

# Ensemble Support Vector Regression: A New Non-parametric Approach for Multiple Imputation

Daria Scacciatelli <sup>1</sup>

Dipartimento di Statistica, Probabilità e Statistiche Applicate,  
e-mail: daria.scacciatelli@gmail.com

**Abstract:** The complex case in which several variables contain missing values, needs to be analyzed by means of an iterative procedure.

In this paper we propose a new non-parametric method for multiple imputation based on Ensemble Support Vectors Regression. This procedure was tested with different simulation schemes. We show that the results obtained in this way are better than the ones obtained with other methods usually employed to get a complete data-set.

**Keywords:** Missing Data, Multiple Imputation, Non-Parametric Methods, Bootstrap

## 1. Introduction

Analyses of multivariate data are frequently hampered by missing values. Partially empty data-set are hard to interpret and should be avoided whenever possible. The source of missingness of some variables can be numerous and its nature can vary from total randomness (Missing Completely at Random, MCAR) to a particular dependence from the variables (Non-Ignorable, NI).

Up to some time ago, the only method available for treating missing data was listwise deletion. Recent advances in theoretical and computational statistics, have produced a new generation of procedures, both parametric and non-parametric. These new procedures employ Multiple Imputation (MI), Rubin (1987), a simulation method for which every missing value is replaced by a set of  $m > 1$  plausible values, which are bayesian draws from the conditional distribution of the missing observations given the observed data. Instead of using bayesian predictive distribution in multiple imputation, some authors, in order to obtain several values for each missing value, use different approaches based on non-parametric methods with bootstrap.

Here, we propose a new non-parametric approach for multiple imputation based on Ensemble Support Vector Regression.

## 2. Multiple Imputation Approaches

By Rubin's Multiple Imputation  $m$  possible alternative versions of the complete data are produced. Then each dataset is analyzed by a complete data method. Rubin proposed a method to combine the results of  $m$  analyses.

If  $Q$  is a generic scalar parameter to estimate, we compute  $m$  different versions of the

---

<sup>1</sup>Address of correspondence: Dipartimento di Statistica, Probabilità e Statistiche Applicate P.le Aldo Moro 5 00185 Roma

estimator  $\hat{Q}$  and its variance  $U: [\hat{Q}^{(j)}, U^{(j)}], j = 1, \dots, m$ . Rubin's estimate is simply the average of  $m$  estimates:

$$\bar{Q} = m^{-1} \sum_{j=1}^m \hat{Q}^{(j)} \quad (1)$$

where the uncertainty in  $\bar{Q}$  has two components: the average *within* the imputations variances  $\bar{U}$  and the *between* imputation variances  $B$ .

The total variance is  $T = \bar{U} + (1 + m^{-1})B$ , where  $T^{-\frac{1}{2}}(\bar{Q} - Q) \sim t_\nu$ , by which we can calculate confidence intervals and hypothesis tests.

The imputations of the incomplete dataset can be generated by employing either the Expectation-Maximization, Dempster *et al.* (1977), or the Data Augmentation, Shafer (1997), algorithms. In order to apply Rubin's Multiple Imputation, some conditions have to be satisfied: we must have proper imputations, multinormal data, mechanism of MAR type. The validity of this approach depends on these assumptions, if the imputations are proper then  $\bar{Q}$  is a consistent, asymptotically normal estimator, and the variance  $T$  is an unbiased estimator of its asymptotic variance. Nevertheless, the evaluation of the properness of MI is analytically intractable, and therefore it is best done via simulations, Brand *et al.* (2003). Besides, the standard multiple imputation approach is based on distributional assumptions and it is parametric. However, in general, there are many applications where fully parametric approaches may not be suitable, then it is important to focus on semi-parametric or non-parametric imputation methods, without taking into account restrictive hypotheses, Durrant (2005). An alternative non-parametric approach, based on Nearest Neighbors, is the Approximate Bayesian Bootstrap, that was considered by Rubin and Schenker (1986). The bootstrap was also employed in missing data problem by Efron (1994), and recently Di Ciaccio proposed a multiple imputation approach based on bootstrap and on a non-parametric model, the Regression Tree with a Bagging algorithm, Di Ciaccio (2008).

### 3. A new approach based on Ensemble Support Vector Regression

Support vector machines are derived from Statistical Learning Theory and were first introduced by Vapnik (1999). In support vector regression (SVR) the input  $\mathbf{x}$  is mapped to a higher dimensional feature space  $\Phi(\mathbf{x})$  in a non linear manner. This is given by (2)

$$f(\mathbf{x}) = (\mathbf{w} \cdot \Phi(\mathbf{x})) + b \quad (2)$$

where  $b$  is the threshold for the support vector equation,  $\mathbf{w}$  is a constant vector that can be estimated by reducing the Empirical Risk and the complexity term. In (3)

$$R(f) = \frac{1}{2} \|\mathbf{w}\|^2 + C \cdot \sum_{i=1}^{\ell} L_{\epsilon}(y_i, f(\mathbf{x})) \quad (3)$$

the first term  $\|\mathbf{w}\|^2$  characterizes the complexity model and the second term represents the Empirical Risk,  $\ell$  is the number of the examples of training set,  $C$  is a constant that describes the trade-off and the  $\epsilon$ -insensitive loss function  $L_{\epsilon}(y)$  is given by:

$$L_{\epsilon}(y, f(\mathbf{x})) = \begin{cases} 0 & \text{se } |y - f(\mathbf{x})| \leq \epsilon \\ |y - f(\mathbf{x})| - \epsilon & \text{otherwise} \end{cases} \quad (4)$$

Some authors considered the use of Ensemble Support Vector Regression, Wang and Yang (2008); Kim *et al.* (2003). The main aspect in constructing the SVR ensemble is that each single SVR has to be as much as possible different one from another. This requirement can be obtained by using different training sets. Some methods to generate the training samples are bagging, boosting, randomization. In the bagging method, several SVRs are trained independently on bootstrap samples and then they are aggregated with an appropriate combination technique. Our non-parametric approach based on Ensemble Support Vector Regression (SVRE) allows us to obtain  $m$  complete datasets by which we can evaluate the additional variability due to missing values. The SVRE method initializes, for example by the use of the mean, the missing values in the dataset, and then, iteratively until convergence, considers each variable with missing data as the target variable. Some details of our method, are described below.

---

**Algorithm 1** The Algorithm

---

1. For each variable: initialize missing data with the mean.
  2. Iterate (max num.of iteration= $T$ ).
  3. Iterate ( $j=1$  to num. variables  $J$ ).
  4. Set the variable  $j$  as the target variable.
  5. Select cases which do not have missing value for the variable  $j$ .
  6. Estimate the predictor model based on Support Vector Regression Ensemble.
  7. Impute the missing values of variable  $j$  by the predictor model estimated on step 6.
  8. Update missing values of variable  $j$ . Go to step 3.
  9. If  $\delta < 0.001$  then STOP or else go to step 2.
- 

To evaluate the convergence of the procedure at the iteration  $t$ , we use the index defined below:

$$\delta_t = \frac{1}{n_{mis}} \sum_{j=1}^J \frac{1}{\sigma_j^2} \sum_{i \in M_j} (\hat{x}_{ij}^t - \hat{x}_{ij}^{t-1})^2 \quad (5)$$

With  $\hat{x}_{ij}^t$  we indicate the estimated value, obtained applying to the missing value  $x_{ij}$  the model at the iteration  $t$ . Then in step 8, to update the value  $\hat{x}_{ij}^t$  we use the formula:

$$\hat{x}_{ij}^t \leftarrow \gamma \hat{x}_{ij}^{t-1} + (1 - \gamma) \hat{x}_{ij}^t \quad (6)$$

## 4. Results

To evaluate the proposed SVRE algorithm we considered different simulation schemes with both multinormal and non-multinormal data, and different correlation levels between variables. Data were generated using copulas. For each simulation scheme we generated 300 random populations with 5 variables, from each population we extracted one sample of 120 units, where we fixed randomly the percentage of missing values on the first two variables (15% and 30%). The predictor model used in step 6, was the Ensemble Support vector regression algorithm, with 30 bootstrap replications. In particular, we fixed  $T = 50$  and  $\gamma = 0.9$ . Based on single value prediction, we compared the behavior of SVRE algorithm with Rubin's Multiple Imputation (MI) ( $m = 30$ ), and, for a lower bound reference, also with the mean imputation (MEAN) naive algorithm. We evaluated the

difference between the imputed data values and the known ones by the weighted sum of square:

$$SSEW = \frac{1}{n_{miss}} \sum_{j=1}^J \frac{1}{\sigma_j^2} \sum_{i \in M_j} (\hat{x}_{ij} - x_{ij})^2 \quad (7)$$

**Table 1::** Results of imputation: percentage of samples favorable to each method.

| Data                    | 15% of missing |      |    | 30% of missing |      |    |
|-------------------------|----------------|------|----|----------------|------|----|
|                         | MEAN           | SVRE | MI | MEAN           | SVRE | MI |
| Multinormal Independent | 47             | 33   | 20 | 59             | 30   | 11 |
| Multinormal Dependent   | 0              | 10   | 90 | 0              | 9    | 91 |
| Non-normal Independent  | 71             | 23   | 6  | 80             | 17   | 3  |
| Non-normal Dependent    | 2              | 67   | 31 | 5              | 63   | 32 |

If our objective is to obtain a complete data matrix, then Rubin's Multiple Imputation works well with multinormal dependent data. As the correlation level decreases, Rubin's method could be substituted by simpler models as, for instance, the mean imputation. In order to obtain a complete data set in case of non-normal dependent data, our non-parametric method SVRE represents an efficient alternative to Rubin's Multiple Imputation. Moreover, our method gives good imputations also when the correlation level decreases slightly.

## References

- Brand J., Buuren S., Groothuis-Oudshoorn K. and Gelsema E.S. (2003) A toolkit in sas for the evaluation of multiple imputation methods, *Statistical Neerlandica*, 57, 36–45.
- Dempster A.P., Laird N. and Rubin D.B. (1977) Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society*, 39, 1–38.
- Di Ciaccio A. (2008) Bootstrap and nonparametric predictors to impute missing data, *Cladag*.
- Durrant G.B. (2005) Imputation methods for handling item-nonresponse in the social sciences: A methodological review, *NCRM Working Paper Series*.
- Efron B. (1994) Missing data, imputation, and the bootstrap, *Journal of the American Statistical Association*, 89, 463–475.
- Kim H.C., Pang S., Je H.M., Kim D. and Yang Bang S. (2003) Constructing support vector machine ensemble, *Pattern Recognition*, 36, 2757–2767.
- Rubin D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, Jhon Wiley & Sons.
- Rubin D.B. and Schenker N. (1986) Multiple imputation for interval estimation from simple random samples with ignorable nonresponse, *Journal of the American Statistical Association*, 81, 366–374.
- Shafer J. (1997) *Analysis of Incomplete Multivariate Data*, Chapman and Hall.
- Vapnik V.N. (1999) *The Nature of Statistical Learning Theory*, Springer Verlag.
- Wang F. and Yang H.Z. (2008) epsilon-insensitive support vector regression ensemble algorithm based on improved adaboost, *Computer Engineering and Applications*, 44, 42–44.