HOW EFFICIENT IS ESTIMATION WITH MISSING DATA?

Seliz G. Karadoğan, Letizia Marchegiani, Lars Kai Hansen, Jan Larsen

[1] DTU Informatics, Technical University of Denmark,
DK-2800, Kgs. Lyngby, Denmark
[2] Department of Computer and System Sciences
Sapienza, University of Rome, 00185 Roma, Italy

ABSTRACT

In this paper, we represent a new evaluation approach for missing data techniques (MDTs) where the efficiency of those are investigated using listwise deletion method as reference. We experiment on classification problems and calculate misclassification rates (MR) for different missing data percentages (MDP). We compare three MDTs: pairwise deletion (PW), mean imputation (MI) and a maximum likelihood method that we call complete expectation maximization (CEM). We use synthetic dataset, Iris dataset and Pima Indians Diabetes dataset. We train a Gaussian mixture model (GMM) with missing at random (MAR) data. We test the trained GMM for two cases, in which test dataset is missing or complete. The results show that CEM is the most efficient method in both cases while MI is the worst of the three. PW and CEM prove to be more stable with respect to especially higher MDP values than MI.

***Index Terms*—** Machine learning, supervised learning, missing data techniques

1 Introduction

The reconstruction of degraded audio and video sequences, the analysis of images with missing pixels or occlusions, the manipulation of distorted signals because of a sensor failure or outliers are just some of the wide range of situations in which it is necessary to face the missing data problem. This issue, in fact, is really common in various studies and in several applications using statistical approaches, such as: psychological and psychometric analyses dealing with surveys without all the requested answers, market researches exploiting incomplete interviews or medical diagnoses based on partial accessible information.

Different strategies have been investigated in different areas to handle the missing data problem and many techniques have been proposed. Basically, it is possible to group these techniques in three big categories: *deletion* methods, *imputation* methods and *model-based* methods. In the first ones, the analysis considers only the present data. The deletion procedure can be executed removing only the missing elements (*pairwise deletion*) or the entire units containing them (*listwise deletion*) [1, 2]. In the imputation methods, the holes in the data set are replaced with other estimates, so that, like in the pairwise deletion, all the available information is kept and utilized. The simplest way to implement an imputation process is to substitute the missing value of a variable for the mean value of the same variable (*mean imputation*) [3]. In [4], Rubin proposes the concept of *multiple imputation (MI)*, which consists of inserting several values, instead of just one, for each missing instance. This process generates many complete imputed data sets and standard complete data methods are, then, used to examine each of them. The model-based methods, instead, are able to perform directly their analysis on the incomplete set, without changing or ignoring part of the available information. In particular, maximum likelihood (ML) approaches are the most representative in this category and Expectation Maximization algorithms (EM) are often used in this perspective ([5], [6]).

The behaviour of these methods has been explored in literature, taking into account of the classification of the distribution of missingness proposed by Rubin in [7]. Specifically, data are *missing at random (MAR)* if the probabilities of missingness could depend on the observed data, but not on the missing ones; *missing completely at random (MCAR)* if the probabilities depend on neither the observed and nor the missing data. In the opposite case, data are *missing not at random (MNAR)*.

Roth in [2] provides a qualitative evaluation of the most common missing data approaches considering scenarios in applied psychology. Allison in [1] analyses advantages and disadvantages of the same methods, on the basis of three criteria: the capability to minimize bias, maximize the use of available information and yield good estimates of uncertainty. Schafer and Graham, perform in [8] an analysis close to the cited work of Allison using means, bias and mean square error to evaluate the model estimation accuracy and the behaviour of the standard error to evaluate the margin of the uncertainty. Myrtveit et al. in [9] investigate missing data methods in the context of software cost modelling. In particular, the work focuses on the possible benefits that could be obtained thanks to the use of maximum likelihood, multiple imputation and similar response pattern imputation (identifying the most similar unit without missing information and replacing the missing part with the correspondent values of this unit) approaches, instead of the listwise deletion one, that is considered the most frequently utilized in their field.

To our knowledge, a standard and general strategy to compare different missing data techniques (MDTs) and to evaluate their performance have not been proposed yet. In this paper, to fill the gap, we propose a specific definition of efficiency that can be used to analyse how an algorithm operates on missing data. The efficiency of MDTs is computed considering the listwise deletion method as a reference. Specifically, we test the behaviour of the maximum likelihood method in [6] (*Complete EM*), the pairwise deletion and the mean imputation ones in a classification problem, using the Gaussian mixture model [10], with different percentage of missing information in the training set. We calculate the efficiency of an MDT, for different missing data percentages (MDP) where train data is MAR in two different contexts . In the first one, we use a complete (no missing values) test set, to evaluate how well the model is estimated. In the second one, we use test set with missing values, to evaluate how robust the estimated model is to missing data . We consider the latter as a more realistic scenario. The analysis is performed using synthetic data, Pima Indians Diabetes and IRIS data sets.

In section 2, we introduce the learning model used and missing data techniques that are evaluated in terms of efficiency, that is defined in section 3. Finally, the experiments and results are discussed in section 4.

2 Modeling Framework and Methods

2.1 Modeling Framework

The model used within this work is the Gaussian mixture model (GMM) that is used and explained in [10]. Define $\boldsymbol{x}$ as the d-dimensional input feature vector and the associated output, $y \in \{1, 2, \cdots, C\}$, of class labels, assuming C mutually exclusive classes. The joint input/output density is modeled as the Gaussian mixture.

$$p(y, \boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} P(y|k)p(\boldsymbol{x}|k)P(k) \tag{1}$$

$$p(\boldsymbol{x}|k) = \tag{2}$$
$$\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)$$

where K is the number of components, $p(\boldsymbol{x}|k)$ are the component Gaussians mixed with the non-negative priors $P(k)$, $\sum_{k=1}^{K} P(k) = 1$ and the class-cluster posteriors $P(y|k)$, $\sum_{y=1}^{C} P(y|k) = 1$. The k'th Gaussian component is described by the mean vector $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}_k$. $\boldsymbol{\theta}$ is the vector of all model parameters, i.e., $\boldsymbol{\theta} \equiv \{P(y|k), \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(k) : \forall k, y\}$. The joint input/output for each components is assumed to factorize, i.e., $p(y, \boldsymbol{x}|k) = P(y|k)p(\boldsymbol{x}|k)$.

The input density associated with Eq. (1) is given by

$$p(\boldsymbol{x}|\boldsymbol{\theta}_u) = \sum_{y=1}^{C} p(y, \boldsymbol{x}) = \sum_{k=1}^{K} p(\boldsymbol{x}|k)P(k),$$

where $\boldsymbol{\theta}_u \equiv \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(k) : \forall k, y\}$. Assuming a 0/1 loss function the optimal Bayes classification rule is $\widehat{y} = \max_y P(y|\boldsymbol{x})$ where[1]

$$P(y|\boldsymbol{x}) = \frac{p(y, \boldsymbol{x})}{p(\boldsymbol{x})} = \sum_{k=1}^{K} P(y|k)P(k|\boldsymbol{x})$$

with $P(k|\boldsymbol{x}) = p(\boldsymbol{x}|k)P(k)/p(\boldsymbol{x})$.

Define data set of labeled examples $\mathcal{D}_l = \{\boldsymbol{x}_n, y_n; n = 1, 2, \cdots, N_l\}$. The negative log-likelihood for the data sets, which are assumed to consist of independent examples, is given by

$$L = -\log p(\mathcal{D}|\boldsymbol{\theta}) = -\sum_{n\in\mathcal{D}_l} \log \sum_{k=1}^{K} P(y_n|k)p(\boldsymbol{x}_n|k)P(k)$$

The model parameters are estimated with an iterative modified EM algorithm [11]:

1. To initialize the mean ($\boldsymbol{\mu}_0$) and covariance ($\boldsymbol{\Sigma}_0$) matrices, all train data set is considered as one normal distribution. In the case of missing data, the calculations are done using only observed data and the $\boldsymbol{\Sigma}_0$ is regularized (see section 2.2). Then, since random points from the distribution can not be taken as cluster center points because of missing data, we draw L random samples using the $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$, and get rid of outliers. Instead of taking random center points from the remaining samples, we use KKZ method assuming the clusters will be distant from each other [12]. The KKZ method is as the following:
    - The first center point is taken as the sample having the largest L2 norm
    - Other center points are calculated as having the largest distance to the closest center points
2. Compute posterior component probability for all $n \in \mathcal{D}_l$:
$$p(k|y_n, \boldsymbol{x}_n) = \frac{P(y_n|k)p(\boldsymbol{x}_n|k)P(k)}{\sum_k P(y_n|k)p(\boldsymbol{x}_n|k)P(k)}. \tag{3}$$
3. For all k update means and covariance matrices
$$\boldsymbol{\mu}_k = \frac{\sum\limits_{n\in\mathcal{D}_l} \boldsymbol{x}_n P(k|y_n, \boldsymbol{x}_n)}{\sum\limits_{n\in\mathcal{D}_l} P(k|y_n, \boldsymbol{x}_n)}, \quad \boldsymbol{\Sigma}_k = \frac{\sum\limits_{n\in\mathcal{D}_l} \boldsymbol{S}_{kn} P(k|y_n, \boldsymbol{x}_n)}{\sum\limits_{n\in\mathcal{D}_l} P(k|y_n, \boldsymbol{x}_n)}$$
where $\boldsymbol{S}_{kn} = (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top$.
4. For all k update cluster priors and class cluster posteriors
$$P(k) = \frac{\sum\limits_{n\in\mathcal{D}_l} P(k|y_n, \boldsymbol{x}_n)}{N_l}, \quad P(y|k) = \frac{\sum\limits_{n\in\mathcal{D}_l} \delta_{y_n} P(k|y_n, \boldsymbol{x}_n)}{\sum\limits_{n\in\mathcal{D}_l} P(k|y_n, \boldsymbol{x}_n)}$$

2.2 Pairwise Deletion

In pairwise (PW) method, the only difference made on the model we use, is the update of posterior input density $p(\boldsymbol{x_n}|k)$, the mean vector μ_k and covariance matrix Σ_k. To update those, observed data for each variable or pair of variables are used. However, the estimated covariance matrix is unbiased and is not guaranteed to be positive semi definite. We regularize the covariance matrix by inflating the diagonal elements by the factor $(1 + h)$ as in Eq. 4 which is commonly used approach [13] given by

$$\boldsymbol{\Sigma}^{'} = \boldsymbol{\Sigma} + h\boldsymbol{I} \tag{4}$$

where $\boldsymbol{I}$ is the identity matrix and h is a regularization parameter. h is determined in the following way:

$$\boldsymbol{\Sigma}^{'} = \boldsymbol{\Sigma} + h\boldsymbol{I} = \boldsymbol{V}\boldsymbol{U}\boldsymbol{V^{-1}} + h\boldsymbol{V}\boldsymbol{V^{-1}} = \boldsymbol{V}(\boldsymbol{h}+\boldsymbol{U})\boldsymbol{V^{-1}} \tag{5}$$

where $\boldsymbol{V}\boldsymbol{U}\boldsymbol{V^{-1}}$ is the eigenvalue decomposition of the covariance matrix $\boldsymbol{\Sigma}$, where V is the square matrix whose ith column is the eigenvector qi of $\boldsymbol{\Sigma}$ and U is the diagonal matrix whose diagonal elements are the corresponding eigenvalues. Then, we choose h such that $(h + U) > 0$ to have nonnegative eigenvalues in regularized covariance matrix.

2.3 Mean Imputation

Mean imputation (MI) method is a replacement technique where a missing variable is replaced by the corresponding mean value [3]. The model we use is not effected in this method, since we have complete data after imputation. This method keeps all data, and is easy to implement. However, the variance estimates are lessened as more means are added.

2.4 Complete Expectation Maximization

This method is a maximum likelihood missing data technique that is proposed in [6]. EM is used both for the estimation of model components and for dealing with missing data. Posterior component probability, $p(k|y_n, \boldsymbol{x}_n)$ is again calculated as in Eq. 3, but only on observed dimensions. To update the mean vector, $E[x_n^m|x_n^o]$ is substituted for missing components of x_n and to update the covariance matrix, $E[x_n^m x_n^{m^T}|x_n^o]$ is substituted for outer product matrices containing missing components:

$$E[x_n^m|x_n^o] = \mu_n^m + \Sigma_n^{mo}\Sigma_n^{oo^{-1}}(x_n^o - \mu_n^o),$$

$$E[x_n^m x_n^{m^T}|x_n^o] = \Sigma_n^{mm} - \Sigma_n^{mo}\Sigma_n^{oo^{-1}}\Sigma_n^{mo^T} + E[x_n^m|x_n^o]E[x_n^m|x_n^o]^T.$$

[1] The dependence on $\boldsymbol{\theta}$ is omitted.

3 Efficiency Definition

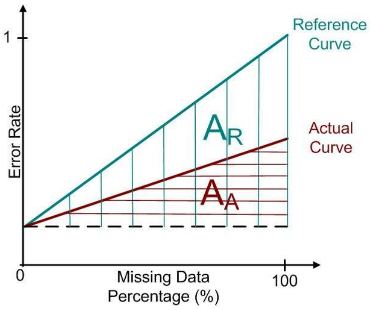

Fig. 1: The illustration for the efficiency calculation method used.

As data have more missing values, the resultant error rate(ER) gets higher due to lack of information. However, the resultant MDP-ER curve is different for different missing data techniques (MDTs). In this work, we use the curve for listwise deletion(LW) method as the reference. In other words, we calculate how efficient a technique makes use of data with missing values instead of simply ignoring them. As seen in Figure 1, the definition of efficiency (Eff) is obtained by calculating the area under the reference and actual curves (curves of MDTs investigated) as in Eq. 6. When the actual curve is the same as the reference curve, the efficiency is 0%, while it is 100%, when it is a straight line (i.e ER is not effected as MDP changes, the method is completely robust to MDP).

$$Eff_{\%} = \frac{A_R - A_A}{A_R} \times 100 \tag{6}$$

4 Experimental Evaluations

The experiments are carried out using MATLAB on synthetically generated data and two datasets from UCI archive, Iris and Pima-Indian-Diabetes [14]. MDP is determined randomly (MAR). The experiment is done such that not all values can be missing in one observation (if all data in all directions are missing it would be equal to deleting it, so reducing training data as in our reference method). We experiment how the misclassification rate (MR) changes with MDP and calculate the efficiency (Eq. 6) using those results for different MDP values. We experiment for two cases, where test dataset also has missing values (case 1) with same MDP, or it is complete (case 2). Case 2 investigates how well the model is estimated, while the case 1 how robust the estimated model is to missing data. We make 100 iterations for each experiment, while changing MDP between 0% and 70%.

4.1 Synthetic Dataset

The algorithm is tested on synthetic data. The multidimensional input data is generated on a Gaussian mixture model. The number of mixtures K, is 3. The difficulty of the problem is determined using the following SNR calculation:

Let ds_{kl} be the distance between μ_k and μ_l, eig_k be a vector consisting of eigenvalues of Σ_k and mean() be the arithmetic mean operator, then

$$\mathrm{SNR_{dB}} = 10\log\left(\frac{(\mathrm{mean}(\sum_{1\leq k\leq K, k\prec l\prec K} \mathrm{ds_{kl}}))^2}{\mathrm{mean}(\sum_{1\leq k\leq K}\mathrm{mean}(\mathrm{eig_k}))}\right)$$

We use SNR of 10 dB, for a 10 dimensional data. 150 observations are generated for both training and test sets. Figure 2 shows first three principal components plotted against each other for data used for this work.

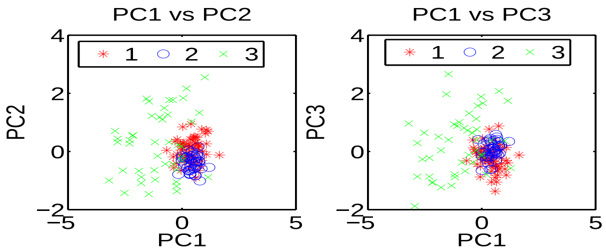

Fig. 2: The principal components(PCs) plot for the data generated with 3 different classes

The figure 3 shows the results for synthetic data generated. In case 1, CEM is the most efficient method, however PW is competitive to it. CEM gives an efficiency of 40%, even at MDP of 70%. MI is clearly the worst method in terms of efficiency. The efficiency of MI decreases as MDP gets higher, while CEM and PW give more stable efficiency results. In case 2, results are similar and still CEM is the best. While MI performs better compared to case 1, CEM is slightly worse.

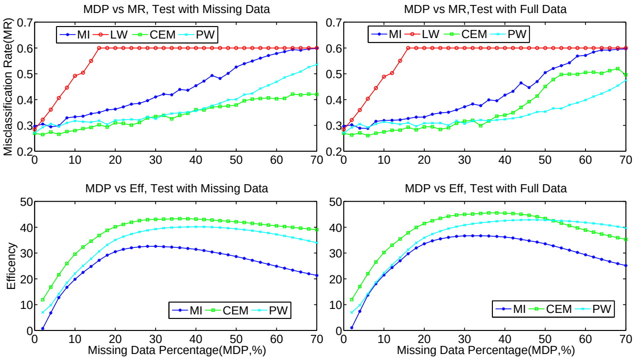

Fig. 3: The results for synthetically generated data **left**: Test set with full data **right**: Test set with missing data **top**: MR plot against MDP **bottom**: Eff plot against MR

4.2 Iris Dataset

Iris dataset is one of the most commonly used datasets in machine learning literature. It consists of 3 classes of 50 instances each referring to a type of iris plant with 4 attributes. One class is linearly separable from the others; the other two are not linearly separable from each other. We use 100 instances for train and 50 instances for test sets.

We show the results for this dataset in Figure 4. In case 1, CEM is still the most efficient method, MI and PW show a similar behaviour. CEM gives an efficiency of 70%, even at MDP of 70%. In case 2, PW is worse than MI and CEM is still the best method. Compared to case 1, the efficiency of CEM and PW is lower while the efficiency of MI is higher.

4.3 Pima Indians Diabetes Dataset

Pima Indians Diabetes Dataset contains 2 classes that are diabetes positive or negative with 7 attributes (age, pregnancy number etc.). We use 200 instances for train and 200 instances for test sets.

The results are shown in Figure 5. Both in case 1 and case 2, CEM overcomes other two methods, whereas PW and MI give similar re-

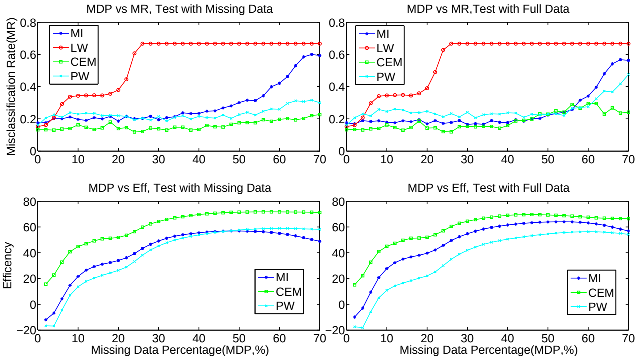

Fig. 4: The results for Iris dataset **left**: Test set with full data **right**: Test set with missing data **top**: MR plot against MDP **bottom**: Eff plot against MR

sults. The efficiency of CEM at MDP of 70% is around 20%, not as high as other datasets, but still giving the highest performance.

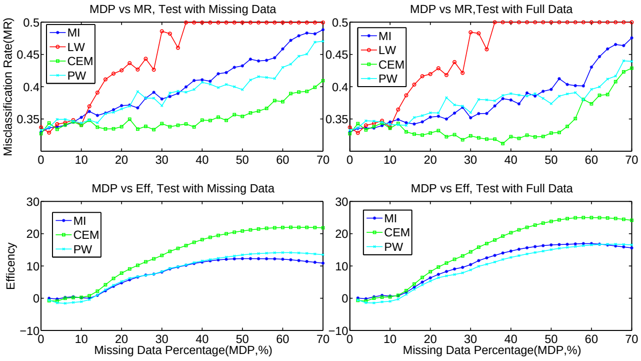

Fig. 5: The results for Pima Indians Diabetes dataset **left**: Test set with full data **right**: Test set with missing data **top**: MR plot against MDP **bottom**: Eff plot against MR

4.4 General Discussion

We observe that, generally CEM is the most efficient missing data method, while PW is worse than CEM, but still slightly better than MI especially for high MDP values. The results coincide with previous work [1, 15]. In [1], where they compare missing data methods using different criteria (the capability to minimize bias, maximize the use of available information and yield good estimates of uncertainty), ML methods are found to be the best. In [15], where they compare 6 different methods including PW and EM methods, the results again support ML approaches. Although CEM and PW perform well for both cases we experimented, we observe that they are more efficient to use when test data set also has missing values. MI is more efficient to use when we have a full test data set. Thus, MI is better at estimating the model, but the estimated model is not that robust to missing data in test set, and vice versa for CEM. Another observation made from the results is that CEM and PW give more stable results for higher MDP values, so it would be more trustworthy to use them in situations where MDP for test set is undetermined. Although MI turned out to be the least efficient approach, it would be still acceptable to use it especially for low MDP values, since it is very easy to implement and clearly computationally less expensive.

5 Conclusion

We proposed a new evaluation approach for MDTs where the efficiency of those are investigated using listwise deletion method as reference. We experimented on classification problems and calculated MR for different MDPs. We compared three different MDTs: pairwise deletion, mean imputation and complete EM. We used synthetic dataset, Iris dataset and Pima Indians Diabetes dataset. We used a Gaussian mixture model (GMM) trained with MAR data. We tested for missing or complete dataset. The results showed that CEM was the most efficient method in both cases while MI was the worst of the three. We observed that PW and CEM are more stable with respect to especially higher MDP values than MI. We also observed that MI performed better with complete test set, so was better at estimating the model, but the estimated model was not that robust to missing data in test set, vice versa for PW and CEM.

6 References

[1] P.D. Allison, *Missing Data*, Quantitative Applications in the Social Sciences. Sage Publications, 2001.

[2] P.L. Roth, “Missing data: A conceptual review for applied psychologists,” *Personnel Psychology*, vol. 47, no. 3, pp. 537–560, 1994.

[3] A.R.T. Donders, G.J.M.G. van der Heijden, T. Stijnen, and K.G.M. Moons, “Review: a gentle introduction to imputation of missing values,” *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[4] D.B. Rubin, *Multiple Imputations for nonresponse in surveys*, New York: Wiley, 1987.

[5] A.P. Dempster, N.M. Laird, and D.B.Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[6] Z. Ghahramani and M.I.Jordan, “Supervised learning from incomplete data via an EM approach,” in *Advances in Neural Information Processing Systems 6*. 1994, pp. 120–127, Morgan Kaufmann.

[7] D.B. Rubin, “Inference and missing data,” *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[8] J.L. Schafer and J.W. Graham, “Missing data: Our view of the state of the art,” *Psychological methods*, vol. 7, no. 2, pp. 147–177, 2002.

[9] I. Myrtveit, E. Stensrud, and U.H. Olsson, “Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods,” *IEEE Transactions on Software Engineering*, vol. 27, no. 11, pp. 999–1013, 2001.

[10] J. Larsen, A. Szymkowiak, and L.K. Hansen, “Probabilistic hierarchical clustering with labeled and unlabeled data,” *International Journal of Knowledge Based Intelligent Engineering Systems*, vol. 6, no. 1, pp. 56–63, 2002.

[11] L.K. Hansen, S. Sigurdsson, T. Kolenda, F.A. Nielsen, U. Kjems, and J. Larsen, “Modeling text with generalizable Gaussian mixtures,” in *ICASSP’00. Proceedings*. IEEE, 2000, vol. 6, pp. 3494–3497.

[12] T. Su and J.G. Dy, “In search of deterministic methods for initializing K-means and Gaussian mixture clustering,” *Intelligent Data Analysis*, vol. 11, no. 4, pp. 319–338, 2007.

[13] T. Schneider, “Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values,” *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[14] UC Irvine Machine Learning Repository, ,” ‘http://archive.ics.uci.edu/ml/.

[15] D.A. Newman, “Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques,” *Organizational Research Methods*, vol. 6, no. 3, pp. 328, 2003.