Model Assisted Statistics and Applications 15 (2020) 263–272
DOI 10.3233/MAS-200493
IOS Press

Missing data imputation in multivariate t distribution with unknown degrees of freedom using expectation maximization algorithm and its stochastic variants

Paul Kimani Kinyanjui, Cox Lwaka Tamba*, Luke Akong'o Orawo and Justin Obwoge Okenye
Department of Mathematics, Egerton University, Egerton, Kenya

Abstract. Many researchers encounter the missing data problem. The phenomenon may be occasioned by data omission, non-response, death of respondents, recording errors, among others. It is important to find an appropriate data imputation technique to fill in the missing positions. In this study, the Expectation Maximization (EM) algorithm and two of its stochastic variants, stochastic EM (SEM) and Monte Carlo EM (MCEM), are employed in missing data imputation and parameter estimation in multivariate t distribution with unknown degrees of freedom. The imputation efficiencies of the three methods are then compared using mean square error (MSE) criterion. SEM yields the lowest MSE, making it the most efficient method in data imputation when the data assumes the multivariate t distribution. The algorithm's stochastic nature enables it to avoid local saddle points and achieve global maxima; ultimately increasing its efficiency. The EM and MCEM techniques yield almost similar results. Large sample draws in the MCEM's E-step yield more or less the same results as the deterministic EM. In parameter estimation, it is observed that the parameter estimates for EM and MCEM are relatively close to the simulated data's maximum likelihood (ML) estimates. This is not the case in SEM, owing to the random nature of the algorithm.

Keywords: Expectation maximization (EM), stochastic EM, Monte Carlo EM, unknown degrees of freedom

1. Introduction

In research studies, data may be characterized by missing values. Since most of the statistical methods cannot be applied directly on such datasets, the data analyst has to pre-treat the data. This may be done by deleting the rows or columns with missing values. However, deletion methods may lead to inadvertent loss of crucial information, which may have negative effects on the inferences. Additionally, the complete cases may not constitute a representative sample of the original dataset (Pigott, 2001; Raghunathan, 2004). Due to such uncertainties, model-based techniques are preferred in remedying missing data problems since in addition to using all the available information, they also preserve the distribution of the original data.

The expectation maximization (EM) algorithm is a model-based iterative technique popularly used for parameter estimation in the presence of missing values. The deterministic method is implemented in two parts namely the expectation (E) step and the maximization (M) step (McKnight et al., 2007). The algorithm iteratively alternates between the two steps until convergence is achieved. One of the major drawbacks of EM is that it may be trapped in local saddle points, preventing it from achieving the desired output. Over the years, a number of EM variants aimed

[*] Corresponding author: Cox Lwaka Tamba, Department of Mathematics, Egerton University, Egerton 536-20115, Kenya. E-mail: cox.tamba@egerton.ac.ke.