

# On the Use of Conceptual Reconstruction for Mining Massively Incomplete Data Sets\*

Srinivasan Parthasarathy  
Dept. of Computer and Information Science  
Ohio State University  
Columbus, Ohio-43210  
srini@cis.ohio-state.edu

Charu C. Aggarwal  
IBM T. J. Watson Research Center  
Yorktown Heights, NY 10598  
charu@us.ibm.com

August 5, 2002

## Abstract

Incomplete data sets have become almost ubiquitous in a wide variety of application domains. Common examples can be found in climate and image data sets, sensor data sets and medical data sets. The incompleteness in these data sets may arise from a number of factors: in some cases it may simply be a reflection of certain measurements not being available at the time; in others the information may be lost due to partial system failure; or it may simply be a result of users being unwilling to specify attributes due to privacy concerns. When a significant fraction of the entries are missing in all of the attributes, it becomes very difficult to perform any kind of reasonable extrapolation on the original data. For such cases, we introduce the novel idea of conceptual reconstruction, in which we create effective conceptual representations on which the data mining algorithms can be directly applied. The attraction behind the idea of conceptual reconstruction is to use the correlation structure of the data in order to express it in terms of concepts rather than the original dimensions. As a result, the reconstruction procedure estimates only those conceptual aspects of the data which can be mined from the incomplete data set, rather than force errors created by extrapolation. We demonstrate the effectiveness of the approach on a variety of real data sets.

---

\*Both authors contributed equally to this work. This is the extended version of the ACM KDD Conference paper [2].

**Keywords:** incomplete data, missing values, data mining

## 1 Introduction

In recent years, a large number of data sets which are available for data mining tasks are incompletely specified. An incompletely specified data set is one in which a certain percentage of the values are missing. This is because the data sets for data mining problems are usually extracted from real world situations in which either not all measurements may be available or not all the entries may be relevant to a given record. In other cases, where data is obtained from users directly, many users may be unwilling to specify all the attributes because of privacy concerns [3, 16]. In many cases, such situations result in data sets in which a large percentage of the entries are missing. This is a problem since most data mining algorithms assume that the data set is completely specified.

There are a variety of solutions which can be used in order to handle this mismatch for mining massively incomplete data sets. For example, if the incompleteness occurs in a small number of rows, then such rows may be ignored. Alternatively, when the incompleteness occurs in a small number of columns, then only these columns may be ignored. In many cases, this reduced data set may suffice for the purpose of a data mining algorithm. None of the above techniques would work for a data set which is massively incomplete, because it would lead to the ignoring of almost all the records and attributes. Common solutions to the missing data problem include the use of imputation, statistical or regression based procedures [4, 5, 10, 11, 19, 20, 15, 17] in order to estimate the entries. Unfortunately, these techniques are also prone to estimation errors with increasing dimensionality and incompleteness. This is because when a large percentage of the entries are missing, each attribute can be estimated to a much lower degree of accuracy. Furthermore, some

attributes can be estimated to a much lower degree of accuracy than others, and there is no way of knowing a-priori which estimations are the most accurate. A discussion and examples of the nature of the bias in using direct imputation based procedures may be found in [7].

We note that any missing data mechanism would rely on the fact that the attributes in a data set are not independent from one another, but that there is some predictive value from one attribute to another. If the attributes in a data set are truly uncorrelated, then any loss in attribute entries leads to a true loss of information. In such cases, missing data mechanisms cannot provide any estimate to the true value of a data entry. Fortunately, this is not the case in most real data sets, in which there are considerable redundancies and correlations across the data representation.

In this paper, we discuss the novel technique of conceptual reconstruction, in which we express the data in terms of the salient concepts of the correlation structure of the data. This conceptual structure is determined using techniques such as Principal Component Analysis [8]. These are the directions in the data along which most of the variance occurs, and are also referred to as the *conceptual directions*. We note that even though a data set may contain thousands of dimensions, the number of concepts in it may be quite small. For example, in text data sets the number of dimensions (words) are over 100,000 but there are often only 200-400 salient concepts [14, 9]. In this paper, we will provide evidence of the claim that even though predicting the data along arbitrary directions (such as the original set of dimensions) is fraught with errors. This problem is especially true in massively incomplete data sets in which the errors caused by successive imputation add up and result in a considerable drift from the true results. On the other hand, the components along the conceptual directions can be predicted quite reliably. This is because the conceptual reconstruction method uses these redundancies in an effective way so as to estimate whatever conceptual representations are reliably possible rather than force extrapolations on the original set of attributes. As the data dimensionality increases, even massively incomplete data sets can be

modeled by using a small number of conceptual directions which capture the overall correlations in the data. Such a strategy is advantageous, since it only tries to derive whatever information is truly available in the data. We note that this results in some loss of interpretability with respect to the original dimensions; however the aim of this paper is to be able to use available data mining algorithms in an effective and accurate way. The results in this paper are presented only for the case when the data is presented in explicit multidimensional form and are not meant for the case of latent variables.

This paper is organized as follows. The remainder of this section provides a formal discussion of the contributions of this paper. In the next section we will discuss the basic conceptual reconstruction procedure, and provide intuition on why it should work well. In section 3, we provide the implementation details. Section 4 contains the empirical results. The conclusions and summary are contained in section 5.

## **1.1 Contributions of this paper**

This paper discusses a technique for mining massively incomplete data sets by exploiting the correlation structure of data sets. We use the correlation behavior in order to create a new representation of the data which predicts only as much information as can be reliably estimated from the data set. This results in a new full dimensional representation of the data which does not have a one-to-one mapping with the original set of attributes. However this new representation reflects the available concepts in the data accurately and can be used for many data mining algorithms, such as clustering, similarity search or classification.

## 2 An Intuitive Understanding of Conceptual Reconstruction

In order to facilitate further discussion, we will define the percentage of attributes missing from a data set as the *incompleteness factor*. The higher the incompleteness factor, the more difficult it is to obtain any meaningful structure from the data set. The conceptual reconstruction technique is tailored towards mining massively incomplete data sets for high dimensional problems. As indicated earlier, the attributes in high dimensional data are often correlated. This results in a natural conceptual structure of the data. For instance, in a market basket application, a concept may consist of groups or sets of closely correlated items. A given customer may be interested in particular kinds of items which are correlated and may vary over time. However, her conceptual behavior may be much clearer at an *aggregate level*, since one can classify the *kinds* of items that she is most interested in. In such cases, even when a large percentage of the attributes are missing, it is possible to obtain an idea of the conceptual behavior of this customer.

A more mathematically exact method for finding the aggregate conceptual directions of a data set is Principal Component Analysis (PCA) [8]. Consider a data set with  $N$  records and dimensionality  $d$ . In the first step of the PCA technique, we generate the covariance matrix of the data set. The covariance matrix is a  $d*d$  matrix in which the  $(i, j)$ th entry is equal to the covariance between the dimensions  $i$  and  $j$ . In the second step we generate the eigenvectors  $\{\bar{e}_1 \dots \bar{e}_d\}$  of this covariance matrix. These are the directions in the data, which are such that when the data is projected along these directions, the second order correlations are zero. Let us assume that the eigenvalue for the eigenvector  $\bar{e}_i$  is equal to  $\lambda_i$ . When the data is transformed to this new axis-system, the value  $\lambda_i$  is also equal to the variance of the data along the axis  $\bar{e}_i$ . The property of this transformation is that most of the variance is retained in a small number of eigenvectors corresponding to the largest values of  $\lambda_i$ . We retain the  $k < d$  eigenvectors which correspond to the largest eigenvalues. An

important point to understand is that the removal of the smaller eigenvalues for highly correlated high dimensional problems results in a new data set in which much of the noise is removed [13], and the qualitative effectiveness of data mining algorithms such as similarity search is improved [1]. This is because these few eigenvectors correspond to the conceptual directions in the data along which the non-noisy aspects of the data are preserved. One of the interesting results that this paper will show is that these relevant directions are also the ones along which the conceptual components can be most accurately predicted by using the data in the neighborhood of the relevant record. We will elucidate this idea with the help of an example. Throughout this paper, we will refer to a retained eigenvector as a *concept* in the data.

## 2.1 On the Effects of conceptual reconstruction: An Example

Let  $\mathcal{Q}$  be a record with some missing attributes denoted by  $B$ . Let the specified attributes be denoted by  $A$ . Note that in order to estimate the conceptual component along a given direction, we find a set of neighborhood records based on the known attributes only. These records are used in order to estimate the corresponding conceptual coordinates. Correspondingly, we define the concept of an  $(\epsilon, A)$ -neighborhood of a data point  $\mathcal{Q}$ .

**Definition 1** *An  $(\epsilon, A)$ -neighborhood of a data point  $\mathcal{Q}$  is the set of records from the data set  $D$  such that the distance of each point in it from  $\mathcal{Q}$  based on only the attributes in  $A$  is at most  $\epsilon$ . We shall denote this neighborhood by  $\mathcal{S}(\mathcal{Q}, \epsilon, A)$ .*

Once we have established the concept of  $(\epsilon, A)$ -neighborhood, we shall define the concept of  $(\epsilon, A, \bar{e})$ -predictability along the eigenvector  $\bar{e}$ . Intuitively, the predictability along an eigenvector  $\bar{e}$  is a measure of how closely the value along the eigenvector  $\bar{e}$  can be predicted using only the behavior of the neighborhood set  $\mathcal{S}(\mathcal{Q}, \epsilon, A)$ .

**Definition 2** For a given eigenvector  $\bar{e}$ , let  $\mathcal{N}$  be the coordinates along  $e$  in the transformed domain for the set  $\mathcal{S}(Q, \epsilon, A)$ . Let  $\mu$  be the mean of the elements in  $\mathcal{N}$  and  $\sigma$  be the standard deviation. The  $(\epsilon, A, e)$ -predictability of a data point  $Q$  is defined as the ratio  $|\mu/\sigma|$ .

Since the above ratio measures the mean to standard deviation ratio, greater amount of certainty in the accuracy of the prediction is obtained when the ratio is high. We note that the value of the predictability has been defined in this way, since we wish to make the definition scale invariant. We

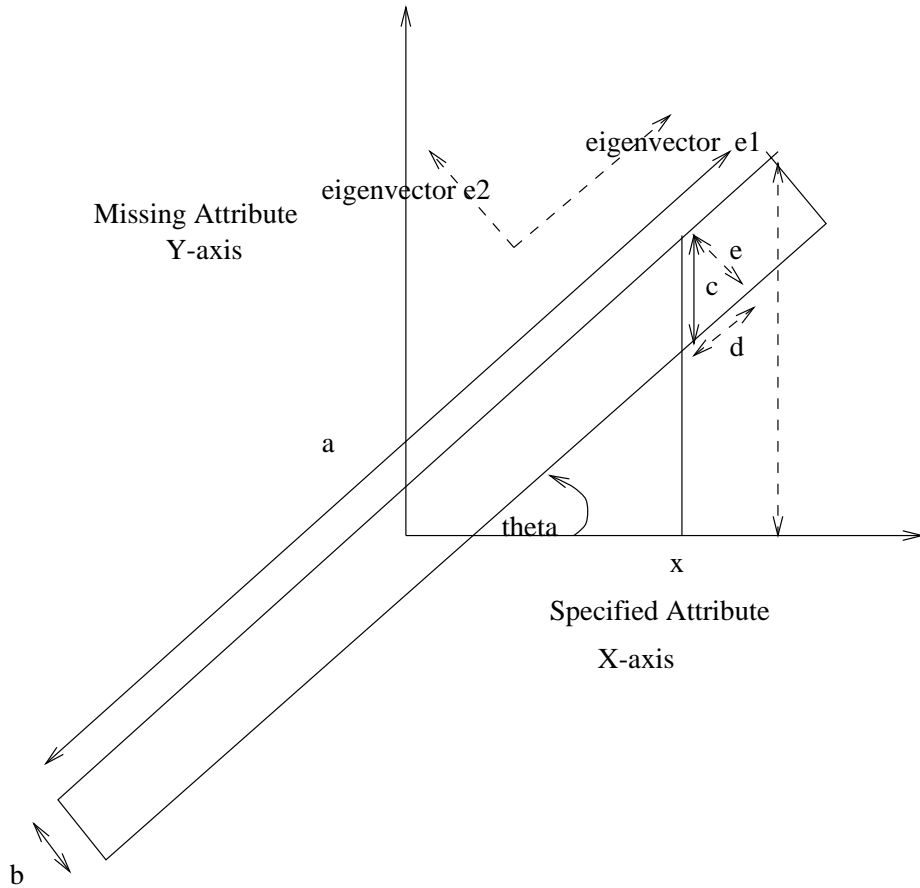


Figure 1: Predictability for a Simple Distribution

shall now illustrate with the help of an example, why  $(\epsilon, A, \bar{e})$ -predictability of eigenvector  $\bar{e}$  is higher when the corresponding eigenvalue is larger. In Figure 1, we have shown a two-dimensional example for the case when a data set is drawn from a uniformly distributed rectangular distribution centered

at the origin. We also assume that this rectangle is banked at an angle  $\theta$  from the  $X$ -axis and the sides of this rectangle are of lengths  $a$  and  $b$  respectively. Since the data is uniformly generated within the rectangle, if we were to perform PCA on the data records, we would obtain eigenvectors parallel to the sides of the rectangle. The corresponding eigenvalues would be proportional to  $a^2$  and  $b^2$  respectively. Without loss of generality, we may assume that  $a > b$ . Let us assume that the eigenvectors in the corresponding directions are  $\bar{e}_1$  and  $\bar{e}_2$  respectively. Since the variance along the eigenvector  $\bar{e}_1$  is larger, it is clear that the corresponding eigenvalue is also larger. Let  $\mathcal{Q}$  be a data point for which the  $X$ -coordinate  $x$  is shown in Figure 1. Now, the set  $\mathcal{S}(\mathcal{Q}, \epsilon, \{X\})$  of data records which is closest to the point  $\mathcal{Q}$  based on the coordinate  $X = x$  is in a thin strip of width  $2\epsilon$  centered at the segment marked with a length of  $c$  in Figure 1. In order to make an intuitive analysis without edge effects, we will assume that  $\epsilon \rightarrow 0$ . Therefore, in the diagram for Figure 1, we have just used a vertical line which is a strip of width zero. Then, the standard deviation of the records in  $\mathcal{S}(\mathcal{Q}, \epsilon, \{X\})$  along the  $Y$  axis is given by  $c/\sqrt{12} = b \cdot \secant(\theta)/\sqrt{12}$  using the formula for a uniform distribution along an interval<sup>1</sup> of length  $c$ . The corresponding components along the eigenvectors  $\bar{e}_1$  and  $\bar{e}_2$  are  $d/\sqrt{12} = |c \cdot \text{sine}(\theta)/\sqrt{12}|$  and  $e/\sqrt{12} = |c \cdot \text{cosine}(\theta)/\sqrt{12}|$  respectively. The corresponding means along the eigenvectors  $\bar{e}_1$  and  $\bar{e}_2$  are given by  $|x \cdot \text{sec}(\theta)|$  and 0 respectively. Now we can substitute for the mean and standard deviation values in Definition 2 in order to obtain the following results:

1. The  $(\epsilon, \{X\}, \bar{e}_1)$ -predictability of the data point  $\mathcal{Q}$  is  $|x/b \cdot \text{sine}(\theta)|$ .
2. The  $(\epsilon, \{X\}, \bar{e}_2)$ -predictability of the data point  $\mathcal{Q}$  is 0.

Thus, this example illustrates that predictability is much better in the direction of the larger eigenvector  $\bar{e}_1$ . Furthermore, with reduced value of  $\theta$ , predictability along this eigenvector (which has an angle  $\theta$  with the specified attribute) improves. We will now proceed to formalize some of

---

<sup>1</sup>Details may be found in [6]

these intuitive results.

## 2.2 Key Intuitions

**Intuition 1** *The larger the value of the eigenvalue  $\lambda_i$  for  $\bar{e}_i$ , the greater the relative predictability of the conceptual component along  $\bar{e}_i$ .*

This intuition summarizes the implications of the example discussed in the previous section. In the previous example, it was also clear that the level of accuracy with which the conceptual component could be predicted along an eigenvector was dependent on the angle with which the eigenvector was banked with the axis. In order to formalize this notion we introduce some additional notations. Let  $(b_1, \dots, b_n)$  correspond to the unit direction vector along a principle component (eigenvector) in a data set with  $n$  attributes. Clearly the larger the value of  $b_i$ , the more the variance of the projection of attribute  $i$  along the principle component  $i$  and vice versa.

**Intuition 2** *For a given vector  $\bar{e}_i$ , the larger the weighted ratio  $\sqrt{\sum_{i \in A} b_i^2} / \sqrt{\sum_{i \in B} b_i^2}$ , the greater the relative predictability of the conceptual component along  $\bar{e}_i$ .*

## 3 Details of the Conceptual Reconstruction Technique

In this section we outline the overall conceptual reconstruction procedure along with key implementation details. More specifically, two fundamental problems with the implementation need to be discussed. In order to find the conceptual directions, we first need to construct the covariance matrix of the data. Since the data is massively incomplete, this matrix cannot be directly computed but only estimated. This needs to be carefully thought out in order to avoid bias in the process of determining the conceptual directions. Second, once the conceptual vectors (principal components) are found, we will work out the best methods for finding the components of records with missing

data along these vectors.

### 3.1 The Conceptual Reconstruction Algorithm

The overall conceptual reconstruction algorithm is illustrated in Figure 2. For the purpose of the following description, we will assume without loss of generality that the data set is centered at the origin.

Step 1: Compute Covariance Matrix  $M$  from data set  $D$ .  
 Step 2: Compute Eigenvectors  $\{e_1 \dots e_d\}$  of Covariance matrix  $M$  with eigenvalues  $\lambda_1 \geq \lambda_2 \dots \geq \lambda_d$ .  
 Step 3: Retain the subset of eigenvectors  $\{\bar{e}_1 \dots \bar{e}_m\}$  with largest values of  $\lambda_i$ .  
 Step 4: For each record  $Q$  in  $D$  with specified attributes  $A$  and missing attributes  $B$   
 Step 4A: For each retained eigenvector  $\bar{e}_i$   
 Step 4A1: Let  $Y_A^i$  be the projection of known attribute set  $A$  of  $Q$  on  $\bar{e}_i$   
 Step 4A2: Compute  $K$  records  $C$  which are closest to  $Q$  using the Euclidean distance on the attribute set  $A$   
 Step 4A3: Let  $Y_B^i$  be the average projection of attribute set  $B$  of the records in  $C$  on  $\bar{e}_i$   
 Step 4A4: Set the conceptual coordinate along  $\bar{e}_i$  of data record  $Q$  to  $Y_A^i + Y_B^i$

Figure 2: Conceptual Reconstruction Procedure

The goal in Step 1 is to compute the covariance matrix  $M$  from the data. Since the records have missing data, the covariance matrix cannot be directly constructed. Therefore, we need methods for estimating this matrix. In a later subsection, we will discuss methods for computing this matrix  $M$ . Next, we compute the eigenvectors of the covariance matrix  $M$ . The covariance matrix for a data set is positive semi-definite and can be expressed in the form  $M = PNP^T$ , where  $N$  is a diagonal matrix containing the eigenvalues  $\lambda_1 \dots \lambda_d$ . The columns of  $P$  are the eigenvectors  $\bar{e}_1 \dots \bar{e}_d$ , which form an orthogonal axis-system. We assume without loss of generality, that the eigenvectors are sorted so that  $\lambda_1 \geq \lambda_2 \geq \dots \lambda_d$ . To find these eigenvectors, we rely on the popular Householder reduction to tridiagonal form and then apply the QL transform [8], which is the fastest known method to compute eigenvectors for symmetric matrices.

Once these eigenvectors have been determined, we decide to retain only those which preserve

the greatest amount of variance from the data. Well known heuristics for deciding the number of eigenvectors to be retained may be found in [8]. Let us assume that a total of  $m \leq d$  eigenvectors  $\bar{e}_1 \dots \bar{e}_m$  are retained. Next we set up a loop for each retained eigenvector  $\bar{e}_i$  and incompletely specified record  $Q$  in the database. We assume that the set of known attributes in  $Q$  are denoted by  $A$ , whereas the set of unknown attributes are denoted by  $B$ . We first find the projection of the specified attribute set  $A$  onto the eigenvector  $\bar{e}_i$ . We denote this projection by  $Y_A^i$ , whereas the projection for the unspecified attribute set  $B$  is denoted by  $Y_B^i$ . Next, the  $K$  nearest records to  $Q$  are determined using the Euclidean distance on the attribute set  $A$ . The value of  $K$  is a user-defined parameter and should typically be fixed to a small percentage of the data. For the purposes of our implementation, we set the value of  $K$  consistently to about 1% of the total number of records, subject to the restriction that  $K$  was at least 5. This representative set of records is denoted by  $C$  in Figure 2. Once the set  $C$  have been computed we estimate the missing component  $Y_B^i$  of the projection of  $Q$  on  $\bar{e}_i$ . For each record in the set  $C$  we compute its projection along  $e_i$  using the attribute set  $B$ . The average value of these projections is then taken to be the estimate  $Y_B^i$  for  $Q$ . Note, that it is possible that the records in  $C$  may also have missing data for the attribute set  $B$ . For such cases, only the components from the specified attributes are used in order to calculate the  $Y_B^i$  values for that record. The conceptual coordinate of the record  $Q$  along the vector  $\bar{e}_i$  is given by  $Y^i = Y_A^i + Y_B^i$ . Thus, the conceptual representation of the record  $Q$  is given by  $(Y^1 \dots Y^m)$ .

### 3.2 Estimating the Covariance Matrix

At first sight, a natural method to find the covariance between a given pair of dimensions  $i$  and  $j$  in the data set is to simply use those entries which are specified for both dimensions  $i$  and  $j$  and compute the covariance. However, this would often lead to considerable bias, since the entries which are missing in the two dimensions are also often correlated with one another. Consequently,

the covariance between the specified entries is not a good representative of the overall covariance in a real data set. This is especially the case for massively incomplete data sets in which the bias may be considerable. By using dimensions on a pairwise basis only, such methods ignore a considerable amount of information that is hidden in the correlations of either of these dimensions with the other dimensions for which fully specified values are available.

In order to harness this hidden information, we use a procedure in which we assume a distribution model for the data and estimate the parameters of this model in terms of which the covariances are expressed. Specifically, we use the technique discussed in [10], which assumes a Gaussian model for the data, and estimates the covariance matrix for this Gaussian model using an Expectation Maximization (EM) algorithm. Even though some inaccuracy is introduced because of this modeling assumption, it is still better than the vanilla approach of pairwise covariance estimation. To highlight some of the advantages of this approach, we conducted the following experiment.

We used the Musk data set from the UCI data set repository to create an incomplete data set in which 20% of the attribute values were missing. We computed the conceptual directions using both the model based approach<sup>2</sup> and the simple pairwise covariance estimation procedure. We computed the unit direction vector (*estimated vector*) along each of the conceptual directions under both estimation methods and compared these direction vectors with the corresponding unit vectors constructed from the fully specified data set (*actual vector*). The dot product of the estimated vector and the actual vector will be in the range [0,1], 1 indicating coincidence (maximum accuracy) and 0 indicating the two vectors are orthogonal (minimal accuracy). Figure 3 describes the results of this experiment on the first 30 eigenvectors. Clearly, the EM estimation method outperforms the pairwise estimation method. The absolute accuracy of the EM-estimation method is also rather high. For example, for the first 13 eigenvectors (which covers more than 87% of the variance in the

---

<sup>2</sup>Note that we did not run the EM algorithm to convergence but only for 30 iterations for this experiment.

data set) the accuracy is typically above 0.94.

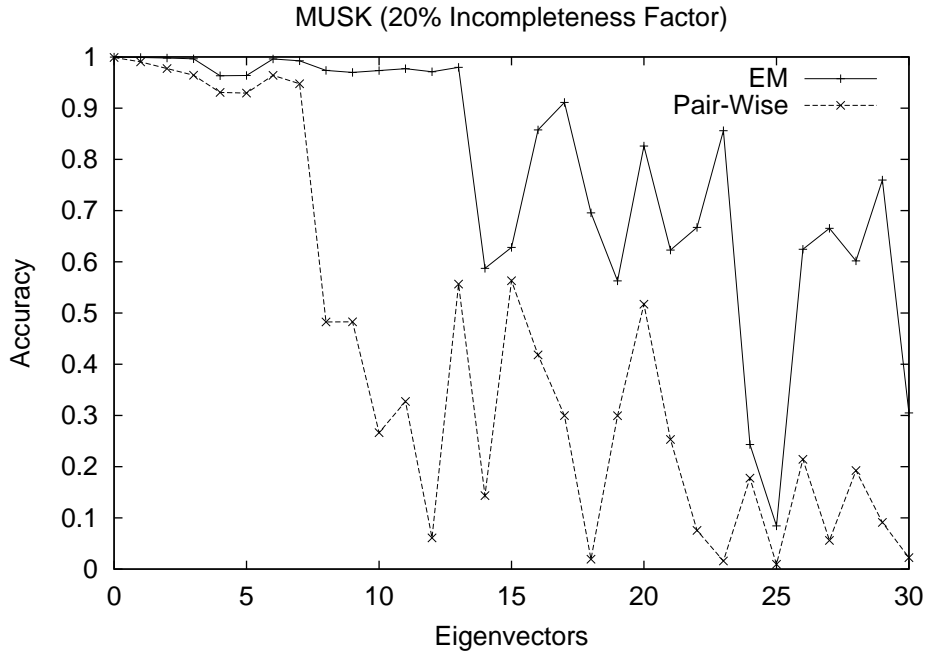


Figure 3: Comparing EM and Pair-Wise Estimation

Once the conceptual vectors have been identified the next step is to estimate the projection of each record  $Q$  onto each conceptual vector. In the previous section, we discussed how a set  $C$  of close records are determined using the known attributes in order to perform the reconstruction. We defined  $C$  to be the set of records in the neighborhood of  $Q$  using the attribute set  $A$ . The  $Y_B^i$  value for  $Q$  is estimated using the records in set  $C$ . It is possible to further refine the performance using the following observation.

The values of  $Y_B$  for the records in  $C$  may often show some clustering behavior. We cluster the  $Y_B$  values in  $C$  in order to create the sets  $C_1 \dots C_r$ , where  $\cup_{i=1}^r C_i = C$ . For each set  $C_i$ , we compute the distance of its centroid to the record  $Q$  using the known attribute set  $A$ . The cluster that is closest to  $Q$  is used to predict the value of  $Y_B$ . The intuition behind this method is obvious.

The time complexity of the method can be obtained by summing the time required for each

step of Figure 2. The first step is the computation of the covariance matrix, which normally (when there is no missing data) requires processing time of  $O(d^2 \cdot N)$ . For the missing data case, since essentially we use the EM procedure to estimate this matrix at each iteration until convergence is achieved, the lower bound on the total cost may be approximated as  $O(d^2 \cdot N \cdot it)$  where  $it$  is the number of iterations for which the EM algorithm is run. For a more exact analysis of the complexity of the EM algorithm and associated guarantees of convergence (to a local maximum of the log-likelihood) we refer the reader elsewhere[18, 12]. The process of step 2 is simply the generation of the eigenvectors which requires a time of  $O(d^3)$ . However, since only  $m$  of these eigenvectors need to be retained, the actual time required for the combination of steps 2 and 3 is  $O(d^2 \cdot m)$ . Finally, step 4 requires  $m$  dot product calculations for each record and requires a total time of  $O(N \cdot d \cdot m)$ .

## 4 Empirical Evaluation

In order to perform the testing, we used several completely specified data sets (Musk(1 & 2), BUPA, Wine, and Letter-Recognition) in the UCI<sup>3</sup> machine learning repository. The Musk 1 data set has 475 instances and 166 dimensions<sup>4</sup>. The Musk 2 data set has 6595 instances and 166 dimensions. The Letter-Recognition data set has 16 dimensions and 20000 instances. The BUPA data set has 6 dimensions and 345 instances. The incomplete records were generated by randomly removing some of the entries from the records. We introduce a notion of incompleteness in these data sets by randomly eliminating values in records of the data set. One of the advantages of this method is that since we already know the original data set, we can compare the effectiveness of the reconstructed data set with the actual data set to validate our approach. We use several evaluation metrics in

---

<sup>3</sup><http://www.cs.uci.edu/~mllearn>

<sup>4</sup>Number of relevant dimensions

order to test the effectiveness of the reconstruction approach. These metrics are designed in various ways to test the robustness of the reconstructed method in preserving the inherent information from the original records.

**Direct Error Metric:** Let  $Y_{estimated}^i(\mathcal{Q})$  be the estimated value of the conceptual component for the eigenvector  $i$  using the reconstruction method. Let  $Y_{actual}^i(\mathcal{Q})$  be the true value of the projection of the record  $\mathcal{Q}$  on to eigenvector  $i$ , if we had an oracle which knew the true projection onto eigenvector  $i$  using the original data set. Obviously, the closer  $Y_{actual}^i(\mathcal{Q})$  is to  $Y_{estimated}^i(\mathcal{Q})$ , the better the quality of the reconstruction. We define the relative error<sup>5</sup> along the eigenvector  $i$  as follows:

$$Error_i = \frac{\sum_{\forall \mathcal{Q} \in D} |Y_{estimated}^i(\mathcal{Q}) - Y_{actual}^i(\mathcal{Q})|}{\sum_{\forall \mathcal{Q} \in D} |Y_{actual}^i(\mathcal{Q})|}$$

Clearly, lower values of the error metric are more desirable. In many cases even when the absolute error in estimation is somewhat high, empirical evidence suggests that the correlations between estimated and actual values continue to be quite high. This indicates that even though the estimated conceptual representation is not the same as the true representation, the estimated and actual components are correlated so highly that the direct application of many data mining algorithms on the reconstructed data set is likely to continue to be effective. To this end, we computed the covariance and correlation of these actual and estimated projections for each eigenvector over different values of  $\mathcal{Q}$  in the database. A validation of our conceptual reconstruction procedure would be if the correlations between the actual and estimated projections are high. Also, if the magnitude of the covariance between the estimated and actual components along the principal eigenvectors were high it would provide further validation of our intuitions that the principle eigenvectors provide the directions of the data which have the greatest predictability.

---

<sup>5</sup>Note that this error metric only takes into account records that have missing data. Complete records (if any) play no role in the computation of this metric.

**Indirect Error Metric:** Since, the thrust of this paper is to compute conceptual representations for indirect use on data mining algorithms rather than actual attribute reconstruction, it is also useful to evaluate the methods with the use of an indirect error metric. In this metric, we build and compare the performance of a data mining algorithm on the reconstructed data set. To this effect, we use classifier trees generated from the original data set and compare it with the performance of the classifier trees generated from the reconstructed data set. Let  $CA_o$  be the classification accuracy with the original data set, and  $CA_r$  be the classification accuracy with the reconstructed data set. This metric also referred to as *Classification Accuracy Metric* (CAM) measures the ratio between the above two classification accuracies. More formally:

$$CAM = \frac{CA_r}{CA_o}$$

Thus, the indirect metric measures how close to the original dataset the reconstructed data set is, in terms of classification accuracy.

#### 4.1 Evaluations with Direct Error Metric

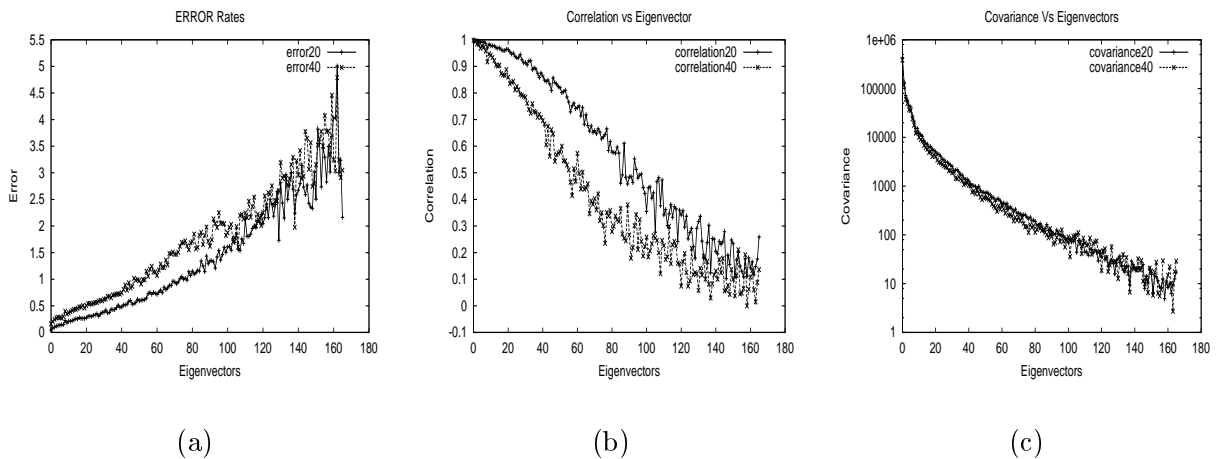


Figure 4: (a) Error (b) Correlation(estimated,actual), and (c) Covariance(estimated,actual) as a function of eigenvectors for the Musk(1) dataset at 20% and 40% missingness

The results for the Musk(1) dataset is shown in Figure 4. In all cases, we plot the results as a function of the eigenvectors ordered by their eigenvalues where eigenvector 0 corresponded to the one with the largest eigenvalue.

Figure 4a offers some empirical evidence for Intuition 1. Clearly, the predictability is better on eigenvectors with larger variance. In this data set we note that the error rapidly increases for the eigenvectors with small variance. For eigenvectors 145-165 the relative error is larger than 3. This is because these are the noise directions in the data along which there are no coherent correlations among the different dimensions. For the same reason, these eigenvectors are not really relevant even in fully specified data sets, and are ignored from the data representation in dimensionality reduction techniques. The removal of such directions is often desirable even in full specified data sets, since it leads to the pruning of noise effects from the data [13].

To further validate our approach, we calculated the covariances and correlations between the actual and estimated components along the different eigenvectors. The results are illustrated in Figures 4b, and 4c. For this data set the largest eigenvectors show a very strong correlation and high covariance between the estimated and actual projections. The correlation value for the largest 20 eigenvectors is greater than 0.95. For the first five eigenvectors, there is about an 8% drop in the average error, while the correlation continues to be extremely significant (around 0.99).

As expected, the average errors are higher for 40% incompleteness factor when compared to 20% incompleteness factor. However, the general trend of variation in error rate with the magnitude of the variance along a principal component is also retained in this case. The correlations between the true and estimated values continue to be quite high. These results are encouraging, and serve to validate our key intuitions, especially given the high level of incompleteness of this data set.

Similar trends were observed for the Musk(2), BUPA and Wine datasets. The results are illustrated in Figure 5, Figure 6 and Figure7 respectively. Once again for these datasets we observed the

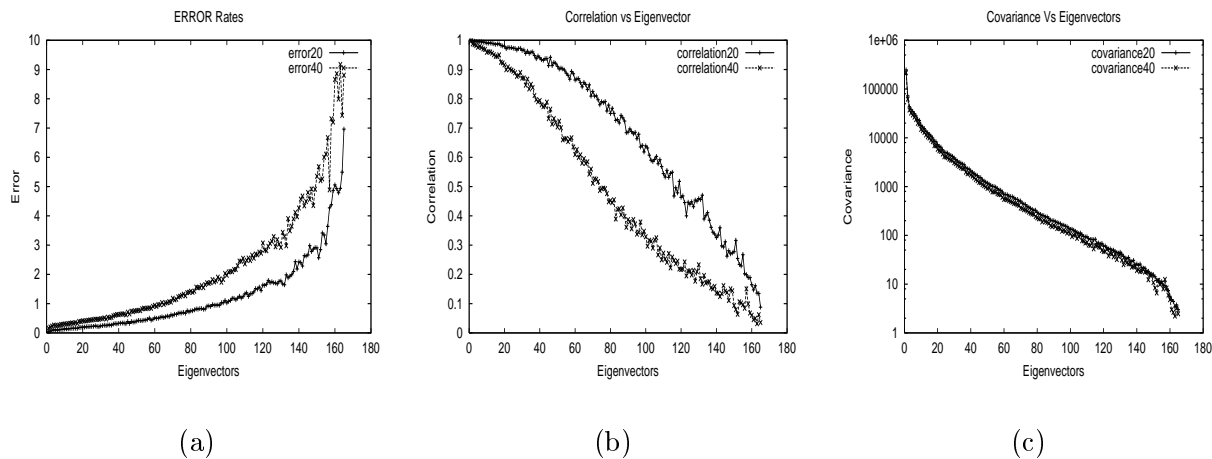


Figure 5: (a) Error (b) Correlation(estimated,actual), and (c) Covariance(estimated,actual) as a function of eigenvectors for the Musk(2) dataset at 20% and 40% missingness

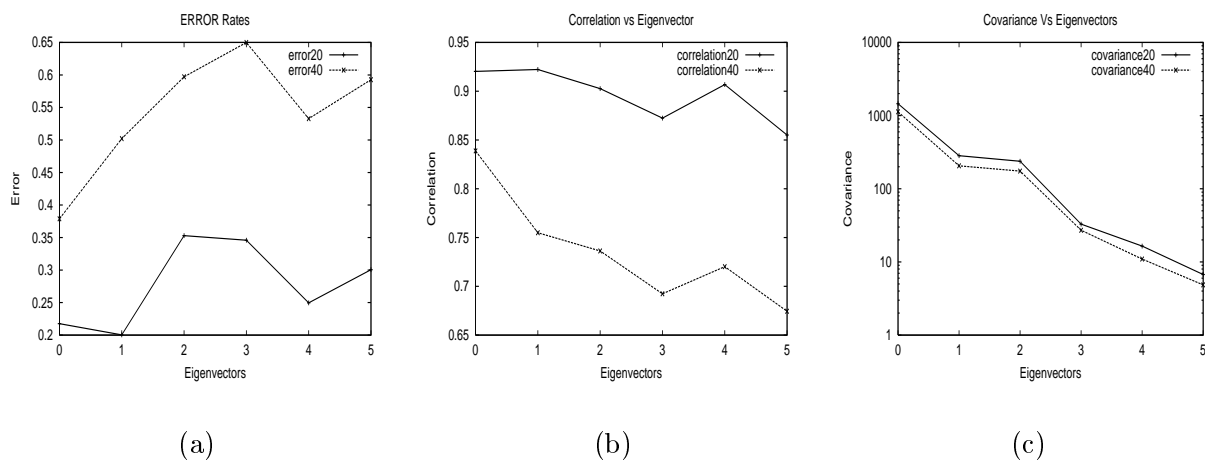


Figure 6: (a) Error (b) Correlation(estimated,actual), and (c) Covariance(estimated,actual) as a function of eigenvectors for the BUPA dataset at 20% and 40% missingness

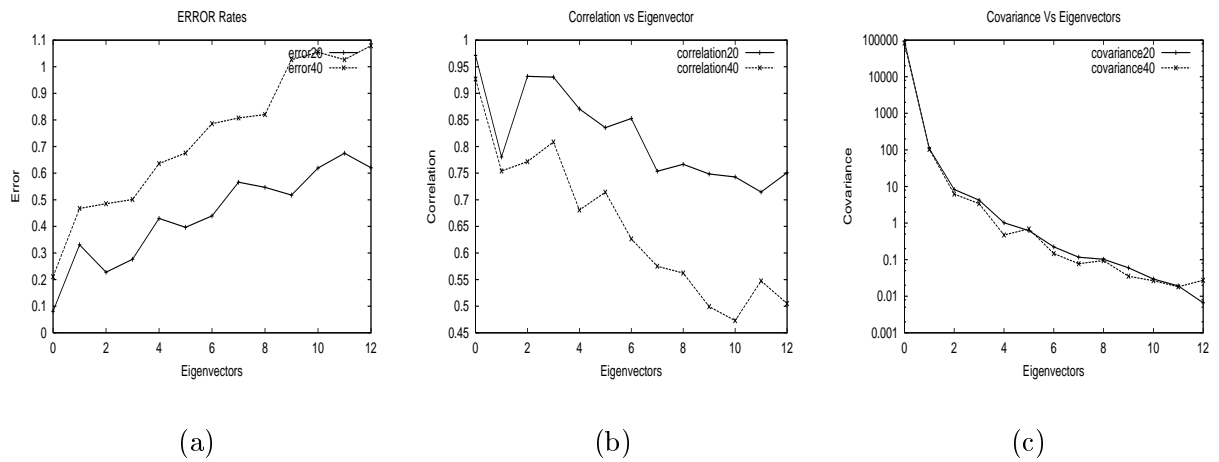


Figure 7: (a) Error (b) Correlation(estimated,actual), and (c) Covariance(estimated,actual) as a function of eigenvectors for the Wine dataset at 20% and 40% missingness

following trends: the eigenvectors with largest variance had the lowest estimation error; there was very high correlation and covariance between the estimated and actual values along the eigenvectors with high variance; and increasing the level of missingness from 20 to 40% resulted in slightly poorer estimation quality (as determined by the direct metrics). The results for the Letter Recognition

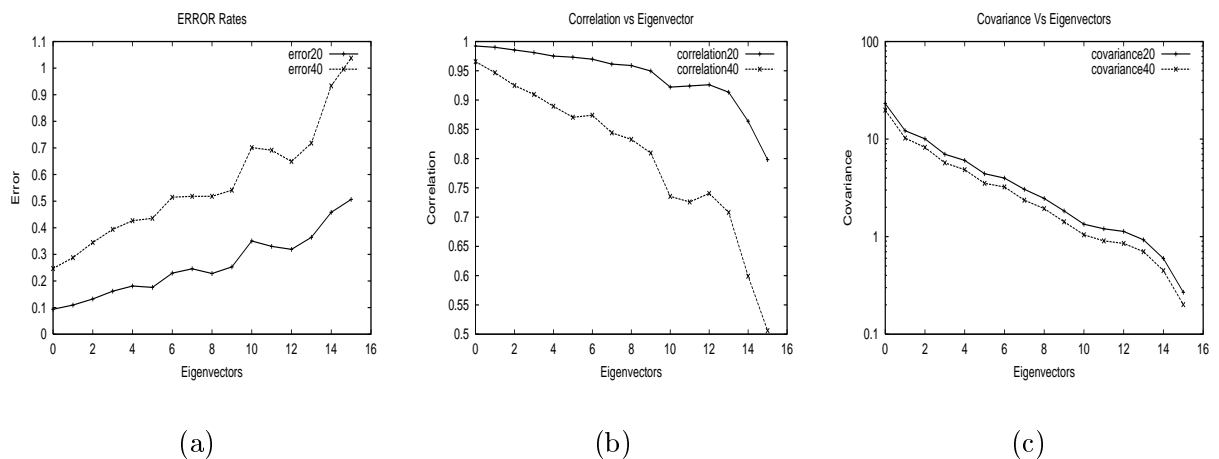


Figure 8: (a) Error (b) Correlation(estimated,actual), and (c) Covariance(estimated,actual) as a function of eigenvectors for the Letter-Recognition dataset at 20% and 40% missingness

data set were slightly different and are illustrated in Figure 8. While the observed correlations

between the actual and estimated projections were reasonably high for the eigenvectors with high variance, the observed covariances were decidedly on the lower side. Furthermore, the correlations were also not quite as high as the other data sets. This is reflective of the fact that this is a data set in which there cross-attribute redundancy in data representation, i.e. the correlation structure of this dataset, is weak. Such a data set is a very difficult case for the conceptual reconstruction approach, or any other missing data mechanism. This is because any removal of attribute values in such a data set would lead to true loss of information, which cannot be compensated for by the inter-attribute correlation redundancy. As we shall see, our experiments with the indirect metric bear this fact out.

However, in general, our observation across a wide variety of data sets was that the correlation between the actual components and re-constructed components tends to be quite high. This robustness of the correlation metric indicates that for a *particular* eigenvector, the error is usually created by *either* a consistent underestimation or a consistent overestimation of the conceptual component. This consistency is quite significant, since it implies that a simple linear translation of the origin along the eigenvector, could reduce the error rate further. Of course, the direction of translation is not known apriori. However, for typical data mining tasks such as clustering and similarity search where the *relative* position of the data records with respect to one another is more relevant, it is not necessary to perform this translation. In such cases, the reconstructed data set would continue to be highly reliable.

## 4.2 Results with Indirect Metric

Since the purpose of the conceptual reconstruction method is to provide a new representation of the data on which data mining algorithms can be directly applied, it is useful to test the effects of using the procedure on one such algorithm. To this effect, we use a decision tree classifier [19],

Dataset	$CA_o$	$CAM_{20\%}(RC)$	$CAM_{20\%}(MI)$	$CAM_{40\%}(RC)$	$CAM_{40\%}(MI)$
BUPA	62.4	0.963	0.89	0.927	0.875
Musk (1)	76.2	0.943	0.937	0.92	0.89
Musk (2)	95.0	0.96	0.95	0.945	0.917
Letter Recognition	84.9	0.825	0.749	0.62	0.54
Wine	91.0	0.98	0.98	0.927	0.88

Table 1: Evaluation of Indirect Metric

which we apply both to the original (complete) representation and the conceptual representation of the missing data.

In Table 1, we have illustrated<sup>6</sup> the accuracy of the classifier on a conceptual representation of the data, when the percentage of incomplete entries varies from 20 % to 40 % respectively ( $CAM(RC)$  columns). We have also illustrated the accuracy on the original representation in the same Table ( $CA_o$  column). In addition we compared the reconstruction approach also with an approach that fills missing values using mean imputation ( $CAM(IM)$  columns).

For all the datasets and at different levels of missingness our approach is clearly superior to the approach based on mean imputation. The only exception to the above is the Wine dataset, where at 20% missingness the two schemes are comparable. In fact in some cases the improvement in accuracy is nearly 10%. This improvement is more apparent in datasets where the correlation structure is weaker (Letter-Recognition, Bupa) than in datasets where the correlation structure is

---

<sup>6</sup>Note that the original classification task for both Musk (1) and Musk (2) is to classify the original molecules into Musk and non-Musk. These data sets represents a multiple-instance classification problem with the total number of instances significantly exceeding the original number of molecules. The classification accuracies reported here are for the case where each instance is treated as an independent entity and is therefore different from the original classification problem, since C4.5 does not support the multiple instance problem.

stronger (Musk, Wine datasets). One possible reason for this is that although mean imputation often results in incorrect estimations, the stronger correlation structure in the Musk datasets enables C4.5 to ignore the incorrectly estimated attribute values thereby ensuring that the classification performance is relatively unaffected. Note also that the improvement of our reconstruction approach over mean imputation is more noticeable as we move from 20% missingness to 40% missingness. This is true of all the datasets including the Wine dataset.

For the case of the BUPA, Musk(1) and Musk(2) data sets, the C4.5 classifier built on the reconstructed dataset (our approach) was *at least* 92% as accurate as the original data set even with 40% incompleteness. In most cases, the accuracy was significantly higher. This is evidence of the robustness of the technique and its applicability as a procedure to transform the data without losing the inherent information available in it.

Out of the five data sets tested, only the letter recognition data set did not show as effective a classification performance as the other three data sets. This difference is especially noticeable at the 40% incompleteness factor. There are three particular characteristics of this data set and the classification algorithm which contribute to this. The first reason is because correlation structure of the data set was not strong enough to account for the loss of information created by the missing attributes. Although our approach outperforms mean imputation, the weak correlation structure of this dataset tends to amplify the errors of the reconstruction approach. We note that any missing data mechanism needs to depend upon inter-attribute redundancy, and such behavior shows that this dataset is not as suitable for missing data mechanisms as the other datasets. Second, on viewing the decision trees that were constructed we noticed that for this particular dataset, the classifier happened to pick the eigenvectors with lower variance first, while selecting the splitting attributes. These lower eigenvectors also are the ones where our estimation procedure results in larger errors. This problem may not however, occur in a classifier in which the higher eigenvectors

are picked first (as in PCA-based classifiers). Finally, in this particular data set, several of the classes are inherently similar to one another, and are distinguished from one another by only small variations in their feature values. Therefore, removal of data values has a severe effect on the retainment of the distinguishing characteristics among different classes. This tends to increase the misclassification rate.

We note that even though the applicability of the general conceptual reconstruction technique applies across the entire spectrum of generic data mining problems, it is possible to further improve the method for particular problems. This can be done by picking or designing the method used to solve that problem more carefully. For example, we are evaluating strategies by which the overall classification performance in such reconstructed data sets can be improved. As mentioned earlier, one strategy under active consideration is to use class-dependent PCA-based classifiers. This has two advantages. First, since these are PCA-based our reconstruction approach naturally fits into the overall model. Secondly, class-dependent approaches are typically better discriminators in data sets with a large number of classes, and will improve the overall classification accuracy in such cases. An interesting line of future research would be to develop conceptual reconstruction approaches which are specially tailored to different data mining algorithms.

## **5 Conclusions and Directions for Future Work**

In this paper, we introduced the novel idea of conceptual reconstruction for mining massively incomplete data sets. The key motivation behind conceptual reconstruction is that by choosing to predict the data along the conceptual directions, we use only that level of knowledge as can be reliably predicted from the incomplete data. This is more flexible than the restrictive approach of predicting along the original attribute directions. We show the effectiveness of the technique on

a wide variety of real data sets. Our results indicate that even though it may not be possible to reconstruct the original data set for an arbitrary feature or vector, the conceptual directions are very amenable for reconstruction. Therefore, it is possible to reliably apply data mining algorithms on the conceptual representation of the reconstructed data sets.

In terms of future work one interesting line is to extend the proposed ideas to work with categorical attributes. Recall that the current approach works well only on continuous attributes since it relies on PCA. Another interesting avenue of future research could involve investigating refinements to the estimation procedure that can improve the efficiency (using sampling) and accuracy (perhaps by evaluating and using the refinements suggested in Section 3.1) of the conceptual reconstruction procedure.

**Acknowledgements:** We would like to thank the people involved in the review process for providing detailed comments that helped improve the quality and readability of the paper.

## References

- [1] C. C. Aggarwal. On the Effects of Dimensionality Reduction on High Dimensional Similarity Search. *ACM PODS Conference*, 2001.
- [2] C. C. Aggarwal, S. Parthasarathy. Mining Massively Incomplete Data Sets by Conceptual Reconstruction. *ACM KDD Conference*, 2001.
- [3] R. Agrawal, R. Srikant. Privacy Preserving Data Mining. *In ACM SIGMOD*, 2000.
- [4] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. *Classification and Regression Trees*, Wadsworth, Belmont, 1984.

- [5] 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*, 39:1-38, 1977.
- [6] A. W. Drake. Fundamentals of Applied Probability Theory. *McGraw-Hill*, 1967.
- [7] Z. Ghahramani, M. I. Jordan. Learning from incomplete data. *Department of Brain and Cognitive Sciences*, Paper No. 108, *MIT*, 1994.
- [8] I. T. Jolliffe. *Principal Component Analysis*, Springer-Verlag, New York, 1986.
- [9] J. Kleinberg, A. Tomkins. Applications of linear algebra to information retrieval and hypertext analysis. *ACM PODS Conference, Tutorial Survey*, 1999.
- [10] R. Little, D. Rubin. Statistical Analysis with Missing Data Values. *Wiley Series in Prob. and Stat.*, 1987.
- [11] R. J. A. Little, M. D. Schluchter. Maximum Likelihood estimate for mixed continuous and categorical data with missing values. *Biometrika*, 72:497-512.
- [12] G. J. McLachlan and T. Krishnan, The EM Algorithm and Extensions, John Wiley and Sons, 1997.
- [13] C. H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala: Latent Semantic Indexing: A Probabilistic Analysis. *ACM PODS Conference*, 1998.
- [14] K. V. Ravikanth, D. Agrawal, A. Singh. Dimensionality Reduction for Similarity Searching in Dynamic Databases. *In ACM SIGMOD*, 1998.
- [15] S. Rowells. EM Algorithms for PCA and SPCA.
- [16] D. B. Rubin. Advances in Neural Information Processing Systems, volume 10, pages 626-631, *Morgan Kaufmann*, 1998. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

- [17] J. Schafer. *Analysis of Incomplete Data Sets by Simulation*. Chapman and Hall, 1994, London.
- [18] J. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall, 1997, London.
- [19] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman, 1993.
- [20] J. R. Quinlan. Unknown Attribute values in Induction. *Proceedings of the Sixth International Conference on Machine Learning*, 1989.