R code for fitting Multimix

Murray A. Jorgensen
Auckland University of Technology
New Zealand
majmurr@gmail.com

January 12, 2019

1 This document

This document is intended to serve two purposes. While the R version of *Multimix* is under development it will describe the **R** objects and scripts used in this project for my benefit whenever I need a refresher and for the benefit of anyone else who may be interested, particularly possible contributors. Later the document will form part of the package documentation, probably after further editing.

2 The Byar and Green Prostate Cancer data

In order to be able to give concrete examples of the data structures that we will be introducing we consider the clustering of patients on the basis of pre-trial variables alone for the Prostate Cancer clinical trial data of Byar and Green (1980) reproduced in Andrews and Herzberg (1985, pp 261– 247).

This data was obtained from a randomized clinical trial comparing four treatments for 506 patients with prostatic cancer. These patients had been grouped by physicians using clinical criteria into Stage 3 and Stage 4 of the disease. Patients were classified as having Stage 3 prostatic cancer when the cancer was contained locally and there was no clinical evidence that the disease had spread elsewhere, whilst they were classified as having Stage 4 prostatic cancer when there was evidence that the disease had spread to other

parts of the body. This evidence could have been obtained from elevated acid phosphatase levels, x-rays, or both of these. Byar and Green (1980) originally considered this data in their investigation to determine whether there may be an optimal treatment for each patient based on his individual characteristics. They approached this question by looking for treatment-covariate interactions in an exponential survival model. They found that younger patients with high grade tumours should have been treated with estrogens, whereas older patients with low grade tumours were likely to be harmed by estrogen treatment.

There are twelve pre-trial covariates measured on each patient, seven may be taken to be continuous, four to be discrete, and one variable (SG) is an index nearly all of whose values lie between 7 and 15, and which could be considered either discrete or continuous. We will treat SG as a continuous variable.

Table 1: Variables in dataset

Variable	Abbreviation	No. of Levels
Age	Age	
Weight	Wt	
Performance rating	PF	4
Cardiovascular disease history	HX	2
Systolic Blood pressure	SBP	
Diastolic blood pressure	DBP	
Electrocardiogram code	EKG	7
Serum haemoglobin	HG	
Size of primary tumour	SZ	
Index of tumour stage and histolic grade	SG	
Serum prostatic acid phosphatase	AP	
Bone metastases	BM	2

A preliminary inspection of the data showed that the size of the primary tumour (SZ) and serum prostatic acid phosphatase (AP) were both skewed variables. These variables have therefore been transformed. A square root transformation was used for SZ, and a logarithmic transformation was used for AP to achieve approximate normality. (As for correlation, skewness over the whole data set does not necessarily mean skewness within clusters but when clusters were formed within-cluster skewness was observed for these

variables.) Observations that had missing values in any of the twelve pre-treatment covariates were omitted from further analysis, leaving 475 out of the original 506 observations available.

3 User inputs

These objects should be created by the user. (In some early versions of `Setup.r`, then known as `Noniter.r`, they are created within that script.)

p is a scalar giving the number of variables to be modeled.

dframe is a data frame containing the p variables to be modeled. Variables to be modeled as continuous should be vectors, variables to be modeled as discrete should be factors. Other columns may be present but the variables to be modeled should be the leftmost p columns. There are no other constraints on the order of the columns. No missing values may occur among the variables to be modeled. It is recommended that continuous variables be standardized so as to mostly lie within the interval $(-3, 3)$ as conventional identities for sums of squares and products are employed in the present version of the scripts and can give numerical difficulties with unstandardized data. (It is intended that later versions will remove the necessity to standardize.)

cdep is a list of vectors. Each vector is a set of positions of columns of `dframe[,1:p]` that may be dependent in a mixture component. All columns must be continuous variables. The sets must be disjoint.

lcdep is a list of vectors. Each vector is a set of positions of columns of `dframe[,1:p]` that may be dependent in a mixture component and which make up the variables of a location model. The first element in each vector corresponds to the discrete variable of the location model, the remaining ≥ 1 elements correspond to the continuous variables. The sets must be disjoint.

Z is an $n \times q$ matrix, where n is the number of rows of `dframe` and q is the number of components in the mixture. During the fitting Z_{ij} holds the currently estimated probability that observation i belongs to component j. Commonly Z is initialized to a matrix of indicator columns for a partition of the data.

W is an $n \times q$ matrix, based on Z but with each column multiplied by a constant so as to sum to 1 over rows.

The objects `dframe[,1:p]`, `cdep`, `lcdep` and Z define the data and the form of each mixture component model.

The script `DataModelZ.r` is an example of an **R** script specifying the user inputs. It loads the **R** workspace file `cancer.RData` which contains a pre-processed version of the Byar-Green Prostate Cancer data. The script `cancer.r` shows how this workspace was created. We may refer to this example later in this document.

Note that to keep the code completely general subsequent scripts do not refer to any user-defined objects other than `dframe[,1:p]`, `cdep`, `lcdep` and Z.

4 Setup.r

This script contains code that is not intended to be iterated. Whenever possible objects are created and given values here rather than in an iteration step in order to reduce the execution time of code which is executed repeatedly. We now discuss the main objects in the code in several groups.

4.1 Functions of data

`Setup.r` organizes the data in `dframe` into new structures such as lists of matrices, depending on the model being fitted. Similarly some functions of the data are also stored in this way.

We have need to store cross-products of variables. Because we have chosen to store these in matrices rather than higher order arrays we introduce a *pairing function* to give an integer index to an pair of unequal integers where the smaller integer is on the left. The function `pair.index`:$(u, v) \mapsto N$ is such a function and the functions `left`:$N \mapsto u$ and `right`:$N \mapsto v$ jointly provide an inverse. The function definitions in the script do not give a clear idea of how the pairs of variables are ordered but Figure 1 should illustrate how the off-diagonal integer pairs are ordered.

Also included in the script are functions `Pair.Index`, `Left` and `Right` which do a similar job when both members of a pair may be the same as when squares are included along with products. I don't need these functions at the moment but you never know when they might come in handy!

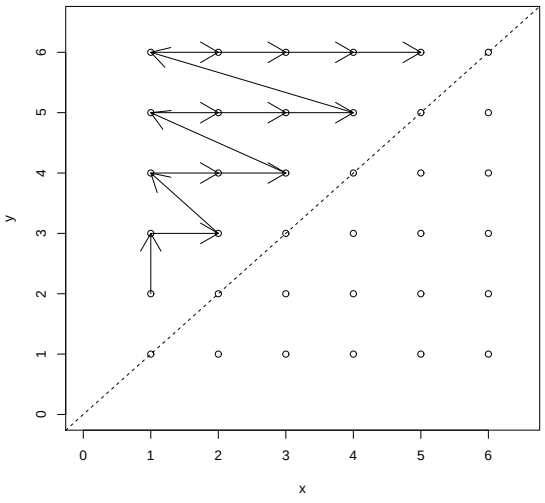

Figure 1: The pairing function *pair.index* orders pairs as shown.

dvals is a list of matrices for each discrete variable *not included* in a location model. The matrix for each discrete variable is made up of an indicator column of length n for each level (value) of the variable. A general value of `dvals` has the form `dvals[[vbl]][obs, level]` and is 1 iff discrete variable `vbl` takes level `level` for observation `obs`. In the cancer example of section 3 `dvals` is a list of two matrices: a 475×2 matrix and a 475×6 matrix corresponding to the variables CVD and EKG respectively.

ldvals is a list of matrices for each discrete variable within a location model. The matrix for each discrete variable is made up of an indicator column of length n for each level (value) of the variable. A general value of `ldvals` has the form `ldvals[[vbl]][obs, level]`. In the cancer example of section 3 `ldvals` is a list of two matrices: a 475×2 matrix and a 475×3 matrix corresponding to the variables BM and Perf respectively.

ovals is the matrix of continuous variables that are outside any location model, and not associated with any other continuous variable in the

model. In the cancer example `ovals` is a 475×3 matrix with the variables Size, Index, and PAP as columns.

ovals2 holds the squares of the elements of `ovals`.

cvals is a list with a member for each fully continuous partition cell (that is a cell not including discrete cells or cells listed in `lcdep`) each member being the matrix of continuous variables in that cell. A general value of `cvals` has the form `cvals[[cell]][obs, vbl]`. In the cancer example there is only a single multivariate normal cell and `cvals` is a list with one member, a 475×2 matrix with the variables SBP and DBP as columns.

cvals2 As for `cvals`, but the data is squared.

lcvals is a list with a member for each location partition cell, each member being the matrix of continuous variables in that cell. A general value of `lcvals` has the form `lcvals[[cell]][obs, vbl]`. In the cancer example `lcvals` is a list of two matrices: a 475×2 matrix with the variables Weight and Haem as columns and a 475×1 matrix with the variable Age as its column.

lcvals2 As for `lcvals`, but the data is squared.

cprods is a list with a member for each fully continuous partition cell containing at least two variables, each member being the matrix of pairwise products of the continuous variables in that cell, as ordered by `pair.index`. A general value of `cprods` has the form `cprods[[cell]][obs, pair.index(vbl1, vbl2)]`. In the cancer example `cprods` is a list with a single member, that being a a 475×1 matrix with the products of SBP and DBP as its column.

lcprods is a list with a member for each location partition cell, each member being the matrix of pairwise products of the continuous variables in that cell, as ordered by `pair.index`. A general value of `lcprods` has the form `lcprods[[cell]][obs, pair.index(vbl1, vbl2)]`. In the cancer example `lcprods` is a list with two members, a 475×1 matrix with the products of Weight and Haem as its column and a 475×0 matrix, there being only a single continuous variable in the second

location model cell. For the consistency of the code it is necessary to admit degenerate matrices such as `lcprods[[2]]` in this example.

ldxc is a list with a member for each location partition cell, each member being a list with a member for each level of the cell's discrete variable that member being a matrix of the products of the level's indicator with the continuous variables of the model. A general value of `ldxc` has the form `ldxc[[cell]][[level]][obs, vbl]`. In the cancer example `ldxc` is a list of two lists: `ldxc[[1]]` has a member for each of the two levels of BM, each member formed as the pointwise product of of the indicator variable for its level with the 475×2 matrix with the variables Weight and Haem as columns; `ldxc[[2]]` has a member for each of the three levels of Perf, each member formed as the pointwise product of of the indicator variable for its level with the 475×1 matrix with the variable Age.

4.2 Setting up structures for storing statistics

The structures in the previous section are built from the columns of the $n \times p$ data frame `dframe[, 1:p]`. The structures to be described now are related to these but the n rows are now replaced by q rows. These rows are formed by weighted averaging over the observation values where the weights are appropriate columns of the matrix Z, This averaging takes place in the E-step module `Estep.r` and all that `Setup.r` does is to create dummy lists of the appropriate length. Just the same we will describe the intended use of these objects now.

dstat is a list of matrices for each discrete variable not included in a location model. The matrix for each discrete variable is made up of a column of length q for each level (value) of the variable giving the expected proportion of each level (column) for each mixture component (row). Rows sum to 1.

ldstat is a list of matrices for each discrete variable within a location model. The matrix for each discrete variable is made up of a column of length q for each level (value) of the variable giving the expected proportion of each level (column) for each mixture component (row). Rows sum to 1.

ostat is the matrix with a column for each continuous variable outside any location mode whose q rows give the current estimated mean for each mixture component.

ostat2 is the matrix with a column for each continuous variable outside any location mode whose q rows give the current estimated mean square for each mixture component.

osvar is the matrix with a column for each continuous variable outside any location mode whose q rows give the current estimated variance for each mixture component.

cstat is a list with a member for each nontrivial, fully continuous, partition cell, that is not including discrete cells or cells listed in `lcdep`, each member being a matrix with a column for each continuous variable in that cell, whose q rows give the current estimated mean for each mixture component.

cstat2 is a list with a member for each nontrivial, fully continuous, partition cell, each member being a matrix with a column for each continuous variable in that cell, whose q rows give the current estimated mean *square* for each mixture component.

cvar is a list with a member for each nontrivial, fully continuous, partition cell, each member being a matrix with a column for each continuous variable in that cell, whose q rows give the current estimated variance for each mixture component.

cpstat is a list with a member for each nontrivial, fully continuous, partition cell, each member being the matrix with rows for each of the q mixture components and columns for each pair of continuous variables in that cell, as ordered by `pair.index`. The matrix elements are the currently expected products of the variable pairs arranged by component and pair.

ccov is a list with a member for each nontrivial, fully continuous, partition cell, each member being the matrix with rows for each of the q mixture components and columns for each pair of continuous variables in that cell, as ordered by `pair.index`. The matrix elements are the currently

expected covariances of the variable pairs arranged by component and pair.

MVMV is a list with a member for each nontrivial, fully continuous, partition cell, each member being a list with members for each of the q mixture components whose values are the covariance matrix estimates for that cell and component.

lcstat is a list with a member for location partition cell, each member being a matrix with a column for each continuous variable in that cell, whose q rows give the current estimated mean for each mixture component.

lcstat2 is a list with a member for location partition cell, each member being a matrix with a column for each continuous variable in that cell, whose q rows give the current estimated mean *square* for each mixture component.

lcpstat is a list with a member for each location cell, each member being the matrix with rows for each of the q mixture components and columns for each pair of continuous variables in that cell, as ordered by `pair.index`. The matrix elements are the currently expected products of the variable pairs arranged by component and pair.

lccov is a list with a member for each location cell, each member being the matrix with rows for each of the q mixture components and columns for each pair of continuous variables in that cell, as ordered by `pair.index`. The matrix elements are the currently estimated covariances of the variable pairs arranged by component and pair.

ldxcstat is a list with a member for each location partition cell, each member being a list with a member for each level of the cell's discrete variable that member being a matrix of mean values of the continuous variables for each level-class combination.

5 Estep.r

This script implements the E-step of the EM algorithm for the *Multimix* model. In other words it calculates the structures holding the complete-data sufficient statistics by forming expectations, these being weighted means of

the data with respect to the weights in the columns of W. Notice the major use of the matrix function `crossprod()` which is an efficiently implemented internal function of R.

6 Mstep.r

After the execution of the E-step we have the expectations of the complete-data sufficient statistics. The maximum likelihood estimates in the complete-data problem are functions of the sufficient statistics; the EM-updated parameter estimates are the same function of the expected complete-data sufficient statistics.

7 likcalc.r

The purpose of this script is to calculate the matrices Z and W from the current parameter estimates and the data. It could be merged with either `Estep.r` or `Mstep.r` but I choose to keep it separate. We should normally use these scripts in the sequence `Mstep.r` $\rightarrow$ `likcalc.r` $\rightarrow$ `Estep.r` because `likcalc.r` needs the parameter estimates calculated in `Mstep.r` and `Estep.r` needs W, which is calculated in `likcalc.r`.

Appendix: Maximum Likelihood Estimation in the Location Model

Because the Location Model of Olkin and Tate [1961] may not be familiar to the reader we review ML estimation in this model. The location model is a probability distribution for $p+1$ random variables, 1 discrete and p continuous. The discrete variable takes values $\lambda_1, \ldots, \lambda_k$ with probabilities $\pi_1, \ldots, \pi_k$. It is convenient to specify the discrete variable via k indicator variables $\boldsymbol{x} = x_1, \ldots, x_k$ such that $x_{ij} = 1$ if the discrete part of the ith observation has value λ_j and is zero otherwise.

We denote the p continuous variables by $\boldsymbol{y} = y_1, \ldots, y_p$. The Location model is specified by requiring that the conditional distribution of $\boldsymbol{y}|\boldsymbol{x}$ is Multivariate Normal:

$$\boldsymbol{y}|\boldsymbol{x} = \boldsymbol{e}_j \sim N_p(\boldsymbol{\mu}_j, \Sigma),$$

where $\boldsymbol{e}_j$ is the vector whose only nonzero element is 1 in position j, $j = 1, \ldots, k$.

Thus the parameters of the Location model are k scalars $\pi_1, \ldots, \pi_k$, constrained to be non-negative and such that $\sum_{j=1}^{k} \pi_j = 1$, k vectors $\boldsymbol{\mu}_j$ of length p and a symmetric, positive-definite $p \times p$ matrix Σ.

Consider now the maximum likelihood estimation of $\pi_1, \ldots, \pi_k$ and $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ from data $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, \ldots, n$, for the moment assuming that Σ is known. To avoid confusion between πs the factor $(2\pi)^{-\frac{n}{2}}$ is omitted from the likelihood L_Σ.

$$L_\Sigma(\boldsymbol{\pi}, \boldsymbol{\mu}) = \prod_{i=1}^{n} |\Sigma|^{-\frac{1}{2}} \prod_{j=1}^{k} \left\{ \pi_j \exp\left(-\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{\mu}_j)' \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_j) \right) \right\}^{x_{ij}}$$

Taking logs

$$\ell_\Sigma(\boldsymbol{\pi}, \boldsymbol{\mu}) = -\frac{n}{2} \log(|\Sigma|) + \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} \left\{ \log(\pi_j) - \frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{\mu}_j)' \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_j) \right\}$$

We may notice at first that ℓ_Σ separates into the sum of terms in $\boldsymbol{\pi}$ and in $\boldsymbol{\mu}$ which may be maximised separately. A simple Lagrange Multiplier derivation shows that $\hat{\boldsymbol{\pi}}_j = n_j/n$, $j = 1, \ldots, k$ where $n_j = \sum_{i=1}^{n} x_{ij}$.

Secondly if we let I_j be the set of i such that $x_{ij} = 1$ then we may separate the part of ℓ_Σ involving $\boldsymbol{\mu}$ into k parts of form

$$\sum_{i \in I_j} \left\{ -\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{\mu}_j)' \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_j) \right\}$$

Again, the maximisation of these terms may be done separately, and the task is the same as that done required when finding the MLEs for the Multivariate Normal Distribution. A good reference for this is Mardia et al. [1979, Section 4.2.2.1]. We find

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^{n} x_{ij} \boldsymbol{y_i}}{\sum_{i=1}^{n} x_{ij}} = \bar{\boldsymbol{y}}_j$$

for $j = 1, \ldots, k$.

Now, allowing Σ to vary we may now form the profile log-likelihood for $V = \Sigma^{-1}$:

$$\ell_p(V) = \frac{n}{2}\log(|V|) + \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}\left\{-\frac{1}{2}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_j)'V(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_j)\right\}$$

omitting a constant term. Now $\ell_p(V) = \sum_{j=1}^{k}\ell_{jp}(V)$ where

$$\ell_{jp}(V) = \frac{n_j}{2}\log(|V|) + \sum_{i\in I_j}\left\{-\frac{1}{2}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_j)'V(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_j)\right\}$$

Now define

$$S_j = \frac{1}{n_j}\sum_{i\in I_j}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_j)'(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_j),$$

for $j = 1, \ldots, k$.

We may then write

$$\ell_{jp}(V) = \frac{n_j}{2}\left(\log(|V|) - tr(VS_j)\right)$$

[Mardia et al., 1979, Section 4.1.1].

Now

$$\frac{\partial \ell_{jp}(V)}{\partial V} = \frac{n_j}{2}\left(2\Sigma - \mathrm{Diag}\Sigma - 2S_j + \mathrm{Diag}S_j\right)$$

and so

$$\frac{\partial \ell_p(V)}{\partial V} = \frac{n}{2}\left(2\Sigma - \mathrm{Diag}\Sigma\right) - \sum_{j=1}^{k}\left(2n_jS_j - \mathrm{Diag}(n_jS_j)\right) = \frac{n}{2}\left(2\Sigma - \mathrm{Diag}\Sigma - 2\tilde{S} + \mathrm{Diag}\tilde{S}\right)$$

where

$$\tilde{S} = \frac{1}{n}\sum_{j=1}^{k} n_jS_j = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_j)'(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_j).$$

This vanishes when and only when $\Sigma = \tilde{S}$, the ML estimate of Σ in the Location model.

References

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

I. Olkin and R. F. Tate. Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Stat.*, 32:448–465, 1961.