Referee's report

STC 0830: On a hybrid data cloning method
and its application in generalized linear mixed models

Authors: Hossein Baghishani, Håvard Rue and Mohsen Mahammadzadeh

Journal: *Statistics and Computing*

Date: *July 19, 2010*

The manuscript describes a combination of the data cloning (DC) method (which normally requires time consuming MCMC) and a novel integrated nested Laplace approximation (INLA, to replace MCMC) to calculate maximum-likelihood estimates in GLMM. Besides computational aspects the paper also discusses asymptotic behavior of the proposed method. The topic of the paper is certainly interesting and suitable for *Statistics and Computing.* Nevertheless, I have some concerns as described below (the first two are highly important for me) .

1. The asymptotic properties are mixed with the description of the algorithm such that it is rather difficult to follow the line of the paper, especially in Sections 2 and 3. I would suggest that all Theorems and Lemmas appearing in Sections 2 and 3 are moved into one "Technical" section (Section 5 now) in a well structured way and the main results of these theorems and lemmas are only used in Sections 2 and 3 to explain how and why the algorithm works. The authors should also better describe why they need asymptotic normality of the hybrid DC-based distribution. I assume that they need it to be able to use the INLA within DC. But this message is somewhat hidden in a bunch of theorems and lemmas.

   I think that it is necessary to divide the paper into two parts

   a) the first part accessible to most of the readers of the journal who are especially interested in relationship between statistics and computing and who are not particularly strong in theoretical statistics. They only need to know how the algorithm works and only briefly why it works;

   b) the second part which rigorously proves why the algorithm works and which can be skipped by readers interested purely in computation (and estimation of their models).

2. The data used in the examples are publicly available (I assume). Hence it would be useful for the readers to prepare well commented scripts with all the analyses shown in the paper such that the reader will be able to reproduce him/herself the results. These scripts should be available on-line (on `http://www.r-inla.org`?) and also to the reviewers of the paper. I do not think that current description of the use of R INLA package on p. 8 is sufficient.

3. In Section 4.2 you use PQL as a benchmark ML algorithm. Why not Gaussian quadrature for which the results can be obtained using R package `lme4` or SAS `proc nlmixed`? In general, PQL can lead to biased results so I would prefer not to use it. Further, in Section 4.3 you report REML estimates. What exactly do you mean by that in the binary data context of this section? Which numerical method was used to calculate the estimates?

Minor remarks

P. 4 (31) Typo *withe*.

P. 5 (53) The second formula is mathematically not correct. You have to move $1/\sqrt{k}$ to left-hand-side. Current version of the formula does not make sense since $k \to \infty$ and hence anything containing k must not appear on the right-hand-side.

P. 7 (35) I would prefer $Y_i \,|\, \boldsymbol{\beta},\, u_i$ (similarly to p. 9 (24)) instead of $Y_i \,|\, \boldsymbol{\beta}, p_i$ which is confusing (p_i depends on $\boldsymbol{\beta}$ and u_i).

P. 19 (54) Typo *therefor*.

References The list of references contains many (not only typographical) errors. Typographically, titles of journals are sometimes typed in lower case letters whereas upper case letters should be used. With respect to the facts, at least the following errors are present:

- ✧ Pages for Rue, Martino and Chopin (2009) are 319–392 (counting also Discussion) and not 1–35.
- ✧ Tierney and Kadane has been published in 1986 and not in 2009 and under different title ("Accurate approximations for posterior moments and marginal densities").