REM algorithm for design of experiments with missing data

Yadolah Dodge and Alice Zoppè
University of Neuchâtel, Statistics Group
P.O. Box 1825
2002 Neuchâtel, Switzerland
yadolah.dodge@unine.ch

1. Introduction

The problem of missing observations in the analysis of designed experiment is an old one, and it has usually been dealt with by replacing methods, i.e. looking for values to susbstitute the missing ones in the incomplete data matrix. Healy and Westmacott (1956) introduced an iterative procedure that, at each iteration, replaces the missing value by its expected value, given the previous estimates. When Dempster, Laird and Rubin (1977) defined the EM algorithm, the procedure by Healy and Westmacott was presented as an example of the EM algorithm in the analysis of variance designs under the normal linear model. Note that the goal of the EM algorithm is not to replace missing values in the incomplete data matrix, but the maximum likelihood estimation of parameters; this is obtained by substituting the expected value of the functions of missing values in the loglikelihood function. As a consequence, before attempting the estimation via the EM algorithm, it is necessary to establish which are the estimable parametric functions. If the EM algorithm is run, it gives numerical estimates for the parameters on the basis of replaced values. When enough data are missing, the design matrix may not be of maximal rank, and these estimates may have no statistical meaning.

In the literature on the analysis of experiments with missing values, the book by Dodge (1985) offers a different viewpoint, where replacement of the missing values has no role to play, the goal of the analysis being the estimation of the estimable parameters or of the parametric functions. Estimable parameters are found by analyzing the incidence matrix: an iterative procedure, called the R process, derives a final matrix which defines which are the estimable parametric functions. Several methods are presented, together with programs, for different kinds of models.

We suppose that that $\{Y_{ijk}\}$ is a collection of independent and normally distributed random variables each having a common unknown variance σ^2 and each having an expectation of the form:

$$E[Y_{ijk}] = \mu + \alpha_i + \beta_j$$

where $i = 1,\ldots,I$, $j = 1,\ldots,J$ and $k = 1,\ldots,n_{ij}$. We are interested in estimating and submitting to hypothesis testing the parameters $\mu, \alpha_i, \beta_j, \sigma^2$. We suppose that there are n_{ij} observations in cell (i,j) and that each row, as well as each column as at leat one observation.

2. REM algorithm for experiments with missing data

We present now a procedure which starts from the matrix of observations and, for a chosen model, establishes the estimable parameters or the estimable parametric functions.

As it applies a modification of the R process followed by the EM algorithm, this algorithm is called REM. We describe it for additive two-way classification models, but it can be generalized for more general additive models. As before, the two factors are indicated by α and β, the factors having respectively I and J levels, so that $\boldsymbol{\theta}' = [\mu, \alpha_1 \ldots, \alpha_I, \beta_1 \ldots, \beta_J]$.

From the observed data the algorithm derives an initial matrix M_0 of dimension $I \times J$: m_{ij}=1 if at least one observation is available for the the i-th level of factor α and for the j-th level of factor β; otherwise m_{ij}=0.

The R process is applied to M_0 and the final matrix M is obtained: M has ones in the cells corresponding to the estimable values. The rows and the columns of M are permuted so that the permuted final matrix M^* is block diagonal; the permutations orders are provided in two vectors, r for rows and c for columns, so that the original matrix can be easily reconstructed.

The procedure then determines the diagonal-block structure of M^*; let Q be the number of blocks and denote them by M_1^*, M_2^*, ..., M_Q^*. Each block is characterized by a set r_q of row indeces (referring to levels of the α-factor) and a set c_q of column indeces (referring to levels of β-factor), which identify the estimable expected values of the form:

$$\mu + \alpha_{iq} + \beta_{jq}$$

where $\alpha_{iq} \in r_q$ and $\beta_{jq} \in c_q$.

If M is not block-diagonal, that is $Q = 1$, then the design is completely connected and all cells are estimable. The design matrix is of maximal rank, and the algorithm derives the complete estimable data matrix Y_e, adding one missing value for any combinations of i and j not present in the available (observed) data matrixY_o; the EM algorithm is applied only once to this matrix.

If M^* is block diagonal, then for each of the Q blocks the procedure extracts from the original data set one subsets, according to the block structure of M^* and the permutation vectors r_q and c_q, such that the corresponding design matrix X_q is of maximal rank, $q = 1, \ldots, Q$. The REM algorithm then derives, for each of the Q blocks, an estimable data matrix $Y_{e,q}$, and the EM algorithm is applied to each subset, giving the maximum likelihood estimates of the cells values.

REFERENCES

Birkes, D., Dodge, Y., and Seely, J. (1976). Spanning sets for estimable contrasts in classification models. Annals of Statistics, 4, 86-107.

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 B, 39, 1-38.

Dodge, Y. (1985). Analysis of experiments with missing data. New York: Wiley.

Healy, M. and Westmacott, M. (1956). Missing values in experiment analysed on automatic computers. Applied Statistics, 5, 203-206.

RESUME

L'application directe de l'algorithme EM pour un jeu de donnée provenant des plans d'expérience avec des données manquantes, peut mener à des estimations de fonctions paramétriques non estimables. Dans cet article nous présentons un ajustement de l'algorithme EM pour des modèles de classification additive qui empêchera l'utilisateur d'obtenir des résultats non fiables. L'algorithme proposé est appelé REM.

Mots-clés: algorithmes EM, données manquantes; plans d'expérience, plans factoriels.