# Open Problems in Mixed Models

- Determining how to deal with a not positive definite covariance matrix of random effects, $\mathbf{D}$, during maximum likelihood estimation algorithms. Several strategies are discussed in Section 2.15. For example, in our own `R` function `lmeFS`, we allow matrix $\mathbf{D}$ to be any (no restriction) but symmetric. As studied in Section 2.16, if matrix $\mathbf{D}$ becomes not positive definite during iterations, function `lme` of library `nlme` in `R` returns an error. Function `lme4` of the library with the same name does not fail and returns a singular nonnegative definite matrix. The question remains how a not nonnegative definite matrix $\mathbf{D}$ can be projected on the space of nonnegative definite matrices, $\mathbb{D}_+$. In particular, shall we benefit from an expensive log-likelihood maximization on the boundary of $\mathbb{D}_+$? This question is closely related to testing what random effects variables are statistically significant.

- Testing the variance-covariance matrix of random effects, particularly testing whether a specific random effect is not statistically significant (variance=0). This question is closely related to difficulties of the numerical implementation described above. The exact $F$-test for a linear mixed model, as a generalization of an ANOVA test, is suggested in Section 3.5, and generalized to a nonlinear mixed model in Section 8.15.2. Tests for overdispersion in the framework of random intercepts in logistic and Poisson models are discussed in Sections 7.3.7 and 7.5.10, respectively, and the test for homogeneity in the meta-analysis model is discussed in Section 5.2.3. However, unlike its linear version most of these tests do not yield the exact/nominal significance level in a small sample, and more work is required to eliminate or reduce this discrepancy. Even the $F$-test by itself may not be very powerful, and a search for a better test is urgent and practically important as a tool for mixed model criticism. Several recent papers study the alternatives, including those of Giampaoli and Singer (2009), and Li and Zhu (2013).

- Testing what variables belong to fixed effects and what variables belong to random effects. Which variables affect the mean function and which variables affect the variance of the dependent variable is not a trivial matter. Existing methods of hypothesis testing work separately with fixed and random effects. We need tests that identify fixed and random parts in a mixed model simultaneously. Practically nothing has been done in this direction. Again, in asymptotic setting, when the number of clusters is large, the information matrix is block diagonal which implies that the choice of the fixed or random can be done separately. For small $N$, this is not true, and therefore simultaneous variable selection is required.

- Development of mixed-model-specific information criteria to address the increasing number of parameters, such as generalization of AIC/BIC, or Mallow's $C_p$. We have sufficient evidence that these criteria are helpful when

adjusting for an increasing number of fixed effects parameters. However, random effects parameters, namely, elements of the random effects covariance matrix, have a different nature and should not be counted in the same way as fixed effects coefficients. In Section 1.6 we suggest some variants of Akaike's criterion treating the mixed model with a large number of parameters as an inverse ill-posed problem, called healthy AIC, but much work remains to be done.

- Variable selection, or more generally, model selection in the framework of mixed models. This topic is closely related to problems formulated previously. Three types of variable selection schema are available: (1) fixed effects variable selection assuming that random effects variables are known, say, random intercepts; (2) random effects variable selection assuming that fixed effects variables are known; and (3) having a set of variables, what variables go to fixed effects, what variables go to random effects, and what variables go nowhere. Only a handful of papers consider the problem, such as a recent paper by Peng and Lu (2012) in an asymptotic setting where selection is much easier. Especially important and difficult is the problem of mixed model selection when the number of potential variables is larger, sometimes much larger than the number of clusters, as in the case of genetics data. Then in addition to the difficulty of the variable selection criterion, a computational burden emerges.

- Power computation and sample size determination for mixed models. An important feature of a mixed model is that two sample sizes should be distinguished: the number of clusters, $N$, and the number of observations per cluster, $n$. Obviously, the number of clusters is more important because when the number of clusters goes to infinity and the number of observations per cluster is fixed, beta-estimates are consistent, but not otherwise. On the other hand, $n$ plays a role in getting the power desired. From asymptotic consideration, the power function of detecting a beta-coefficient $\delta$ versus the zero null hypothesis is equal: $P = \Phi(-Z_{1-\alpha/2} + \delta/\sqrt{V(N,n)})$, where $\Phi$ is the cumulative distribution function of the standard normal distribution, $\alpha$ is the size of the test (typically, $\alpha = 0.05$), $Z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$, and $V(N,n)$ is the variance of the beta estimate. In a particular case when $n = 1$, we arrive at the standard formula for sample size determination in the double-sided Wald test, $n = (Z_P + Z_{1-\alpha/2})^2 V/\delta^2$ (Demidenko, 2007b). For example, in the case of a linear balanced model, the variance of the beta-coefficient is a diagonal element of the matrix $N^{-1}\sigma^2 \left(\mathbf{X}'(\mathbf{I} + \mathbf{ZDZ}')^{-1}\mathbf{X}\right)^{-1}$, which is a function of $N$ and $n$. A similar formula for the variance can be applied to generalized and nonlinear mixed models, but then it would require integration to obtain the Fisher information matrix. We need to extend these computations to the case of unbalanced data where the distribution of the number of observations per cluster is a part of the statistical design.

- Design of optimal experiments with mixed models. In engineering and industrial settings, fixed and random effects design matrices may be chosen as a part of experimental design. Although the theory of optimal design of experiments is well studied for linear and nonlinear regression models, not many theoretical developments exist in the mixed model framework. Similar to sample size

determination, a mixed model leads to a nontrivial choice between design for fixed and random effects. Only a handful of papers exist on the topic (e.g., Dette et al., 2010) and more research is needed to address this important application of mixed models. The idea of adaptive design seems attractive (Glaholt et al., 2012).

- Statistical hypothesis testing using noniterative quadratic estimators of $\sigma^2$ and matrix $\mathbf{D}$ are discussed in Sections 3.10 to 3.12. The closed-form expression for these estimators makes it possible to study the small-sample properties and development of new statistical tests. In particular, we need extensive simulations; how these estimates behave for small number of clusters, $N$; and non-Gaussian, possibly skewed, distributions with long tails. The importance of noniterative estimates is explained by the possibility of using them for testing a statistical hypothesis on $\mathbf{D}$ that creates an opportunity to testing the statistical significance of the random effects.

- Studying the small-sample-size properties of the beta-estimates and the respective statistical hypothesis tests. A paramount question regarding nonlinear statistical models involves small-sample properties of estimators. The linear mixed effects model is the simplest nonlinear statistical model in which advances can be achieved. Currently, the $t$-test is used for the statistical significance of fixed effects coefficients assuming that the covariance matrix of random effects is fixed and known. We can adjust for the fact that an estimate of matrix $\mathbf{D}$ is used that would lead to widening the confidence intervals. More research should be done in studying how the confidence intervals and hypothesis testing can be improved using the profile likelihood; see Section 3.4 as an introduction.

- The Gauss-Markov theorem for mixed model or estimated GLS. The Gauss-Markov theorem is the cornerstone of linear models. If the scaled covariance matrix of random effects, $\mathbf{D}$, is known, the estimated generalized least squares estimator for fixed effects coefficients, $\boldsymbol{\beta}$, is BLUE (best linear unbiased estimator) and has a minimum covariance matrix (the estimator is efficient) among all (linear and nonlinear) unbiased estimators with normal observations. In Section 3.6.1 it is shown that the maximum likelihood estimator of fixed effects is unbiased in a small sample, and it remains unbiased with many other quadratic estimates of $\mathbf{D}$, such as MINQUE, MM, and VLS, discussed in Chapter 3. Thus, the set of unbiased estimator for $\boldsymbol{\beta}$ is nonempty, and therefore the question as to which is the most efficient unbiased estimator is valid. Several avenues can be taken to tackle this problem. For example, one may seek an estimate of $\mathbf{D}$ as a quadratic function of the observations that minimizes the covariance matrix of the fixed effects coefficients or its derivatives at $\mathbf{D} = \mathbf{0}$. A good start may be the simplest random effect model, the meta-analysis model, discussed in Chapter 5.

- Develop better computational algorithms for generalized linear and nonlinear mixed models, including maximum likelihood estimation based on numerical integration. Three types of quantities are computed in traditional (or approximate) log-likelihood maximization: the values of the log-likelihood function,

its derivatives (the score equations), and the Hessian (or information) matrix. It should be noted that score equations are most important because the MLE is defined as the solution of these equations. The Hessian estimate is less important because any positive definite matrix provides the convergence of the maximization algorithm. While the existing methods concentrate on the log-likelihood approximation via integration, we should pay more attention to score equations. The improved Laplace approximation suggested in Section 7.1.2 can be used to approximate the integral or the Gauss-Hermite quadrature. To speed up the convergence, one can increase the number of nodes while iterations progress.

- Improve computational algorithms by recognizing that the beta-parameters, $\boldsymbol{\beta}$, and Cholesky factor elements, $\boldsymbol{\delta}$, can be combined in a linear combination $\mathbf{A}_i\boldsymbol{\beta} + \mathbf{U}_s\boldsymbol{\delta}$, as outlined in Section 8.14. The prototype of the algorithm is implemented in our function `nlmeFSL`, but more efficient C/FOTRAN code is required to see the full advantage.

- Starting values for linear and nonlinear mixed model estimation algorithms. A good choice for starting values may be crucial for a successful run, especially for generalized and nonlinear mixed models with a large number of random effects or complicated variance-covariance structure. It seems that the most important is the choice of matrix $\mathbf{D}$. Several recommendations may be explored: (1) a few iterations of the fixed-point algorithm, as discussed in Section 2.13; and (2) noniterative quadratic estimates of matrix $\mathbf{D}_*$ and $\sigma^2$, as discussed in Sections 3.10 to 3.12. Less obvious and yet more important is the choice for starting values for nonlinear mixed models. Beta-parameters may be estimated using `glm` and `nls,` and matrix $\mathbf{D}_*$ may be estimated based on the residuals. Definitely, more work is required to study theoretical properties of these starting values and to test these suggestions via extensive simulations involving 'difficult' data sets.

- Development of the criterion that the MLE of $\mathbf{D}$ is a positive definite matrix. We have developed such criterion for a linear mixed model in Section 2.6 and for a meta-analysis model in Section 5.1.2. A similar criterion is needed for generalized linear and nonlinear mixed models. This criterion can serve as a preliminary test for the adequateness of the mixed model and random effects against overspecification.

- Development of an adequate stopping criterion (criteria) for maximization of the log-likelihood function, especially with generalized linear and nonlinear models. The log-likelihood maximization may be a complex problem, especially when variables are close to collinear or when the nonlinear model has a complicated variance-covariance structure. Proximity between iterations defined as $\|\mathbf{a}_s - \mathbf{a}_{s-1}\| < \varepsilon$, where $\varepsilon$ is a small number, does not guarantee that $\mathbf{a}_s$ is the point of the global maximum. In order to claim that iterations converged to a local maximum, the gradient of the log-likelihood function at $\mathbf{a}_s$ must be zero. A question arises: What small is small? For instance, is $10^{-1}$ or $10^{-8}$ a small gradient? The interpretable stopping criteria were developed for

the nonlinear least squares, and discussed in Section 13.3.5. Similar criteria should be developed for mixed model maximization algorithms.

- Existence of the maximum likelihood estimate (MLE) for generalized linear and nonlinear mixed models. The MLE may not exist; thus, before starting a maximization algorithm one has to be sure that the maximizer exists. You may spend a lot of time on model testing and playing with the start values but eventually fail because the MLE, simply does not exist—the criteria for MLE existence are important. For a linear mixed effects model, MLE exists with probability 1, as discussed in Section 2.5. Things become more complicated for generalized and nonlinear mixed models. For example, in the case of binary dependent data, the conditions for the data separation must be fulfilled, as presented in Section 7.10. You may generalize the existence criteria developed for nonlinear regression by the author (Demidenko, 1989, 2000, 2008) to the existence of the MLE in the mixed model.

- Uniqueness of the log-likelihood maximum. The log-likelihood function is not a quadratic function even for a linear mixed model because of the presence of variances and covariances. Thus, the possibility exists of converging to a local maximum. As proven by Demidenko (2000), for many nonlinear regressions the probability that the normal equation has two or more distinct solutions is positive. We need criteria by which one can test whether the maximum log-likelihood found is global, as suggested by Demidenko (2008). As a conjecture, the log-likelihood function for a linear mixed model is unimodal (local maximum=global maximum). For a generalized linear mixed model, such as the logistic or Poisson model, this question is open A good start is to investigate the uniqueness of the maximum likelihood estimate for a Poisson model with random intercepts. The uniqueness criteria for a general nonlinear mixed model are even more difficult than those for a nonlinear regression but are not completely intractable. In general, criteria for uniqueness are model-dependent and mathematically challenging.

It should be noted that some literature exists that deals with some of the problems outlined above. We have deliberately not tried to mention all existing publications in these directions because it would require much more space. Therefore, an important part of advancing along the lines of these problems will be a careful review of work already done.