Mixture of Gaussians for Distance Estimation with Missing Data

Emil Eirola[a,*], Amaury Lendasse[a,b,c,d], Vincent Vandewalle[e,f], Christophe Biernacki[d,f]

[a] *Department of Information and Computer Science, Aalto University, FI–00076 Aalto, Finland*
[b] *IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain*
[c] *Computational Intelligence Group, Computer Science Faculty, University of the Basque Country, Paseo Manuel Lardizabal 1, Donostia/San Sebastián, Spain*
[d] *Laboratoire P. Painlevé, UMR 8524 CNRS Université Lille I, Bât M2, Cité Scientifique, F–59655 Villeneuve d'Ascq Cedex, France*
[e] *EA 2694, Université Lille 2, 1 place de Verdun, 59045 Lille Cedex, France*
[f] *INRIA Lille – Nord Europe Parc Scientifique de la Haute Borne Park Plazza – Bât A – 40 avenue Halley 59650 Villeneuve d'Ascq, France*

Abstract

The majority of all commonly used machine learning methods can not be applied directly to data sets with missing values. However, most such methods only depend on the relative differences between samples instead of their particular values, and thus one useful approach is to directly estimate the pairwise distances between all samples in the data set. This is accomplished by fitting a Gaussian mixture model to the data, and using it to derive estimates for the distances. Experimental simulations confirm that the proposed method provides accurate estimates compared to alternative methods for estimating distances.

Keywords: missing data, distance estimation, mixture model

1. Introduction

In many real world machine learning tasks, data sets with missing values (also referred to as incomplete data) are all too common to be easily ignored.

*Corresponding author
Email address: emil.eirola@aalto.fi (Emil Eirola)

Values could be missing for a variety of reasons depending on the source of the data, including measurement error, device malfunction, operator failure, etc. However, many modelling approaches start with the assumption of a data set with a certain number of samples, and a fixed set of measurements for each sample. Such methods can not be applied directly if some measurements are missing. Simply discarding the samples or variables which have missing components often means throwing out a large part of data that could be useful for the model. It is relevant to look for better ways of dealing with missing values in such scenarios.

In this paper, the particular problem of estimating distances between samples in a data set with missing values is studied. Being able to appropriately estimate distances between samples, or between samples and prototypes, has numerous applications. It directly enables the use of several powerful statistical and machine learning methods which are based only on distances and not the direct values, e.g.: nearest neighbours (k-NN), support vector machines (SVM), or radial basis function (RBF) neural networks [1].

There are alternative paradigms for dealing with missing data which could be used in conjunction with a machine learning method. *Conditional mean imputation*, which is optimal in terms of minimising the mean squared error of the imputed values, suffers from leading to biased derived statistics of the data. For instance, estimates of variance or distances are negatively biased. *Conditional random draw* is more appropriate for generating a representative example of a fully imputed data set, but has too much variability in estimates of any single values, or distances between particular samples, to be accurate. *Multiple imputation* (drawing several representative imputations of the data, analysing each set separately, and combining the results) results in unbiased and accurate estimates after a sufficiently high number of draws. In the context of machine learning, repeating the analysis several times is however impractical as training and analyzing a sophisticated model tends to be computationally expensive.

Finite mixture models have proven to be a versatile and powerful modelling tool in a wide variety of applications. Particularly mixture models of Gaussians have been studied extensively to describe the distributions of data sets. The general approach to estimating the model from data is maximum likelihood (ML) estimation by the EM algorithm [2]. This has been extended to estimating Gaussian mixture models for data sets with missing values [3, 4].

Using a Gaussian mixture model is appropriate for estimating pairwise

distance between samples, as it 1) can be optimised efficiently even in the presence of missing values, 2) allows one to derive estimates of pairwise distances, 3) is flexible enough to cover any distribution of samples, and 4) is sufficiently sophisticated to provide non-linear imputation.

An important consideration when dealing with missing data is the missing-data mechanism. We assume that a missing value represents a value which is defined and exists, but for an unspecified reason is not known. Following the conventions of [5], the assumption here is that data are Missing-at-Random (MAR):

$$P(M|x_{\text{obs}}, x_{\text{mis}}) = P(M|x_{\text{obs}}),$$

i.e., the event of a measurement being missing is independent from the value it would take, conditional on the observed data. The stronger assumption of Missing-Completely-at-Random (MCAR) is not necessary, as MAR is an ignorable missing-data mechanism in the sense that maximum likelihood estimation still provides a consistent estimator [5].

The paper is organised as follows. Section 2 reviews the EM algorithm for mixtures of Gaussians, and introduces the extension to missing data. Section 3 presents the estimation of pairwise distances. Section 4 includes comparison experiments on simulations of data with missing values.

2. EM for mixture of Gaussians with missing data

2.1. The standard EM algorithm

Before studying the case with missing data, we present the conventional EM algorithm for fitting a mixture of Gaussians [1, Section 9.2]. Given data $\mathbf{X}$ consisting of a set of N observations $\{\boldsymbol{x}_i\}_{i=1}^N$, we wish to model the data using a mixture of K Gaussian distributions. The log-likelihood is given by

$$\log \mathcal{L}(\theta) = \log p(\mathbf{X} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_i \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right), \quad (1)$$

where $\mathcal{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the probability density function of the multivariate normal distribution. The log-likelihood can be maximised by applying the EM-algorithm. Initialisation consists in choosing values for the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$, and mixing coefficients π_k for each component k. The E-step is to evaluate the probabilities t_{ik} using the current parameter values:

$$t_{ik} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_i \,|\, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (2)$$

In the M-step, the parameters are re-estimated with the updated probabilities:

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{i=1}^{N} t_{ik}\boldsymbol{x}_i, \quad \boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{i=1}^{N} t_{ik}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T, \quad \pi_k = \frac{N_k}{N}, \quad (3)$$

where $N_k = \sum_{i=1}^{N} t_{ik}$. The E and M-steps are alternated repeatedly until convergence is observed in the log-likelihood or parameter values.

2.2. Missing data extension of the EM algorithm

The standard EM algorithm for fitting Gaussian mixture models has been extended to handle data with missing values [3, 4]. The input data $\mathbf{X}$ is now a set of observations $\{\boldsymbol{x}_i\}_{i=1}^{N}$ such that for each sample there is an index set $O_i \subseteq \{1, \ldots, d\}$ enumerating the observed samples. The indices in the complement set M_i correspond to missing values in the data sample $\boldsymbol{x}_i$. In the case with missing values, the log-likelihood can be written as

$$\log \mathcal{L}(\theta) = \log p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}_i^{O_i} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right) \quad (4)$$

where as a shorthand of notation, $\mathcal{N}(\boldsymbol{x}_i^{O_i} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is also used for the *marginal* multivariate normal distribution probability density of the observed values of $\boldsymbol{x}_i$.

In the EM algorithm, in order to account for the missing data, some additional expectations need to be computed in the E-step. These include the conditional expectations of the missing components of a sample ($\tilde{\boldsymbol{\mu}}_{ik}^{M_i}$) with respect to each Gaussian component k, and their conditional covariance matrices ($\tilde{\boldsymbol{\Sigma}}_{ik}^{MM_i}$). For convenience, we also define corresponding imputed data vectors $\tilde{\boldsymbol{x}}_{ik}$ and full covariance matrices $\tilde{\boldsymbol{\Sigma}}_{ik}$ which are padded with zeros for the known components. Then the E-step is:

$$t_{ik} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i^{O_i} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_i^{O_i} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (5)$$

$$\tilde{\boldsymbol{\mu}}_{ik}^{M_i} = \boldsymbol{\mu}_k^{M_i} + \boldsymbol{\Sigma}_k^{MO_i}(\boldsymbol{\Sigma}_k^{OO_i})^{-1}(\boldsymbol{x}_i^{O_i} - \boldsymbol{\mu}_k^{O_i}), \quad \tilde{\boldsymbol{x}}_{ik} = \begin{pmatrix} \boldsymbol{x}_i^{O_i} \\ \tilde{\boldsymbol{\mu}}_{ik}^{M_i} \end{pmatrix}, \quad (6)$$

$$\tilde{\boldsymbol{\Sigma}}_{ik}^{MM_i} = \boldsymbol{\Sigma}_k^{MM_i} - \boldsymbol{\Sigma}_k^{MO_i}(\boldsymbol{\Sigma}_k^{OO_i})^{-1}\boldsymbol{\Sigma}_k^{OM_i}, \quad \tilde{\boldsymbol{\Sigma}}_{ik} = \begin{pmatrix} \mathbf{0}^{OO_i} & \mathbf{0}^{OM_i} \\ \mathbf{0}^{MO_i} & \tilde{\boldsymbol{\Sigma}}_{ik}^{MM_i} \end{pmatrix} \quad (7)$$

The notation $\boldsymbol{\mu}_k^{M_i}$ refers to using only the elements from the vector $\boldsymbol{\mu}_k$ specified by the index set M_i, and similarly for $\boldsymbol{x}_i^{O_i}$, etc. For matrices, $\boldsymbol{\Sigma}_k^{MO_i}$ refers to elements in the rows specified by M_i and columns by O_i, and so on. The expressions for the parameters in equations (6) and (7) originate from the observation that the conditional distribution of the missing components also follows a multivariate normal distribution, with these parameters [6, Thm. 2.5.1].

The M-step remains functionally equivalent, the only changes being that the component means are estimated from the imputed data vectors and the covariance matrix estimate requires an additional term to include the covariances concerning the imputed values.

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{N} t_{ik} \tilde{\boldsymbol{x}}_{ik}, \;\; \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{i=1}^{N} t_{ik} \left[(\tilde{\boldsymbol{x}}_{ik} - \boldsymbol{\mu}_k)(\tilde{\boldsymbol{x}}_{ik} - \boldsymbol{\mu}_k)^T + \tilde{\boldsymbol{\Sigma}}_{ik} \right], \;\; \pi_k = \frac{N_k}{N} \tag{8}$$

2.3. High-dimensional data

As the number of free parameters grows with the square of the data dimension, in high-dimensional cases it is often not possible to fit a conventional Gaussian mixture model, or even a model with a single Gaussian component. A version of Gaussian mixture models, *high-dimensional data clustering (HDDC)*, has been proposed for this scenario [7] where in the M-step, the covariances matrices are replaced by a reduced representation. Applying this idea to the case of missing data is possible, by modifying the covariance matrices of each component after calculating them in the M-step. However, the computational gains obtained from having a reduced representation are not available, as the complete covariance matrices are still needed in order to calculate the conditional parameters in the following E-step.

From the different variants presented in [7], the experiments in this paper use the $[a_{ij}b_iQ_id_i]$ model. The number of significant eigenvalues is selected separately for each component by the scree test, where the dimension is selected when the subsequent eigenvalues have a difference smaller than a specified threshold (0.001 of the trace of the covariance matrix).

2.4. Initialisation

In our implementation of the EM algorithm with missing data, the means $\boldsymbol{\mu}_k$ are initialised by a random selection of complete samples from the data, if available. When there are insufficient complete samples, some components

are initialised by incomplete samples, with the missing components imputed by the sample mean. The covariances $\mathbf{\Sigma}_k$ are initialised with the sample covariance of the data (ignoring samples with missing values). Alternatively, the covariances can be initialised as diagonal matrices, using only the sample variance of each variable.

2.5. Efficiency

The sweep operator can be used to efficiently perform the necessary calculation involving the inverse of the submatrix of the covariance matrix in the E-step; sweeping over the observed values for each sample to find the parameters of the conditional distribution [5]. Alternatively, an equivalent process is to calculate the inverse of each covariance matrix, and reverse sweep only over the missing values for each sample. This is more efficient for larger data sets with few missing samples.

2.6. Model selection

The number of components is selected according to the Akaike information criterion [8] with the small sample (second-order) bias adjustment [9]. Using the corrected version is crucial, as the number of parameters grows relatively large when increasing the number of components. The corrected Akaike information criterion is a function of the log-likelihood:

$$\mathrm{AIC}_C = -2 \log \mathcal{L}(\theta) + 2P + \frac{2P(P+1)}{N - P - 1} \tag{9}$$

where P is the number of free parameters. $P = Kd + K - 1 + \frac{1}{2}Kd(d+1)$ in the case of full, separate, covariance matrices for each of the K components. With high-dimensional data sets, the number of parameters quickly tends to become larger than the number of available samples when increasing the number of components, and the criterion would not be valid anymore. This effect can be mitigated by imposing restrictions on the structure of the covariance matrices, but this would also make the model less powerful.

3. Distance estimation with missing data

The indended application of the mixture of Gaussians model is to use it for distance estimation. The problem of estimating distances between samples with missing data is non-trivial, since even perfect imputation (by the

conditional expectation) results in biased estimates for the distance. Using additional knowledge about the distribution of the data leads to more accurate estimates.

In the following, we focus on calculating the expectation of the *squared* Euclidean (ℓ^2) distance. Estimating the ℓ^2-norm itself could be feasible, but due to the square-root, the expressions do not simplify and separate as cleanly. Another motivation for directly estimating the squared distance is that many methods for further processing of the distance matrix actually only make use of the squared distances (e.g., RBF and SVM), while others only consider the ranking of the distances (nearest neighbours).

Given two samples $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^d$ with components $x_{i,l}$, $x_{j,l}$ $(1 \leq l \leq d)$, which may contain missing values, denote by $M_i \subseteq \{1, \ldots, d\}$ the set of indices of the missing components for each sample $\boldsymbol{x}_i$. Partition the index set into four parts based on the missing components, and the expression for the squared distance $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$ can be split accordingly:

$$\begin{aligned}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = \sum_{l=1}^{d}(x_{i,l} - x_{j,l})^2 = \sum_{l \notin M_i \cup M_j}(x_{i,l} - x_{j,l})^2 + \sum_{l \in M_j \setminus M_i}(x_{i,l} - x_{j,l})^2 \\ + \sum_{l \in M_i \setminus M_j}(x_{i,l} - x_{j,l})^2 + \sum_{l \in M_i \cap M_j}(x_{i,l} - x_{j,l})^2\end{aligned}$$

The missing values can be modelled as random variables, $X_{i,l}$, $l \in M_i$. Taking the expectation of the above expression, by the linearity of expectation:

$$\begin{aligned}\mathrm{E}\left[\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right] = \sum_{l \notin M_i \cup M_j}(x_{i,l} - x_{j,l})^2 + \sum_{l \in M_j \setminus M_i}\left((x_{i,l} - \mathrm{E}[X_{j,l}])^2 + \mathrm{Var}[X_{j,l}]\right) \\ + \sum_{l \in M_i \setminus M_j}\left((\mathrm{E}[X_{i,l}] - x_{j,l})^2 + \mathrm{Var}[X_{i,l}]\right) \\ + \sum_{l \in M_i \cap M_j}\left((\mathrm{E}[X_{i,l}] - \mathrm{E}[X_{j,l}])^2 + \mathrm{Var}[X_{i,l}] + \mathrm{Var}[X_{j,l}]\right)\end{aligned}$$

In the final summation, it is necessary to consider $X_{i,l}$ and $X_{j,l}$ to be uncorrelated, given the known values of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. This assumption is not restrictive, and follows directly from the common approach that samples are independent draws from an unknown multivariate distribution.

It thus suffices to find the expectation and variance of each random variable separately. If the original samples $\boldsymbol{x}_i$ are thought to originate as independent draws from a multivariate distribution, the distributions of the

random variables $X_{i,l}$ can be found as the conditional distribution when conditioning their joint distribution on the observed values. Then finding the expected squared distance between two samples reduces to finding the (conditional on the observed values) expectation and variance of each missing component separately. Define $\tilde{\boldsymbol{x}}_i$ to be an imputed version of $\boldsymbol{x}_i$ where each missing value has been replaced by its conditional mean, and define $\sigma^2_{i,l}$ as the corresponding conditional variance:

$$\tilde{x}_{i,l} = \begin{cases} \mathrm{E}[X_{i,l} \,|\, \boldsymbol{x}_i^{O_i}] & \text{if } l \in M_i, \\ x_{i,l} & \text{otherwise} \end{cases} \quad \sigma^2_{i,l} = \begin{cases} \mathrm{Var}[X_{i,l} \,|\, \boldsymbol{x}_i^{O_i}] & \text{if } l \in M_i, \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

With these notations, the expectation of the squared distance can conveniently be expressed as:

$$\mathrm{E}\left[\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right] = \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j\|^2 + s_i^2 + s_j^2, \quad \text{where} \quad s_i^2 = \sum_{l \in M_i} \sigma^2_{i,l} \tag{11}$$

This form of the expression particularly emphasises how the uncertainty of the missing values is accounted for. The first term – the distance between imputed samples – already provides an estimate of the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, but including the variances of each imputed component is the deciding factor.

The conditional means and covariances can be calculated using the Gaussian mixture model. These are calculated separately for each component in the M-step, and it only remains to determine the overall conditional mean and covariance matrix. These are found weighted by the memberships as follows:

$$\tilde{\boldsymbol{x}}_i = \sum_{k=1}^{K} t_{ik} \tilde{\boldsymbol{x}}_{ik}, \qquad \tilde{\boldsymbol{\Sigma}}_i = \sum_{k=1}^{K} t_{ik} \left(\tilde{\boldsymbol{\Sigma}}_{ik} + \tilde{\boldsymbol{x}}_{ik} \tilde{\boldsymbol{x}}_{ik}^T \right) - \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^T \tag{12}$$

The expression for the covariance is found by direct calculation of the second moments. In order to estimate pairwise distances, the conditional variances $\sigma^2_{i,l} = \tilde{\boldsymbol{\Sigma}}_{i,ll}$ can be extracted from the diagonal of the conditional covariance matrix.

In summary, the complete procedure for estimating distances consists of the following steps:

1. For integers K from 1 to a chosen maximum:
2. Fit a Gaussian mixture model of K components by the EM algorithm
3. Calculate AIC_C
4. Choose the model which minimises AIC_C
5. Apply the chosen model to estimate distances

4. Experiments

To study the effectiveness of the proposed approach, some simulated experiments are conducted to compare the algorithm to alternative methods on several data sets with two different performance criteria. Starting with a complete data set, values are removed at random with a fixed probability. As the true distances between samples are known, the methods can then be compared on how well they estimate the distances after values have been removed.

4.1. Data

Several different data sets are used for the experiments, and they are listed in Table 1. As the problem of pairwise distance estimation is unsupervised, any regression outputs or class labels for the samples are ignored.

To make distances meaningful, the variables in each data set are standardised to zero mean and unit variance before values are removed. This standardisation is conducted only in order to have comparable error rates between repeated experiments, and the methods used do not depend on the variables being standardised.

4.2. Methods

The Gaussian mixture model approach is compared to two other methods for estimating distances:

PDS The Partial Distance Strategy [11]. Calculate the sum of squared differences of the mutually known components and scale to the missing components:

$$\hat{d}_{ij}^2 = \frac{d}{d - |M_i \cup M_j|} \sum_{l \notin M_i \cup M_j} (x_{i,l} - x_{j,l})^2 \tag{13}$$

Table 1: Data sets used for the experiments, with the number of samples (N), number of variables (D), and source.

Name	N	D	Source
Computer Hardware	209	6	[10]
Glass Identification	214	9	[10]
Housing	506	9	[10]
Iris	150	4	[10]
Pima Indians Diabetes	768	8	[10]
Servo	167	4	[10]
Stocks	950	9	†
Statlog (Vehicle Silhouettes)	846	18	[10]
Automobile (with some categorical variables discarded)	159	15	[10]
Breast Cancer Wisconsin (Diagnostic)	569	30	[10]
Breast Cancer Wisconsin (Prognostic)	194	32	[10]
Breast Tissue	106	9	[10]
Connectionist Bench (Sonar, Mines vs. Rocks)	208	60	[10]
Ecoli	336	7	[10]
Ionosphere	351	33	[10]
Parkinsons	195	22	[10]
Spambase (preprocessed by taking the logarithm of each value)	4601	57	[10]
SPECTF Heart	267	44	[10]
Wine	178	13	[10]
Finance	500	41	†
Image Segmentation	2310	18	[10]
Tecator	240	100	‡

† Available on our website
‡ http://www.dm.unibo.it/~simoncin/tecator

For samples which have no common known components, the method is not defined. For such pairs, the average of the pairwise distances which were possible to estimate is returned instead.

ICkNNI Incomplete-case k-NN imputation [12]. An improvement of complete-case k-NN imputation, here any sample with a valid pattern of missing values is viable nearest neighbour. In accordance to the suggestions in [12], up to $k = 5$ neighbours are considered. The imputation fails whenever there are no samples with such valid patterns. For these cases, the missing value is imputed by the sample mean for that variable.

In addition, the mixture model is compared to estimating the distribution by a single Gaussian. As another alternative, the distances are estimated after using the mixture model for imputation by the conditional mean – equivalent to discarding the variance terms of Eq. (11). Using a single Gaussian

for imputation by the conditional mean is an interesting special case, as it is equivalent to a least-squares linear regression.

For data sets where the Gaussian mixture model can not be estimated with full covariance matrices, the HDDC approach is used to regularise the covariance matrices.

4.3. Performance criteria

The methods are evaluated by three different performance criteria. First, the methods are compared by the root mean squared error (RMSE) of all the estimated pairwise distances in the data set,

$$C_1 = \Big(\frac{1}{\lambda}\sum_{i>j}(\hat{d}_{ij} - d_{ij})^2\Big)^{1/2}$$

where, d_{ij} is the true Euclidean distance between samples i and j calculated without any missing data, and $\hat{d}_{ij}$ is the estimate of the distance provided by each method after removing data. The scaling factor λ is determined so that the average is calculated only over those distances which are estimates, discarding all the cases where the distance can be calculated exactly because neither sample has any missing components: $\lambda = MN - \frac{M(M+1)}{2}$.

A common application for pairwise distances is a nearest neighbour search, and thus we also consider the average (true) distance to the predicted nearest neighbour,

$$C_2 = \frac{1}{N}\sum_{i=1}^{N} d_{i,\mathrm{NN}(i)}, \quad \text{where} \quad \mathrm{NN}(i) = \arg\min_{j\neq i} \hat{d}_{ij}$$

Here, $\mathrm{NN}(i)$ is the nearest neighbour of the ith sample as estimated by the method, and $d_{i,\mathrm{NN}(i)}$ is the true Euclidean distance between the samples as calculated without any missing data. The criterion measures how well the method can identify samples which actually are close in the real data.

Mean relative error of all pairwise distances:

$$C_4 = \frac{1}{\lambda}\sum_{i>j}\frac{\left|\hat{d}_{ij} - d_{ij}\right|}{d_{ij}}$$

This criterion gives more weight to small distances, and is also an ℓ^1-type error. (As some data sets contain duplicate samples, sample pairs i, j where $d_{ij} = 0$ are ignored when calculating this criterion.)

4.4. Procedure

Values are removed from the data set independently at a fixed probability p. For each value of p, 100 repetitions are conducted for the Monte Carlo simulation, and simulations are run for value of p of 0.05 (low ratio of missing values) and 0.20 (relatively high ratio of missing values). The EM algorithm is run for 200 iterations, and repeated for a total of 5 times for each number of components. Runs are aborted if a covariance matrix becomes too poorly conditioned (condition number over 10^{12}). The best solution in terms of log-likelihood is selected, and the number of components is selected by the AIC_C criterion.

Having 100 repetitions of the same set-up enables the use of statistical significance testing to assess the difference between the mean errors of different methods. The testing is conducted as a two-tailed paired t-test, with a significance level of $\alpha = 0.01$. Comparing the performance of the best method to that of every other method results in a multiple hypothesis scenario, and thus the Bonferroni correction [13] is used to control the error rate.

4.5. Results and Discussion

The average RMSE values for the methods are presented in Table 2. The data sets are grouped into three categories as follows, depending on which mixture model is applicable. The first category includes data for which the EM algorithm converges appropriately with at least two components, and here the number of components is selected by the AIC_C criterion. For data in the second group, the EM algorithm either did not converge with two components, or the AIC_C indicated that a single component is clearly sufficient. In the the third group are those data sets for which the EM algorithm would not converge even with one component, and we apply the HDDC instead.

The most immediate observation is that using a single Gaussian or the mixture model tends to give the most accurate results for most data sets. For certain data (computer hardware, ecoli, image segmentation), it appears that ICkNNI leads to better estimates.

For many data sets in the first category (Stocks, Iris, Statlog), the mixture model provides a clear advantage over a using a single Gaussian. In other cases the differences are not significant, meaning that it may be sufficient to model the data with one Gaussian for this purpose, and indeed several runs resulted in mixture models of consisting of a single Gaussian, as evidenced by the mean K values of less than two.

For most data sets in all three groups, it can also be seen that including the variance terms of equation (11) tends to lead to an improvement in the accuracy compared to only conducting imputation.

Table 3 shows the corresponding performances in terms of the true distance to the predicted nearest neighbour. Although this criterion is much more emphasised on the accuracy of small distances, it reveals similar behaviour between the methods on all data sets. Table 4 shows the mean relative error of each method. The relative accuracies are otherwise similar, except that in nearly all cases, using the mixture model for imputation is more accurate in terms of average relative error than directly estimating distances.

In terms of computational resources, the mixture model is more expensive than PDS or ICkNNI, but correspondingly delivers more accurate results. As such, the mixture model can be recommended in any situation where the computational task is feasible. Comparing imputation by the mixture model to using the model for directly estimating distances, there is no difference in computational cost, but the accuracy is higher with directly estimated distances

5. Conclusions

The problem of estimating distances in a data set with missing values can be reduced to finding the conditional means and variances separately for each missing value. Having a Gaussian mixture model of the distribution of the data enables these quantities to be estimated. In order to fit the mixture model, certain extensions to the standard EM algorithm are presented in Section 2.

The combination of these ideas provides for a method to estimate distances, and the simulations in Section 4 show that the method is competitive, if not better, than alternative methods in terms of accuracy.

For future work, it remains to investigate the most effective ways to extend the method to high-dimensional cases where the number of parameters would exceed the number of samples. Furthermore, it will be interesting to study the influence of the distance estimation when used with machine learning methods such as SVM and RBF neural networks.

References

[1] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[2] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B (Methodological) 39 (1977) pp. 1–38.

[3] Z. Ghahramani, M. Jordan, Learning From Incomplete Data, Technical Report, Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab, 1995.

[4] L. Hunt, M. Jorgensen, Mixture model clustering for mixed data with missing information, Computational Statistics & Data Analysis 41 (2003) 429–440.

[5] R. J. A. Little, D. B. Rubin, Statistical Analysis with Missing Data, Wiley-Interscience, second edition, 2002.

[6] T. W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley-Interscience, New York, third edition, 2003.

[7] C. Bouveyron, S. Girard, C. Schmid, High-dimensional data clustering, Computational Statistics & Data Analysis 52 (2007) 502–519.

[8] H. Akaike, A new look at the statistical model identification, Automatic Control, IEEE Transactions on 19 (1974) 716–723.

[9] C. M. Hurvich, C.-L. Tsai, Regression and time series model selection in small samples, Biometrika 76 (1989) 297–307.

[10] A. Asuncion, D. J. Newman, UCI machine learning repository, 2011. University of California, Irvine, School of Information and Computer Sciences.

[11] J. K. Dixon, Pattern recognition with partly missing data, Systems, Man and Cybernetics, IEEE Transactions on 9 (1979) 617–621.

[12] J. Van Hulse, T. M. Khoshgoftaar, Incomplete-case nearest neighbor imputation in software measurement data, Information Sciences (2011). In Press, Corrected Proof.

[13] J. Shaffer, Multiple hypothesis testing, Annual review of psychology 46 (1995) 561–584.

Table 2: 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, $\alpha = 0.05$) 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.

		PDS	ICkNNI	Single Gaussian		Mixture model		Mean K
Computer Hardware	5%	0.689	**0.430**	0.477	(0.461)	0.478	(0.463)	3.70
$N = 209$, $D = 6$	20%	1.109	**0.707**	0.732	(0.730)	**0.727**	(**0.719**)	3.42
Glass Identification	5%	0.528	0.339	0.228	(**0.224**)	**0.219**	(**0.217**)	1.81
$N = 214$, $D = 9$	20%	0.964	0.682	**0.522**	(**0.516**)	**0.514**	(**0.517**)	1.75
Housing	5%	0.506	**0.314**	0.327	(0.335)	0.324	(0.331)	1.33
$N = 506$, $D = 13$	20%	0.997	0.668	**0.594**	(0.647)	**0.594**	(0.642)	1.27
Iris	5%	0.447	0.256	0.253	(0.258)	**0.227**	(0.233)	2.55
$N = 150$, $D = 4$	20%	0.669	0.378	0.369	(0.385)	**0.330**	(0.349)	2.46
Pima Indians	5%	0.537	0.408	**0.384**	(0.411)	**0.382**	(0.406)	2.20
$N = 768$, $D = 8$	20%	0.952	0.691	**0.599**	(0.712)	**0.599**	(0.697)	2.20
Servo	5%	0.567	0.452	**0.379**	(0.418)	**0.379**	(0.417)	1.75
$N = 167$, $D = 4$	20%	0.868	0.637	**0.513**	(0.618)	0.515	(0.617)	1.68
Stocks	5%	0.348	**0.061**	0.184	(0.184)	0.085	(0.085)	7.45
$N = 950$, $D = 9$	20%	0.651	0.180	0.345	(0.353)	**0.158**	(0.161)	7.58
Statlog	5%	0.344	0.198	0.153	(0.158)	**0.137**	(0.141)	2.89
$N = 846$, $D = 18$	20%	0.727	0.567	**0.310**	(0.332)	**0.296**	(0.314)	2.69

		PDS	ICkNNI	Single Gaussian	
Automobile	5%	0.359	0.257	**0.228**	(**0.229**)
$N = 159$, $D = 15$	20%	0.750	0.732	**0.458**	(0.488)
Breast Cance (Diag.)	5%	0.355	0.258	**0.140**	(0.141)
$N = 569$, $D = 30$	20%	0.802	1.054	**0.342**	(0.351)
Breast Cance (Prog.)	5%	0.349	0.294	**0.161**	(0.163)
$N = 194$, $D = 32$	20%	0.787	1.071	**0.387**	(0.399)
Breast Tissue	5%	0.430	0.255	**0.215**	(**0.213**)
$N = 106$, $D = 9$	20%	0.787	0.571	**0.421**	(**0.422**)
Connectionist Bench	5%	0.343	0.372	**0.192**	(**0.192**)
$N = 208$, $D = 60$	20%	0.793	1.356	0.532	(**0.515**)
Ecoli	5%	0.745	**0.458**	**0.465**	(**0.466**)
$N = 336$, $D = 7$	20%	1.254	**0.763**	**0.771**	(0.799)
Ionosphere	5%	0.275	0.253	0.241	(**0.221**)
$N = 351$, $D = 33$	20%	0.623	1.089	0.586	(**0.516**)
Parkinsons	5%	0.337	0.258	**0.181**	(0.186)
$N = 195$, $D = 22$	20%	0.727	0.961	**0.369**	(0.392)
Spambase	5%	1.004	0.721	**0.682**	(0.715)
$N = 4601$, $D = 57$	20%	2.209	1.839	**1.369**	(1.614)
SPECTF Heart	5%	0.332	0.330	**0.202**	(0.206)
$N = 267$, $D = 44$	20%	0.766	1.218	**0.492**	(0.497)
Wine	5%	0.369	0.274	**0.255**	(0.271)
$N = 178$, $D = 13$	20%	0.740	0.629	**0.474**	(0.546)

		PDS	ICkNNI	HDDC mixture		Mean K
Finance	5%	0.717	0.521	**0.491**	(**0.491**)	1.00
$N = 500$, $D = 41$	20%	1.539	1.457	**1.142**	(1.152)	1.00
Image Segmentation	5%	0.568	**0.338**	0.456	(0.450)	3.30
$N = 2310$, $D = 18$	20%	1.183	**0.756**	0.992	(0.960)	3.25
Tecator	5%	0.052	1.181	0.004	(**0.004**)	1.00
$N = 240$, $D = 100$	20%	0.121	2.370	0.032	(**0.031**)	1.00

Table 3: Average of the mean distance to the estimated nearest neighbour. The best result for each row is underlined, and any results which are not statistically significantly different (two-tailed paired t-test, $\alpha = 0.05$) 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.

		PDS	ICkNNI	Single Gaussian		Mixture model		Mean K
Computer Hardware	5%	0.687	0.521	0.514	(0.523)	**0.504**	(0.524)	3.70
$N = 209$, $D = 6$	20%	1.237	0.848	0.809	(0.848)	**0.776**	(0.844)	3.42
Glass Identification	5%	1.071	0.917	0.882	(0.889)	**0.880**	(0.887)	1.81
$N = 214$, $D = 9$	20%	1.713	1.243	1.101	(1.129)	**1.092**	(1.127)	1.75
Housing	5%	1.041	**0.891**	0.908	(0.903)	0.903	(0.899)	1.33
$N = 506$, $D = 13$	20%	1.779	1.368	1.296	(1.312)	**1.287**	(1.308)	1.27
Iris	5%	0.486	0.366	0.347	(0.364)	**0.343**	(0.360)	2.55
$N = 150$, $D = 4$	20%	1.020	0.543	0.473	(0.537)	**0.458**	(0.525)	2.46
Pima Indians	5%	1.530	1.224	1.176	(1.210)	**1.173**	(1.209)	2.20
$N = 768$, $D = 8$	20%	2.655	1.721	1.544	(1.676)	**1.535**	(1.678)	2.20
Servo	5%	1.077	0.844	0.755	(0.819)	**0.753**	(0.821)	1.75
$N = 167$, $D = 4$	20%	1.781	1.197	**0.967**	(1.152)	**0.966**	(1.151)	1.68
Stocks	5%	**0.241**	**0.242**	0.294	(0.286)	0.252	(0.249)	7.45
$N = 950$, $D = 9$	20%	0.443	0.347	0.487	(0.463)	0.330	(**0.328**)	7.58
Statlog	5%	1.312	1.243	1.217	(1.223)	**1.210**	(1.216)	2.89
$N = 846$, $D = 18$	20%	1.741	1.587	1.379	(1.405)	**1.355**	(1.384)	2.69

		PDS	ICkNNI	Single Gaussian	
Automobile	5%	0.975	0.973	0.977	(**0.959**)
$N = 159$, $D = 15$	20%	**1.193**	1.513	1.259	(1.223)
Breast Cance (Diag.)	5%	2.481	2.438	**2.406**	(**2.405**)
$N = 569$, $D = 30$	20%	2.832	3.041	**2.485**	(2.488)
Breast Cance (Prog.)	5%	3.457	3.405	**3.371**	(**3.371**)
$N = 194$, $D = 32$	20%	3.801	3.929	**3.489**	(3.497)
Breast Tissue	5%	0.858	0.783	**0.773**	(**0.773**)
$N = 106$, $D = 9$	20%	1.227	1.028	**0.901**	(0.911)
Connectionist Bench	5%	5.331	5.342	**5.292**	(**5.292**)
$N = 208$, $D = 60$	20%	5.542	5.639	**5.433**	(**5.433**)
Ecoli	5%	1.098	0.775	**0.743**	(0.770)
$N = 336$, $D = 7$	20%	2.368	1.162	**1.040**	(1.149)
Ionosphere	5%	2.775	2.745	2.776	(**2.738**)
$N = 351$, $D = 33$	20%	2.985	3.279	2.961	(**2.888**)
Parkinsons	5%	1.684	1.666	**1.642**	(**1.643**)
$N = 195$, $D = 22$	20%	1.970	2.293	**1.806**	(**1.808**)
Spambase	5%	3.311	2.688	2.660	(**2.640**)
$N = 4601$, $D = 57$	20%	4.712	3.869	3.500	(**3.467**)
SPECTF Heart	5%	4.624	4.599	**4.567**	(**4.568**)
$N = 267$, $D = 44$	20%	4.939	4.953	**4.723**	(**4.725**)
Wine	5%	1.984	1.928	**1.921**	(1.925)
$N = 178$, $D = 13$	20%	2.477	2.295	**2.163**	(2.199)

		PDS	ICkNNI	HDDC mixture		Mean K
Finance	5%	3.348	3.183	3.168	(**3.157**)	1.00
$N = 500$, $D = 41$	20%	3.785	3.777	3.411	(**3.381**)	1.00
Image Segmentation	5%	0.780	**0.553**	0.568	(0.574)	3.30
$N = 2310$, $D = 18$	20%	1.278	**1.037**	**1.026**	(1.050)	3.25
Tecator	5%	0.592	1.172	**0.592**	(**0.592**)	1.00
$N = 240$, $D = 100$	20%	**0.594**	1.867	0.595	(0.595)	1.00

Table 4: Average relative error 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, $\alpha = 0.05$) 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.

		PDS	ICkNNI	Single Gaussian		Mixture model		Mean K
Computer Hardware	5%	0.143	**0.101**	0.235	(0.108)	0.157	(0.109)	3.70
$N = 209$, $D = 6$	20%	0.252	**0.184**	0.388	(**0.180**)	0.279	(0.189)	3.42
Glass Identification	5%	0.092	0.054	0.057	(0.040)	0.050	(**0.038**)	1.81
$N = 214$, $D = 9$	20%	0.180	0.127	0.169	(**0.101**)	0.142	(**0.100**)	1.75
Housing	5%	0.070	**0.035**	0.054	(0.039)	0.052	(0.039)	1.33
$N = 506$, $D = 13$	20%	0.158	0.102	0.122	(**0.095**)	0.119	(**0.094**)	1.27
Iris	5%	0.142	0.090	0.114	(0.091)	0.097	(**0.084**)	2.55
$N = 150$, $D = 4$	20%	0.215	0.134	0.180	(0.136)	0.147	(**0.125**)	2.46
Pima Indians	5%	0.092	0.065	0.078	(0.064)	0.073	(**0.063**)	2.20
$N = 768$, $D = 8$	20%	0.173	0.122	0.135	(0.123)	0.130	(**0.121**)	2.20
Servo	5%	0.166	0.137	0.140	(0.125)	0.137	(**0.123**)	1.75
$N = 167$, $D = 4$	20%	0.251	0.193	0.194	(0.182)	0.193	(**0.181**)	1.68
Stocks	5%	0.066	**0.012**	0.054	(0.043)	0.020	(0.018)	7.45
$N = 950$, $D = 9$	20%	0.124	0.042	0.111	(0.087)	0.040	(**0.037**)	7.58
Statlog	5%	0.040	0.021	0.018	(0.017)	0.016	(**0.015**)	2.89
$N = 846$, $D = 18$	20%	0.093	0.074	0.046	(0.044)	0.041	(**0.041**)	2.69

		PDS	ICkNNI	Single Gaussian	
Automobile	5%	**0.051**	0.053	0.086	(0.056)
$N = 159$, $D = 15$	20%	**0.117**	0.205	0.185	(0.131)
Breast Cance (Diag.)	5%	0.033	0.021	0.011	(**0.010**)
$N = 569$, $D = 30$	20%	0.080	0.122	0.033	(**0.031**)
Breast Cance (Prog.)	5%	0.032	0.024	0.012	(**0.012**)
$N = 194$, $D = 32$	20%	0.079	0.113	0.038	(**0.037**)
Breast Tissue	5%	0.084	0.044	0.046	(**0.035**)
$N = 106$, $D = 9$	20%	0.169	0.121	0.112	(**0.084**)
Connectionist Bench	5%	0.024	0.025	0.013	(**0.012**)
$N = 208$, $D = 60$	20%	0.056	0.105	0.040	(**0.037**)
Ecoli	5%	0.120	**0.071**	0.115	(**0.070**)
$N = 336$, $D = 7$	20%	0.218	**0.133**	0.206	(**0.133**)
Ionosphere	5%	0.027	0.021	0.038	(**0.018**)
$N = 351$, $D = 33$	20%	0.064	0.177	0.103	(**0.049**)
Parkinsons	5%	0.042	0.028	0.022	(**0.020**)
$N = 195$, $D = 22$	20%	0.101	0.139	0.058	(**0.055**)
Spambase	5%	0.053	0.035	0.048	(**0.032**)
$N = 4601$, $D = 57$	20%	0.144	0.137	0.124	(**0.105**)
SPECTF Heart	5%	0.027	0.025	0.017	(**0.017**)
$N = 267$, $D = 44$	20%	0.066	0.106	0.049	(**0.045**)
Wine	5%	0.054	**0.037**	0.039	(**0.037**)
$N = 178$, $D = 13$	20%	0.114	0.097	**0.080**	(0.082)

		PDS	ICkNNI	HDDC mixture		Mean K
Finance	5%	0.039	0.025	0.026	(**0.019**)	1.00
$N = 500$, $D = 41$	20%	0.100	0.117	0.081	(**0.059**)	1.00
Image Segmentation	5%	0.059	**0.022**	0.038	(0.028)	3.30
$N = 2310$, $D = 18$	20%	0.149	**0.082**	0.134	(0.109)	3.25
Tecator	5%	0.005	0.242	0.000	(**0.000**)	1.00
$N = 240$, $D = 100$	20%	0.012	0.370	0.005	(**0.004**)	1.00