2010

Imputation methods for dealing with missing scores in biometric fusion

Yaohui Ding
West Virginia University

Imputation Methods for Dealing with Missing Scores in Biometric Fusion

Yaohui Ding

Thesis submitted to the
Eberly College of Arts and Sciences
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Master of Science
in
Statistics

Committee members

Arun Ross, Ph.D., Committee Co-Chair
E. James Harner, Ph.D., Committee Co-Chair
Mark Culp, Ph.D., Committee Member

Department of Statistics

Morgantown, West Virginia
2010

Keywords: Biometric Fusion; Missing Data; Imputation; Maximum Likelihood Estimation; Expectation-Maximization; Gaussian Mixture Model

ABSTRACT

Imputation Methods For Dealing with Missing Scores in Biometric Fusion

Yaohui Ding

Biometrics refers to the automatic recognition of individuals based on their physical or behavioral characteristics. Multimodal biometric systems, which consolidate multiple biometric characteristics of the same person, can overcome several practical problems that occur in single modality biometric systems. While fusion can be accomplished at various levels in a multimodal biometric system, score level fusion is commonly used as it offers a good trade-off between fusion complexity and data availability. However, missing scores affect the implementation of most biometric fusion rules. While there are several techniques for handling missing data, the imputation scheme, which replaces missing values with predicted values, is preferred since this scheme can be followed by a standard fusion scheme designed for complete data. Performance of the following imputation methods are compared: Mean/Median Imputation, K-Nearest Neighbor (KNN) Imputation and Imputation via Maximum Likelihood Estimation (MLE). A novel imputation method based on Gaussian Mixture Model (GMM) assumption is also introduced and it exhibits markedly better fusion performance than the other methods because of its ability to preserve the local structure of the score distribution. Experiments on the MSU database assess the robustness of the schemes in handling missing scores at different training set sizes and various missing rates.

Acknowledgement

I would like to express my sincere gratitude to Dr. Arun Ross for his technical and financial support throughout my study. The topic of this study is based on his idea. Without his wise suggestions and constant encouragement, I don't think I can enter and make any progress in this exciting field.

I would also like to express my deep appreciation to the chair of the Department of Statistics, Dr. James Harner, for his guidance and help throughout my whole period of graduate education in Department of Statistics. It has been a great privilege to be his student.

I would also like to acknowledge Dr. Mark Culp for serving in my thesis committee and providing me with valuable suggestions to improve this work. Several methods used in this thesis were based on his lectures in linear regression and categorical regression.

I extend my special gratitude to all my colleagues, especially Brian, Aglika, Asem, Raghav and Raghu for their friendship and help to improve my work. Special thanks to my wife Jiajia. Without her understanding and care, this thesis would not have been possible.

This work was supported by the NSF Center for Identification Technology Research (CITeR). I also thank Dr. Anil Jain at Michigan State University for granting me access to their database.

Contents

List of Figures

1 Introduction

Biometrics, or biometric recognition, refers to the automatic recognition of individuals based on their physical and behavioral characteristics, such as face, fingerprint, iris and voice [1]. Biometric systems that consolidate multiple biometric characteristics of the same identity are known as multimodal biometric systems. Multimodal biometric systems overcome many practical problems like noisy sensor data, non-universality and/or lack of distinctiveness of the biometric trait, unacceptable error rates, and spoof attacks [2].

The consolidation of different biometric sources is called biometric fusion. Biometric fusion can be implemented at various levels, such as image level, feature level, rank level, score level and decision level. Fusion at the score level is the most popular approach discussed in the literature [1, 2].

Most techniques for score level fusion are designed for a complete *score vector*[1] where the scores to be fused are assumed to be available. These techniques cannot be invoked when *score vectors* are incomplete.

Deletion methods, which omit all incomplete vectors, have negative implications for parameter bias and inefficiency [3, 4, 5], and are not suitable for use in biometric systems [6]. Certain "strong" classification methods can get fair results without deletion, especially when working with decision trees methods, such as Dynamic Path Generation [7] and the Lazy Decision Tree approach [8, 9]. Imputation methods, on the other hand, which substitute the missing scores with predicted values are better since (a) they do not delete any of the *score vectors* which may contain useful information for identification, and (b) their application could be followed by a standard score fusion scheme.

Many imputation methods are widely known. The Mean Imputation is one of the most frequently used methods. The Median Imputation seems to be more robust than Mean Imputation, since the mean can be affected by the presence of outliers. In microarrays data analysis, missing values are sometimes replaced by zero. The shortcomings of these simple imputation methods have been discussed in the literature [10, 11, 12].

Dixon [13] introduced the K-Nearest Neighbor (KNN) imputation technique for dealing with missing values in supervised classification problems. One significant advantage of KNN imputation is that the correlation structure of the data can be taken into consideration without any strict model assumption. In the Hot Deck (HD) Imputation method [14], a missing value (the recipient) of an attribute

[1]Here, the elements of the vector are the scores generated by the individual matchers

is filled in with a value (the donor) from the current data by using an estimated distribution for the missing attribute. The simplest way to implement HD is to randomly draw an observed value in the corresponding attribute as the donor. In order to incorporate the uncertainty caused by the estimation process, Multiple Imputation (MI) methods [15, 16, 17] estimate the missing values several times with the values drawn from a fitted distribution.

Another popular method is based on Maximum Likelihood Estimation (MLE) to handle the parameter estimation problem in the case of missing data [15, 27]. Variants of the Expectation-Maximization (EM) algorithm are used in these Maximum Likelihood (ML) procedures. They are generally superior to deletion methods, as MLE utilizes all the observed attributes of the data and can incorporate the probability mechanism leading to the missing data. However, the original MLE is a parameter estimation method rather than an imputation method. Therefore this method cannot be applied directly in a classification environment.

Besides several parametric (Mean, MLE) and nonparametric (KNN, HD) techniques which have been stated above, there are some semi-parametric techniques that allow to control the trade-off between parsimony of sample size and flexibility of model assumption. One approach is based on the use of Gaussian Mixture Models (GMMs) [18, 19, 21]. GMMs are not constrained to a specific functional form, but allow for a large class of distributions. Priebe [22] shows that, with 10,000 observations, a log-normal density can be well approximated by a mixture of 30 Gaussian components. An empirical study by DiZio et al. [23] shows that, for the preservation of the covariance structure, a random draw method is preferable over a conditional mean method when the GMM is used.

The missing score problem in multimodal biometrics has received limited attention. Nandakumar et al. [28] designed a Bayesian approach utilizing both ranks and scores to perform fusion in an identification system. The proposed method can handle missing information by assigning a fixed rank value to the marginal likelihood ratio corresponding to the missing entity. Fatukasi et al. [6] compared three different variants of the KNN imputation method in biometric fusion.

In this work, two existing methods (Mean imputation and KNN imputation) for handling missing scores are analyzed. Further, the MLE method is used as an imputation scheme in biometric fusion. Next, a novel imputation method called Hot Deck via the Gaussian Mixture Model (HD GMM) is also introduced. This can be viewed as an extension of the Hot Deck imputation scheme, but instead of using the current values in the training dataset, a simulated dataset is employed as the pool of donors. The MSU database [1] containing scores of three modalities (face, fingerprint and hand-geometry) is used in the experiments.

2 Multimodal Biometric System

2.1 Overview of Biometrics

A biometric system is essentially a pattern recognition system consisting of the following main modules [24]:

- Sensor module, which captures the biometric data of an individual. An example is a fingerprint sensor that images the ridge and valley structure of a user's finger.
- Feature extraction module, in which the acquired biometric data is processed to extract a set of salient or discriminatory features.
- Matcher module, in which the features extracted during recognition are compared against the stored templates to generate match scores. For example, in the matcher module of a fingerprint-based biometric system, the number of matching minutiae between the input and the template fingerprint images is determined and a match score is reported.
- System database module, which is used to store the biometric templates of the enrolled users.

A number of biometric characteristics exist and are broadly used in various applications, such as DNA, fingerprint, iris, face, hand geometry and ear. Each biometric has its strengths and weaknesses, and a single biometric is not expected to effectively meet the requirements of all applications. In other words, no single biometrics is superior under all circumstances.

2.2 Multimodal Biometrics

Multimodal biometric systems overcome several practical problems of single-biometric systems, like noisy sensor data, non-universality and/or lack of distinctiveness of the biometric trait, unacceptable error rates, and spoof attacks [2]. The procedure by which information from multiple biometric traits is consolidated is called biometric fusion, which is the critical component in multimodal biometrics.

2.2.1 Fusion of Biometrics

The layout of a bimodal biometric system is shown in Figure 1, and the purpose is to illustrate the various levels of fusion for combining two (or more) biometric

systems. The three possible levels of fusion are: (a) fusion at the feature level, (b) fusion at the match score level, (c) fusion at the decision level.

- Feature level: The raw data captured from each sensor will be used to build a feature vector, which uniquely identifies a given person in the feature space. Combining more feature vectors results in one vector with higher dimensionality and may increase the probability of correctly identifying a person.

- Match Score level: Fusion at the match score level is typically more effective than fusion at the decision level. Each single-modal biometric system measures and calculates its own match score. Match scores are a measure of the similarity or distance between features derived from a presented sample and a stored template. A match or non-match decision is made based on a certain decision threshold. For example, one approach is to construct a *score vector* using the match scores from each biometric modality, then a trained classifier will decide one of two classes: "Accept" (genuine user) or "Reject" (imposter user) based on the *score vector*.

- Decision level: Fusion at this level is the least informative. Each biometric system makes a decision and then those decisions are combined, usually using majority voting scheme. Some methods to weight the decisions from each biometrics are also used.

Apart from the above, fusion is possible at the raw data level or the rank level. Fusion at the score level is considered to be the most common approach due to the ease in accessing and combining the scores generated by different matchers [1].

2.2.2 Fusion at Score Level

Fusion techniques at the score level can be divided into three categories. The transformation-based score fusion will normalize the match scores to a common domain prior to combining them. Choice of the normalization scheme and combination weights are data-dependent and require extensive empirical evaluation. In a classifier-based score fusion, scores from multiple matchers will be treated as a *score vector*, therefore, a classifier can be constructed to discriminate genuine and imposter scores. In this case, biometric fusion is considered as a typical classification problem. Density-based score fusion is usually based on the likelihood ratio test. Based on the Neyman-Pearson theorem, if the underlying densities of genuine and imposter scores are explicitly known, the likelihood ratio fusion

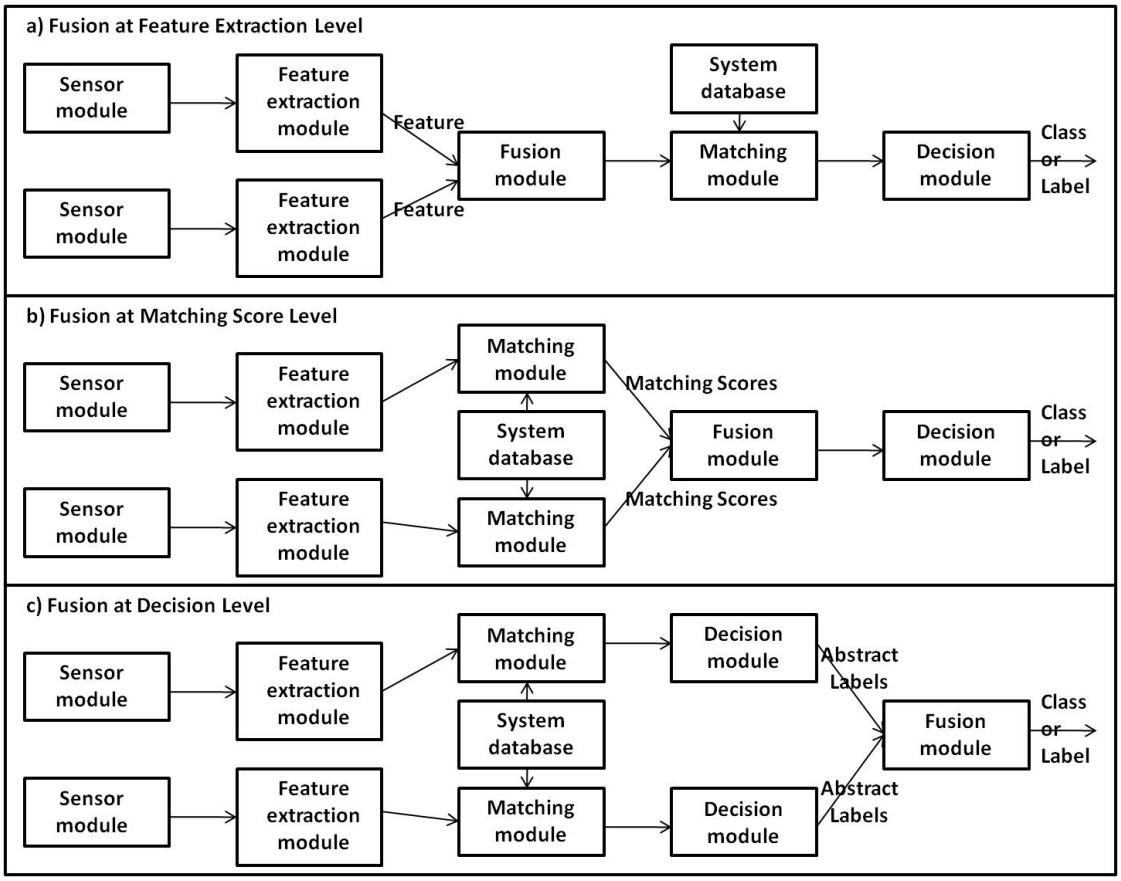

Figure 1: The biometric fusion could be implemented at various levels: a) fusion at feature level; b) fusion at match score or rank level; c) fusion at decision level. This figure is based on [24].

technique will provide the highest Genuine Accept Rate (GAR) for a fixed False Accept Rate (FAR) [29]. However, the underlying densities of scores cannot be exactly estimated in practice.

In this work, a common transformation-based score fusion technique, namely, the sum rule has been used to obtain all the Receiver Operating Characteristic (ROC) curves summarizing the fusion performance. As mentioned earlier, a score normalization scheme is required prior to merging the scores from different modalities into a single scalar score. Based on an empirical evaluation, Jain et al. [1] found that the min-max normalization scheme followed by a simple sum of scores fusion resulted in a superior GAR than other normalization and fusion techniques for the dataset used here. So the same process is used in this work.

2.2.3 Performance Measures

Usually, the performance of a biometric system can be measured in terms of two error rates, False Accept Rate (FAR) and False Reject Rate (FRR) [25]. The FAR refers to the errors that occur when a system mistakes the biometric measurements from two different individuals to be from the same person. In statistics, FAR is the probability of a type-II error. The FRR refers to the errors that the biometric system mistakes two biometric measurements from the same person to be from two different people. FRR is the probability of a type-I error. FAR and FRR are also called as False Match Rate (FMR) and False Non-Match Rate (FNMR), respectively, in some literature.

To understand the performance of a biometric system, a plot of FAR vs. FRR is usually used. This is known as a Receiver Operating Characteristic (ROC) curve. ROC curves present a non-dimensional, basic technical performance measure for comparing two or more biometric systems. It can also display the trade-offs between FAR and FRR over a wide range of thresholds. In this study, ROC curves are plotted as GAR (Genuine Accept Rate) vs. FAR, where GAR is the complement of FRR (GAR = 1-FRR).

Equal Error Rate (EER) can also be used to give a threshold independent performance measure of a biometric system [2]. It is the point where the FAR equals the FRR. In the other words, EER is the error rate occurring when the decision threshold of a system is set so that the proportion of false rejections are approximately equal to the proportion of false acceptances. In Figure 2, the EER of the hand-geometry scores in the MSU database is about 10.7%.

Figure 3 provides an example of a multimodal biometric system that can demonstrate a better recognition performance than using a single biometric. This

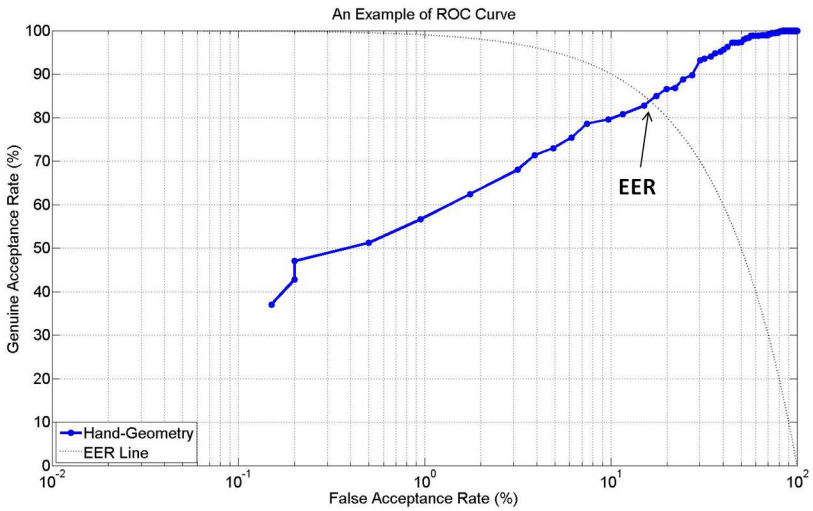

Figure 2: Example of the ROC Curve (GAR vs FAR) for Hand-Geometry scores in the MSU database.

example employs the simple sum of scores as the fusion scheme, and the min-max normalization technique is used for transforming the match scores from face, fingerprint and hand-geometry into a common domain before fusion.

2.3 When Score Goes Missing in Biometric Fusion

Most techniques for score level fusion are designed for a complete *score vector*, where the scores to be fused are assumed to be available. When any of the scores are missing, these techniques cannot be invoked.

Incomplete *score vectors* can occur under different conditions. There are many causes for missing data such as the failure of a matcher to generate a score (e.g., a fingerprint matcher may be unable to generate a score when the input image is of inferior quality), the absence of a trait during image acquisition (e.g., a surveillance multibiometric system may be unable to obtain the iris of an individual), and sensor malfunction, where the sensor pertaining to a modality may not be operational (e.g., failure of a fingerprint sensor due to wear and tear of the device).

There are several methods to deal with missing data as pointed out earlier. However, some practical constraints have to be dealt with when it comes to biometrics.

Compared with most studies that adopt the entire dataset for the analysis, a

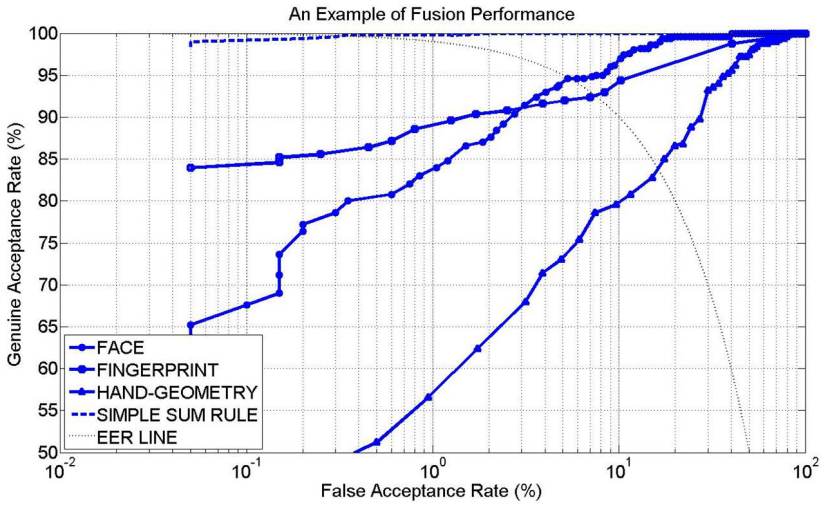

Figure 3: Example of ROC curves for three single-modal biometrics in the MSU database and the simple sum fusion.

fixed and complete training set is preferred in the context of biometrics because (a) imputation is based on this entire fixed set rather than a dynamically changing set, and (b) one can easily handle both complete and incomplete *score vectors* in the test set.

The independence between *score vectors* requires a vector-by-vector imputation process rather than a batch process where all missing scores are imputed at the same time. With this understanding, only the observed part of this *score vector* and the training set can be used to perform the imputation. Any information from the other independent vectors cannot be incorporated.

Unlike the missing data problem in Gene-expression or other data mining applications where usually a large number of variables are used, current multimodal biometric systems involve less than 5 modalities. Therefore, exhaustive methods which consider all possible combinations of missing patterns, such as the exhaustive fusion framework [6], are likely to be more efficient in multimodal biometrics. On the other hand, some imputation methods like Bayesian Network (BN) [30] which require more variables to compute probabilistic relationships between them, might be unusable in a biometrics environment.

3 Patterns of Missing Data

Distinguishing between different patterns of missing data is important because it will impact the choice of method used for handling the problem. Rubin [27] defines a taxonomy for different patterns of missing data.

- Missing Completely At Random (MCAR): the probability of an observation being missing does not depend on the value of the observed or unobserved data. In mathematical terms, this is written as:

$$Pr(X^m|X^{mis},X^{obs}) = Pr(X^m), \tag{1}$$

  where X^m denotes the missingness mechanism[2] of data, and X^{mis} and X^{obs} denote the unobserved part and observed part, respectively.

- Missing At Random (MAR): given the observed data, the missingness mechanism does not depend on the unobserved data. Mathematically,

$$Pr(X^m|X^{mis},X^{obs}) = Pr(X^m|X^{obs}). \tag{2}$$

- Missing Not At Random: When neither MCAR nor MAR hold, we say the data are Missing Not At Random, abbreviated as MNAR. In other words, the mechanism of missing data does depend on the unobserved data.

In order to analyze the performance of various methods dealing with missing data, researchers usually randomly remove some entries from a complete dataset to generate missing data artificially. In this case, the generated missing data will follow the MCAR pattern. In reality, researchers may not know the real reason behind the missing data in most cases. Therefore, the generated dataset can be assumed to conform to a MAR pattern. In this work, either MCAR or MAR is assumed, suggesting that the missing data is not dependent on the value which is missing.

It must be noted that, methods for distinguishing between MCAR and MAR are computationally expensive. Ramoni and Sebastiani [32] describe a novel method, called the Robust Bayesian Estimator (RBE), which does not depend on any assumption of the missing patterns discussed above. The robustness is achieved from bounding the incomplete data with the set of all possible estimates, which are constrained by the incomplete data itself.

[2]In statistical literature, the cause for the missing data is sometimes called the missingness mechanism

4 Imputation Methods

4.1 Criteria

As stated by Marker et al. [33], two major criteria should be employed in assessing the performance of imputation methods: firstly, a good imputation method should preserve the natural relationship between variables in a multivariate dataset (in our case, the variables correspond to scores originating from multiple classifiers); secondly, a good imputation method should embody the uncertainty caused by the imputed data by deriving variance estimates.

These two criteria are applicable for imputation in a biometric score dataset. Additionally, the use of imputed data should result in comparable performance to that of the original data containing no missing data. Some imputation methods may not result in good performance if they overstate or understate the relationship between variables, or if they omit the uncertainty in the imputed data.

4.2 Notation

In the context of multimodal biometric systems, a user i offers p biometric modalities. The system will generate a vector of match scores, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, where each match score corresponds to one modality. Suppose there are n users, then the score matrix with n observations and p attributes can be written as:

$$\mathbf{D} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & x_{ij} & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix},$$

where x_{ij} denotes the match score from the j-th modality of the i-th user. Similarly, the training set can be expressed as $\mathbf{D}^{\mathbf{tr}}$.

If there is no missing data, the conventional fusion techniques can be implemented on each observation (row) separately, and then make the decision whether each observation belongs to a genuine user or an imposter. For any observation $\boldsymbol{x}_i$ containing missing scores, it can be written in the form $(\boldsymbol{x}_i^{obs}, \boldsymbol{x}_i^{mis})$, where $\boldsymbol{x}_i^{obs}$ and $\boldsymbol{x}_i^{mis}$, respectively, denote the observed and missing attributes for observation i. The missing values $\boldsymbol{x}_i^{mis}$ can be replaced with the imputed value $\boldsymbol{x}_i^{imp}$ using the methods considered below.

Different multivariate distributions will be assumed in the following methods. Let Θ denote all the parameters to be estimated in a particular model. Take the

MLE method as an example. The dataset $\mathbf{D}$ will be assumed to have a p-variate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ and covariance matrix Σ, so here $\Theta = (\boldsymbol{\mu}, \Sigma)$ corresponds to the parameters of the multivariate normal distribution. Since both MLE and GMM methods use iterative algorithms for estimation, let $\Theta^{(t)}$ denote all the parameters to be estimated at the t-th iteration.

4.3 Mean Imputation

In mean imputation, missing values $\boldsymbol{x}_i^{mis}$ are filled by the average of scores from the corresponding attributes in the training set $\mathbf{D^{tr}}$. For example, suppose the second attribute of $\boldsymbol{x}_i$ is missing, then this missing score will be imputed by the average of the second attribute in $\mathbf{D^{tr}}$:

$$x_{i2}^{imp} = \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \xi_{j2}, \tag{3}$$

where $\xi_{j2} \in \mathbf{D^{tr}}$ is the j-th observation from the training set $\mathbf{D^{tr}}$, and n_{tr} is the total number of samples in the training set.

Several drawbacks of the mean imputation scheme have been pointed out by Little and Rubin [10]. Obviously, the variance is underestimated. Besides that, replacing all missing values in a modality with a single value will artificially distort the shape of the distribution of original scores, which will cause a bias for our classification purposes. Similar disadvantages occur in the Median Imputation method, although the median is less affected by the presence of outliers in the distribution.

4.4 K-Nearest Neighbor Imputation

In a classical KNN imputation, the missing values of an observation are imputed based on a given number of instances (k) in $\mathbf{D^{tr}}$ that are most similar to the instance of interest. A measure of distance d between two instances should be determined. In this work, a Euclidean distance function is considered. Let $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ be two observations; then d is defined as:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{h \in O_i \cap O_j} (x_{ih} - x_{jh})^2,$$

where $O_i = \{\text{h}|\text{the h-th attribute of the i-th observation is observed}\}$. In other words, only the mutually observed attributes are used to calculate the distance between observations.

The KNN algorithm is described as follows:

1) For each observation $\boldsymbol{x}_i$, apply the distance function d to find the k nearest neighbor vectors in the training set $\mathbf{D}^{\mathbf{tr}}$;
2) The missing attributes $\boldsymbol{x}_i^{mis}$ are imputed by the average of the corresponding attributes from those k nearest neighbors.

KNN imputation does not require the creation of a predictive model for each attribute, and so it can easily treat instances with multiple missing values. However, there are some concerns with respect to KNN imputation. Firstly, which distance function should be used for a particular dataset? The choice could be Euclidean, Manhattan, Mahalanobis, Pearson, etc. In this work the Euclidean distance is employed. Secondly, the KNN algorithm searches through the entire dataset looking for the most similar instances, and can therefore be a very time consuming process. Thirdly, the choice of k, will impact the results. The choice of a small k may produce a deterioration in the performance of the classifier after imputation due to overemphasis on a few dominant instances in the estimation process of the missing values. On the other hand, a neighborhood of large size would include instances that are significantly different from the instance containing missing values thereby hurting the estimation process, and the classifier's performance declines. According to our analysis (not shown here), we found $k = 5$ to provide the best imputation accuracy on our relatively small dataset.

4.5 Imputation through MLE

The theoretical benefits of Maximum Likelihood Estimation (MLE) are widely known. After incorporating the Expectation Maximization (EM) algorithm, the MLE via EM method can be used to handle the problem of parameter estimation in an incomplete dataset, even under the MAR assumption [11]. In order to explain this algorithm, a simple example using a bivariate normal dataset with missing data will be first introduced. Then we will bring in the use of the *sweep operator* [10] and show how this operator provides a simple and convenient way of performing the ML calculations for incomplete normal data. Finally, we will introduce the EM algorithm for incomplete multivariate normal data and its implementation in a biometrics environment.

4.5.1 An Example of Bivariate Normal Data

In a bivariate normal dataset, the maximizer of the log-likelihood of incomplete data can be computed directly [10]. Considering a bivariate normal sample with m complete bivariate observations $\{(x_{i1}, x_{i2}); i = 1, \ldots, m\}$ and $n-m$ univariate observations $\{x_{i1}; i = m+1, \ldots, n\}$, the parameters of the joint distribution of x_{i1} and x_{i2}, $\boldsymbol{\theta}$, can be written as

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma) = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22}),$$

and the log-likelihood of $\boldsymbol{\theta}$ can be computed by:

$$\begin{aligned} l(\boldsymbol{\mu}, \Sigma | \boldsymbol{x}^{obs}) = & -\frac{1}{2}\{m \ln |\Sigma| + \sum_{i=1}^{m} (x_i - \boldsymbol{\mu}) \Sigma^{-1} (x_i - \boldsymbol{\mu})^T\} \\ & -\frac{1}{2}\{(n-m) \ln \sigma_{11} + \sum_{l=m+1}^{n} \frac{(x_{i1} - \mu_1)^2}{\sigma_{11}}\}. \end{aligned}$$

This likelihood equation, however, does not have an obvious solution. But the joint distribution of x_{i1} and x_{i2} can be factored into a marginal distribution of x_{i1} and a conditional distribution of x_{i2} given x_{i1}:

$$f((x_{i1}, x_{i2} | \boldsymbol{\mu}, \Sigma) = f(x_{i1} | \mu_1, \sigma_{11}) f(x_{i2} | x_{i1}, \beta_{0,2\cdot1}, \beta_{1,2\cdot1}, \sigma_{22\cdot1}),$$

where, $f(x_{i1} | \mu_1, \sigma_{11})$ is the normal distribution with mean μ_1 and variance σ_{11}, and $f(x_{i2} | x_{i1}, \beta_{0,2\cdot1}, \beta_{1,2\cdot1}, \sigma_{22\cdot1})$ is the normal distribution with mean $\beta_{0,2\cdot1} + \beta_{1,2\cdot1}\mu_1$ and the variance $\sigma_{22\cdot1}$. "$2 \cdot 1$" denotes the regression of x_{i2} on x_{i1}. This parameterization can be denoted by $\boldsymbol{\phi}$:

$$\boldsymbol{\phi} = (\mu_1, \sigma_{11}, \beta_{0,2\cdot1}, \beta_{1,2\cdot1}, \sigma_{22\cdot1}).$$

Here, we express the relationship between these 2 different parameterizations $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$:

$$\begin{aligned} \beta_{0,2\cdot1} &= \mu_2 - \beta_{1,2\cdot1}\mu_1, \\ \beta_{1,2\cdot1} &= \sigma_{12}/\sigma_{11}, \\ \sigma_{22\cdot1} &= \sigma_{22} - \sigma_{12}^2/\sigma_{11}. \end{aligned} \tag{4}$$

Similarly, the components of $\boldsymbol{\theta}$, other than μ_1 and σ_{11}, can be expressed as follows:

$$\begin{aligned}\mu_2 &= \beta_{0,2\cdot1} + \beta_{1,2\cdot1}\mu_1,\\ \sigma_{12} &= \beta_{1,2\cdot1}\sigma_{11},\\ \sigma_{22} &= \sigma_{22\cdot1} + \beta_{1,2\cdot1}^2\sigma_{11}.\end{aligned} \tag{5}$$

The density of the data $\boldsymbol{x}^{obs}$ factors in the following way:

$$\begin{aligned}f(\boldsymbol{x}^{obs}|\boldsymbol{\theta}) &= \prod_{i=1}^{m} f(x_{i1},x_{i2}|\boldsymbol{\theta}) \prod_{i=m+1}^{n} f(x_{i1}|\boldsymbol{\theta})\\ &= [\prod_{i=1}^{n} f(x_{i1}|\mu_1,\sigma_{11})][\prod_{i=1}^{m} f(x_{i2}|x_{i1},\beta_{0,2\cdot1},\beta_{1,2\cdot1},\sigma_{22\cdot1})].\end{aligned} \tag{6}$$

The first bracketed factor in the above equation is the density of an independent sample of size n from the normal distribution with mean μ_1 and variance σ_{11}. The second factor is the density for m observations from the conditional normal distribution with mean $\beta_{0,2\cdot1} + \beta_{1,2\cdot1}\mu_1$ and the variance $\sigma_{22\cdot1}$.

ML estimates can be obtained by independently maximizing the likelihoods corresponding to these two components:

$$\begin{aligned}\hat{\mu}_1 &= n^{-1}\sum_{i=1}^{n} x_{i1},\\ \hat{\sigma}_{11} &= n^{-1}\sum_{i=1}^{n}(x_{i1}-\hat{\mu}_1)^2.\end{aligned} \tag{7}$$

Maximizing the second factor uses standard regression results and yields:

$$\begin{aligned}\hat{\beta}_{0,2\cdot1} &= \bar{x}_2 - \hat{\beta}_{1,2\cdot1}\bar{x}_1,\\ \hat{\beta}_{1,2\cdot1} &= s_{12}/s_{11},\\ \hat{\sigma}_{22\cdot1} &= s_{22} - s_{12}^2/s_{11},\end{aligned} \tag{8}$$

where

$$\bar{x}_j = m^{-1}\sum_{i=1}^{m} x_{ij},$$

and

$$s_{jk} = m^{-1}\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k), \quad j,k = 1,2.$$

The ML estimates of other parameters can now be obtained using (5):

$$\begin{aligned}\hat{\mu}_2 &= \hat{\beta}_{0,2\cdot 1} + \hat{\beta}_{1,2\cdot 1}\mu_1 \\ &= \bar{x}_2 + \hat{\beta}_{1,2\cdot 1}(\hat{\mu}_1 - \bar{x}_1), \qquad (9)\end{aligned}$$

$$\hat{\sigma}_{22} = \hat{\sigma}_{22\cdot 1} + \hat{\beta}^2_{1,2\cdot 1}(\hat{\sigma}_{11} - s_{11}). \qquad (10)$$

4.5.2 The Sweep Operator

The *sweep operator* provides a simple and convenient way of performing the ML calculations for incomplete normal data, and this ML calculation is critical for the EM algorithm which will be introduced in the following subsection.

The *sweep operator* is defined for symmetric matrices as follows [10]:

A *pxp* symmetric matrix G is said to be swept on row and column k if it is replaced by another symmetric *pxp* matrix H with elements defined as follows:

$$\begin{aligned} h_{kk} &= -1/g_{kk}, \\ h_{jk} = h_{kj} &= g_{jk}/g_{kk}, \quad k \neq j, \\ h_{jl} &= g_{jl} - g_{jk}g_{kl}/g_{kk} \quad k \neq j, k \neq l. \end{aligned} \qquad (11)$$

Considering the $3x3$ case:

$$\mathbf{G} = \begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{bmatrix},$$

then:

$$\mathbf{H} = \mathbf{SWP[1]G} \begin{bmatrix} -1/g_{11} & g_{12}/g_{11} & g_{13}/g_{11} \\ g_{12}/g_{11} & g_{22} - g^2_{12}/g_{11} & g_{23} - g_{13}g_{12}/g_{11} \\ g_{13}/g_{11} & g_{23} - g_{13}g_{12}/g_{11} & g_{33} - g^2_{13}/g_{11} \end{bmatrix}. \qquad (12)$$

We use the notation $\mathbf{SWP[k]G}$ to denote the matrix $\mathbf{H}$ defined by (11). Also, the result of successively applying the operations $\mathbf{SWP[k_1]G}$, $\mathbf{SWP[k_2]G}$, ..., $\mathbf{SWP[k_i]G}$ to the matrix $\mathbf{G}$ will be denoted by $\mathbf{SWP[k_1, k_2, \ldots, k_i]G}$. The *sweep operator* is commutative.

Suppose we arrange the original parameters of the joint distribution of x_{i1} and x_{i2}, $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})$, in the following symmetric matrix:

$$\boldsymbol{\theta}^* = \begin{bmatrix} -1 & \mu_1 & \mu_2 \\ \mu_1 & \sigma_{11} & \sigma_{12} \\ \mu_2 & \sigma_{21} & \sigma_{22} \end{bmatrix}.$$

If $\boldsymbol{\theta}^*$ is swept on row and column 1, we obtain from (12):

$$\mathbf{SWP}[\mathbf{1}]\boldsymbol{\theta}^* = \begin{bmatrix} -(1+\mu_1^2/\sigma_{11}) & \mu_1/\sigma_{11} & \mu_2-\mu_1\sigma_{12}/\sigma_{11} \\ \mu_1/\sigma_{11} & -\sigma_{11}^{-1} & \sigma_{12}/\sigma_{11} \\ \mu_2-\mu_1\sigma_{12}/\sigma_{11} & \sigma_{12}/\sigma_{11} & \sigma_{22}-\sigma_{12}^2/\sigma_{11} \end{bmatrix}. \tag{13}$$

Comparing with the equation (4), the above equation (13) can be rewritten as:

$$\boldsymbol{\phi}^* = \mathbf{SWP}[\mathbf{1}]\boldsymbol{\theta}^* = \begin{bmatrix} -(1+\mu_1^2/\sigma_{11}) & \mu_1/\sigma_{11} & \beta_{0,2\cdot1} \\ \mu_1/\sigma_{11} & -\sigma_{11}^{-1} & \beta_{1,2\cdot1} \\ \beta_{0,2\cdot1} & \beta_{1,2\cdot1} & \sigma_{22\cdot1} \end{bmatrix} \tag{14}$$

and

$$\begin{bmatrix} -(1+\mu_1^2/\sigma_{11}) & \mu_1/\sigma_{11} \\ \mu_1/\sigma_{11} & -\sigma_{11}^{-1} \end{bmatrix} = \mathbf{SWP}[\mathbf{1}] \begin{bmatrix} -1 & \mu_1 \\ \mu_1 & \sigma_{11} \end{bmatrix}. \tag{15}$$

The operation of sweeping on a variable in effect turns that variable from an outcome variable into a predictor variable. Similarly, there is also an operator inverse to sweep, that turns predictor variables into outcome variables:

$$\mathbf{H} = \mathbf{RSW}[\mathbf{k}]\mathbf{G},$$

where,

$$\begin{aligned} h_{kk} &= -1/g_{kk}, \\ h_{jk} = h_{kj} &= -g_{jk}/g_{kk}, \quad k \neq j, \\ h_{jl} &= g_{jl} - g_{jk}g_{kl}/g_{kk} \quad k \neq j, k \neq l. \end{aligned} \tag{16}$$

This *reverse sweep* (RSW) is commutative and is the inverse operator to sweep; that is:

$$(\mathbf{RSW}[\mathbf{k}])(\mathbf{SWP}[\mathbf{k}])G = (\mathbf{SWP}[\mathbf{k}])(\mathbf{RSW}[\mathbf{k}])G = G.$$

After defining the RSW, we can introduce the transformation between $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\phi}} = \mathbf{SWP}[\mathbf{1}]\hat{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{SWP}[\mathbf{1}] \begin{bmatrix} -1 & \hat{\mu}_1 \\ \hat{\mu}_1 & \hat{\sigma}_{11} \end{bmatrix} & \begin{matrix} \hat{\beta}_{0,2\cdot1} \\ \hat{\beta}_{1,2\cdot1} \end{matrix} \\ \begin{matrix} \hat{\beta}_{0,2\cdot1} & \hat{\beta}_{1,2\cdot1} \end{matrix} & \hat{\sigma}_{22\cdot1} \end{bmatrix}, \tag{17}$$

and

$$\hat{\boldsymbol{\theta}} = \mathbf{RSW}[\mathbf{1}]\hat{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{SWP}[\mathbf{1}] \begin{bmatrix} -1 & \hat{\mu}_1 \\ \hat{\mu}_1 & \hat{\sigma}_{11} \end{bmatrix} & \begin{matrix} \hat{\beta}_{0,2\cdot 1} \\ \hat{\beta}_{1,2\cdot 1} \end{matrix} \\ \begin{matrix} \hat{\beta}_{0,2\cdot 1} & \hat{\beta}_{1,2\cdot 1} \end{matrix} & \hat{\sigma}_{22\cdot 1} \end{bmatrix}. \tag{18}$$

4.5.3 MLE via EM in Multivariate Normal Data

A general method for using MLE in missing data imputation was described by Dempster et al. [18] in their influential article on the EM algorithm. The key idea of EM is to solve a difficult incomplete-data estimation problem by iteratively solving an easier complete-data problem. Intuitively,"fill" in the missing data with the best guess under the current estimate of the unknown parameters (E-STEP), then re-estimate the parameters from the observed and filled-in data (M-STEP). An overview of EM has been given in [10, 16, 20].

In order to obtain the correct answer, Dempster et al. [18] showed that, rather than filling in the missing data values per se, the complete-data sufficient statistics should be computed in every iteration. The form of these statistics depends on the model under consideration. With the assumption of K-variate normal distribution, the hypothetical complete dataset $\mathbf{D}$ belongs to the regular exponential family. So $\sum_{i=1}^{n} x_{ik}$ and $\sum_{i=1}^{n} x_{ik}x_{ij}$ are sufficient statistics of samples from this distribution $(j,k = 1,\ldots,K)$. The modified t-th iteration of E-STEP can then be written as:

$$E\left(\sum_{i=1}^{n} x_{ik} | \mathbf{D}^{\mathbf{tr}}, \boldsymbol{x}_i^{obs}, \Theta^{(t)}\right) = \sum_{i=1}^{n} x_{ik}^{(t)}, \quad k = 1,\ldots,K,$$

$$E\left(\sum_{i=1}^{n} x_{ik}x_{ij} | \mathbf{D}^{\mathbf{tr}}, \boldsymbol{x}_i^{obs}, \Theta^{(t)}\right) = \sum_{i=1}^{n} \left(x_{ik}^{(t)} x_{ij}^{(t)} + c_{ijk}^{(t)}\right),$$

where,

$$x_{ik}^{(t)} = \begin{cases} x_{ik}, & \text{if } x_{ik} \text{ is observed,} \\ E\left(x_{ik} | \mathbf{D}^{\mathbf{tr}}, \boldsymbol{x}_i^{obs}, \Theta^{(t)}\right), & \text{if } x_{ik} \text{ is missing,} \end{cases} \tag{19}$$

and

$$c_{ijk}^{(t)} = \begin{cases} 0, & \text{if } x_{ik} \text{ or } x_{ij} \text{ is observed,} \\ Cov\left(x_{ik}, x_{ij} | \mathbf{D^{tr}}, \boldsymbol{x}_i^{obs}, \Theta^{(t)}\right), & \text{if } x_{ik} \text{ and } x_{ij} \text{ are missing.} \end{cases} \tag{20}$$

Missing values x_{ik} are thus replaced by the conditional mean of x_{ik} given the set of values $\boldsymbol{x}_i^{obs}$, available for that observation. These conditional means and the nonzero conditional covariances are easily found from the current parameter estimates by sweeping the augmented covariance matrix so that the variables $\boldsymbol{x}_i^{obs}$ are predictors in the regression equation and the remaining variables are outcome variables.

The M-STEP of the EM algorithm is straightforward and is a standard MLE process, i.e.,

$$\mu_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} x_{ik}^{(t)}, \quad k = 1, \ldots, K, \tag{21}$$

$$\sigma_{jk}^{(t+1)} = \frac{1}{n} E\left(\sum_{i=1}^{n} x_{ik} x_{ij} | \mathbf{D^{tr}}, \boldsymbol{x}_i^{obs} \right) - \mu_k^{(t+1)} \mu_j^{(t+1)}. \tag{22}$$

The algorithm will iterate repeatedly between the two steps until the difference between covariance matrices in subsequent M-STEPs falls below some specified convergence criterion. Although the classical EM algorithm will stop at this M-STEP, it is straightforward to get the imputed values by performing the E-STEP one more time, which means using the sweep operator and the regression equations with $\boldsymbol{x}_i^{obs}$ as predictors one more time.

4.5.4 Some Comments about MLE

When this method is used in the biometric scenario involving match scores, some additional constraints are required. As mentioned in Section 2.3, the different observations (vectors) should be assumed to be independent, and this assumption should be maintained as much as possible during the estimation and imputation procedure. In order to accommodate this assumption, when calculating the ML estimation for the incomplete vector $\boldsymbol{x}_i = (\boldsymbol{x}_i^{obs}, \boldsymbol{x}_i^{mis})$, only the training set $\mathbf{D^{tr}}$ and the observed part of this vector $\boldsymbol{x}_i^{obs}$ will be included. This strategy is relevant in a practical situation where the training set is relatively fixed.

A notable drawback of EM algorithm should be pointed out. The imputed scores from the EM algorithm lack the residual variability which are present in

the training set with complete data because they fall exactly on the regression line when using the parameters estimated by the iterations [17]. The Multiple Imputation (MI) method proposed by Rubin [27] accounts for missing data by restoring not only the natural variability in the missing-data, but also by incorporating the uncertainty caused by the estimation process.

The general strategy of MI can be summarized as follows: impute missing values using an appropriate model which can plausibly represent the data with random variation, repeat this $m > 1$ times to produce m data sets with complete data, and then combine the results to obtain overall estimates using Rubin's Rules [27]: the overall estimate is the simple average of the m estimates, and the overall estimate of the standard error is a combination of the within-imputation variability, W, and the between-imputation variability, B:

$$T = W + [(1 + 1/m) * B].$$

NORM, a very useful program proposed by Schafer [16], creates multiple imputation for incomplete data with arbitrary patterns of missing values under the multivariate normal model. Although it is designed for multiple imputation, a similar algorithm has been employed to calculate the ML estimates of an incomplete dataset. NORM is used as an auxiliary software in this study.

4.6 Imputation via the GMM Estimation

As mentioned earlier, the MLE method is based on the multivariate normal assumption to determine the likelihood function form and sufficient statistics. Although this assumption is mild, an obvious violation of normality often happens in biometrics because of the inherent discrimination between genuine and imposter scores.

Finite mixture models allow more flexibility, because they are not constrained to one specific functional form. As shown in Fraley and Raftery [34, 38], many probability distributions can be well approximated by mixture models. At the same time, in contrast to nonparametric schemes, mixture models do not require a large number of observations to obtain a good estimate [22, 23].

Let observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be a random sample from a finite mixture model with K underlying components in unknown proportions $\pi_1, \ldots, \pi_K$. Let the density of $\boldsymbol{x}_i$ in the k-th component be $f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k$ is the parameter vector for component k. In this case, $\Theta = (\pi_1, \ldots, \pi_K; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) = (\boldsymbol{\pi}, \boldsymbol{\theta})$, and then the

density of $\boldsymbol{x}_i$ can be written as:

$$f(\boldsymbol{x}_i;\Theta) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k),$$

where $\sum_{k=1}^{K} \pi_k = 1, \pi_k \geq 0$, for $k = 1,\ldots,K$.

Finite mixture models are frequently used when the component densities $f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k)$ are taken to be p-variate normal distributions $\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}_k,\Sigma_k)$, where observation i belongs to component k. This model has been studied by Titterington et al. [35], and by McLachlan & Basford [19]. Further details on the maximum likelihood estimates of the components of Θ can be found in McLachlan and Peel [21].

When Gaussian Mixture Models are used in imputation, two main steps will be essential: the density estimation using the GMM assumption and the imputation itself based on this estimated density.

4.6.1 Density Estimation using GMM

The EM algorithm of Dempster et al. [18] is applied to the finite mixture model for density estimation. Let the vector of indicator variables, $\boldsymbol{z}_i = (z_{i1},\ldots,z_{iK})$, be defined by:

$$z_{ik} = \begin{cases} 1 & \text{if observation } i \in \text{component } k, \\ 0 & \text{if observation } i \notin \text{component } k. \end{cases}$$

where $\boldsymbol{z}_i, i = 1,\ldots,n$, are independently and identically distributed according to a multinomial distribution generated by a single trial of an experiment with K mutually exclusive outcomes having probabilities $\pi_1,\ldots,\pi_K$.

Let $\hat{\Theta}$ denote the maximum likelihood estimate of Θ. Then each observation, $\boldsymbol{x}_i$, can be allocated to component k on the basis of the estimated posterior probabilities. The estimated posterior probability that observation $\boldsymbol{x}_i$, belongs to component k, is given by:

$$\hat{z}_{ik} = pr(\text{observation } i \in \text{component } k|\boldsymbol{x}_i;\hat{\Theta}) = \frac{\hat{\pi}_k f_k(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_k)}{\sum_{k=1}^{K} \hat{\pi}_k f_k(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_k)}.$$

and $\boldsymbol{x}_i$ is assigned to component k if:

$$\hat{z}_{ik} > \hat{z}_{ik'} \quad \text{for } k = 1,\ldots,K \quad k \neq k'.$$

The EM algorithm consists of defining an initial guess for the parameters to be estimated, and iteratively estimating the parameters until convergence of the Expectation step (E-step) and the Maximization step (M-step).

The E step requires calculating the expectation of the log-likelihood of the complete data conditioned on the observed data and the current value of the parameters:

$$\hat{z}_{ik} = \hat{z}_{ik}^{(t)} = E(z_{ik}|\boldsymbol{x}_i^{obs}; \Theta^{(t)}) = \frac{\pi_k f_k(\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)})}{\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)})}.$$

That is, z_{ik} is replaced by $\hat{z}_{ik}$, the estimate of the posterior probability that observation i belongs to component k. The remaining calculations in the E step are analogous to those required in the standard EM algorithm for incomplete normal data:

$$E(z_{ik}x_{ij}|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)}) = \begin{cases} \hat{z}_{ik}x_{ij}, & x_{ij} \text{ observed}, \\ \hat{z}_{ik}E(x_{ij}|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)}), & x_{ij} \text{ missing}. \end{cases}$$

$$E(z_{ik}x_{ij}^2|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)}) = \begin{cases} \hat{z}_{ik}x_{ij}^2, & x_{ij} \text{ observed}, \\ \hat{z}_{ik}[(E(x_{ij}|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)}))^2 + Var(x_{ij}|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)})], & x_{ij} \text{ missing}. \end{cases}$$

For $j \neq j'$,

$$E(z_{ik}x_{ij}x_{ij'}|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)}) = \begin{cases} \hat{z}_{ik}x_{ij}x_{ij'}, & x_{ij} \text{ and } x_{ij'} \text{ observed}, \\ \hat{z}_{ik}x_{ij}E(x_{ij'}|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)}), & x_{ij} \text{ observed}, x_{ij'} \text{ missing}, \\ \hat{z}_{ik}E(x_{ij}|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)})x_{ij'}, & x_{ij'} \text{ observed}, x_{ij} \text{ missing}, \\ \hat{z}_{ik}[E(x_{ij}|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)})E(x_{ij'}|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)}) & \\ +Cor(x_{ij}, x_{ij'}|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)})], & x_{ij} \text{ and } x_{ij'} \text{ missing}. \end{cases}$$

In the M step of the algorithm, the new parameters $\boldsymbol{\theta}^{(t+1)}$ are estimated from the sufficient statistics of the complete data:

$$\hat{\pi}_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \quad \text{for } k = 1, \ldots, K,$$

$$\hat{\mu}_{kj}^{(t+1)} = \frac{1}{n\hat{\pi}_k} E(\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} x_{ij}|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)}),$$

$$\hat{\Sigma}_{kjj'}^{(t+1)} = \frac{1}{n\hat{\pi}_k} E(\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} x_{ij}x_{ij'}|\boldsymbol{x}_i^{obs}; \boldsymbol{\theta}_k^{(t)}) - \hat{\mu}_{kj}^{(t+1)}\hat{\mu}_{kj'}^{(t+1)}.$$

Although a mixture model has great flexibility in modeling, a restriction on the number of components K is still required because, along with an increase in the number of parameters, the estimation of these parameters from the training data might imply a greater variance for each of the parameters. In this study, the Bayesian Information Criterion (BIC) [36] is employed. The BIC can be written as

$$BIC \equiv -2L(\hat{\Theta}|\boldsymbol{x}^{obs}) + \nu_K \log(n_{tr})$$

where $L(\hat{\Theta}|\boldsymbol{x}^{obs})$ is the maximized log-likelihood function given the observed data, ν_K is the number of parameters to be estimated in the assumed model, and n_{tr} is the number of observations in training set. The target is to find that ν_K which minimizes BIC, and then a reasonable number of components K is obtained.

4.6.2 Two Imputation Methods via the GMM

With a reasonable density estimation method, various imputation schemes are possible. DiZio et al. [23] point out that for the preservation of the covariance structure, the Random Draw (RD) method is preferable over the Conditional Mean method (introduced by Nielsen [37]) based on the GMM assumption. The estimates of the Gaussian mixture model parameters are obtained as:

$$f(\boldsymbol{x}_i;\Theta) = \sum_{k=1}^{K} \pi_{ik} N_p(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\Sigma_k). \tag{23}$$

In practice, the random drawing of a value $\boldsymbol{x}_i^{mis}$ from the distribution of

$$f(\boldsymbol{x}_i^{mis}|\boldsymbol{x}_i^{obs};\Theta) = \sum_{k=1}^{K} \pi_{ik} N_p(\boldsymbol{x}_i^{mis}|\boldsymbol{x}_i^{obs};\Theta), \tag{24}$$

could be accomplished in two simple steps: First, draw a value k from the multinomial distribution $Multi(1;\hat{\pi}_{i1},\ldots,\hat{\pi}_{iK})$; then, given k, generate a random value from the p-variate conditional Gaussian distribution $N_p(\boldsymbol{x}_i^{mis}|\boldsymbol{x}_i^{obs};\Theta)$ as the imputation of the missing value.

HD methods are generally preferred over other imputation techniques because of low operational cost, reduced nonresponse bias on univariate statistics, and univariate plausibility (i.e., use of existing values in current dataset). On the other hand, donor-based imputation can produce attenuation of associations [14]. The main principle of the HD method is to use the current set of existing

scores (donors) to provide imputation values for the incomplete vectors (recipients), based on some reasonable rules like conditional distribution or distance measurement. However, if the current set is not large enough, recipients would only have limited donors to choose from, and this will reduce the accuracy of imputation. In the experiments, a bigger simulated dataset ($n_{sim} = 10n_{tr}$) based on the estimation of the mixture model parameters is used as the "imputation pool".

The rule of choosing donors can be either random or based on some distance function. In this work the Euclidean distance measurement d is employed to find the best donor for an incomplete observation. Recall that the distance measure d between two observations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ has been defined as:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{h \in O_i \cap O_j} (x_{ih} - x_{jh})^2. \tag{25}$$

The Hot Deck Imputation procedure can be described in the following steps:

1) Use the estimated parameters, Θ, of GMM to simulate a dataset $\mathbf{D}^{\mathbf{sim}}$, having a larger size than $\mathbf{D}^{\mathbf{tr}}$;
2) For each observation $\boldsymbol{x}_i$, apply the distance function d to find the nearest neighbor in the simulated set $\mathbf{D}^{\mathbf{sim}}$;
3) The missing attributes $\boldsymbol{x}_i^{mis}$ are imputed by the corresponding attributes from the nearest neighbors taken from $\mathbf{D}^{\mathbf{sim}}$.

5 Experiments and Results

5.1 The MSU Database

The Michigan State University (MSU) database used in this study, contains 500 genuine and 12,250 imposter *score vectors*. Take the i-th *score vector* as an example, It is a 3-tuple: (x_{i1}, x_{i2}, x_{i3}), where x_{i1}, x_{i2} and x_{i3} correspond to the match scores obtained from face, fingerprint and hand-geometry matchers, respectively. The detail of the database has been described by Ross and Jain [25]. The fingerprint and face data were obtained from user set I consisting of 50 users. Each user was asked to provide five face images and five fingerprint impressions (of the same finger). This data was used to generate 500 (50*10) genuine scores and 12,250 (50*5*49) imposter scores for each modality. The hand geometry data was collected separately from user set II which also consists of 50 users. This also resulted in 500 genuine scores and 12,250 imposter scores for this modality. Each user in set I was randomly paired with a user in set II. Thus the corresponding genuine and imposter scores for all three modalities were available for testing.

It should be noted that the sample sizes of genuine scores and imposter scores are highly imbalanced in this database. Byon et al. [26] demonstrate that, when the class sizes are highly imbalanced, classification methods tend to strongly favor the majority class, resulting in very low detection accuracy of the minority class.

In order to simplify the problem and retain generality, the proportion of genuine score and imposter score is fixed at 1:4 in this study. This means a total of 500 genuine scores and 2000 imposter scores are randomly selected from the original database. Figure 4 shows the density of the selected dataset and the recognition performance of each modality.

5.2 Generation of Missing Data

In order to evaluate the performance of imputation methods, missing entries were synthetically introduced into a complete (that has no missing data) match score matrix. There are two different ways that are widely used to introduce missing data: the histogram-based scheme and the rate-based scheme [41].

In the histogram-based scheme, histograms are produced for each attribute, and then entries are removed from the complete matrix based on these histograms. In this case, the histogram of the artificially missing entries is similar to that of the original matrix. In the rate-based scheme, a specific percentage of the entries are randomly selected and then removed from the complete score matrix.

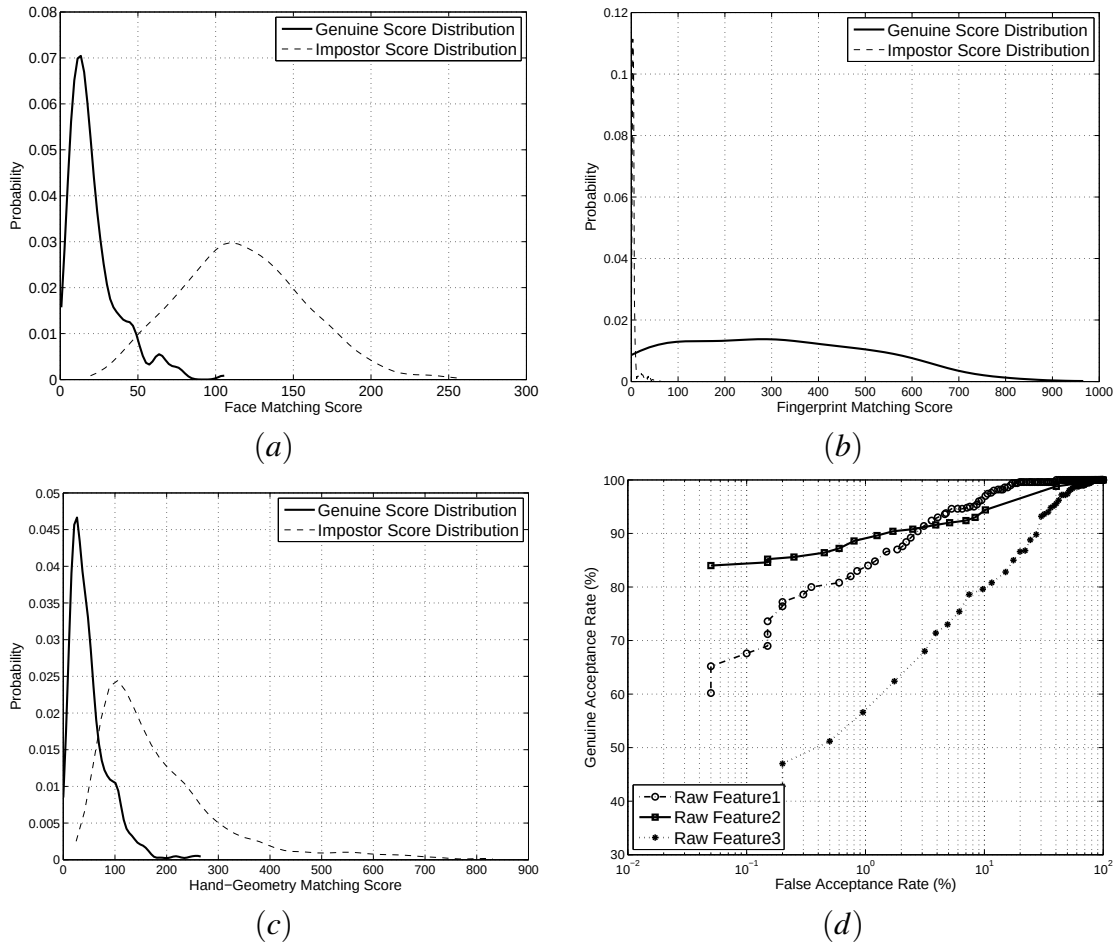

Figure 4: Density plots of the genuine and imposter scores in the selected dataset: (a) Face; (b) Fingerprint; (c) Hand-Geometry; (d) ROC curves for the 3 modalities.

The former cannot be used in this work because the histograms or the estimates of densities are also used by some of the imputation methods, such as the GMM-based methods. If the histogram from the original score matrix fits the model assumed by an imputation method, the artificially missing data will also fit the assumed model well, and this imputation method will result in an optimized performance. Therefore, the rate-based scheme was used to generate missing data in the following experiments.

Figure 6 illustrates the construction of training sets and test sets used in this study. 50% *score vectors* were first randomly selected from the dataset as the training set. The proportion of genuine scores to imposter scores was set to 1:4. The remaining *score vectors* were used as the test set. Next, for each modality, 10% of the scores were randomly removed from the test set, in order to artificially generate the missing data while making sure that each observation contained at least one observed score. As a result, a dataset with 50% of the observations as the training set and a 10% missing rate for the test set was generated.

Two different sizes for the training set, that were 10% and 50% of the entire dataset, were used for comparison. Similarly, two missing rates, 10% and 50%, were specified for the test set. As a result, four different datasets encompassing different training rates and missing rates were generated. Figure 5 illustrates the scatter plots of two such test sets. It indicates that the two classes are reasonably separated in three dimensional space and therefore, a relatively simple fusion method can perform well on this dataset.

As mentioned in the previous sections, although the above procedure to generate missing data is completely random and the datasets appear to conform to the MCAR scenario, the MAR assumption or the MNAR assumption may be more appropriate in operational data.

5.3 Transformation before Imputation

From the density plots of each attribute in Figure 4, it is noted that the dataset has an obvious deviation from the multivariate normal assumption. So certain transformations have to be applied to accommodate methods which assume normality. However, according to our study (not shown here), the transformation of the score matrix will distort the Euclidean distance between *score vectors*, and degrade the accuracy of the density estimation process. This is also true for any row-wise or column-wise normalization performed before imputation. Nevertheless, transformation does bring some benefits to the MLE method.

After applying the MLE imputation on the generated test set with missing

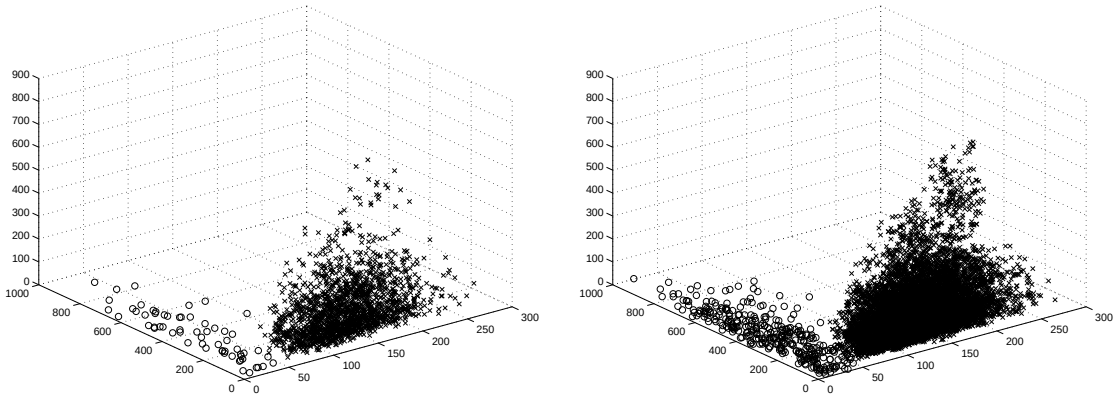

Figure 5: Scatter plots of the original test set ('o': Genuine and 'x': Imposter): (a) When 50% of the MSU database is used for testing; (b) When 90% of the MSU database is used for testing.

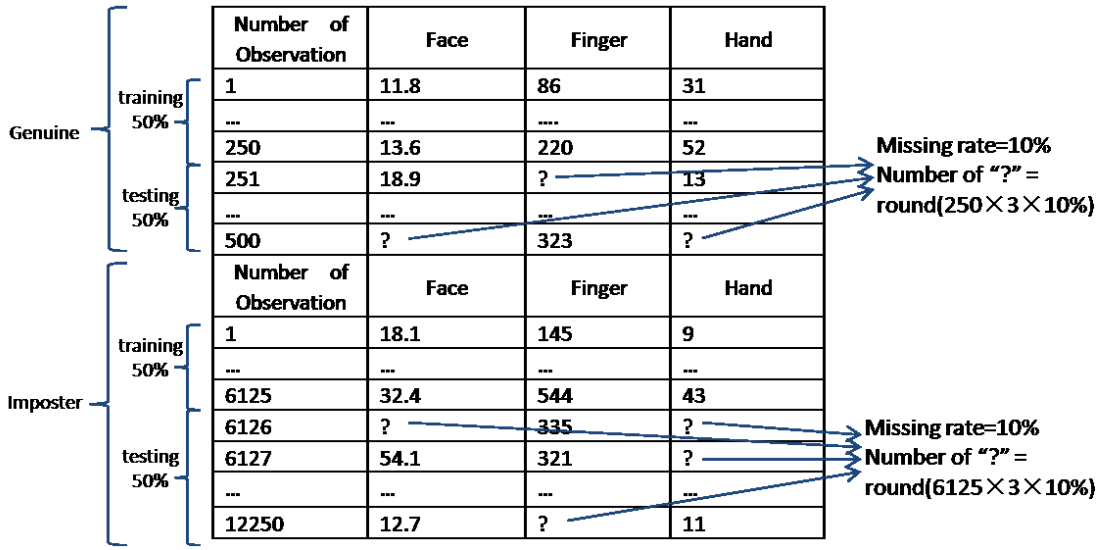

Figure 6: Generation of the datasets used in the experiments. Here, 50% of the dataset is used as training data, and a missing rate of 10% is specified for the test set.

data, we observed that some imputed values were negative, although all the match scores in the original dataset are positive. These unexpected negative scores impact the recognition performance. This phenomenon does not occur when other methods are used. Since the MLE method uses the regression coefficients from ML estimation to predict the final imputation, some of those coefficients could be negative resulting in non-positive predictions.

In order to solve this problem, a square root transformation was applied before processing using the MLE method. After the imputation procedure, the imputed datasets were transformed back by squaring all the values. Figures 7 and 8 show the improved performance by using square root transformation on different training sets, especially at a lower FAR level.

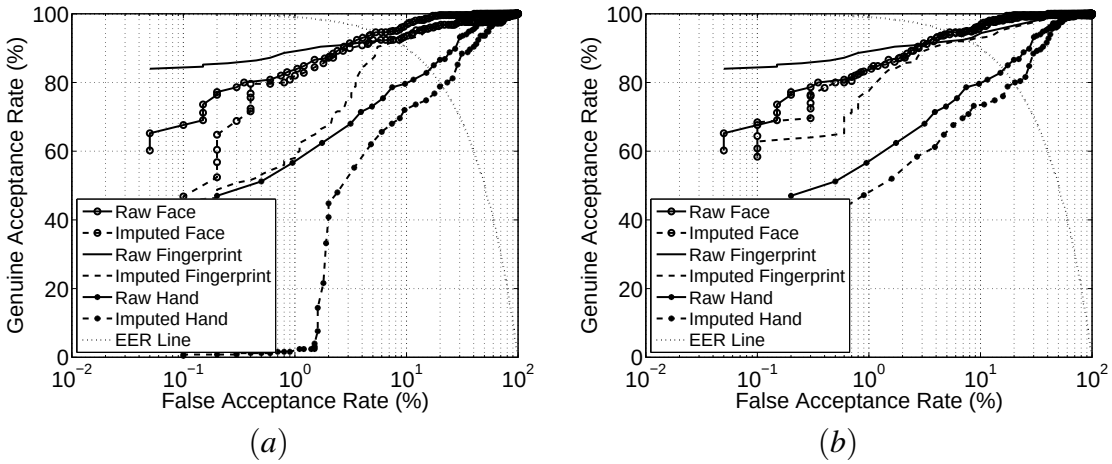

Figure 7: ROC curves after using MLE imputation. Here, 50% of the dataset is employed as training set, and a missing rate of 10% is specified for the test set: (a) before transformation; (b) after transformation.

5.4 Comments on Mean Imputation

Figures 9 (a) and (b) show a surprisingly good performance at a lower missing rate (10%) when the mean imputation scheme is used. However, when the missing rate becomes larger, the performance decreases sharply.

Unlike those studies using the weighted mean, the overall mean was used in this study which ignored the different sample sizes of the two classes on purpose.

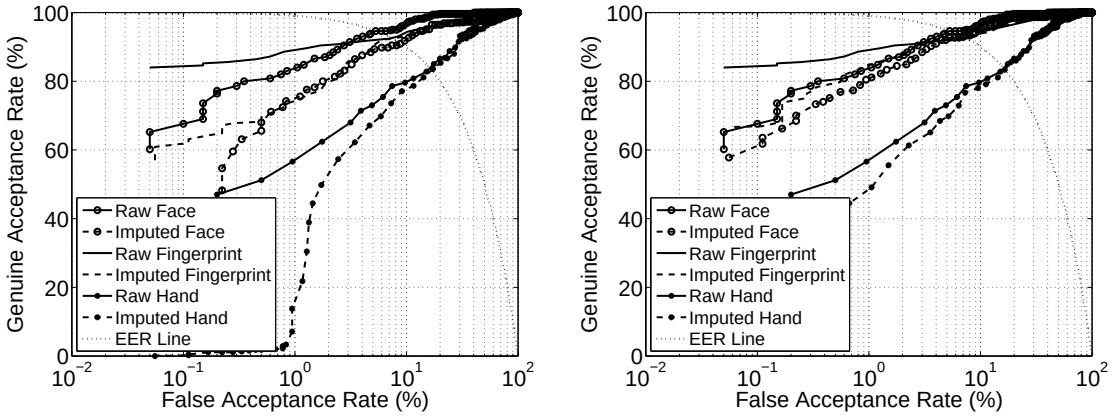

Figure 8: ROC curves after using MLE imputation. Here, 10% of the dataset is employed as training set, and a missing rate of 10% is specified for the test set: (a) before transformation; (b) after transformation.

It is because, in practice, researchers cannot guarantee the proportion of the genuine users and imposters. The scale of scores from the two classes also impacts the mean value significantly, and it greatly depends on which biometric classifier is being used to generate the match scores.

In a complete training set, the target/label value of each vector is available, so it is possible to build two different models using the scores from two classes, separately. Then the average of the results from those two models can be used as the final imputation. However, this approach has similar weaknesses as the mean imputation scheme, and from the ROC curves, it can be concluded that this kind of imputation is not beneficial.

5.5 Random Draw and Hot Deck via the GMM

In Figures 10 and 11, it is observed that both the imputation methods based on the Gaussian mixture model perform fairly well at a 50% training rate. But when the training sample size reduces, the performance of random draw (RD) imputation decreases sharply. The possible reason has to do with the nuances of the imputation process of RD. Recall the value k which was drawn from $Multi(1; \hat{\pi}_{i1}, \ldots, \hat{\pi}_{iK})$, that played a critical role in the process, because the final imputed value depended upon the component that was chosen. A slightly biased value for k will cause an

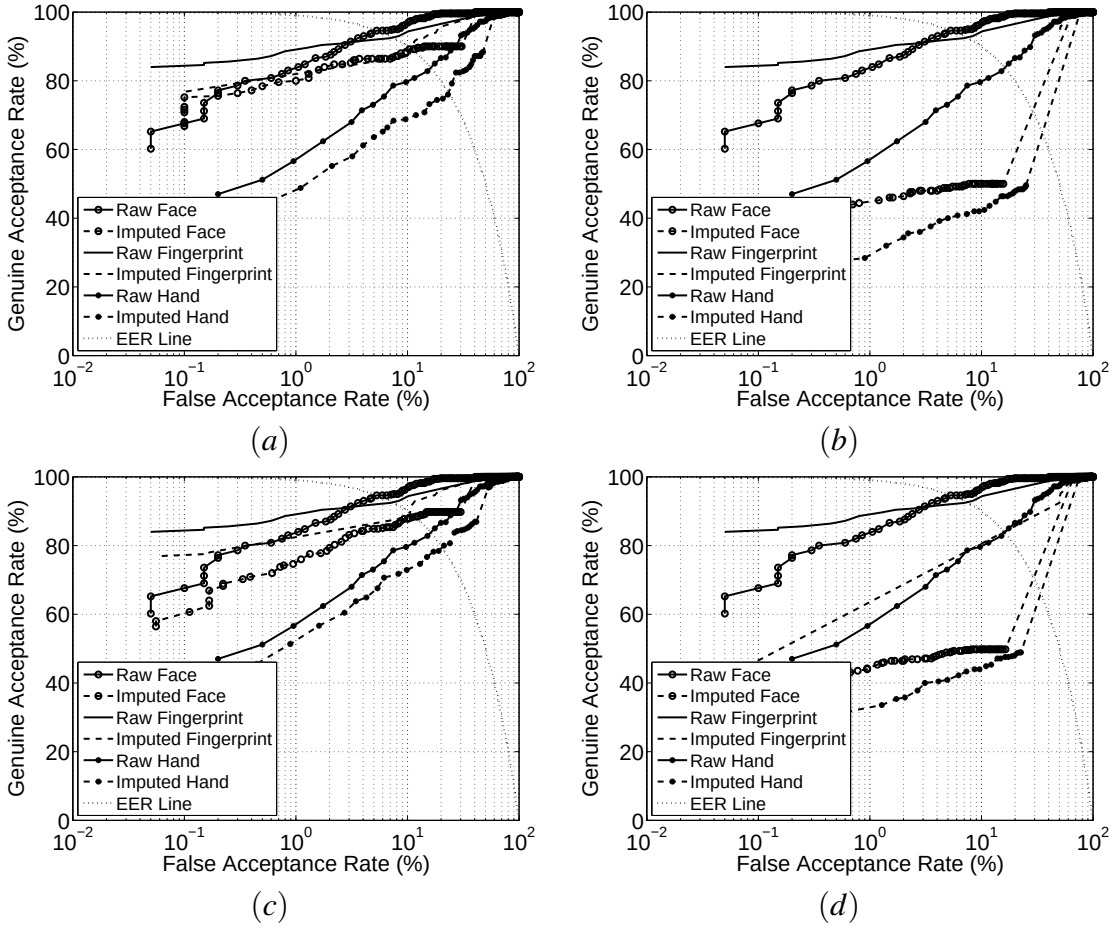

Figure 9: ROC curves after using Mean imputation: (a) a larger training set (50%) and a smaller missing rate (10%); (b) a larger training set (50%) and a larger missing rate (50%); (c) a smaller training set (10%) and a lower missing rate (10%); (d) a smaller training set (10%) and a larger missing rate (50%).

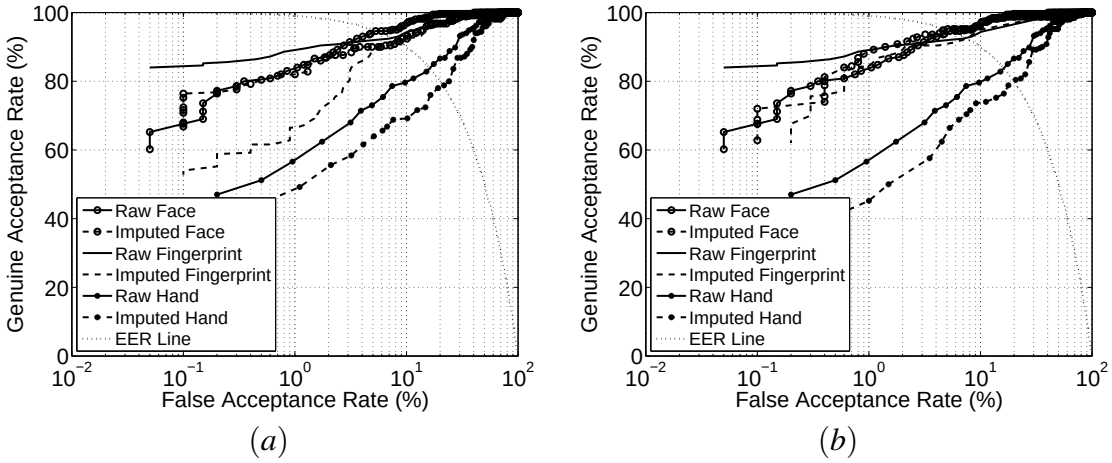

Figure 10: Comparison of RD imputation and HD imputation based on Gaussian mixture models. Here, a larger training set (50%) and a lower missing rate (10%) are specified: (a) RD GMM; (b) HD GMM.

enormous deviation from the true distribution corresponding to the missing score. Therefore, if the size of the training set is not large enough, the RD GMM method is more likely to generate a large bias. In contrast, the Hot Deck method does not rely on the value k, but uses the distance corresponding to the observed part to choose the "closest" neighbors in the simulated data.

5.6 Fusion Results

The min-max normalization scheme followed by the simple sum of scores has been observed to result in reasonable improvement in matching accuracy of a multimodal biometric system [1]. This scheme was used to present the fusion results of various imputation methods in Figures 13 and 14.

From the ROC curves, it is observed that it is difficult for all the methods to maintain good performance when the missing rate is 50%. Although the performance of all schemes decrease sharply at a higher missing rate, the HD GMM shows consistently good performance among the various methods. Even when the training set is small (10%), HD GMM still provides an acceptable EER of less than 10%. Figure 12 is the scatter plot of the imputed dataset after using the HD GMM method. The line patterns are due to multiple observations sharing the same value from the corresponding donor.

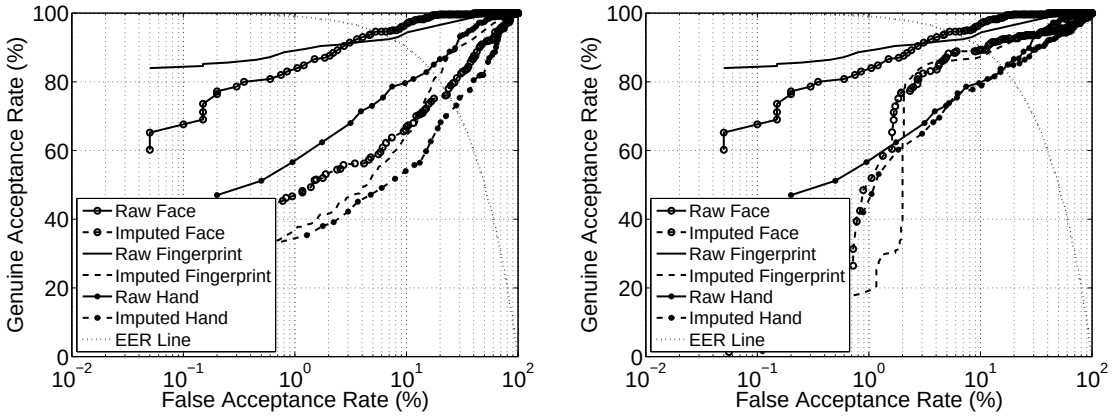

Figure 11: Comparison of RD imputation and HD imputation based on Gaussian mixture models. Here, a smaller training set (10%) and a larger missing rate (50%) are specified: (a) RD GMM; (b) HD GMM.

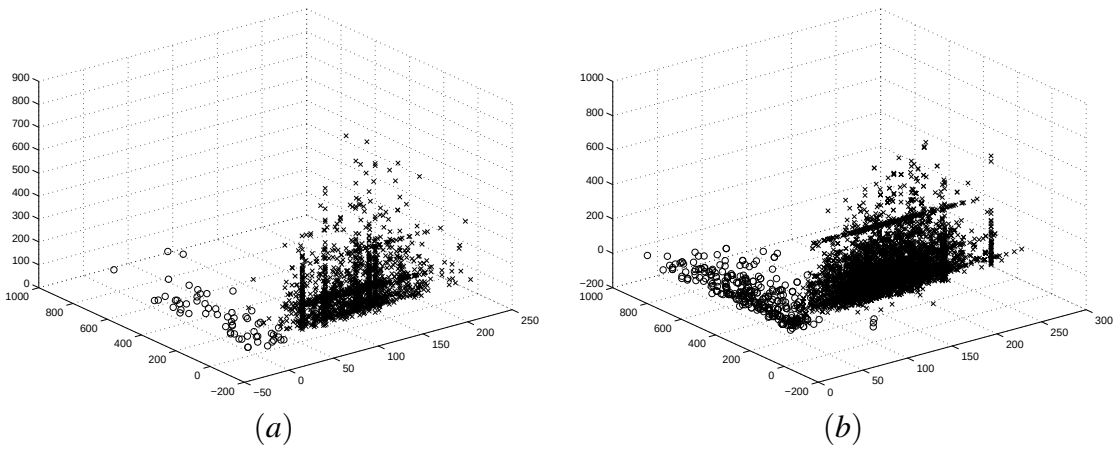

Figure 12: Scatter plots after Hot Deck via the GMM at different training sets and missing rates: (a) 50% as training set, and a missing rate of 10% is specified for the test set; (b) 10% as training set, and a missing rate of 50% is specified for the test set.

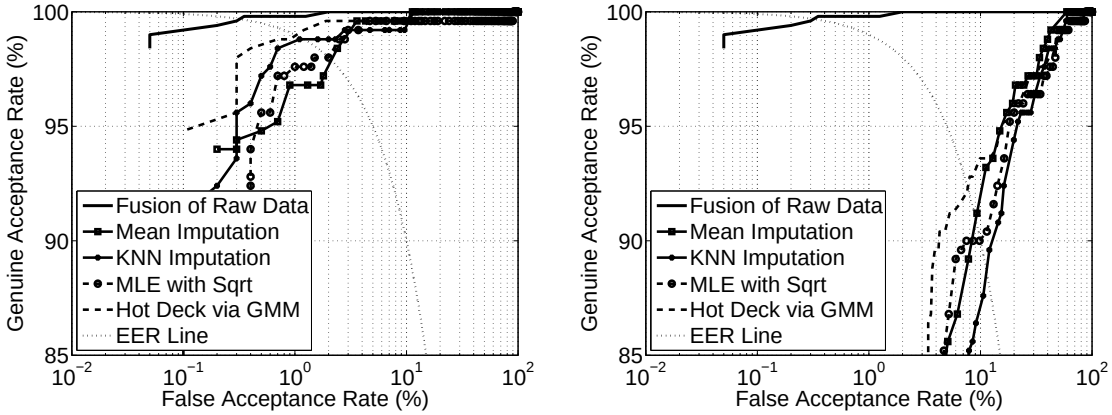

Figure 13: Fusion performance after using different imputation methods. Here, 50% of the dataset is employed as training data: (a) a missing rate of 10% is specified for the test set; (b) a missing rate of 50% is specified for the test set.

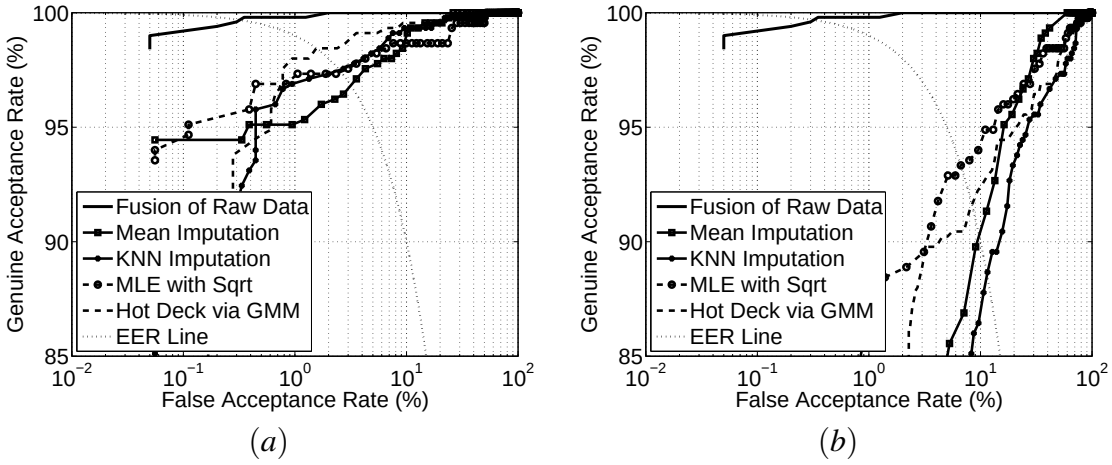

Figure 14: Comparison of different imputation methods. Here, 10% of the dataset is employed as training data: (a) a missing rate of 10% is specified for the test set; (b) a missing rate of 50% is specified for the test set.

Table 1:
The preservation of data structure in MLE method. Here, 10% of the dataset is employed as training data, and a missing rate of 10% is specified for the test set.

	Training set			Test set			Imputed set		
$\hat{\mu}$	98.1	67.0	143.1	97.9	66.3	161.4	89.1	55.3	153.9
$\hat{\Sigma}$	3175	-4402	1525	2783	-4603	2337	2700	-708	1009
	-4402	23901	-5512	-4603	22715	-6503	-708	19452	-3050
	1525	-5512	12816	2337	-6503	16518	1009	-3050	14277

Table 2:
The preservation of data structure in MLE method. Here, 10% of the dataset is employed as training data, and a missing rate of 50% is specified for the test set.

	Training set			Test set			Imputed set		
$\hat{\mu}$	98.1	67.0	143.1	97.9	66.3	161.4	30.4	288.2	59.5
$\hat{\Sigma}$	3175	-4402	1525	2783	-4603	2337	1002	-2767	285
	-4402	23901	-5512	-4603	22715	-6503	-2767	43324	-1448
	1525	-5512	12816	2337	-6503	16518	285	-1448	2499

MLE imputation gives the best performance when the training set is small (at a 10% training rate) in Figure 14. This is because an accurate ML estimate has been obtained, which leads to a good preservation of the covariance matrix in the original dataset. From Tables 1 and 2, it is observed that when the missing rate increases, the accuracy of MLE degrades sharply, and the EER of the imputed set increases from 5% to 8%.

6 Summary

The results in the previous sections indicate that the imputation of missing data through the GMM is a powerful scheme for multimodal biometric fusion. Particularly, imputation via Hot Deck from the simulated dataset generated using the estimated GMM, results in a better recognition performance than the others. Imputation by randomly drawing from the estimated GMM is also a viable option when the training set is large enough to obtain an accurate estimation. Imputation based on Maximum Likelihood Estimation provides an alternative way when the training set is relatively small. In addition, in order to preserve the original scale of the score matrix, certain transformations can be applied when normality is violated.

In the future, the robustness of the assumptions of every method will be further analyzed. This is expected to offer additional guidance on how to choose imputation methods for a particular dataset. Also, more work which combine the imputation methods with score normalization and fusion will be conducted. Finally, these experiments will be repeated on large operational datasets.

References

[1] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems", *Pattern Recognition*, **38**(12), pp. 2270–2285, 2005.

[2] A. Ross, K. Nandakumar, and A. Jain, *Handbook of Multibiometrics*, Springer, Secaucus, NJ, USA, 2006.

[3] G. King, J. Honaker, A. Joseph, and K. Scheve, "Analyzing incomplete political science data: An alternative algorithm for multiple imputation", in *American Political Science Review*, **95**, pp. 49–69, 2001

[4] R. Brown, "Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods", in *Structural Equation Modeling*, **1**, pp. 287–316, 1994

[5] J. Graham, S. Hofer, and D. MacKinnon, "Maimizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures", in *Multivariate Behavioral Research*, **31**, pp.197–218, 1996

[6] O. Fatukasi, J. Kittler, and N. Poh, "Estimation of missing values in multimodal biometric fusion", in *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2008.

[7] J. Quinlan, "Induction of decision trees", in *Machine Learning*, **1**, pp. 81–106, 1986.

[8] O. Lobo, and M. Noneao, "Ordered estimation of missing values for propositional learning", in *Journal of the Japanese Society for Artificial Intelligence*, **1**, pp. 499–503, 2000.

[9] H. Friedman, R. Kohavi, and Y. Yun, "Lazy decision trees", in *the 13th national conference on artificial intelligence*, pp. 717–724, 1996.

[10] R. Little and D. Rubin, in *Statistical Analysis with Missing Data*, Wiley, New York, 1st ed., 1987.

[11] J. Schafer, and J. Graham, "Missing data: Our view of the state of the art", in *Psychological Methods* , 2002.

[12] J. Schafer, and J. Schenker, "Inference with imputed conditional means", in *Journal of the American Statistical Association*, **95**, 144–154, 2000.

[13] J. Dixon, "Pattern recognition with partly missing data", in *IEEE Transactions on Systems, Man and Cybernetics*, **9**(10), pp. 617–621, 1979.

[14] G. Kalton, and D. Kasprzyk, "The treatment of missing survey data", in *Survey Methodology*, **12**(1), pp. 1–16, 1986.

[15] R. Little, "Regression with missing x's: A review", in *Journal of the American Statistical Association*, **87**(420), pp. 1227–1237, 1992.

[16] J. Schafer, *Analysis of incomplete multivariate data*, London, 1997.

[17] D. Rubin, and N. Schenker, "Multiple imputation for interval estimation from simple random samples with ignorable nonresponse", in *Journal of the American Statistical Association*, **81**, pp. 366–374, 1986.

[18] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", in *Journal of the Royal Statistical Society. Series B (Methodological)*, **1**(39), pp. 1–38, 1977.

[19] G. McLachlan, and K. Basford, *Mixture Models: Inference and Applications to Clustering*, Marcel-Dekker, New York, 1988.

[20] G. McLachlan, and T. Krishnan, *The EM algorithm and extensions*, New York: Wiley, 1996

[21] G. McLachlan, and D. Peel, *Finite Mixture Models*, Wiley, NewYork, 2000.

[22] C. Priebe, "Adaptive mixtures", in *Journal of the American Statistical Association*, **89**, pp. 796–806,1994.

[23] M. DiZio, U. Guarnera, and O. Luzi, "Imputation through finite gaussian mixture models", in *Computational Statistics & Data Analysis*, **51**(11), pp. 5305–5316, 2007.

[24] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition", in *IEEE Transactions on Circuits and Systems for Video Technology*, **14**, pp. 4–20, 2004.

[25] A. Ross, and A. Jain, "Information fusion in biometrics", in *Pattern Recognition*, **13**, pp. 2115–2125, 2003.

[26] E. Byon, A. Shrivastava, and Y. Ding, "A classification procedure for highly imbalanced class sizes", in *IIE Transactions*, **4**(42), pp. 288–303,2010.

[27] D. Rubin, *Multiple imputation for nonresponse in surveys*, Wiley, 1987.

[28] K. Nandakumar, A. Jain, and A. Ross, "Fusion in multibiometric identification systems: What about the missing data?", in *IEEE/IAPR International Conference on Biometrics*, pp. 743–752, 2009.

[29] K. Nandakumar, Y. Chen, S. Dass, and A. Jain, "Likelihood Ratio Based Biometric Score Fusion", in *IEEE Transactionson on Pattern Analysis and Machine Intelligence*, **30**(2), pp. 342–347, 2008.

[30] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers", in *Machine Learning*, **29**(2-3), pp. 131–163, 1997.

[31] D. Williams, X. Liao, Y. Xue, L. Carin, and B. Krishnapuram, "On classification with incomplete data", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**(3), pp. 427–436, 2007 .

[32] M. Ramoni, and P. Sebastiani, "Robust learning with missing data", in *Machine Learning*, **45**(2), pp. 147–170, 2001.

[33] D. A. Marker, D. Judkins, and M. Winglee, "Large-scale imputation for complex surveys", in *Survey Nonresponse*, John Wiley and Sons, 1999.

[34] C. Fraley, and A. Raftery, "Model-based clustering, discriminant analysis, and density estimation", in *Journal of the American Statistical Association*, **97**, pp. 611–631, 2002.

[35] D. Titterington, A. Smith, and U. Makov, *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York, 1985.

[36] G. Schwarz, "Estimating the dimension of a model", in *Annals of Statistics*, **6**, pp. 46–464, 1978.

[37] S. Nielsen, "Nonparametric conditional mean imputation", *Journal of Statistical Planning and Inference*, **99**, pp. 129–150,2001 .

[38] J. Marron, and M. Wand, "Exact mean integrated squared error", in *Annals of Statistics*, **20**, pp. 712–736, 1992.

[39] L. Hunt, and M. Jorgensen, "Mixture model clustering for mixed data with missing information", in *Computational Statistics & Data Analysis*, **41**(3-4), pp. 429–440, 2003.

[40] J. Kittler, M. Hatef, and R. Duin, "On combining classifiers", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, pp. 226–239, 1998.

[41] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A bayesian missing value estimation method for gene expression profile data", in *Bioinformatics*, **19**(16), pp. 2088–2096, 2003.