

Use of Mixture Models in Editing and Imputation of Survey Data

Ugo Guarnera - guarnera@istat.it
Italian National Institute of Statistics

Marco Di Zio - Italian National Institute of Statistics
Orietta Luzi - Italian National Institute of Statistics

Cardiff, 24-26 April 2006

Finite Mixture Models

Let $f_1(\cdot; \theta_1), \dots, f_K(\cdot; \theta_K)$ be K *p.d.f.s* of the same parametric family and \mathbf{Y} a p -dimensional random vector whose *p.d.f.* can be written in the form:

$$f(\mathbf{y}; \Phi) = \sum_{k=1}^K \pi_k f_k(\mathbf{y}; \theta_k)$$

where $\sum_k \pi_k = 1, \pi_k \geq 0$ for $k = 1, \dots, K$

and $\Phi = (\theta_1, \dots, \theta_K; \pi_1, \dots, \pi_K)$.

The density f is called a *mixture* of the distributions f_1, \dots, f_K , the densities f_k are the *mixture components*, and the parameters π_k the *mixing proportions*. When each f_k is a gaussian density, $f_k(y) = N_k(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, f is called a *gaussian mixture*.

Gaussian Mixture Models

Mixtures of Gaussian distributions are a powerful tool mainly in two fields:

- Classification
- Density estimation

Gaussian Mixture Models in Official Statistics

In the context of Official Statistics, when quantitative variables have to be analysed, Gaussian Mixtures Models (GMM) can be useful for classification in the editing phase, and for density estimation in the imputation phase.

- Classification \Rightarrow Identification of Unity Measure Errors
- Density Estimation \Rightarrow Imputation of Item Non-Responses

Editing Unity Measure Errors (1)

- A unity measure error (UME) is a systematic error occurring when the 'true' value of a variable X is reported in a wrong scale. Typically it acts as multiplication by a constant factor: $X \rightarrow X \cdot C$ (e.g. $C = 1,000$).
- Due to the UME, the final estimates can be strongly biased. We need to identify responding units affected by UME in one or more survey variables $X_j \quad j = 1, \dots, p$
- When data variability is large, identification of unity measure errors is difficult because of the overlapping between distributions corresponding to true and erroneous data.

Editing Unity Measure Errors (2)

The problem of identifying UMEs can be approached from a probabilistic clustering perspective by assuming that observations are from a mixture of a finite number of populations each corresponding to a specific error pattern (i.e. a specific sub-set of variables affected by UME).

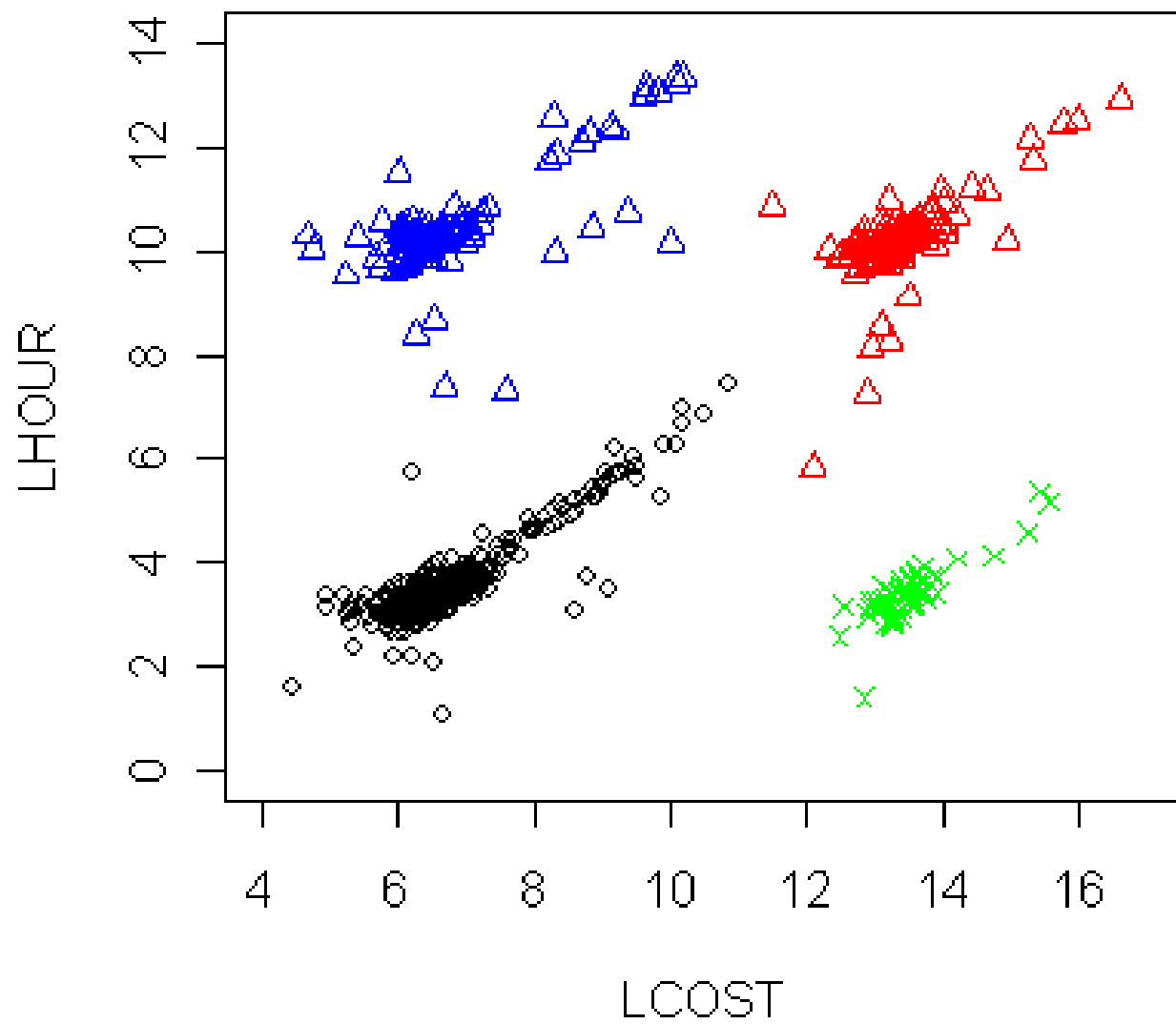
Once a suitable mixture model have been estimated, the model parameters can be used to estimate, for each unit, the **posterior probabilities** (PP) of belonging to the different clusters (error patterns). Each unit is then assigned to the cluster corresponding to the highest PP.

Details can be found in Di Zio et al. (CSDA,2006)

Editing Unity Measure Errors (3)

Posterior probabilities can be used to prioritize units to be re-contacted according to:

1. the degree of belief in their assignment (if the highest PP is not very close to one, there is a non-negligible risk of mis-classification)
2. the potential expected impact of misclassification on the final estimates (*selective editing*)



Imputation

In the context of Official Statistics it is common practice to manage partial non-response through imputation, i.e. filling in missing values with 'plausible' ones. Under Missing at Random (MAR) assumption, missing items can be imputed on the base of observed data. Two main classes of methods are generally used:

1. parametric methods (EM, Regression...)
2. non parametric methods (Hot-Deck, Nearest Neighbor, ...);

Pros and cons

- Non-parametric techniques are generally substantiated by asymptotic arguments. Thus they require a large number of observations. Application of non-parametric methods may produce attenuation of association between variables.
- Using parametric methods, associations can be better preserved, but results depend strongly on the model specification. Therefore a great care has to be taken in model building.

Imputation via GMM

Finite Mixture Models provide a semiparametric imputation method that can be considered in between of the two previous approaches. It allows to handle data from unknown distributions trying, at the same time, to be as parsimonious as possible.

The idea is to estimate the data distribution through a suitable mixture model and to use the estimated model to impute missing items.

Proposed method

The main steps of the proposed procedure are the following:

1. estimate mixture models with different number of components;
2. choose the model that 'best' fits data;
3. use the estimated parameters of the selected model to derive the conditional distributions corresponding to the different error patterns;
4. for each error pattern, impute missing items through the appropriate conditional distribution.

Estimation and Model Selection

- Model parameters are estimated via an appropriate Expectation Maximization (EM) algorithm that performs the ML estimates in presence of missing data (Hunt and Jorgensen (2002, CSDA)). It basically combines the standard EM algorithm for Gaussian mixtures with the EM algorithm for incomplete normal data.
- EM is initialized through an ordinary K-Means algorithm;
- The 'best' model is selected according to the Bayesian Information Criterion (BIC);

Imputation step (1)

Let $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from the p -dimensional random vector \mathbf{Y} distributed as a finite mixture of K Gaussian distributions:

$$f(\mathbf{y}_i; \Phi) = \sum_{k=1}^K \pi_k N_k(\mathbf{y}_i; \boldsymbol{\theta}_k).$$

Let $(\mathbf{y}_{obs,i})$ denote the observed values of \mathbf{y}_i , and $(\mathbf{y}_{mis,i})$ the missing ones:

$$\mathbf{y}_i = (\mathbf{y}_{mis,i}, \mathbf{y}_{obs,i}).$$

Imputation step (2)

Once the parameters $\Phi = (\pi_k, \theta_k)_{k=1, \dots, K}$ have been estimated, the posterior probabilities τ_{ik} for each observation i ($i = 1, \dots, n$) of belonging to group k can be estimated as:

$$\hat{\tau}_{ik} = \frac{\hat{\pi}_k N_k(\mathbf{y}_{obs,i}; \hat{\theta}_k)}{\sum_{k=1}^K \hat{\pi}_k N_k(\mathbf{y}_{obs,i}; \hat{\theta}_k)};$$

The conditional distributions $g_k(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i}; \hat{\theta}_k)$ corresponding to group k can be also estimated via standard formulas of the normal distribution.

Imputation step (3)

Imputation of $\mathbf{y}_{mis,i}$ ($i = 1, \dots, n$) can be performed by means of two strategies:

- **Random draw:** draw a value $\mathbf{y}_{mis,i}$ from the conditional distribution $g(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i}, \hat{\Phi}) = \sum_{k=1}^K \hat{\tau}_{ik} g_k(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i}, \hat{\theta}_k)$;
- **Conditional expectation:** impute each missing vector $\mathbf{y}_{mis,i}$ with the conditional expectation of the r.v. $\mathbf{Y}_{mis,i} | \mathbf{Y}_{obs,i}$ w.r.t. $g(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i}, \hat{\Phi})$.

Experiments(1)

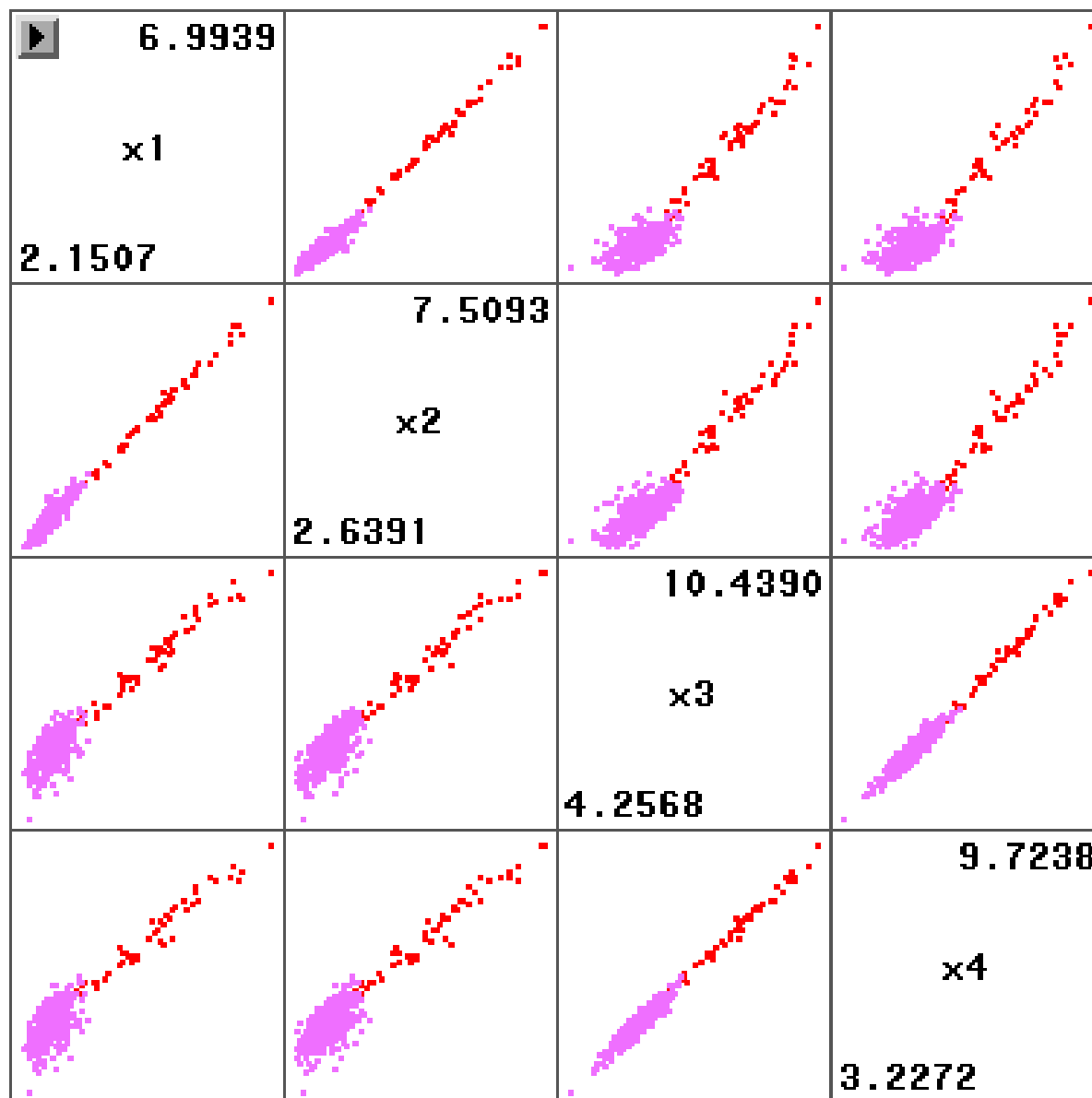
Main goal: evaluating imputation via GMM through a comparison with nearest neighbor donor (NND)imputation. The simulation study has been performed on data from the Italian Labour Cost (LCS) survey 1997. The following steps have been repeated 100 times:

- a simple random sample with replacement of n units is extracted from the LCS dataset and the four variables *employees* (Y_1), *worked hours* (Y_2), *gross income* (Y_3), and *social contributions* (Y_4) are analyzed;
- for the first three variables missing values (MCAR) are introduced according to different rates: 20% for Y_1 , 30% for Y_2 , 40% for Y_3 . No missing values are introduced for Y_4 ;

Experiments(2)

- mixture Models are estimated with different number K of components on the incomplete dataset. The best model is selected according to BIC;
- missing items are imputed using the best mixture model (both random draw and conditional expectation), and NND;
- for each imputation method, indicators are computed comparing the sample mean and the sample variance-covariance matrix of the complete original sample with those of the imputed dataset.

A final evaluation is obtained by averaging the indicators values over the 100 repetitions for each group of experiments.



Results (nobs=500)

After each iteration differences are computed between means, variances/covariances and individual values of the original dataset and the corresponding quantities of the imputed dataset. Indicators are then built by averaging over the 100 iterations. $m_j \rightarrow$ sample mean of Y_j ; $p_j \rightarrow$ predictive accuracy for Y_j ; $var / cov \rightarrow$ diagonal/off-diagonal elements of the Covariance matrix.

	<i>var</i>	<i>cov</i>	<i>p</i> ₁	<i>p</i> ₂	<i>p</i> ₃	<i>m</i> ₁	<i>m</i> ₂	<i>m</i> ₃
NND	7.06	10.95	1.12	1.21	0.82	0.16	0.13	0.10
MM0	3.44	3.81	0.88	0.92	0.63	0.10	0.08	0.05
MM1	3.27	5.51	1.25	1.31	0.88	0.11	0.11	0.07

NND= Nearest Neighbor; MM0=Mixture-Conditional Mean; MM1=Mixture-Random Draw; indicators are relative differences (%) w.r.t. the corresponding quantities in the original data-set

Comments

- In general, imputation via mixture model seems better than NND imputation.
- MM0 is better than MM1 except for preservation of variances.
- Experiments with simulated data show that Gaussian Mixtures overperform NND mainly when some of the analyzed variables are not strongly associated.

Problems

- Likelihood is not bounded for heteroscedastic gaussian mixtures.
- In general the likelihood surface has several local maxima: the EM algorithm might converge to any one of them.
- In some cases, estimates from EM depend strongly on the initialization procedure: the starting point of the EM must be chosen with great care.
- Gaussian mixture can hardly model data from semicontinuous distributions (e.g. zeros inflation in survey data). In these cases NND imputation seems more appropriate.