Is it a computing algorithm or a statistical procedure: Can you tell or do you care?

Xiao-Li Meng*
Harvard University, Cambridge, MA, U.S.A.
meng@stat.harvard.edu

For years, it irritated me whenever someone called the EM algorithm an "estimation' procedure. I'd argue passionately that EM merely is an algorithm designed to compute a maximum likelihood estimator (MLE), which can be computed by many other methods. Therefore the estimation principle/procedure is MLE, not EM, and it is dangerous to mix the two, for example by introducing modifications to EM steps without understanding how they would alter MLE as a statistical procedure. The reality, however, is that the line between computing algorithms and statistical procedures is becoming increasingly blurred. As a matter of the fact, practitioners are now typically given a black box, which turns data into an "answer". Is such a black box a computing algorithm or a statistical procedure? Does it matter that we know which is which? Should I continue to be irritated by the mixing of the two? This talk reports my contemplations of these questions that originated in my taking part in a team that investigated the self-consistency principle introduced by Efron (1967). We will start with a simple regression problem to illustrate a self-consistency method that apparently can accomplish something that seems impossible at first sight, and the audience will be invited to contemplate whether it is a magical computing algorithm or a powerful statistical procedure. We will then discuss how such contemplations have played critical roles in developing the self-consistency principle into a full bloom generalization of EM for semi/non-parametric estimation with incomplete data and under an arbitrary loss function, capable of addressing wavelets de-noising with irregularly spaced data as well as variable selection via LASSO-type of methods with incomplete data. Throughout the talk, the audience will also be invited to contemplate a widely open problem: how to formulate in general the trade-off between statistical efficiency and computational efficiency? (This talk is based on joint work with Thomas Lee and Zhan Li.)

Keywords: Incomplete data; Self-consistency; Variable Selection; Wavelets.