AE Comments/Suggestions

General Comments

1. There's a general problem here with the assessment of normality from QQ plots and from standard hypothesis tests (AD, CVM, KS). Even for the classical LM, the latter are overly sensitive to minor departures from strict normality which does not generally compromise scientific inference, while standard QQ plots tend to be sensitive mainly to differences in the tails, and it is harder to distinguish between a general pattern of departure and one compromised by a few aberrant points. One simple way to correct this last is to plot a detrended version of the QQ plot, plotting the difference between the data quantile and the standard normal quantile on the ordinate, which has the effect of making the reference line horizontal and magnifying the vertical scale, making the pattern of departure more visible. As well, there is no discussion of the visual criteria for rejecting normality based on the QQ plots, and the simulation studies make use only of the AD tests without further comment.

   > It sounds like the AE wants the detrended QQ plots used, or at least discussed, in the paper. We can certainly do this, but it brings up the question of where to the MTurk study. Also, this comment has made me think that perhaps we need to add more to the MTurk study to look into the type of departure along with the type of QQ plot. Since we focused on a *t* distribution we focused on the tails of the distribution. Adding other differences in distributional shape to the study would help it seem more complete in its comparison of the QQ plot designs.

   > How can we discuss the visual criteria for rejection? I was assuming that the reader had enough experience with QQ plots, but wanted to point out that experience was not enough.

   > We do comment on the use of CVM and KS tests in the simulation study, including their omission from the formal discussion (p. 16 l. 19), so perhaps the AE overlooked this.

2. More generally in connection with scientific inference—tests and confidence intervals for the random effects themselves—there needs to be some discussion or results related to the impact of non-normality on these tests, for which normality may be necessary theoretically but perhaps of lesser practical importance. In connection with the simulation study, including cases with $\sigma^2_{b_0} = 0$, $\sigma^2_{b_0} = 0$ might be one way to address this in the current study; or else cite some literature on this question. This is similar to the point raised by Reviewer 2 regarding consequences of violation of normality.

We can address this comment in two ways: (1) by running a simulation study; (2) citing literature. Some relevant citations are:

Jiang, J. (2001). Goodness-of-fit tests for mixed model diagnostics. The Annals of Statistics, 29(4), 1137–1164. doi:10.1214/aos/1013699997. Especially the discussion in paragraph 2 of section 1.

The impact of distributional misspecification is felt in the prediction of unobserved units, specifically in the construction of prediction intervals. There are nonparametric approaches to this problem (c.f., Jiang and Zhang, 2002), but we don't need to focus on these at all.

3. In the discussion (or earlier), you need to provide some more insight about why reduced-rank rotated effects are a reasonable thing to do.

Specific Comments

1. p3 / l3: define "shrinkage"

2. p6 / l5: observed Q-Q plot (panel $12 + 2^2$) – why 2^2? Anyway, this panel does seem to stand out vs. the rest, with a few exceptions.

   More resistance on the way we obscure the true plot in a lineup. I think we have responses to this in the first two papers we submitted.

3. p6 / l -2 & fig 4: "none of the null plots conform ..." Well, all are + skewed, but some panels (1, 2, 17, 18, 19) seem quite close, at least with this representation. There seems to be an implicit, but unstated criterion here that none of the points can be outside the (undefined, but probably pointwise) envelope used in this figure.

   I supposed we did abuse the pointwise envelopes a bit with that statement. We could rein back that statement a little—15 out of 20 plots, or 15 our of 19 nulls, still comprises a lot of evidence of something fishy.

4. p 7 / Table 1 & discussion: The table provides a caution against using these standard tests for random effects, but it might be well to give some comparative results for alternatives (Lange-Ryan) designed to overcome these problems.

   We could add results for Lange and Ryan's proposal. This would allow us to point out that it doesn't fix the problem that we are interested in resolving.

   I will review what simulations I ran, and see if I already have that situation completed for the example displayed in Table 1, or if I need to run something else.

5. p8 / l -4 : "equations reveals the inherent dependence" – need to be a bit more explicit here

   We can just extend this sentence. I thought the statement was obvious enough, but that is a trap of being familiar with you're own research!

6. p 11/ Algorithm 1: This is the standard canonical transformation used widely in multivariate regression, MANOVA, canonical correlation, and need not take up space by being stated as

an algorithm.

I suppose that I am not very familiar with the finer details of MANOVA since it did not occur to me that this was the 'standard canonical transformation'. I would be fine saving space if I could find a reference to help the reader.

7. p 11 / l -1: eigenvalues of A^*, U_s, i.e., minimizing (8), ~~making~~ taking

   Small edit.

8. p 14 / Fig 6 caption: "calculated where Q for a simulated random ..." — something wrong here.

   Yes, something certainly is wrong with that sentence. I will reread and fix that.

9. Figs 7-9: The symbols in the legend include λ (which doesn't appear in Fig 7). Perhaps this is a font error?

   I don't see what the AE is referring to here when I look at the submission on my screen. This leads me to believe it is an error on the journal's side of things and not ours.

10. Supplementary materials: Perhaps mention HLMdiag here

    We mention HLMdiag in the list of supplementary materials already. Does the AE want us to include something in the actual supplement? I am unclear what is being asked for here.

Review 1

The discussion is of procedures to test the normality of random effects. But there seems to be a big confusion here. The distribution of the actual (unobserved) random effects is not the same as the distribution of the estimated random effects.

Consider an extreme case, with 50 groups with 1000 observations each and 50 groups with 1 observation each. For the groups with 1000 observations, the estimated random effect will be approximately equal to the true random effect. For the groups with 1 observation, the estimated random effect will be approximately 0. Thus, if the underlying distribution of random effects is normal, the distribution of estimated random effects will look like mixture of a normal with a spike at 0 containing 50% of the mass.

This was not very constructive. The paper acknowledges the difference in distributions, which is why we are arguing to do something different. We don't work with the mixture distributions, but we are still altering the process from what is done all over the literature...

Review 2

General Revisions

1. It should briefly contextualize the role of distributional tests among other types of diagnostics based on residuals.

Is this asking for a brief lit review of these tests?

2. It would be desirable to address the issue of the consequences of violating normality on inference in the model – not just testing normality for its own sake. Under what circumstances would it be crucial to verify normality and when would its violation be relatively innocuous.

   We will do this – the AE also made a similar comment.

3. The theory in this paper is based on the linear model

   $$Y = X\beta + Zb + \varepsilon,\ b \sim N(0, D) \perp \varepsilon \sim N(0, R)$$

   The implementation of diagnostics in this paper is based on the lme4 package whose particular strength, for models with normal errors, is that it can fit crossed random effects. Apart from this, however, it is very restrictive in the structures of D and R matrices it can model.

   The lme function in the nlme package is much more general. On the R side, the lme4 package only allows $R = \sigma^2 I$ while the nlme package allows very rich R-side modeling including ARMA models, spatial correlation and heteroskedastic models, etc. Also, lme allows a great deal of flexibility in the G(D)-side model (non-parametric smoothing splines, heteroscedastic models, etc). Tantalizingly, the methods developed in this paper appear to be fully applicable to these lme models. It's a pity that the diagnostics are not extended to lme since whitening would work even for non-independent observations within clusters. Normality diagnostics for the random effects generating smoothing splines would be particularly interesting where they could serve as a diagnostic for the appropriateness of the spline basis. Thus implementing the methods for lme as well as lmer would extend their applicability considerably.

   I think this reviewer really wants us to do this. The implementation of methods that I have written based on lme4 should be easily extended to lme, we just chose lme4 since it was the 'up and coming' package. I will work to implement what we have done for models fit using lme before we resubmit.

4. The paper is quite well written except that, I think, it would need extensive restructuring to introduce concepts logically in a way that requires much less second guessing. There are too many places where the purpose of concept becomes obvious a page or two later. For example, the idea and the purpose of considering W'b should be discussed before introducing Definition 1.

   This is going to take some careful thought and time, and maybe some fresh eyes (I don't know where we can get these). I think we can focus on this once we address the other concerns.

5. Definition 1 and equation (6) require some statement of the conditions required. The discussion preceding equation (6) is unclear. If the s-dimensional space is 'given' then $J_1(s)$ is a constant, i.e. any two matrices W that span the same column space will yield identical values for $J_1(s)$. What is meant is 'for any given dimension'.

6. There are many ways of defining a fraction of confounding using different symmetric functions of the relative eigenvalues of WAW' to WBW'. There should be a discussion of the decision to use the trace and the possibility of extending the notion using different symmetric functions.

   Something for the discussion section?

7. The approach takes D and R as known although, generally, there might be very little information on some aspects of D. What are the consequences of using this approach given that D and R are estimated?

   I was hoping that we wouldn't be asked this question. In the current paper we acknowledge that we are estimating the covariance matrices, but we leave the rest (and potential use of robust estimates) for future work. One way we could address this is to actually run the simulations using fixed R and D, then we would have something to compare the case where we estimate the matrices (which is what we currently present).

Specific Comments

1. p. 8 l. 52: V has not been previously defined.

   That is a good catch by the reviews. I will have to define V.

2. p. 9, l. 15: This might not be true with a small cluster size.

   I think the issue is that I was too general here. I investigated numerous situations and did not find problems when I used standardized predicted error terms. Of course it was not an exhaustive study so adding in some caveat should make this clear.

3. p. 9, l. 38ff: The exposition would benefit greatly from some motivation at this point. Introducing the random vector $W'\widehat{b}$ whose properties are the subject of the concept of fraction of confounding, would make the Definition 1 much easier to understand.

   This will take some time.

4. p.9,l.52: the fact that $Trace((W'BW)^{-}W'AW)$ is well defined is not trivial but it would follow from subsequent discussions. I think this should be mentioned.

   Do we just add a sentence similar to what the reviewer said?

5. p. 10 l. 4: 'by definition' is misused here.

   OK

6. p. 10 l. 34: The development here is not rigorous. Depending on the value of s and on the rank of B, $(W'BW)^{-1}$ might not exist. I don't think that the development makes it clear why this all does work. The key fact is that $span(A) \subseteq span(B)$ (indeed B-A is non-negative definite) which makes everything work, following de Leeuw (1982), even though A, B, $W'AW$ and $W'BW$ might be singular.

   I did forget to put in this important point, though I was very aware of it from de Leeuw's work.

7. p. 11 l. 16: Here too, the fact that $span(A) \subseteq span(B)$ makes things work but that is not clear in the description of the algorithm.

8. p. 15 l.14: It might be helpful for many readers to explain why this is reasonable: i.e. that the predictor variables can be transformed affinely to make this true. However, this also

suggests that these methods would work best if X is transformed to achieve a nearly diagonal D matrix. Perhaps this should be discussed.

Oh my, do we need to investigate our method with a transformed X?