Response: Did Newton–Raphson really fail?

Xiao-Li Meng

Stat Methods Med Res 2014 23: 312
DOI: 10.1177/0962280213508866

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2014, Vol. 23(3) 312–314

DOI: 10.1177/0962280213508866
smm.sagepub.com
SAGE

Response: Did Newton–Raphson really fail?

Xiao-Li Meng

1 Yes, it did

I gather any author would be grateful when an article published more than 15 years ago still attracts comments, regardless of their sign. I therefore thank Professor MacDonald for bringing back my fond memories of working on my first ''history of science'' article, though that article also reminds me of an unpleasant surprise I experienced when I received the final printed copy. The word ''link'' in the title somehow became ''linik,'' and to this date I am still wondering how could such an obvious and visible typo have escaped the printer's attention?

A reader of Professor MacDonald's letter (hereafter the ''Letter'')—but who has not read my original article (hereafter ''the Original'')—might have a similar question for me: how could the obvious non-negative constraint of a Poisson mean λ have escaped my attention? However, the same reader might not have posed the question if she or he had read the entire paragraph on page 14 of the Original, from which the Letter quoted. It is reproduced below for readers' convenience (the figure and equation numbers in it refer to those in the Original):

> The right plot in the fourth row of Figure 1 illustrates how the Newton-Raphson iterates converged to the wrong limit. Because all the iterates are outside the parameter space (i.e., $\lambda^{(t)} < 0$ for $t \geq 1$), a mirror image (with respect to $\lambda = 0$) of the log-likelihood surface was created to plot the iterates. In contrast, EM iterates can never escape from the parameter space as long as the initial value is inside the space. It is worthwhile to point out that the (original) log-likelihood surface here is as simple and smooth as it can be, and $\lambda^{(0)} = 0.4$ is not an impossible choice of starting value in practice for such problems. For more complicated problems, especially multi-dimensional ones, the Newton-Raphson algorithm can be very sensitive to the starting value, and sometimes fail to converge (which is less harmful then converging to a wrong limit). Of course, when using (3.10), a careful user will not choose $\lambda^{(0)} \leq \log(\bar{x}_{obs}) = 0.447$, which makes its denominator negative (or even zero), the reason for negative iterates and convergence to the wrong limit. (For a careful user, (3.10) is not even a Newton-Raphson iteration once iterates move outside the space where the original log-likelihood was defined. In that sense, the Newton-Raphson iteration failed at the first iteration when $\lambda^{(0)} = 0.4$.) In general, however, it may not be a trivial task to detect such a problem before running the Newton-Raphson algorithm. The general point is that the fast convergence of the Newton-Raphson algorithm often comes at the expense of more human investment in terms of delicate choice of the starting value and careful monitoring of convergence.

The central point of comparison, therefore, was between the EM iterative sequence and the Newton–Raphson sequence, both starting at the same initial point inside the parameter space. The EM sequence always stays inside because the output from its M-step is an MLE, which must

Department of Statistics, Harvard University, Cambridge, UK

Corresponding author:
Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge MA 02138, UK.
Email: meng@stat.harvard.edu

be a plausible parameter value by definition. This property not only contributes to the stability of EM but it can also contribute to its slowness in convergence. In contrast, the Newton–Raphson sequence has no such automatically guaranteed constraint, a property that again can be both a plus or minus for the algorithm: the freedom to move anywhere helps to increase its speed and also potentially its instability. As I concluded in the paragraph following the one quoted above in the Original, "This is a computational example of the common tradeoff between 'efficiency' and 'robustness', a central issue in statistical inference." Obviously for illustrating this tradeoff, my comparison would provide little insight if I had used the log transformation of, or the constrained optimization for, λ, as the Letter suggested.

2 No, it didn't

For practical implementation, the transformation method given in the Letter obviously should be adopted if one wants to use the Newton–Raphson algorithm, which is indeed highly recommended for (very) low-dimensional problems. And the Letter is absolutely correct that it would be an error to claim Newton–Raphson algorithm fails *in general* because we applied it in a naive way, just as it would not be very wise to claim EM is too slow because we adopted only one particular form of the data augmentation in its construction. Indeed, as illustrated in the Original, for this simple problem, there are at least two authentic EM algorithms but with very different convergence rates.

Another point made in the Original helps to highlight this point further. If a numerical analyst were to claim that the normal approximation to a confidence interval fails for λ because the resulting interval encompasses negative values, many reputable statisticians would jump on him to point out that what has failed is not the approximation itself but how it was applied. This was explicitly discussed in the Appendix of the Original, because it is well known to statisticians that the normal approximation should be applied on the $\log \lambda$ scale. Retrospectively, I did commit an "error of omission" in the Original, that is, the missed pedagogical opportunity to link the need for transforming λ when applying the normal approximation to that when implementing the Newton–Raphson algorithm. This is particularly unfortunate because both essentially address the same mathematical concern: quadratic approximations are more accurate on the $\log \lambda$ scale than on the original λ scale.

3 Well, it depends . . .

I of course have no explanation for this omission other than blaming it on my youth. But the additional 15 years of seniority have only made the matter muddier for me. Could there ever be a truly fair comparison between EM and Newton–Raphson, or for that matter, any two types of algorithms? The answer seems to be a hopeless "NO," for at least two reasons.

The obvious one is there has not been, and perhaps never will be, a universally accepted criterion for comparison. Both the Original and the Letter used the number of iterations, which of course is an obvious and useful criterion. But for anyone who has implemented any algorithm, the matter is typically far more involved. What is the CPU time of each iteration? How time consuming is it to program and debug each algorithm? How frequently does each algorithm actually converge? How easy is it to adapt each algorithm to a slightly different context? How easy is it to use each algorithm in an online form, that is, with data arriving sequentially? I am sure that the reader can supply many more questions, all of which may suggest that it is meaningless to claim one algorithm is better than another without specifying what criteria are used, and that it is fruitless to seek one universal criterion that will address everything about which we care.

The multiplicity of criteria turns out to be less troublesome (for comparison purposes), at least in theory, than the multiplicity in constructions. That is, both EM and Newton–Raphson are general recipes for constructing algorithms, not specific algorithms in themselves. Even for the vanilla version of EM, there are essentially infinitely many constructions, one for each data augmentation scheme, and the speed can be made as fast as possible if one does not take into account the difficulty of implementing either its E-step or M-step. A good illustration of this issue is the working parameter approach that David van Dyk and I have investigated in the last 15 years or so, where each value of the working parameter indexes a data augmentation scheme and hence corresponds to a different EM construction, with its own tradeoff of speed and simplicity; see for example, Meng and van Dyk[1–3] and van Dyk and Meng.[4,5]

For Newton–Raphson, the construction issue is to which equation should the iterative scheme be applied? As the Letter illustrated well, by transforming $s(\lambda) = 0$ to the equivalent $t(b) = s(\exp(b)) = 0$, we can already see a significant difference in convergence behavior. But there are infinitely many one-to-one transformations $\lambda = f(b)$ to choose from, just as there are infinitely many data argumentation schemes for constructing EM to solve the same problem. Does then make sense to ask if Newton–Raphson or EM really fails when we can try only a few of each? How do we know there is no clever choice of the transformation (for Newton–Raphson) or data augmentation (for EM) available that could do a much better job than what we have tried?

Not knowing how to answer such a question other than invoking the common wisdom "Well, it depends...," it is time for me to make a toast to Professor MacDonald and thank him for stimulating me to revisit the intriguing issue of comparing Newton–Raphson and EM. Cheers!

Funding

This work was partially supported by US National Science Foundation.

Acknowledgements

I thank the editor for giving me the opportunity to respond and Steven Finch for proofreading.

References

1. Meng X-L and van Dyk DA. The EM algorithm – an old folk-song sung to a fast new tune [with discussion and rejoinder]. *J R Stat Soc Ser B* 1997; **59**: 511–567.
2. Meng X-L and van Dyk DA. Fast EM-type implementations for mixed effects models. *J R Stat Soc Ser B* 1998; **60**: 559–578.
3. Meng X-L and van Dyk DA. Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* 1999; **86**: 301–320.
4. van Dyk DA and Meng X-L. The art of data augmentation [with discussion and rejoinder]. *J Comput Graph Stat* 2001; **10**: 1–111.
5. van Dyk DA and Meng X-L. Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: a graphical guide book. *Stat Sci* 2010; **25**: 429–449.