Dealing with missing data through mixture models

Vincent Vandewalle, Christophe Biernacki

DEALING WITH MISSING DATA THROUGH MIXTURE MODELS

V. Vandewalle[1,2] & C. Biernacki[2,3]

[1]*Univ. Lille, CHU Lille, EA 2694 - Santé publique : épidémiologie et qualité des soins, F-59000 Lille, France*

[2]*Inria*

[3]*Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France*

Abstract

Many data sets have missing values, however the majority of statistical methods need a complete dataset to work. Thus, practitioners often use imputation or multiple imputations to complete the data as a pre-processing step. In this talk it will be shown how mixture models can be used to naturally deal with missing data in an integrated way depending on the purpose. Especially, it will be shown how they can be used to classify the data or derive estimates for the distances. Results on real data will be shown.

I. INTRODUCTION

Missing data are a well-known issue in statistics and have been extensively studied (see for instance [1]). The most generic method to deal with missing data consists in making imputation. It permits to obtain a complete data table on which any standard statistical analysis can be performed. The simplest imputation method consists in imputing missing values by the average, but more sophisticated methods such as the NIPALS algorithm [2] can be used to take into account correlation between variables to input more accurate values. However, imputation tends to underestimate the data variability, and it is often recommended to use multiple imputations, i.e. to replace missing values by many possible values. Thus, producing several possible versions of the completed dataset. A well known multiple imputation method is for instance the fully conditional specification [3] implemented in the R package *mice,* which models the conditional distribution of missing values given the observed ones.

In this communication another solution to deal with missing data will be presented. It consists in modelling the whole distribution of the data by a probabilistic model, also called generative model. In this setting the data distribution is approximated by a mixture of distributions. This approach is generally used to perform model based clustering [4], but it can also take into account missing data in an integrated way through the EM algorithm [5]. In section II, details will be given about the use of mixture models with missing data and an example will be given on how they can be used to estimate distance. In section III, an illustration on real data will be given.

II. METHODOLOGY

For the sake of simplicity it will be supposed that all the data are continuous but extension can be given in the heterogeneous data case at the price of more restrictive assumption on the dependency inside a cluster, e.g. cluster conditional independence assumption. Let consider a sample $\mathbf{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ coming from a mixture of K Gaussian components in dimension

d. Let z_{ik} denote the class of i : $z_{ik} = 1$ if $\boldsymbol{x}_i$ comes from class k and 0 otherwise. Let $\boldsymbol{x}_i^o$ denotes the observed variables for unit i, and $\mathbf{x}^o$ the observed dataset. From the model point of view, $\pi = (\pi_1, \ldots, \pi_K)$ are the proportions of each component, μ_k and Σ_k are the mean and variance covariance matrix of the class k. Let $\lambda_k = (\mu_k, \Sigma_k)$ and θ is the global parameter of the mixture π included. Parameters θ are estimated by maximum likelihood based on the observed data $\mathbf{x}^o$. It is supposed that data are missing at random, i.e. that the probability to be missing does not depends on the value of missing variables given the value the observed ones [1].

In order to maximise the likelihood the EM algorithm is used [5]. The version of the EM that allows taking into account missing values in the dataset is detailed in [6]. At iteration r, the E step consists in computing the conditional expectation $z_{ik}^{(r)} = E[Z_{ik} \mid \boldsymbol{x}_i^o;\ \theta^{(r)}]$ and the conditional expectation $\boldsymbol{x}_{ik}^{m(r)} = E[\boldsymbol{X}_i^m \mid \boldsymbol{x}_i^o, Z_{ik} = 1\ ;\ \theta^{(r)}]$. The M step consists in updating the parameters by maximising the expectation of the completed likelihood. Let notice that the computation of $\boldsymbol{x}_{ik}^{m(r)}$ can be interpreted as an imputation of the missing data given the class and the observed variables. While the M step can be interpreted as parameters estimation on the completed dataset but taking into account the under-estimation of the variance due to the imputation of missing value. At convergence, the estimated model allows us to obtain the probabilities of missing values given the observed one $P(\boldsymbol{X}_i^m \mid \boldsymbol{x}_i^o;\ \hat{\theta})$ which can be used for any purpose such as distance estimation by computing $E[d(\boldsymbol{x}_i,\boldsymbol{x}_{i'})\mid \boldsymbol{x}_i^o,\ \boldsymbol{x}_{i'}^o;\ \hat{\theta}]$ like in [6]. Moreover, it allows to obtain $P(\boldsymbol{Z}_{ik} = 1 \mid \boldsymbol{x}_i^o;\ \hat{\theta})$ which can be directly used for a clustering purpose.

III. RESULTS

Table 1 compares the performances of some pairwise distances strategy such as Partial Distance Strategy (PDS) or Incomplete-case k-NN Imputation (ICkNNI) on three UCI dataset according to the rate of missing data, for more details see [6].

	PDS	ICkNNI	Single Gaussian	Mixture model	Mean K
Computer hardware, $N=209$, $d=6$					
5%	0.676	**0.408**	0.459 (0.443)	0.451 (**0.441**)	3.82
20%	1.102	**0.710**	0.736 (0.737)	**0.732** (**0.730**)	3.70
50%	1.938	1.340	**1.273** (1.327)	1.331 (1.341)	3.52
Glass identification, $N=214$, $d=9$					
5%	0.526	0.337	0.226 (**0.221**)	**0.231** (**0.228**)	2.31
20%	0.971	0.706	**0.524** (**0.522**)	**0.519** (0.532)	2.51
50%	2.098	1.540	**1.197** (1.281)	**1.197** (1.265)	2.45
Housing, $N=506$, $d=13$					
5%	0.514	**0.329**	0.338 (0.348)	**0.331** (0.338)	3.34
20%	1.001	0.672	0.597 (0.650)	**0.587** (0.619)	3.26
50%	2.269	1.593	**1.066** (1.330)	1.104 (1.245)	3.21

Tab.1. Average RMSE of estimated pairwise distances. The best result for each row is underlined, and any results which are not statistically significantly different (two- tailed paired t-test, α1⁄40:05) from the best result are bolded. The values in parenthesis represent the accuracy when the distances are calculated using the particular model for imputation only. The final column shows the mean number of Gaussian components K as selected by the AIC_C criterion, and the mean number of distinct eigenvalues dk for HDDC

IV. DISCUSSION

We see on Table 1 that fitting a mixture of Gaussian distribution often leads to better results than others standards methods, and improve the accuracy of the estimation compared with using only a single Gaussian component. But as the rate of missing data increases it can be advantageous to use only a single Gaussian since not enough data are available to accurately estimate the mixture distribution.

V. CONCLUSION

It has been presented how mixture models can be used to deal with missing data in a very flexible and integrated way. Results of this method have been illustrated is the distance pairwise estimation setting.

References

1. Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.
2. Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate analysis, 1*, 391-420.
3. Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, *76*(12), 1049-1064.
4. McLachlan, G., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.
5. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.
6. Eirola, E., Lendasse, A., Vandewalle, V., & Biernacki, C. (2014). Mixture of gaussians for distance estimation with missing data. *Neurocomputing*, *131*, 32-42.

40 avenue de Halley, 59650 Villeneuve d'Ascq, France
DEALING WITH MISSING DATA THROUGH MIXTURE MODELS, Vincent Vandewalle
vincent.vandewalle@univ-lille2.fr