A SEMIPARAMETRIC PREDICTIVE MEAN MATCHING: AN EMPIRICAL EVALUATION

Marco Di Zio - dizio@istat.it
Italian National Institute of Statistics

Ugo Guarnera - Italian National Institute of Statistics

UN/ECE, Bonn, 27 September 2006

Imputation

In the context of Official Statistics it is common practice to manage partial nonresponse through imputation, i.e. filling in missing values with ‘plausible’ ones.
Two main classes of methods are generally used:

1. *parametric methods* (EM, Regression, ratio imputation,...) ;

2. *nonparametric methods* (Random Hot-Deck, Nearest Neighbor, ...);

Pros and cons

- Nonparametric techniques are generally substantiated by asymptotic arguments. They require a large number of observations. Application of nonparametric methods may produce attenuation of association between variables.

- Using parametric methods, associations can be better preserved, but results strongly depend on the model specification. So a great care must be taken in model building.

Imputation via Gaussian Mixture Models (GMM)

Finite Mixture Models provide a semiparametric imputation method that can be considered in between of the two previous approaches. It allows handling 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.

Finite Gaussian Mixture model with missing values

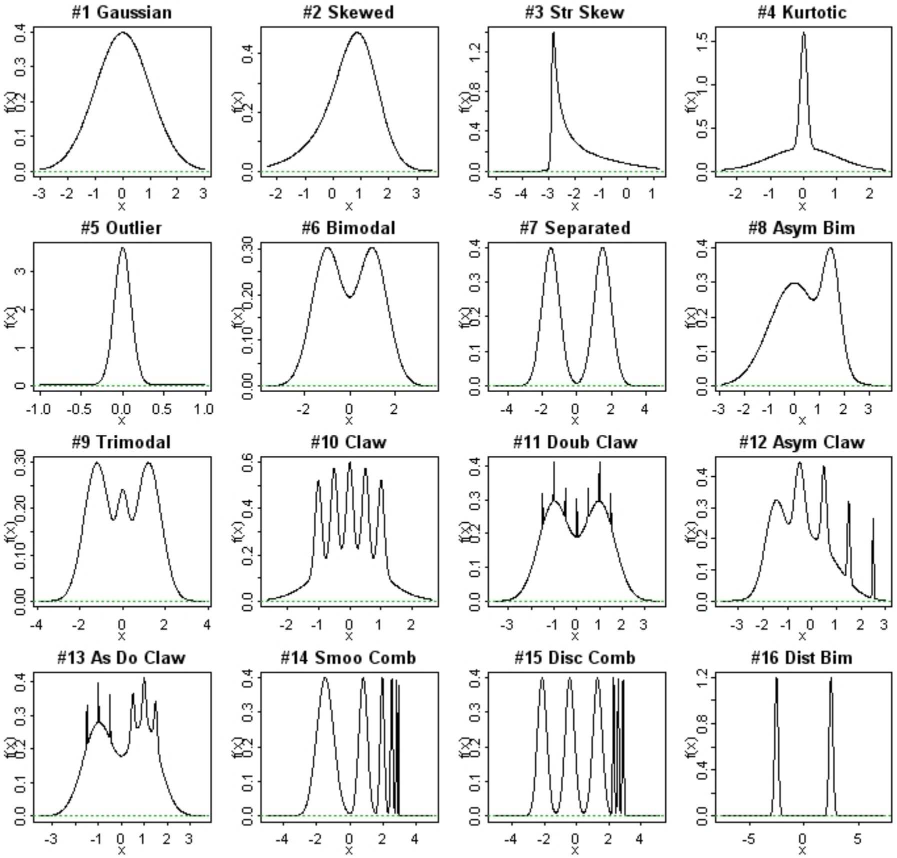
Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from the r.v. $\mathbf{Y}$ (p-dimensional) distributed as a finite mixture of K Gaussian distributions:

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

where $\sum_k \pi_k = 1, \pi_k \geq 0$ for $k = 1, \ldots, K$, and $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$. Let us introduce the vector of indicator variables $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})'$ where z_{ij} is 1 if the individual $i \in$ group k, and 0 otherwise.

For each unit $\mathbf{y}_i = (\mathbf{y}_{mis,i}, \mathbf{y}_{obs,i})$, where $(\mathbf{y}_{obs,i})$ are the observed variables, and $(\mathbf{y}_{mis,i})$ the missing.

Density estimation flexibility



Imputation via GMM

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

- Random draw (MRD): draw a value $\mathbf{y}_{mis,i}$ from the conditional distribution $g(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i}, \hat{\mathbf{\Phi}}) = \sum_{k=1}^{K} \hat{\tau}_{ik} g_k(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i}, \hat{\boldsymbol{\theta}}_k)$;

- Conditional expectation (MCM): 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. $g(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i}, \hat{\mathbf{\Phi}})$.

Comments

- Some experiments showed that imputation through random drawing (MRD) from a finite Gaussian mixture model has a better behavior than MCM, NND (similar preservation of means, better preservation covariances)

- Some experiments showed that Gaussian mixtures overperform NND especially when some of the analyzed variables are not strongly associated.

but...

Comments (2)

The imputed value is synthetic, and some strange situations at micro level can happen, i.e. negative values when data are nonnegative.

Can an hybrid technique maintain the properties of a Gaussian mixture model and the NND? That is: good behavior in the preservation of associations (model) but using ‘live’ values.

A known hybrid method, based on Gaussian distribution and NND, is the Predictive Mean Matching.

Predictive Mean Matching

- parameter estimation of a multivariate Gaussian model (EM-algorithm)

- for each incomplete record $y_{miss,i}$, the conditional expected value $\widehat{\mu}_i = E(\mathbf{Y}_{miss,i}|\mathbf{y}_{obs,i})$ is estimated

- for each incomplete record i, impute through NND with distance $d^2(i,j) = (\widehat{\mu}_i - \widehat{\mu}_j)^T S^{-1}_{\mathbf{Y}_{miss,i}|\mathbf{Y}_{obs,i}}(\widehat{\mu}_i - \widehat{\mu}_j)$

Little, 1988: '...one might tolerate greater matching error for variables that are subject to greater prediction error...' ($\Rightarrow S^{-1}_{\mathbf{Y}_{miss,i}|\mathbf{Y}_{obs,i}}$)

Predictive Mean Matching via Gaussian Mixture models (SPMM)

- parameter estimation of a Gaussian mixture (penalised EM-algorithm)
- for each incomplete record $y_{miss,i}$, the conditional expected value $\widehat{\mu}_i = E(\mathbf{Y}_{miss,i}|\mathbf{y}_{obs,i})$ is estimated
- for each incomplete record i, impute through NND with distance $d^2(i,j) = (\widehat{\mu}_i - \widehat{\mu}_j)^T S^{-1}_{\mathbf{Y}_{miss,i}|\mathbf{Y}_{obs,i}}(\widehat{\mu}_i - \widehat{\mu}_j)$

Remark: for $K = 1$ (only one mixture component) this corresponds to the original Predictive Mean Matching

Experiments (1)

Main goal: *evaluating imputation via SPMM w.r.t. NND, MRD, MCM*.

1) artificial generation of a sample from a given multivariate (5 r.v.) probability distribution, and a set of real data (4 r.v.);

2) introduction of missing values in the sample;

3) estimation of mixture model and imputation;

4) comparison of the imputed with the original dataset through indices

Repeat 100 simulations and averaging the results

Experiments (2)

Sample data generation. Multivariate log-normal distribution, multivariate Gamma (2 par. setting). For each one: 500, 1000 sample size. Bootstrap sampling from a subset of Labour Cost data.

Missing generation (MAR). for 4 r.v. prob. miss. is 0.10 if $y_5 < q_1$, 0.20 if $y_5 \in [q_1, q_3)$, 0.30 if $y5 \geq q_3$. No missing values are introduced in the variable Y_5

Evaluation through the comparison of the mean and covariance matrix computed on original and imputed data set. The indicators are the relative root mean square error, for each single component of the mean and covariance matrix.

Multivariate Gamma (Block Correlations)

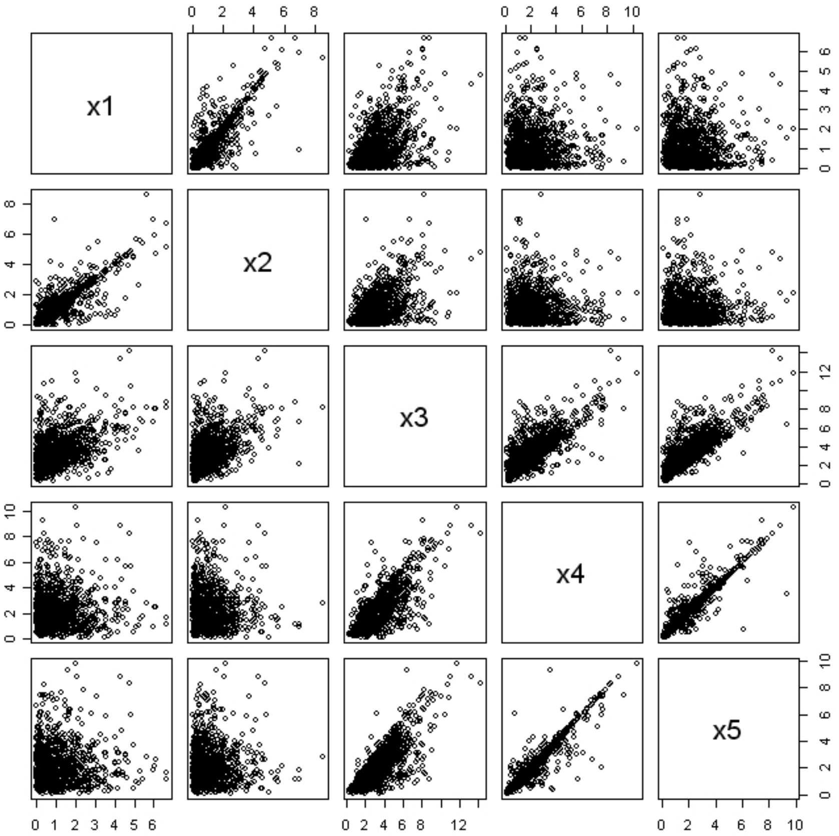


Table 1. Lognormal data with sample size 500 (LN 500) and 1000 (LN 1000)

LN 500

	Dm	DS
SPMM	0.0002	0.0037
NND	0.0003	0.0045
MCM	0.0001	0.0049
MRD	0.0002	0.0028

LN 1000

	Dm	DS
SPMM	0.0001	0.0018
NND	0.0001	0.0023
MDM	0.0000	0.0045
MRD	0.0001	0.0011

Table 2. Multivariate Gamma (MGL) sample size 500 (MGL 500) and 1000 (MGL 1000)

MGL 500

	Dm	DS
SPMM	0.0008	0.0813
NND	0.0007	0.0871
MCM	0.0004	0.0573
MRD	0.0006	0.0499

MGL 1000

	Dm	DS
SPMM	0.0004	0.0348
NND	0.0004	0.0434
MCM	0.0002	0.0422
MRD	0.0003	0.0190

Table 3. Multivariate Gamma (MGH) sample size 500 (MGH 500) and 1000 (MGH 1000)

MGH 500

	Dm	DS
SPMM	0.0014	0.0881
NND	0.0013	0.0715
MCM	0.0006	0.0545
MRD	0.0009	0.0434

MGH 1000

	Dm	DS
SPMM	0.0007	0.0332
NND	0.0007	0.0319
MCM	0.0003	0.0659
MRD	0.0005	0.0250

Table 4. Labour cost data experiment (CLAV)

CLAV

	Dm	DS
SPMM	0.0000	0.0003
NND	0.0000	0.0003
MCM	0.0000	0.0001
MRD	0.0000	0.0001

Results

Preservation of the mean. It is similar in all the methods (slightly better MCM).

Preservation of the covariance matrix. The best one is random draw from the estimated conditional probability distribution (MRD).

NND vs SPMM.
When there are some variables with low correlations: SPMM is better (Tables 1 and 2).

When all the variables are highly correlated, the behavior is similar with a slight preference for NND.

Why?

- The distance in SPMM is based on the conditional expected values estimated through an explicit model. SPMM takes into account the different influences of the covariates on the response variables, while NND treats all the covariates at the same way

- Moreover, weights to variables could be used in NND. But in SPMM the weights vary according to the missing data pattern.

- When all the variables are highly correlated, (all the var. are useful for prediction) imputation based on a real distance function, instead of predictive means, results in a better estimation of the conditional probability distribution.

Final Remarks

As Marker et al. (2000) say: "*There are two main challenges in the field of imputation*":

1. *maximize the use of available data to minimize mean square error for univariate statistics and to preserve covariance structure*

2. *reflect the uncertainty caused by item non-response in variance estimates from imputed data sets*

We have considered only the first, for the second...future works

A question: donor can handle complex situations like the presence of zero values, but are we sure about the statistical properties of NND in this context?