# Referee's report

**bimj–201500110:** Mixture of multivariate $t$ linear mixed models
for multi-outcome longitudinal data with heterogeneity

**Authors:** Wan-Lun Wang

**Journal:** *Biometrical Journal*

**Date:** *August 1, 2015*

---

The author considers an extension of a classical linear mixed model in several respects:

1. Multivariate $t$ distribution is assumed for both random effects and the errors.
2. A finite mixture of the proposed models is considered to allow for situation when study subjects arise from $G$ heterogeneous populations. This additionally allows for model-based clustering or discriminant analysis based on longitudinal data.

Additionally, "multi-outcome" situation (multiple longitudinal outcomes gathered at each visit) is explicitely considered. To estimate the model parameters, the author derives necessary expressions to apply so called AECM algorithm (variant of the EM algorithm) of Meng and Dyk (1997). The method is illustrated on the analysis of a historical dataset on patients with a primary biliary cirrhosis.

I have some major comments to the current paper.

1. In principle, linear mixed models with a multivariate $t$ distribution for random effects and/or errors appear for already some time in the literature as this is also mentioned in the Introduction to current manuscript. The same is true for finite mixtures of the mixed models (also mentioned in the Introduction to current manuscript). A novelty of the current manuscript could then be the combination of both – $t$ **distribution** for random effects/errors and **mixture** of the models (which is done in the current paper) and especially derivation of the algorithm for maximum-likelihood based estimation of such models (which is also done in the current paper) since most methods for dealing with finite mixtures of mixed models are Bayesian based on the MCMC inference. Nevertheless, I have some concerns on how much novelty the current manuscript actually brings (even though I must stress that it is perhaps sufficiently novel, it is only not clear to me). It is also mentioned in the Introduction to the current paper that a **mixture** of the linear mixed models with the $t$ **distribution** is considered (recently) by Bai, Chen and Yao (2015, *Journal of Statistical Computation and Simulation*). I do not have the full paper of Bai et al. available, nevertheless, according to the abstract, they consider in principle the same model as author of the current manuscript (or am I wrong?). Also the estimation method (a certain variant of the EM algorithm) seems to be similar (but perhaps the current paper uses something somehow different). An apparent difference between Bai et al. and the current manuscript is then in the fact that only a single longitudinal marker is considered by Bai et al., whereas multiple longitudinal outcomes are explicitely considered by author of the current manuscript. This certainly brings increased computational difficulties, nevertheless, from a theoretical point of view, the main formulas are the same. With the "multi" version of the model, the only difference to the "single" version is that the model matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are "bigger" (having a sparse structure) and

also the scale matrix of the $t$ distribution is "bigger". It is only necessary to put all observed $Y$'s on $i$th subject into one "long" $\boldsymbol{Y}_i$ for which the mixed model is specified. It should be clearly stated in the manuscript how the proposed methods differ from those of Bai et al. and whether it is a contribution of the method that it is capable to handle the "multi" outcome whereas that if Bai et al. would have mostly difficulties.

2. As I write above, the proposed model does not seem to be fully new. Hence the main contribution of the paper seems to be a computational algorithm to calculate the parameter estimates (even though it should be clearly stated how the computational approach differs from that of Bai et al.). At least a small simulation study should be included to show at least empirically that the algorithm really does what it should do – in terms of both parameter estimation and possibly also correct classification if the model is used for clustering or discriminant analysis.

3. The two groups for the practical analysis are defined at the end of p. 5 as follows: "*At the end point of this cohort study, 140 of the patients had died and are referred to as prognostic Group 1, while 172 were known to be alive and are referred to as prognostic Group 0*". I am not sure whether such division into the two groups is sensible. The problem is with the "end point of this cohort study". The PBC study was no cohort study. It is even written on p. 5 that "*312 patients were recruited between January 1974 and May 1984 . . . and participated . . . until April 1988.*" This was a classical "survival" study where different patients have different follow-up length given by not only the event occurrence but also by the recruitment time. I assume that the author considers April 1988 as the "the end point of the study". Nevertheless, being event free on April 1988 means something different for a patient being recruited in January 1974 and for a patient being recruited in May 1984. Hence I am afraid that the two groups are not well definined while making a classical error that our colleagues – medical doctors – often make when they put on one heap all event free patients by some calendar date irrespective of their actual follow-up time.

Less serious comments include:

1. Even though I am neither English native, nor English language specialist, I feel that the language needs quite considerable improvement. Some sentences are really gramatically incorrect, e.g., *. . . participants must be repeatedly measured their serum bilirubin . . .* on p. 2. There are also numerous apparent typos in the paper, e.g., reference to Booth et al. has a year 2008 in the text (p. 2) but a year 2007 in the reference list. Mclachlan should be McLachlan on p. 4 etc.

2. On p. 4, it is stated that "*. . . do not offer explicitly analytical solutions for model parameters.*" I think that there is no analytical solution for the (ML) estimates of the parameters, not to parameters themselves.

3. On p. 8 "autoregressive process of order 1" is considered for the error terms. Does this make sense if the visit times are irregular, not equidistant in time? Perhaps, some continuous time AR process is needed but certainly not a classical (discrete equidistant time) AR process.