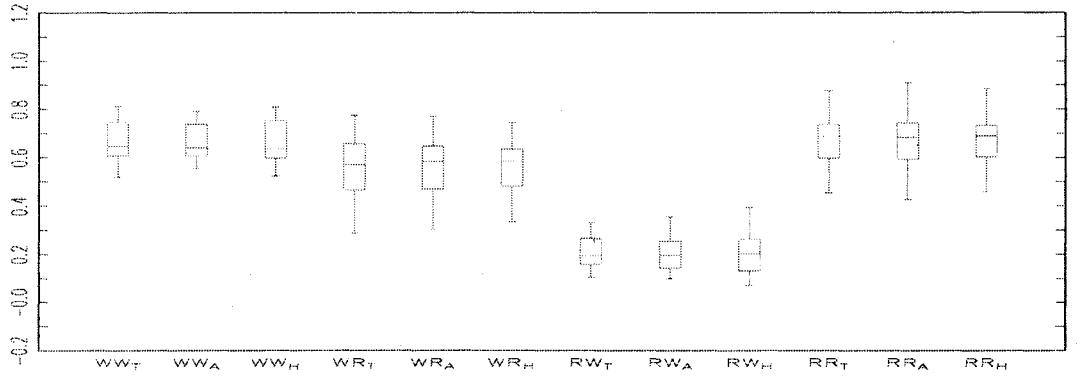


Figure 2. Predicted conditional means for sampled pairs (pooled, with replication)



Jiang, J. (1998). Consistent estimators in generalized linear mixed models. **JASA, 93,** 720-729.
Jorgenson, B., and Labouriau, R.S. (1995). Exponential families and theoretical inference. Lecture Notes, University of British Columbia, Vancouver.
Lee, Y., and Nelder, J.A. (1997). Hierarchical Generalized linear models. **JRSSB, 58,** 619-678.
Lin, X. (1997). Variance components testing in generalized linear models with random effects. **Biometrika, 84,** 309-326.
McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. **JASA, 92,** 162-170.
McCullagh, P., and Nelder, J.A. (1989). **Generalized linear models** (2nd ed.), London: Chapman and Hall.
McGilchrist, C.A. (1994). Estimation in generalized mixed models. **JRSSB, 56,** 61-69.
Prentice, R.L. (1988). Correlated binary regression with covariate specific to each binary observation. **Biometrics,** 44, 1033-1048.
Schall, R. (1991). Estimation in generalized linear models with random effects. **Biometrika, 78,** 719-727.
Singh, A.C. (1995). Predicting functions for generalization of BLUP to mixed nonlinear models. **ASA Proc. Biom. Sec.,** 300-305.
Singh, A.C., Stukel, D.M., and Pfeffermann, D.(1998). Bayesian versus frequentist measures of error in small area estimation. **JRSS (B), 60,** 377-396.
Sutradhar, B.C., and Rao, R.P. (1996). On joint estimation of regression and overdispersion parameters in generalized linear models for longitudinal data. **J. Multivariate Analysis,** 56, No 1, 90-119.
Sutradhar, B.C., and Qu, Z. (1998). On approximate likelihood inference in Poisson mixed model. **Can. J. Statist.,** 26, 169-186.
Vonesh, E.F., and Carter, R.L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. **Biometrics,** 48, 1-17.
Waclawiw, M.A., and Liang, K.-Y. (1993). Prediction of random effects in the generalized linear model. **JASA, 88,** 171-178.
Zeger, S.L., and Karim, M.R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. **JASA, 86,** 79-86.

Table 1. Distinct Offdiagonal Elements of Covariance Matrix (Pooled, without replication)

	Animal Type		Method*		
	1st pair	2nd pair	Taylor	Additive	Hierarchical
Common	WW	WW	0.035	0.029	0.032
Female	WW	WR	0.038	0.030	0.031
	WR	WR	0.043	0.032	0.030
	RW	RW	0.019	0.000	0.036
	RW	RR	0.026	0.000	0.034
	RR	RR	0.035	0.000	0.032
Common	WW	WW	0.031	0.043	0.029
Male	WR	WR	0.038	0.023	0.028
	WW	RW	0.023	0.037	0.029
	WR	RR	0.035	0.022	0.029
	RW	RW	0.018	0.032	0.029
	RR	RR	0.031	0.021	0.029

* Taylor = First order Taylor linearization
Additive = Mixed nonlinear model with additive random effects
hierarchical = Hierarchical covariance model

Table 2. Distinct Offdiagonal Elements of the Covariance Matrix (Pooled, with replication)

	Animal Type		Method*		
	1st pair	2nd pair	Taylor	Additive	Hierarchical
Common	WW	WW	0.004	0.004	0.004
Female	WW	WR	0.005	0.004	0.003
	WR	WR	0.006	0.004	0.003
	RW	RW	0.002	0.004	0.008
	RW	RR	0.003	0.005	0.005
	RR	RR	0.004	0.006	0.004
Common	WW	WW	0.015	0.01	0.015
Male	WR	WR	0.019	0.02	0.016
	WW	RW	0.01	0.008	0.011
	WR	RR	0.016	0.019	0.016
	RW	RW	0.008	0.007	0.009
	RR	RR	0.015	0.018	0.015
Common	WW		0.019	0.013	0.018
Male &	WR		0.024	0.024	0.019
Female	RW		0.01	0.011	0.016
	RR		0.019	0.024	0.018

* Taylor = Second order Taylor linearization
Additive = Mixed nonlinear model with additive random effects
Hierarchical = Hierarchical covariance model