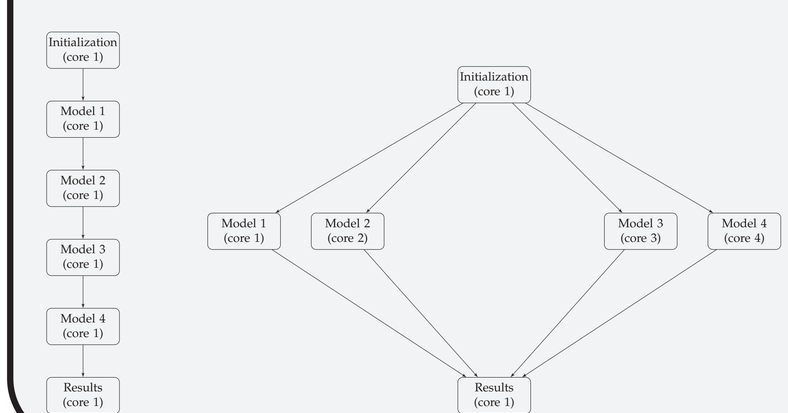
Parallelization of the Mixtures of Multivariate t-Factor Analyzers Software

Mathieu Chalifour and Jeffrey L. Andrews
ChalifourM@mymacewan.ca AndrewsJ24@macewan.ca

Introduction

Mixtures of multivariate t-factor analyzers (MMtFA Andrews and McNicholas, 2011b,a) are a family of statistical models that are used to find inherent groups in data. They assume the data arises from the density $f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_t(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g, \nu_g)$, where $f_t(\cdot)$ is the multivariate t-distribution with mean vector $\boldsymbol{\mu}_g$, factor loading matrix $\boldsymbol{\Lambda}_g$, error variance matrix ψ_g, and degrees of freedom ν_g. Note that G is the number of groups. These models are especially useful since they are robust to outliers and applicable to high dimensional data. Unfortunately, MMtFA models take a significant amount of time to fit even on most modern computers. This is problematic since some applied researchers and scientists may be unwilling to adopt this approach to cluster analysis simply because of the time expense. The main goal of the project was to reduce the runtime for the MMtFA algorithm. This was done by optimizing the software for the MMtFA family by parallelization, allowing each model of the MMtFA family software to be estimated independently across multiple processing cores of a computer simultaneously, instead of a single core in serial order (See Fig 1 for a graphical representation of the difference).

Fig 1: Parallelization



The Software

The software is being developed for the free and open source computing platform R (R Core Team, 2015). R is mainly used for statistical analysis and applied research with many different software packages that can be easily downloaded and installed. Our code relies on the `parallel` package, which allows the user to have the processing cores act semi-independently and simultaneously. Below is a condensed version of the parallelized MMtFA function developed during this project:

```
mmtfa.parallel <- function(...){
  ...                                          # Error-catches and other miscellaneous house-keeping not shown
  numcores <- detectCores()                    # Detects the total number of cores in the computer.
  clus <- makeCluster(numcores)                # Opens copies of R in all the available cores

  clusterEvalQ(clus, library(mclust))
  clusterEvalQ(clus, library(e1071))           # The four clusterEvalQ commands load the R packages needed onto each of the cores.
  clusterEvalQ(clus, library(parallel))
  clusterEvalQ(clus, library(mmtfa))

  clusterExport(clus, ls(environment()),       # This command loads the current working R environment which gives
                envir = environment())         # each core access to the data for which the MMTFA models will be fitted.
  runlist <- clusterApplyLB(clus, runvec,      # Sends a specified number of MMtFA models (dependent on number of groups
                function(g) mmtfa(...))        # and latent factors) to be fitted, using the serial version of MMtFA software on each
                                               # independent processor. This command balances the workload across all processors.
  stopCluster(clus)                            # Closes the multiple copies of R open on the child processors
  ...                                          # Organization of function outputs not shown
}
```

Data Analysis

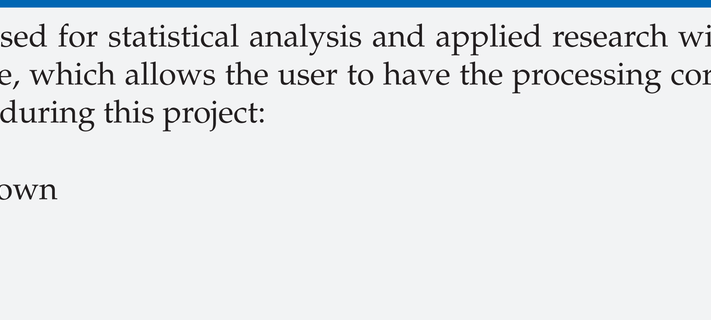
We now apply both the serial and parallel MMtFA functions to three data sets, and additionally compare the results to two other common cluster analyses: k-means and hierarchical clustering.

Wisconsin Breast Cancer: Variables describe characteristics of the cell nuclei, which are computed from a fine needle aspirate (FNA) digitized image of the breast mass. There are a total of 569 observations and 30 continuous measurements. The known groups for classification are Malignant(212) and Benign(357) tumours.

Crabs: The known groups for classification are a cross between the gender and the two colours of the crab, creating four groups each with 50 observations in each. There are 5 morphological measurements for each crab.

Italian Wine: This data set contains 178 observations for 27 continuous measurements on the chemical properties of red wine. The known groups for classification are the three types of wine Barolo(59), Grignolino(71), Barbera(48).

Fig 2: Wine data



Clustering Results: Adjusted Rand

Clustering results are provided via the adjusted Rand index. This index measures pairwise agreement between the known (or 'true') groups present in the data and the groups that the respective algorithm detects. An adjusted Rand of 1 suggests perfect classification, while an adjusted Rand of 0 suggests doing no better than expected from randomly assigning groups. We compare with two popular (and computationally fast) clustering algorithms: hierarchical clustering and k-means.

	MMtFA	Hier-Clus	K-Means
Cancer	0.63	0.05	0.49
Crabs	0.74	0.03	0.02
Wine	0.96	0.00	0.37

Computation Minutes: Serial vs Parallel

This table provides the amount of time (in minutes) required to fit MMtFA on the three datasets. Note that for both the cancer and wine data sets a total of 2,400 models are fit, while for the crabs data only 480 models are fit. This is related to the number of variables present in each data set.

	Serial	Parallel
Cancer	1394.28	300.21
Crabs	193.15	33.40
Wine	164.21	30.40

Results obtained from an 8-core, 4.0 GHz processor (AMD Phenom FX-8350).

Fig 3: Crabs data

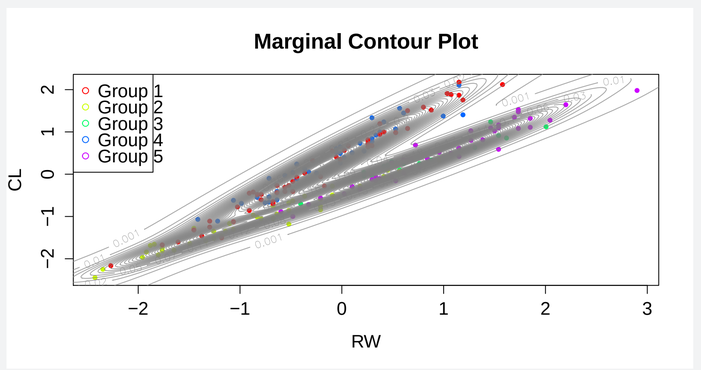
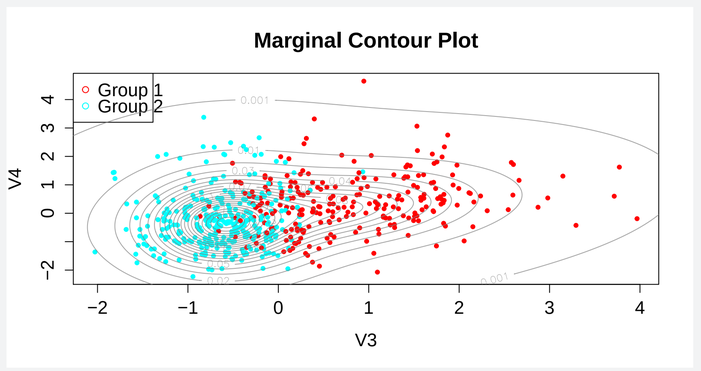


Fig 4: Cancer data



Discussion

There is clearly an advantage to using MMtFA on these three data sets, as it significantly outperforms both of the other methods. Graphical representations of the MMtFA results can be found in Figures 2–4. The parallelization of the MMtFA software was shown to be extremely effective in reducing computation time. We saw the parallelized function fit the models between 4.6–5.8 times faster — thus making the software much more feasible for researchers.

References

Andrews, J. L. and P. D. McNicholas (2011a). Extending mixtures of multivariate t factor analyzers. *Statistics and Computing 21*(3), 361–373.

Andrews, J. L. and P. D. McNicholas (2011b). Mixtures of modified t factor analyzers for model-based clustering, classification, and discriminant analysis. *Journal of Statistical Planning and Inference 141*(4), 1479–1486.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Future Work

A beta version of the MMtFA software is slated to be publicly released by the end of May. Work will continue through collaborations at MacEwan to increase the efficiency of the code over the summer.

Funding

This work is supported by the Natural Sciences and Engineering Research Council of Canada through a Discovery Grant (Andrews) and an Undergraduate Student Research Award (Chalifour). Additional support was provided by a grant from MacEwan University's Research, Scholarly Activity, and Creative Achievement Fund (Andrews).