Stochastic Approximated EM for logistic regression with missing data

Wei. Jiang[a] **Julie. Josse**[a] and **Marc. Lavielle**[b]

[a] CMAP, Ecole Polytechnique
Route de Saclay, 91128 Palaiseau, France
wei.jiang@poltechnique.edu
julie.josse@poltechnique.edu

[b] Xpop, Inria Saclay
Route de Saclay, 91128 Palaiseau, France
marc.lavielle@inria.fr

Mots clefs : Missing data, EM algorithm, trauma major, public health

Abstract

Logistic regression is a reference method in supervised learning but as surprising as it may seem, there are very few solutions for performing a logistic regression and selecting variables with missing data. We suggest a stochastic approximated version of EM algorithm based on Metropolis-Hasting sampling, in order to do statistical inference for the logistic regression model with incomplete data. We propose a complete approach including the estimation of parameters and their variance to derive confidence interval, as well as a model selection procedure and how to handle missing values in testing set. The method is computationally efficient, and its good coverage and properties of variable selection are demonstrated through a simulation study. We illustrate the method on a register from Paris's hospitals on polytraumatized patients to predict the occurrence of an hemorrhagic shock, a leading cause of early preventable death in severe trauma. The aim is to consolidate the existing Red Flag procedure, a binary alert identifying patients with high risk of severe hemorrhage. The methodology is implemented in an R package *misaem*.

1 Introduction

Missing data exist in almost all areas of empirical research. There might be various reasons for missing data to occur, including the survey non-response, unavailability of measurements and loss of data.

One popular approach to handle missing values, consists of modifying the estimation process so that it can be applied on incomplete data. For instance, one can use the EM algorithm [Dempster et al., 1977] to obtain the maximum likelihood estimate (MLE) despite missing values and a supplemented EM algorithm (SEM) [Meng and Rubin, 1991] or Louis' formula [Louis, 1982] for their variance. This strategy is valid under a missing at random (MAR) mechanism [Rubin, 1976, Little and Rubin, 2002], in which missingness of the data is independent of the missing values given the observed data. Even though this approach is perfectly well-fitted towards a specific inference problem with missing values, it turns out that, as surprising as it sounds, there aren't many solutions nor implementation available even for simple models such as logistic regression model which is the focus of this paper.

This could be explained because it is often the case that, the expectation step, in the EM algorithm for logistic regression, involves unfeasible computations. One solution suggested in

[Claeskens and Consentino, 2008, Gilks and Wild, 1992, Ibrahim et al., 1999, Ibrahim et al., 2005] in the framework of generalized linear models, is to use a Monte Carlo EM (MCEM) algorithm [Wei and Tanner, 1990, McLachlan and Krishnan, 2008] replacing the integral by its empirical sum with Monte Carlo sampling. Then, they also estimated the variance using a Monte Carlo version of Louis' formula. For sampling, they used Gibbs samplers along with an adaptive rejection sampling scheme. Still, their approach is much computationally expensive and they considered implementations only for a monotone pattern of missing values, or for missing values on only 2 variables in a dataset.

In this paper, we develop an alternative to MCEM, a stochastic approximation EM (SAEM) [Lavielle, 2014] which uses a stochastic approximation procedure to estimate the conditional expectation of the complete-data likelihood, instead of generating a large number of Monte Carlo samples. SAEM has an undeniable computational advantage over MCEM. In addition, it takes great advantage of allowing easy establishment of model selection criterion based on penalized observed likelihood. This latter characteristic is very useful in practice as only few methods are available to select a model when there are missing values. For example, [Claeskens and Consentino, 2008, Consentino and Claeskens, 2011] considered approximation of AIC while [Jiang et al., 2015] defined a generalized information criteria (GIC) and adaptive fence (AF) and [Liu et al., 2016] in the framework of imputation with Random Lasso (mirl) proposed to combine penalized regression techniques with multiple imputation and stability selection.

References

[Claeskens and Consentino, 2008] Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data. 64:1062–9.

[Consentino and Claeskens, 2011] Consentino, F. and Claeskens, G. (2011). Missing covariates in logistic regression, estimation and distribution selection. *Statistical Modelling*, 11(2):159–183.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

[Gilks and Wild, 1992] Gilks, W. R. and Wild, P. P. (1992). Adaptive rejection sampling for gibbs sampling. *Appl. Statist*, 41(2):337–348.

[Ibrahim et al., 1999] Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (1999). Monte carlo em for missing covariates in parametric regression models. *BIOMETRICS*, 55:591–596.

[Ibrahim et al., 2005] Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346.

[Jiang et al., 2015] Jiang, J., Nguyen, T., and Rao, J. S. (2015). The e-ms algorithm: Model selection with incomplete data. *Journal of the American Statistical Association*, 110(511):1136–1147.

[Lavielle, 2014] Lavielle, M. (2014). *Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools.* Chapman and Hall/CRC.

[Little and Rubin, 2002] Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.

[Liu et al., 2016] Liu, Y., Wang, Y., Feng, Y., and Wall, M. M. (2016). Variable selection and prediction with incomplete high-dimensional data. *Ann. Appl. Stat.*, 10(1):418–450.

[Louis, 1982] Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233.

[McLachlan and Krishnan, 2008] McLachlan, G. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2. ed edition.

[Meng and Rubin, 1991] Meng, X.-L. and Rubin, D. B. (1991). Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86(416):899–909.

[Rubin, 1976] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

[Wei and Tanner, 1990] Wei, G. C. G. and Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.