

Stochastic EM algorithm for incomplete longitudinal data

Caroline Beunckens, Cristina Sotro and Geert Molenberghs

Center for Statistics, Hasselt University, Belgium

Missingness often occurs in data arising from longitudinal studies, inducing imbalance in the sense that not all planned observations are actually made. More specifically, missingness in longitudinal studies usually appears in the form of dropouts, in which subjects fail to complete the study for some reason or another. In his 1976 paper, Rubin provided a formal framework for the field of incomplete data by introducing the important taxonomy of missing data mechanisms, consisting of missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). A missingness process is said to be MCAR if the missingness is independent of both unobserved and observed outcomes, but potentially depends on covariates. An MAR mechanism depends on the observed outcomes and perhaps also on the covariates, but not further on unobserved measurements. Finally, when an MNAR mechanism is operating, missingness does depend on unobserved measurements, maybe in addition to dependencies on covariates and/or on observed outcomes. At the same time, the selection model, the pattern-mixture model, and the shared-parameter model frameworks have been established.

In this talk we focus on full selection models to deal with incomplete longitudinal continuous data. Such models are valid under the most general assumption of MNAR, and are most useful within a sensitivity analysis. Diggle and Kenward (1994) proposed such a selection model, combining the multivariate model for longitudinal continuous data with a logistic regression for dropout. The resulting likelihood is maximized using integration over the missing data.

A popular alternative to such a likelihood method for missing data is based on the EM algorithm (Dempster *et al.*, 1977). The EM algorithm is an iterative algorithm that is relatively easy to program and that produces maximum likelihood estimates. However, it has the disadvantage of possibly converging to local maxima or saddle points of the log-likelihood function, and its limiting position is often sensitive to starting values. Stochastic EM (Celeux and Diebolt, 1985) provides an attractive alternative to EM, which also involves two iterating steps. At the S-step, the missing data are imputed with plausible values, given the observed data and a current estimate of the parameters. At the M-step, the maximum likelihood estimate of the parameters is computed, based on the pseudo-complete data. The final output of the stochastic EM algorithm is a sample from a stationary distribution whose mean is close to the maximum likelihood estimate and whose variance reflects the information loss due to missing data.

We will compare the direct-likelihood method for the Diggle-Kenward model as well as the stochastic EM algorithm using simulations and applying both methods to a dataset.

References

- Celeux, G. and Diebolt, J. (1985) The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational Statistics Quarterly*, **2**, 73–82.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.
- Diggle, P. J. and Kenward, M. G. (1994) Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49–93.
- Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592. With comments by R. J. A. Little and a reply by the author.