

https://dspace.org/

An empirical analysis and application of the expectation-maximization and matrix completion algorithms for varying degrees of missing data

Thulare, Evans Molahlegi

2020

https://hdl.handle.net/10539/35819

Thulare, Evans Molahlegi (2020) An empirical analysis and application of the expectation-maximization and matrix completion algorithms for varying degrees of missing data, University of the Witwatersrand, Johannesburg, <http://hdl.handle.net/10539/35819>

Downloaded from a DSpace repository. https://dspace.org/

An Empirical Analysis and Application of the Expectation-Maximization and Matrix Completion Algorithms for Varying Degrees of Missing Data

---

Evans Molahlegi Thulare

1799336

Supervisors:

Dr. Ritesh Ajoodha

Dr. Ashwini Jadhav



A mini-dissertation submitted in partial fulfilment of the requirements for the degree of Master of Science in the field of e-Science

in the

School of Computer Science and Applied Mathematics

University of the Witwatersrand, Johannesburg

3 November 2020

Declaration

I, Evans Molahlegi Thulare, declare that this research report is my own, unaided work. It is being submitted for the degree of Master of Science in the field of e-Science at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.



Evans Molahlegi Thulare

3 November 2020

Abstract

Incomplete data sets have been a problem in most studies, however, few studies have come to realise that imputation is a solution to this problem. Incomplete data can have a significant effect on the conclusion drawn and decision made. To solve the problem of incomplete data, one should use techniques to recover those missing values, depending on how much the data is missing, how big is the data, how the data has gone missing, etc. In this report, we aimed to compare the performance of the EM algorithm and matrix completion when imputing the missing values for varying degrees of missing data.

Kullback-Leibler (KL) divergence was used as an evaluation metric to observe the performance of Expectation-Maximization (EM) algorithm and matrix completion when estimating missing values relative to the ground-truth distribution. The findings of this research shows that the EM algorithm outperformed matrix completion in both the theoretical (the simulated scenarios of learning from varying degrees of missing data) and the application (the application of theoretical model on real-world data on credit card fraud) models. Few similarities of the algorithms were observed when recovering missing values such as the increasing trend of error as missing values increases and the impact of increasing number of variables in a data set.

Matrix completion only performed better when missing values were beyond approximately 77%. Therefore, from our findings, we conclude that when less than 50% of the data is missing, EM algorithm produces accurate predictions. The EM

algorithm performed better compared to the matrix completion since it first learned the data itself and used maximum likelihood procedures to estimate the parameters of the model while the matrix completion analysed the existing pattern from rows and columns and imputes them using the pattern learned in the data.

Acknowledgements

I thank the Almighty for giving me the opportunity to successfully complete this research report and his mercy is acknowledged.
I would also like to thank my supervisor most sincerely, Dr. Ritesh Ajoodha (Ritso) and I really appreciate my co-supervisor's time and support, Dr. Ashwini Jadhav. The support from Casey Sparkes and Dr. Helen Robertson is highly appreciated. I appreciate also friends who have proofread my research report, Precious Makganoto (MSc candidate from University of the Witwatersrand), Precious Kgabi (PhD candidate from University of Johannesburg), Shanil Ramlall (MSc candidate from University of KwaZulu-Natal) and Olaperi Okuboyejo (PhD candidate from Witwatersrand University). I would like to thank my family for their precious support especially my mother Nelly Thulare. Lastly, the support of the DST-CSIR National e-Science Postgraduate Teaching and Training Platform (NEPTTP) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NEPTTP.

Contents

Declaration i

Abstract ii

Acknowledgements iv

List of Figures vii

List of Tables viii

1 Introduction 1

1.1 Contributions . . . 3

1.2 Research Aim and Objectives . . . 3

1.3 Research Questions . . . 4

2 Literature Review 5

2.1 Research Question 1- Comparison Between EM Algorithm and Matrix Completion . . . 5

2.1.1 Data Used . . . 8

2.1.2 Models Compared to Compute Missing Data . . . 9

2.1.3 Literature Approved Model . . . 10

2.2 Research Question 2- Reliability to Make Accurate Predictions with Incomplete Data Set . . . 11

2.2.1 Data Used . . . 13

2.2.2 Models . . . 13

2.2.3 Literature Approved Model . . . 14

3 Imputing Missing Data 15

3.1 Introduction . . . 15

3.2 Motivation 15
3.3 Overview of the Research Methodology 16
3.4 Methods 18
3.4.1 Matrix Completion 18
3.4.2 Expectation-Maximization Algorithm 19
3.4.3 Summary of Methods 21
3.4.4 Evaluation 22
3.5 Simulated Data 23
3.6 Results For Simulated Data 26
3.7 Conclusion 28

4 Application: Imputing Missing Values for Credit Card Fraud Detection Data **30**
4.1 Introduction 30
4.1.1 Application Data 30

5 Conclusion and Future Work **35**
5.1 Conclusion 35
5.2 Future Work 36

Bibliography **38**

List of Figures

3.1 A figure illustrating the six-phased experimental design with each phase labelled respectively. 25
3.2 Comparison of the EM algorithm and matrix completion with 4 variables 27
3.3 Comparison of the EM algorithm and matrix completion with 6 variables 28
3.4 Comparison of the EM algorithm and matrix completion with 8 variables 28
3.5 Comparison of the EM algorithm and matrix completion with 10 variables 29

4.1 A Sample of the Credit Card Fraud Detection Data Set with 10 Variables. 31
4.2 A figure illustrating principle analysis of the application model . . . 32
4.3 Comparison of the EM algorithm and matrix completion on a credit card default data 34

List of Tables

2.1 Comparison between EM and matrix completion 6
2.2 Reliability to make accurate predictions 11

3.1 Different variables with respect to the data points generated 17
3.2 An example of the data generated from a Bayesian network with four variables modelling a Bernoulli distribution. 24
3.3 An example of a modified data set after 50% of the samples in the complete data set from Table 3.2 were removed (NaN represent missing data). 26

Chapter 1

Introduction

Incomplete data is a problem in many research fields because it has a significant effect on decision making and conclusions drawn (Liu and Gopalakrishnan, 2017). Numerous researchers often tend to avoid imputation and they discard all the rows or variables that consist of missing values (White and Carlin, 2010). This is a risky step to take because one may find that the observable data is not a representative of the whole data set. Algorithms should be used to predict missing values (Bae et al., 2016). This study aims to help people who depends on data to draw conclusions (decision making).

The purpose of this study is to explore the importance of imputing missing values instead of ignoring them. It shows the capability or potential of the Expectation-Maximization (EM) and matrix completion algorithms to estimate missing data for varying degrees of missing values in data sets. Two different kinds of data sets, a simulation and a credit card fraud detection data sets were deployed for this purpose.

Due to high rate of incomplete data sets owing to data privacy, incorrect measurement and other reasons, this research may contribute to fields that rely on data or information to make decisions by estimating the missing values in a data sets using the models or algorithms used in this research. The study sought to answer questions such as: What is the minimum number of samples which can be missing in the data set to get an acceptable entropy to the ground-truth distribution?

Recently, numerous studies have shown interest in solving the issue of estimating missing values by developing, optimising and comparing algorithms that impute

missing values. Juhola and Laurikkala (2013) presented the imputation of missing values on a classification problem and discovered that for a two-class data set, even if the data set consists about 20% to 30% of missing values, good results were still produced. However, for more than two-class data sets, 10% to 20% of missing data can badly affect the conclusion. A guide on handling missing values of counselling psychology information presented by Schlomer et al. (2010), compared three different kinds of imputation techniques. Multiple imputations and full information maximum likelihood were recommended over mean substitution.

In this research project, we aimed to observe the performance of the EM algorithm through a generated data set that follows a Binomial distribution. Several data sets were generated with varying degrees of missing data using a Bayesian network. The ratio of missing data ranges from 10% to 90% with a step of 10% was hidden (treated as missing values) and the EM algorithm in Bayesian network and the matrix completion were used to recover them. The metric used to compare the learned distribution to the ground-truth distribution was the Kullback-Leibler (KL) divergence. The ground-truth distribution was compared with estimated distributions (the distributions that consist of estimated data) after imputing the missing values. The above model is called a theoretical model because it meant to simulate the scenarios of learning from varying degrees of missing data.

Different sets of variables with different data set sizes were used to train the theoretical model. We observed the distance between the ground-truth distribution and the learned distribution using KL divergence. From our observation, it was noted that the EM algorithm outperformed the matrix completion from as little as 10% of missing values until approximately 77% for all different number of variables.

The performance of both algorithms were studied using a real-world data set (the credit card fraud detection data set) of size 307497 with 23 variables obtained from a public data set (*Kaggle*). We noted a dramatic difference between the algorithms

when over 40% of the data is missing. The EM algorithm performed well with a lower KL divergence value compared to the matrix completion method. The EM algorithm is therefore better than the matrix completion in the context explored by this research project.

1.1 Contributions

Contributions of this report to literature are outlined as follows:

- An empirical analysis between the EM model and matrix completion model to recover various degrees of missing values over different sets of variables.
- A framework to estimate how much missing information will provide reliable results (when comparing the learned distribution to the ground-truth distribution).
- An application of this case study in the financial sector to assist in estimating missing values in a credit card fraud detection data.

1.2 Research Aim and Objectives

This research aims to compare the most widely used techniques (the EM algorithm and matrix completion) in the recovery of missing values.

The objectives of this research are to:

1. Evaluate the performance of the EM algorithm versus matrix completion when recovering missing values.
2. Generate a theoretical model that will be used to recover missing values.
3. Use KL divergence to observe the error rate when imputing missing data both on a synthetic and real-world data set.

1.3 Research Questions

This research report has answered the following questions:

1. Does the EM algorithm or matrix completion perform better in recovering different degrees of missing values with different numbers of variables?
2. How much data is necessary to reconstruct the ground-truth distribution?
3. In the case of the best algorithm, does the type of distribution of a data set affect its performance?

This research report is structured as follows: the first chapter consists of an introduction of study, aim and objectives, problem definition and research questions. Related work is outlined in Chapter 2, followed by a detailed methodology (imputing missing data) of the study in Chapter 3. In Chapter 4, we presented the results of our application model. Lastly, we present the conclusion and future work in Chapter 5.

Chapter 2

Literature Review

The presence of missing values occurs when there is no observation for a variable in a data set (Nelwamondo et al., 2013). The incidence of missing values may be caused by a respondent's failure to answer questions, incorrect calculation or human error etc. Agarwal and Tangirala (2017) indicated that there is no algorithm that can perfectly imputes missing values. For missing data to be imputed, the nature of the data set must be understood (Schafer and Graham, 2002). Replacing missing values with plausible values is a common solution to a problem of coping with missing values. This chapter introduces related work of this study.

2.1 Research Question 1- Comparison Between EM Algorithm and Matrix Completion

A summary of the literature review is presented on the Table 2.1 based on the research question of comparing the performance of EM algorithm and matrix completion when estimating missing values. Each column presents the dimension of each paper. The author column presents author names for each paper, data is the data set used by a particular paper, models compared present the various models that were compared, best model is the model that performed better than the other ones on that paper and lastly, the error column is the error for the best performing model.

TABLE 2.1: Comparison between EM and matrix completion

Authors	Data	Models compared	Best models	Error
Emmanuel and Recht (2009)	Generated random positive data (841)	Matrix completion and Minimum Mean Square Error (MMSE)	Matrix completion	0.091 MSE
Schmitt et al. (2015)	Iris (100), E. coli (129), Breast cancer 1(80), Breast cancer 2(89)	Singular Value Decomposition (SVD), K-Nearest Neighbor (KNN), Mean imputation, bayesian Principle Component analysis (bPCA), Fuzzy K-Means, and multiple imputations by chained equations (MICE)	bPCA with Fuzzy K-means	0.057 RMSE
Yu et al. (2013)	Ailerons (4752), Elevators (6344), Bank (2999), Stocks (633), Boston Housing (337)	1-nearest neighbor (1-NN) imputation, mean imputation, Extreme Learning Machine (ELM)	Extreme Learning Machine (ELM)	0.03 MSE

Cai et al. (2010)	geodesic distance (10000 and 1000)	Matrix completion and Singular Value Thresholding (SVT)	Matrix completion	1.7310^4 relative error
Jerez et al. (2010)	"El Álamo I" breast cancer data set (3679 records)	Multiple Imputation, Mean (Mode), hot-Deck, K-Nearest Neighbour Impute (K-NNI), Self-Organising Maps (SOM), and Multi-layer Perceptron (MLP)	multi-layer perceptron (MLP), self-organisation maps (SOM) and k-nearest neighbour (KNN)	0.0091 MSE
Combettes and Pesquet (2011)	Netflix (500)	Matrix completion and nuclear norm combined with the graph regularization	Matrix completion and nuclear norm combined with the graph regularization	0.0063 MSE
Demirtas (2004)	psychiatric data	Multiple imputation, ANOVA and Last observation carried forward (LOCF)	Multiple imputation	0.08 MSE
Wang and Wang (2010)	Home Mortgage Disclosure Act (HMDA)	linear discriminant analysis, back-propagation neural networks (BPNN) and Last observation carried forward (LOCF)	back-propagation neural networks (BPNN)	68.5% accuracy

Fekade et al. (2017)	data collected in IoT systems (sensors)	Probabilistic matrix factorization (PMF), SVM and Deep Neural Network	Probabilistic matrix factorization (PMF)	81% accuracy

2.1.1 Data Used

Several data sets were used to develop the algorithms to estimate missing values in a data set. Jerez et al. (2010) used "El Álamo I" breast cancer data set that consists of 3 679 records from one of the biggest databases in Spain that focuses on breast cancer. Similarly, Schmitt et al. (2015) used 4 different data sets such as the popular Iris data set with only 100 flowers (records) and 4 variables, the E.coli data set that was obtained from UCI machine learning repository with 129 E.coli proteins (records) and 5 variables, lastly breast cancer 1 data set represents 146 records with 4 variables.

Recently, Fekade et al. (2017) used Internet of Things (IoT) sensed data set from different sensors deployed in the Intel Berkeley Research Laboratory. The data set started from February 28th to April 5th, 2004 with measures or data collected every 30 seconds (2,313,682 records). Yu et al. (2013) also used 5 different data sets to observe a best technique that imputes missing values with a small error. The data sets are: Ailerons (7129,5), Elevators (9517,6), Bank (4499,8), Stocks (950,9) and lastly Boston Housing (506,13).

All of the above data sets are found from the UCI machine learning repository which are publicly available. The above data sets were used in different studies to learn different models that can potentially estimate missing values in data sets.

Some data sets deployed were too small to define a model that can be used to impute missing values as it is known that in most cases, the real world data sets are massive.

2.1.2 Models Compared to Compute Missing Data

Different authors compared different types of techniques to estimate missing values in a data set. Jerez et al. (2010) compared statistical techniques (Multiple Imputation, Mean (Mode), hot-Deck) together with machine learning techniques (K-Nearest Neighbour Impute (K-NNI), Self-Organising Maps (SOM), and Multilayer perceptron (MLP)) and it was found that machine learning techniques outperform statistical techniques.

Schmitt et al. (2015) used a breast cancer data set to compare various techniques that imputes missing values such as Singular Value Decomposition (SVD), K-Nearest Neighbor (KNN), Mean imputation, bayesian Principle Component Analysis (bPCA) with Fuzzy K-Means, and multiple imputations by chained equations (MICE) and it was found that bPCA with Fuzzy K-means performs better than the other techniques. Extreme Learning Machine(ELM) performed well with less Mean Square Error (MSE) compared to 1-nearest neighbor (1-NN) imputation and mean imputation (Yu et al., 2013).

Emmanuel and Recht (2009) and Cai et al. (2010) found that matrix completion outperformed Singular Value Thresholding (SVT) when geodesic distance data set was used, and it outperformed Minimum Mean Square Error (MMSE) when generated random positive data was deployed. Similarly, a small error was obtained by Combettes and Pesquet (2011) when matrix completion and nuclear norm combined with graph regularization was used.

Demirtas (2004) compared multiple imputation, Analysis of Variance (ANOVA) and Last Observation Carried Forward (LOCF) to impute missing values on psychiatric data. Multiple imputation was found to be the best model. Back-propagation Neural Networks (BPNN) was found to be the best model with the highest accuracy compared to linear discriminant analysis and LOCF to estimate missing values (Wang and Wang, 2010). Furthermore, Probabilistic Matrix Factorization (PMF), Support Vector Machine (SVM) and Deep Neural Network were compared to estimate missing values by Fekade et al. (2017) and PMF outperformed other techniques.

2.1.3 Literature Approved Model

BPCA with Fuzzy K-means was presented by Schmitt et al. (2015) as the best model with 0.057 RMSE when imputing missing values. The best model was found to be Extreme Learning Machine with an MSE of 0.003 (Yu et al., 2013). Emmanuel and Recht (2009) and Cai et al. (2010) found that matrix completion outperformed other models with MSE of 0.091 and 0.00017310. PMF was found to be the best model when estimating missing values with 81% accuracy. Similarly, 65.5% accuracy was obtained by BPNN as the best model to impute missing values (Wang and Wang, 2010). Lastly, Combettes and Pesquet (2011) presented the smallest error of 0.0063 MSE obtained by matrix completion and nuclear norm combined with the graph regularization which makes it the best model. All the best model mentioned above are with respect to the findings of the author of each paper.

This study investigated the performance of the EM algorithm and matrix completion when estimating missing values. Major takeaway from this section is the illustration and the importance of missing value imputations. The effectiveness of techniques that imputes missing values are presented. The behavior of each technique is observed together with the sizes of the data sets used. EM algorithm and

matrix completion were chosen to be compared in this report since they can better handle a large dimension of data sets when forecasting missing values.

2.2 Research Question 2- Reliability to Make Accurate Predictions with Incomplete Data Set

Table 2.2 presents papers that have used data recovery mechanisms to forecast how much data is needed to make reliable predictions. The author column presents the different authors for each paper, data is the data sets that were used on that particular research, % missing presents the ratio of data that is missing in a data set, the models that were deployed for each research are presented in the models and the evaluation shows the accuracy or the error of the model when predicting missing values.

TABLE 2.2: Reliability to make accurate predictions

Authors	Data	% missing	Models	Evaluation
Geng et al. (2010)	face images	performance on different ratio from 10% to 90%	Multilinear subspace analysis (MSA), M^2SA,kernel-based PCA (KPCA),laplacianface, eigenface, fisherface, KNN	M^2SA with 5.36(MAE) (over 80% accuracy)
Sehgal et al. (2005)	microarray data (time series and non-time series)	modeled on 1 to 20% and applied to 1.7%	KNN, LSImpute, BPCA,and collateral missing value estimation (CMVE)	CMVE with 0.0064 MSE

Heijden et al. (2006)	diagnosis of pulmonary embolism data	38%	complete case analysis, missing-indicator method, single imputation of unconditional and conditional mean, and multiple imputation (MI)	MI with the receiver operating characteristic curve area of above 0.75
Stefanakos and Athanassoulis (2001)	simulation of an incomplete nonstationary time series of wave	16.5% and 33%	Autoregressive–moving-average [ARMA(2,2)]	0.108 MSE
Purwar and Singh (2015)	Pima Indians Diabetes, Wisconsin Breast Cancer, Hepatitis	Trained on 25%, 50% and 75%	hybrid prediction model with missing value imputation (HPM-MI)	HPM-MI performed well with 99.82% for Pima, 99.39% breast cancer and 99.08% for Hepatitis accuracies

2.2.1 Data Used

The data set used by Geng et al. (2010) is 'illum', a subset of the CMU PIE database (face image data set). The images differ by poses and illumination for each individual. The data set consists of 18 564 images, each person captured 13 different poses and 21 illuminations. Sehgal et al. (2005) performed an analysis of imputing missing values on 4 different types of microarray data set (time series and non-time series) with 6445 and 6179 records respectively. Similarly, the diagnosis of pulmonary embolism data with 398 patients (records) was deployed (Heijden et al., 2006).

2.2.2 Models

Purwar and Singh (2015) performed an analysis of missing values using hybrid prediction models with missing value imputation (HPM-MI) technique on various degrees of missing values (25%, 50% and 75%) with 4 different types of data sets. ARMA(2,2) model performed well on simulation of an incomplete non-stationary time series of wave data with 16.5% and 33% of missing values (Stefanakos and Athanassoulis, 2001). Heijden et al. (2006) presented that multiple imputation (MI) outperformed complete case analysis, missing-indicator method, and single imputation of the unconditional and conditional mean. on a data set that presented 38% of missing values.

Multilinear subspace analysis (MSA), Multilinear Subspace Analysis with Missing values (M^2SA), kernel-based PCA (KPCA), laplacianface, eigenface, fisherface and KNN imputation models were compared to impute missing values on different degrees of missing values that range from 10% through 90% with a step of 10% and M^2SA outperformed other models with high accuracy (Geng et al., 2010). Collateral missing value estimation (CMVE) outperformed KNN, LSImpute and BPCA on a data set that was trained on 1% to 20% of missing values and tested on 1.7% of missing values (Sehgal et al., 2005).

2.2.3 Literature Approved Model

Purwar and Singh (2015) presented the performance of HPM-MI on 3 different data sets and the accuracies were as follows: 99.82% on Pima, 99.39% Breast cancer and lastly, 99.08% for Hepatitis data set. The ARMA(2,2) model performed well with 0.108 MSE and CMVE outperformed other techniques with 0.0064 MSE. Geng et al. (2010) demonstrated that M^2SA outperformed other models with an accuracy of 80% and MI performed well with an accuracy of 79%.

No matter how hard the missing values are prevented in a data set, missing data occurs. Researchers are finding a way to design and optimise algorithms that can better learn with the presence of missing values. A key takeaway from this section is to show that the reliable conclusion can still be made with a certain ratio of missing values (Friedman, 2013). The EM algorithm and matrix completion performs better with the presence of missing values and hence were chosen to be compared.

Chapter 3

Imputing Missing Data

3.1 Introduction

This report aimed to evaluate the performance of the Expectation-Maximization (EM) algorithm and the matrix completion when imputing missing values in a data set with varying degrees of missing data. This section explicitly explains the overall methodology that had been used in this report together with data sets and evaluation metric.

3.2 Motivation

The study was motivated by massive inadequacy of techniques that are deployed to estimate missing data. Imputing missing values is challenging especially when no one understands the missing mechanism (Missing at Random, Missing Not at Random and Missing Completely at Random) and when the missing ratio is above 10% (Stuart et al., 2009).

Studies by Nigam et al. (2000) and Celeux et al. (2001) indicated that the EM algorithm in a Bayesian network performs better than a lot of well-known imputation techniques such as listwise deletion, mean (mode) imputation, regression imputation, etc. The EM algorithm unlike other techniques that are used to impute missing values, it learns the data and attributes before imputing the missing data and it guarantees the convergence (Zhang et al., 2011). On the other hand, matrix completion recovers missing information in a data set by analysing the existing pattern

from rows and columns and impute the missing entries with the pattern learned in a data set.

We chose the Bayesian network because of its ability to create distributions among variables in the data set by simply changing its structure and parameters. We chose the EM algorithm to impute missing values since Solanki et al. (2006) highlighted the uniqueness of the EM algorithm when imputing missing values these includes the use of the probabilities to approximate the missing values and the preservation of the correlation with other variables. Ajoodha and Rosman (2017) have also indicated the power of EM algorithm by learning the influence between variables in the incomplete data set using EM algorithm. Stephens and Scheet (2005) have shown that Kullback–Leibler (KL) divergence is useful in measuring the difference between two distributions or comparing distributions. Lastly, Matrix completion technique was chosen as our baseline method.

3.3 Overview of the Research Methodology

The first step was to use the Bayesian network to generate a data set that is Binomially distributed with different variables. We then hid some of the data set in order to estimate them later on. We first removed 10% of the data, which means we were left with 90% observed data. We then estimated that particular 10% of the incomplete data set and observed how good or bad is the model compared to the ground-truth distribution using KL divergence.

The second step was to hide 20% and remain with 80% of the data. We then estimated that 20% of the data that was hidden, we later combined it with the 80% that was remaining and then stored the data in a Bayesian network structure and determined its distribution. We continued with the same approach of hiding values from 10% until 90% with the step of 10%.

On every step of hiding values, we compared the performance of the EM algorithm with matrix completion and stored the results. The purpose of having distributions on each step is to compare the distributions that estimate the data set with the ground-truth distribution (the distribution that generated a data set). The process of imputing missing values was repeated 15 times in both algorithms and an average was plotted. Standard deviation was used as the error bar. The KL divergence was used to compare the estimated distribution with the ground-truth distribution.

We performed a simulation study of a discrete data set that follows a binomial distribution. A total of 40 000 to 100 000 data points were simulated using Bayesian network with different number of variables and different distributions, namely 4, 6, 8, 10 using MATLAB. Bayesian network structures were used to randomly generate the data sets. At each trial we randomised the structure and parameters of the ground-truth model to ensure that we used a randomised distribution for each trial.

The smallest Bayesian structure consists of 4 variables and the largest structure had 10 variables. These structures are treated as ground-truth distributions because they are the ones that are being deployed to randomly generate the data set. Forward sampling methods were used to generate data from the ground-truth distribution. Table 3.1 represents clearly the number of variables with respect to their data points generated.

TABLE 3.1: Different variables with respect to the data points generated

	4 variables	**6 variables**	**8 variables**	**10 variables**
Data Points	40 000	60 000	80 000	100 000

3.4 Methods

Recently, Balakrishnan et al. (2017) drew special attention showing that the EM algorithm in a Bayesian network uses maximum-likelihood estimators to impute missing data and now researchers are learning a lot around its behaviour. The EM algorithm learns the data itself before imputing the missing values and it uses maximum-likelihood procedures to estimate the model parameters. The matrix completion is one of the famous techniques that is used to recover missing records in data sets. The Netflix problem is the famous example that the matrix completion has solved, where movies were suggested for users based on what they have watched before (Takács et al., 2008). We will firstly explore the matrix completion algorithm in the next section and then we will explore the EM algorithm in the next section.

3.4.1 Matrix Completion

The matrix completion method is an alternative approach that has been deployed to recover missing data. Given a matrix of data $\mathbf{P} \in \mathrm{R}^{m \times n}$ and is known to have a rank $r < m, n$, where $m > n$, let $\mathbf{P}_{k,l}$ with $(k, l) \in \mathbf{D}$ be the set of observations where $\mathbf{D}$ is an observe data. Our task is to recover latent values in $\mathbf{P}$ given the observable values. Let the indices of r measured entries be $H \subset \{1, \ldots, m\}\{1, \ldots, n\}$. One of the steps in the matrix completion is to find the matrix with the lowest rank P that will be equal to a matrix A, that we are looking to recover, for all entries in H. The mathematical formula is as follows:

$$\begin{aligned} \min_{P} \quad & rank(P) \\ \text{subject to} \quad & P_{k,l} = A_{k,l} \forall k, l \in H \end{aligned} \tag{3.1}$$

Generally, missing entries cannot be recovered if one cannot assume an additional structure of data matrix $\mathbf{P}$. It is highlighted by Genes et al. (2016) that most of

the low-rank matrix can be estimated only if the number of entries sampled obeys $l \geq Cn^{1.25}rlogn$ where r is the rank of **P** and C is a constant that is positive with a probability of $1 - Cn^{-3}logn$ to recover missing entries. In deploying the matrix completion method missing values are estimated by minimising the decision variable subject to an orthogonal projector of a variable equal to an orthogonal projector of a data matrix **P** (Genes et al., 2016).

3.4.2 Expectation-Maximization Algorithm

The EM algorithm uses maximum-likelihood estimators to impute the missing values. For EM algorithm to impute missing values, it first estimates the parameter by maximizing the log-likelihood function. The EM algorithm iterates between the E-step and M-step to learn the parameter of the data, whenever the likelihood starts having a small difference it stops (Griffiths and Yuille, 2008).

The EM algorithm's first step is where the likelihood is calculated and is called the E-step. The formula is given by the log-likelihood function of the parameters given the data set. The data set must be partitioned into missing values and observed values.

Let our complete data set be Y, we then have $Y = (Y_{obs}, Y_{mis})$. The distribution of Y depends on the unknown parameter θ, *i.e*.

$$P(Y|\theta) = P(Y_{\text{obs}}, Y_{\text{mis}}|\theta) = P(Y_{\text{obs}}|\theta)P(Y_{\text{mis}}|Y_{\text{obs}}, \theta) \tag{3.2}$$

Equation 3.2 can be written as likelihood function,

$$\begin{aligned} L(\theta|Y) &= L(\theta|Y_{obs}, Y_{mis}) \\ &= aL(\theta|Y_{obs})P(Y_{obs}|Y_{mis}, \theta) \end{aligned} \tag{3.3}$$

Where a represents the constant term in the missing data mechanism and under the assumption of Missing at Random (MAR) it may be ignored. In our case a was ignored since we work under the MAR assumption. We get the following equation when we take the log both sides in equation 1.4.

$$l(\theta|Y) = l(\theta|Y_{\text{obs}}) + logP(Y_{\text{mis}}|Y_{\text{obs}}, \theta) + log(a) \tag{3.4}$$

Where $l(\theta|Y)$ is defined as the log-likelihood of complete data set, $l(\theta|Y_{\text{obs}})$ is the log likelihood of the observed data, $log(a)$ is the constant and $P(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$ is the predictive distribution of the missing data, given θ (Schafer, 1997).
Because Y_{mis} cannot be calculated directly since it is unknown, we have to guess θ denoted as $\theta^{(t)}$ to compute the expectation of $l(\theta|Y)$ with respect to $P(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$ *i.e.*

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E[l(\theta|Y)|Y_{\text{obs}}, \theta^{(t)}] \\ &= \int l(\theta|Y)P(Y_{mis}|Y_{obs}, \theta^{(t)})dY_{mis} \\ &= l(\theta|Y_{obs}) + \int logP(Y_{mis}|Y_{obs}, \theta)P(Y_{mis}|Y_{obs}, \theta^{(t)})dY_{mis} \end{aligned} \tag{3.5}$$

On the M-step of the EM algorithm, we obtain θ by maximizing the expectation of the log-likelihood of complete-data from the E-step. This can be expressed mathematically as:

$$\theta^{(t+1)} = \arg\max_{\theta} \mathrm{Q}(\theta|\theta^{(t)}) \tag{3.6}$$

When the EM algorithm finds the parameters of the estimates, it starts with the guess of θ^0 and then iterates between the E-step and M-step until it reaches its convergence step. It only converges when θ estimates differ with a small probability or nearly identical (Hershey and Olsen, 2007).

It is highlighted by Dempster et al. (1977) that the EM algorithm is an iterative technique that iterates between E-step and M-step. It aims to find the estimates of maximum likelihood of models with missing values. The four steps below briefly describe how the EM algorithms behave when it estimates unobserved values.

1. Initialization step: the step is computed by guessing the parameter estimate of θ^0.

2. Expectation step: the E-step, the assumption is made to fix the parameters $\theta^{(t-1)}$ from the first step, the expectation of missing values is computed.

3. Maximization step: the M-step, in this step we deploy the values obtained from the E-step and compute new parameter θ^t estimates that maximizes the variation of the likelihood function

4. Exit condition: the convergence step, if the probability of the observations is almost similar, it breaks or else it iterates from E-step

3.4.3 Summary of Methods

In this study, we compared the performance of the EM algorithm with the matrix completion method when recovering missing values in a data set. The performance of the above two techniques is observed through various degrees of missing values in a data set from 10% through 90% of missing values. We deployed a credit card fraud detection data set from a public data set (*Kaggle*).

When deploying the matrix completion, the missing values were imputed via Templates for First-Order Conic Solvers (TFOCS). Keshavan et al. (2010) indicated that one of the underlying assumptions of the matrix completion is to have the rank that is smaller than the number of instances and variables $(r << m, n)$, where r is rank, m is the number of instances and n is the number of variables. The EM

algorithm was deployed in this study due to its magnificent performance when estimating missing values. It uses the probabilities to imputes the missing values. It preserve the relationship with other variables and which is vital other than most techniques that imputes missing values (McLachlan and Krishnan, 2007). All of the above mentioned techniques were implemented in MATLAB.

3.4.4 Evaluation

KL divergence was used in this study to measure the performance of each model. Two distributions which are the original distribution and estimated distribution were compared using KL divergence to assess the performance of each model. The small value (approximately zero) of KL divergence obtained after comparing the two distributions illustrates a good performance of the model and a very large value presents a poor performance of the model.

KL divergence is simply the distance between observed probability $p(x)$ and estimated probability $q(x)$, it is used in machine learning and statistics to observe the information loss when the approximation is done. This evaluation accuracy is used mostly in fields like machine learning and Deep Learning, to be more precise, in image, pattern and speech recognition (Dong and Peng, 2013).

$$D(p||q) = \int p(y) log \frac{p(y)}{q(y)} dy \tag{3.7}$$

Below are the properties of the KL divergence.

1. If $D(p||p) = 0$ then it said to be self-similar.

2. $D(p||q) = 0$ if and only if $p = q$, it simply means the distance between those two probability functions is zero and the probability functions are identical.

3. If $D(p||q) \geq 0$ for all p, q, meaning there is a difference between the two probability functions. This property is called positivity.

A major take away from this evaluation metric is that the smaller the value of $D(p||q)$, the better. If a large number of KL divergence $D(p||q)$ is obtained, this simply means there is a huge difference between the true probability distribution function and the estimated distribution function. In this evaluation, $D(p||q) \neq D(q||p)$. The importance of KL divergence is to quantify the error between the estimation of distribution and the true distribution. In this research project, we check how far does the estimated probability $q(x)$ differ from the true probability $p(x)$ as we increase the ratio of missing values in the data set using two different imputation techniques. KL divergence was used in this study because we wanted mainly to find the difference between the ground-truth distribution and estimated distribution. KL divergence is easy to interpret as it tells how close the two distributions are.

3.5 Simulated Data

We present a six-phased experimental design to analyse the empirical information loss from recovering a ground-truth distribution from data sets with different degrees of missing values.

The six experimental phases are:

1. Construction of a ground-truth distribution with a fixed number of variables (4,6,8, and 10) using a Bayesian network to decompose the distribution into a set of independence assumptions;

2. Sample the data set from this distribution using forward sampling (a sampling technique where values are generated from a set of topologically ordered variables);

3. Remove degrees of missing values from multiple copies of the original data set;

TABLE 3.2: An example of the data generated from a Bayesian network with four variables modelling a Bernoulli distribution.

	Col_1	Col_2	Col_3	Col_4
1	2	2	1	2
2	2	2	1	1
3	1	1	1	2
...	...	...	...	...
40000	2	1	1	1

4. Store modified data sets per model (EM algorithm as EM and matrix completion as MC) on each level;

5. Store the estimated data set in to a distribution so it can be easy to compare it with ground-truth model;

6. Assess learned distribution by using the metric KL divergence to calculate the information loss from learning the modified data sets.

Figure 3.1 indicates an overview of the six-phased experimental design with each phase labelled respectively.
Table 3.2 provides an example of the generated data from a Bayesian network with four variables modelling a Binomial distribution. The rows in Table 3.2 indicate data points, and the columns indicate variables. In this case we only present four variables as a sample.

In (c) 9 copies of each complete data set are made, and a different percentage of values are randomly removed from each copy 10%, 20%, 30%, ... , 90% . More specifically, from the complete data set in (b) different percentages of observations were randomly removed. Table 3.3 illustrates an example of a modified data set after 50% of the samples from the complete data set (Table 3.2) were removed (NaN indicates the missing value).

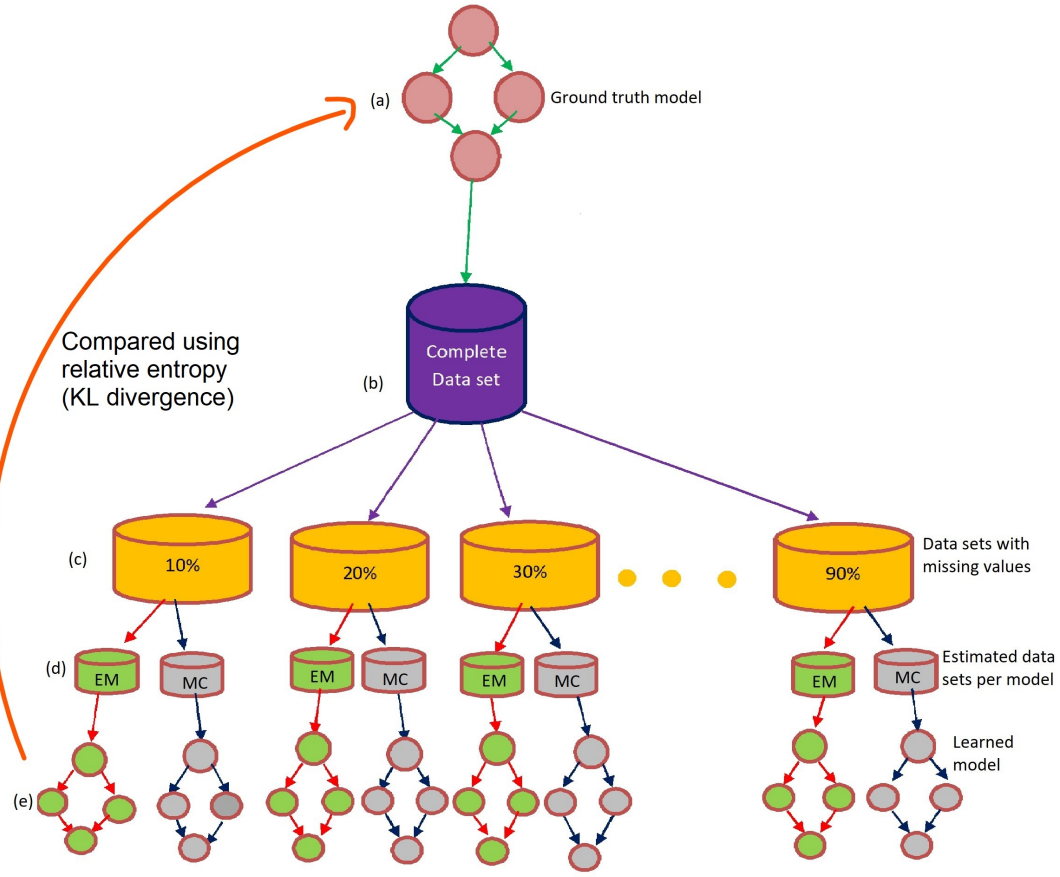


FIGURE 3.1: A figure illustrating the six-phased experimental design with each phase labelled respectively.

In (e) we now learn a new distribution from each of the modified data sets using Bayesian estimation (with a prior of $\alpha = 5$), the EM algorithm and matrix completion to learn the missing values and the original networks DAG. Bayesian learning was used to learn our parameters. In (f), to assess the amount of information lost

TABLE 3.3: An example of a modified data set after 50% of the samples in the complete data set from Table 3.2 were removed (NaN represent missing data).

	Col_1	Col_2	Col_3	Col_4
1	1	NaN	1	2
2	2	1	NaN	NaN
3	1	NaN	2	1
...	...	...	...	...
40000	2	1	NaN	2

learning from missing data the learned distribution is compared to the ground-truth distribution using a KL divergence measure in both the EM algorithm and matrix completion.

3.6 Results For Simulated Data

Figure 3.2 shows the performance of the EM algorithm and matrix completion on a Bayesian network that consists of 4 variables. On the x-axis, the plot shows the percentage of missing data and on the y-axis shows the relative entropy (KL Divergence) of the learned distribution with the EM algorithm and matrix completion with respect to the ground-truth distribution. As the legend shows, the blue line with an asterisk represents the EM algorithm and the red line with squares represents the matrix completion.

The data plotted in this figure are the mean we obtained after running our experiment with 15 trials and the error bars are used to indicate the standard deviation. From Figure 3.2, we observe that when we have approximately 70% of missing data, using the EM algorithm provides a better approximation of the ground distribution and above 77% of missing data we better of using the matrix completion than the EM algorithm for this ground-truth distribution. What we also observed from this plot is that, for up to 50% of the missing values, both models produce similar results as an approximation.

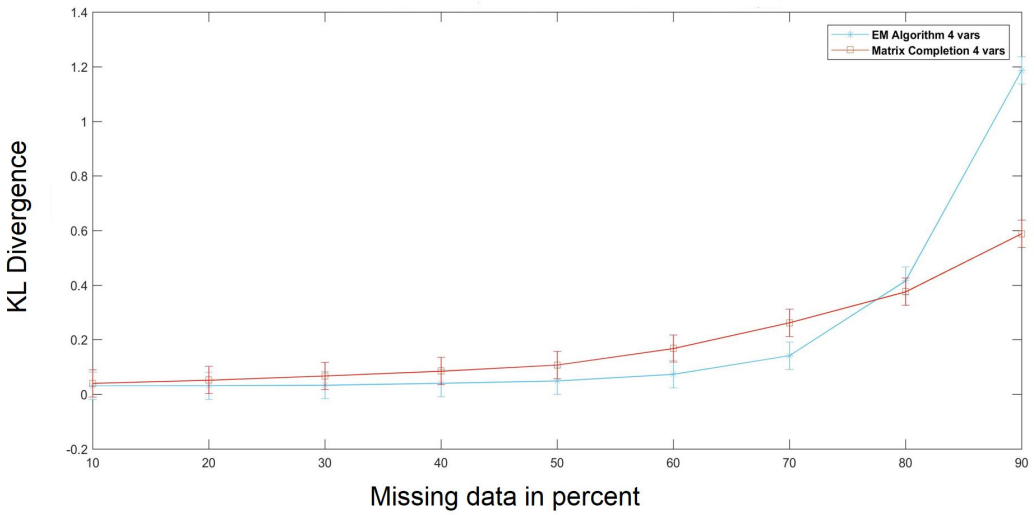


FIGURE 3.2: Comparison of the EM algorithm and matrix completion with 4 variables

Figure 3.3 presents the performance of the EM algorithm together with matrix completion on 6 variables. As compared to the performance of the 2 models on 4 variables, we observed that the model performed in a similar form there is not much difference in the trends. The crucial difference is the increase in the error rate of the matrix completion when 90% of the data is observed. The performance of the EM algorithm and matrix completion with 8 variables is presented in Figure 3.4. After 15 trials of our experiments, only the mean and standard deviation were plotted.

Consistent performance from both models is observed as we increase the number of variables. As we increase variables the EM algorithm produces better approximations than the matrix completion. A major difference between the performance of the EM algorithm and matrix completion is the increase in the error rate as missing values increases. Figure 3.5 presents the estimation of missing values with 10 variables.

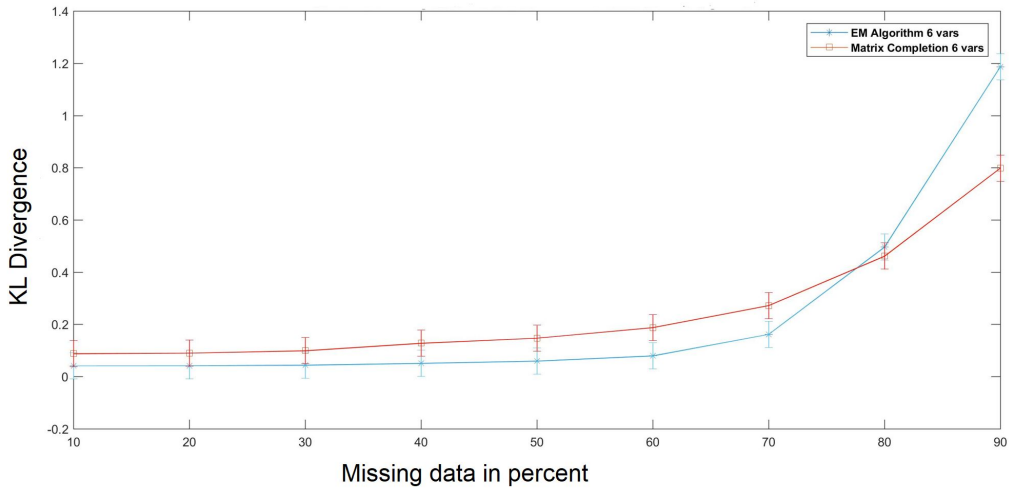


FIGURE 3.3: Comparison of the EM algorithm and matrix completion with 6 variables

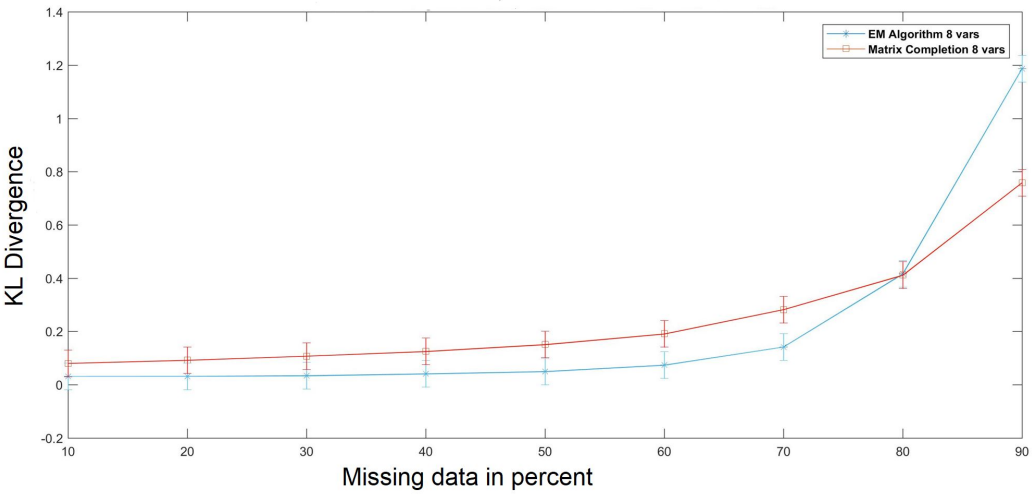


FIGURE 3.4: Comparison of the EM algorithm and matrix completion with 8 variables

3.7 Conclusion

In this method design, the execution of the EM algorithm together with matrix completion was observed. 2 different data sets were deployed. The first data set was binary data which follows Binomial distribution generated randomly using the

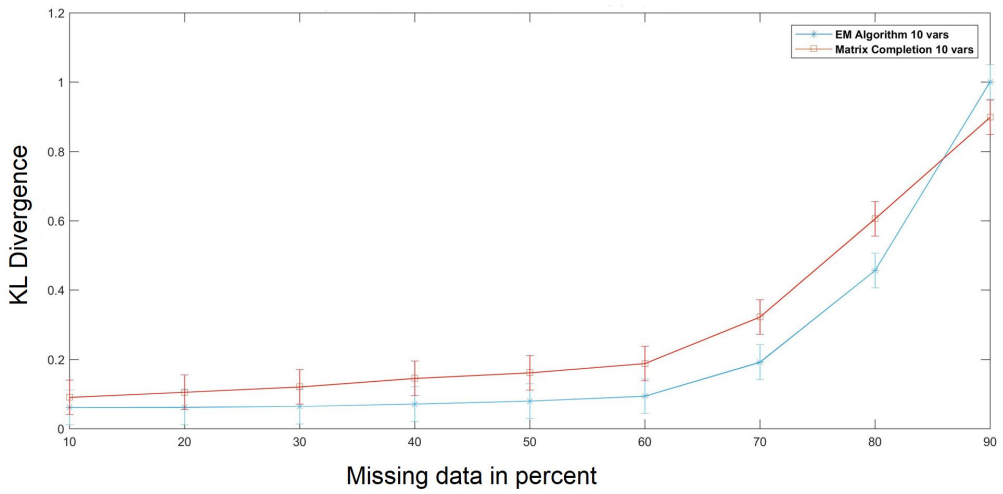


FIGURE 3.5: Comparison of the EM algorithm and matrix completion with 10 variables

Bayesian network structure which is regarded as a theoretical model. The second data set is a credit card fraud detection obtained from a public data set (*Kaggle*).

We randomly hid a certain percentage of the data set and then later estimated them using the EM algorithm and matrix completion. We used KL divergence to assess the performance of both the EM algorithm and matrix completion. The KL divergence in this case was used as an evaluation metric to compare 2 distributions, which are ground-truth and estimated distributions. The ground-truth distribution consist of the data that has been simulated, and the estimated distribution consist of simulated data plus estimated data based on the ratio of missing data. We then measured the difference between the two distributions using KL divergence to find the error.

Now that we had a theoretical understanding of the performance of the EM algorithm and matrix completion on a various number of variables and ratio of missing values, we applied this theory to a practical example using credit card fraud detection with missing data in the next chapter.

Chapter 4

Application: Imputing Missing Values for Credit Card Fraud Detection Data

4.1 Introduction

The concern of incomplete data is relatively common in many fields. Incomplete data introduces lack of quality into the analysis. Missing data affects the properties of statistical estimators and descriptions. Discarding instances that consist of missing values makes the data to lose its structure. In this study, we have imputed the missing values using the EM algorithm and matrix completion, under the missing at random (MAR) assumption.

This section discusses the application of our theoretical model learned in the previous chapter in order to recover missing values from a real-world data set (credit card fraud detection data) obtained from a public data set (*Kaggle*). The performance of both the EM algorithm and matrix completion will be assessed using KL divergence.

4.1.1 Application Data

One financial data set of credit card fraud detection from *Kaggle* is used for this experiment. The data set had 13 rows that consists of missing values, all the rows that had missing values were dropped using a Complete Case (CC) analysis. The shape of the data set was 307511×122. We manipulated the data to get 307498

data points and 23 variables. The selection of variables was based on the correlation among other variables and the absence of missing values. Python language was used for data manipulation. All the implementation of the EM algorithm and matrix completion was done in MATLAB.

This credit card fraud detection data is a mixed variable (categorical and continuous) data set with the following variables (mentioning only the variables that do not appear in Figure 4.1): Region population relative, flag mobile, flag employment phone, flag phone, flag email, region rating client, region rating client, region rating client work city, Hour app process start, reg city not work city, live city not work city, amount credited, amount annuity, and amount income total. Figure 4.1 represent a sample of top 10 variables.

Out[5]:

	0	1	2	3	4	5	6	
TARGET	1	0	0	0	0	0	0	
NAME_CONTRACT_TYPE	Cash loans	Cash loans	Revolving loans	Cash loans	Cash loans	Cash loans	Cash loans	Ca lo
FLAG_OWN_CAR	N	N	Y	N	N	N	Y	
FLAG_OWN_REALTY	Y	N	Y	Y	Y	Y	Y	
NAME_INCOME_TYPE	Working	State servant	Working	Working	Working	State servant	Commercial associate	St serv
NAME_EDUCATION_TYPE	Secondary / secondary special	Higher education	Secondary / secondary special	Secondary / secondary special	Secondary / secondary special	Secondary / secondary special	Higher education	Hig educat
NAME_FAMILY_STATUS	Single / not married	Married	Single / not married	Civil marriage	Single / not married	Married	Married	Marr
CODE_GENDER	M	F	M	F	M	M	F	
CNT_FAM_MEMBERS	1	2	1	2	1	2	3	
CNT_CHILDREN	0	0	0	0	0	0	1	

FIGURE 4.1: A Sample of the Credit Card Fraud Detection Data Set with 10 Variables.

We present the principle analysis of our application model (the application of our theoretical model on real-world data set) in Figure 4.2. The Figure 4.2 shows clearly how the real-world data set was applied in our theoretical model and how the conclusion came about. We performed Complete Case (CC) analysis and variable deletion on our data set deliberately, knowing well that the data set shall be reduced, we may throw away valuable variables and biases may be introduced. However,

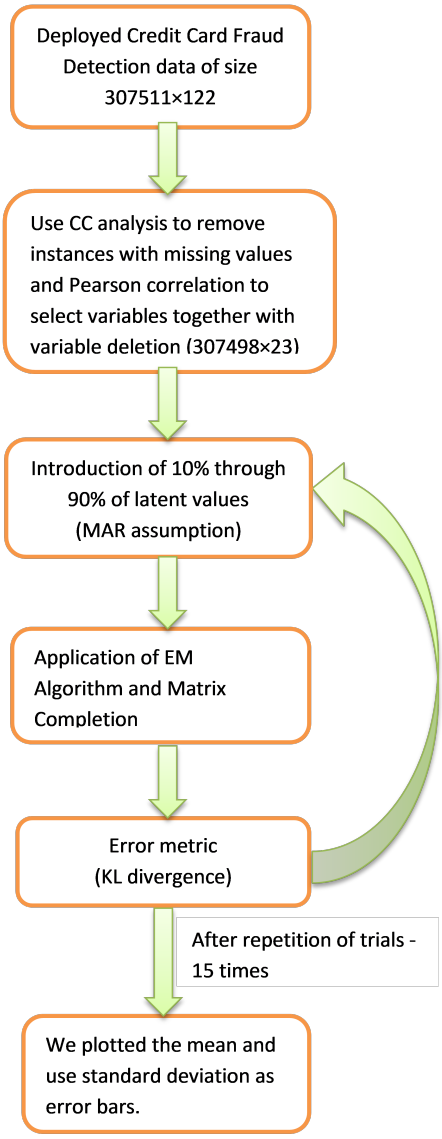


FIGURE 4.2: A figure illustrating principle analysis of the application model

the above-mentioned steps can not be avoided because we must use the complete data set so that we can be able to observe the performance of each algorithm.

Data transformation was applied on the data set so it can be in the range of 0 and 1

inclusively before working with it to avoid the confusion and for easy error observations. Because we used KL divergence to assess the error between the distributions, we had to learn a distribution on our data set and since the data ranged between 0 and 1, we used a continuous distribution. We grouped our data set into 4 bins, since we could not use only one bin (0 to 1) to observe the structure of our data.

To show how the 4 bins were created, if we let d_p to be our data points, then 4 bins that were used are $0 \leq d_p \leq 0.25$, $0.25 < d_p \leq 0.5$, $0.5 < d_p \leq 0.75$, and $0.75 < d_p \leq 1$. We decided to go with 4 bins because we have noticed that as we increase the bins beyond 4, some bins become empty and our distribution shall be discontinuous.

Figure 4.3 manifests clearly the performance of the EM algorithm and matrix completion method on a real-world data set, the credit card fraud detection data. The x-axis presents the ratio of missing values in a data set from 10% through 90% and the y-axis shows our error metric (KL Divergence). Two-line plots are shown in Figure 4.3 in one plot, the blue line with asterisk presents the EM algorithm while the red line with square boxes is for the matrix completion.

We performed 15 trials of estimating missing values for each missing ratio on a data and the mean was plotted, standard deviations were used as our error bars on the plot. Figure 4.3 presents a dramatic difference in the performance of the 2 models from as little as 10% of missing values, especially from 40% to approximately 80%. We can conclude from this plot that the EM algorithm provides better approximation when less than 70% of the data is missing.

The most interesting thing observed from this comparison is that the EM algorithm only has an exponential increase in the error when more than 60% of the data is missing. When less than 60% of the data is missing, we observe a small error. On the other hand, the matrix completion shows a monotonic and linear increase in error from as little as 10%. The only time when the matrix completion performs better is when the missing ratio is above 80%.

Another interesting observation is the gap in performance between 80% and 90%, EM algorithm performed poorly compared to matrix completion. The exponential performance of the EM algorithm explains the gap. The EM algorithm performed more exponential and the matrix completion performed more linear, this happend because of the massive amount of data used in the model. We then can conclude that EM algorithm performs well with a huge amount of data. A trivial answer we get from this plot is that, the increase in missing values implies the increase in the error rate.

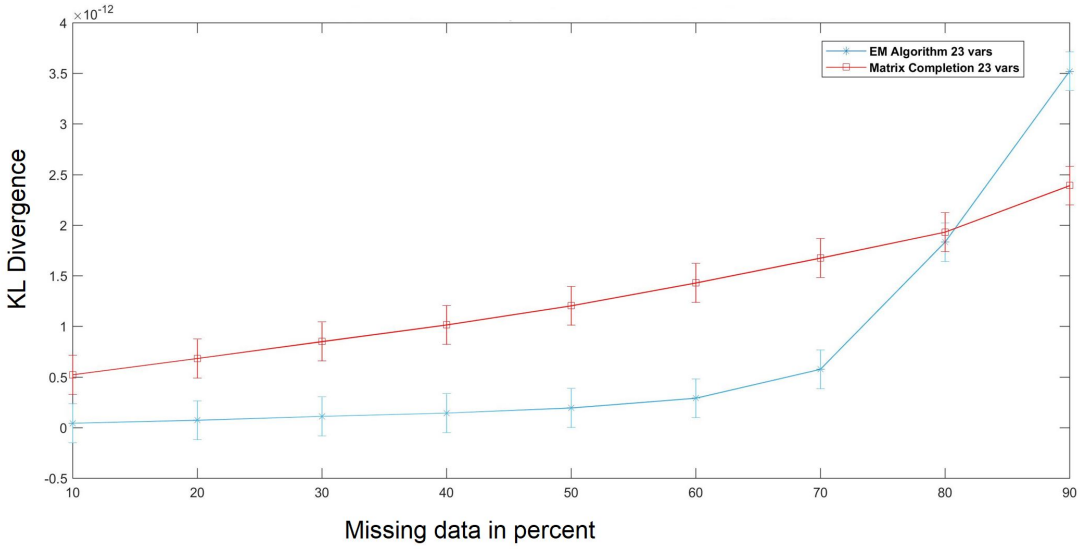


FIGURE 4.3: Comparison of the EM algorithm and matrix completion on a credit card default data

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Incomplete data sets have been a problem in many research fields that deal with data for decision making. Researchers such as statisticians, policymakers and analysts are facing this problem on a day to day basis. If researchers fail to deal with missing values properly, then their conclusions may be inaccurate and biased.

This research report explored the performance of the EM algorithm and the matrix completion to impute the missing values on a data set that follows a Binomial distribution for a theoretical model and practical data set obtained from *Kaggle*. The task was to find the best technique between the EM algorithm and the matrix completion when imputing missing values by observing the error estimations of each model using KL divergence.

The main contribution of this research report was to demonstrate the importance of imputing missing values and to rigorously represent an empirical analysis between the EM algorithm and matrix completion when recovering missing values of different degrees in a data set. Lastly, the theoretical model was applied to recover missing values in a financial data set that is used to detect fraud in credit cards.

Figure 3.2, Figure 3.3, Figure 3.4, and Figure 3.5 presented the performance of the EM algorithm and matrix completion on simulated data with different variables (4, 6, 8, and 10) and various degree of missing values. The error was recorded after the estimation of a certain ratio of missing values for each model.

After we performed 15 trials of the experiment, we plotted the mean and standard deviation as error bars. We noted from the plots that the EM algorithm generally outperformed the matrix completion when 10% of the data is missing through approximately when 75% of the data is missing. From the above-mentioned plots, we observed that the only time when the matrix completion performs better than the EM is when more than 75% of the data is missing. Based on the findings, it was then concluded that the EM algorithm is better for approximating missing values than the matrix completion. Generally, we conclude that to make accurate predictions we need to have at most 60% of missing values.

We observed from our theoretical and application models (A model that was trained using a real world data set) that we can improve the performance of the EM algorithm by increasing the training data set, decrease the number of missing values, and run algorithm for a longer time. On the other hand, matrix completion performs better with small sample and few missing values.

From both the theoretical and application model, we observe consistent performance from both models. The EM algorithm is almost every time above the matrix completion on all the plots. The EM algorithm is better when estimating missing values of up to approximately 75%. We noticed that the type of distribution does not affect the performance of the 2 algorithms since we can observe a similar pattern on both discrete and continuous distributions. The KL divergence was used to compare the 2 distributions which are ground-truth and estimated distribution. The difference was then recorded as the error of the model.

5.2 Future Work

For future work, different kinds of data sets (at least 2 structured and unstructured data) may be deployed. In this study, we used a matrix of rank 12, one may be curious and want to compare the EM algorithm's efficiency with the matrix completion

technique with a smaller rank and also use a matrix of $N \times N$ since the matrix used in this study was 307497×23. Cai et al. (2010) noted a great performance of matrix completion on $N \times N$ matrices. Lastly, the computation of the EM algorithm on the financial data set used takes about 2 to 3 days, for future work the EM algorithm can be optimised.

Bibliography

Agarwal, Piyush and Arun K Tangirala (2017). "Reconstruction of missing data in multivariate processes with applications to causality analysis". In: *International Journal of Advances in Engineering Sciences and Applied Mathematics* 9.4, pp. 196–213.

Ajoodha, Ritesh and Benjamin Rosman (2017). "Tracking influence between naïve Bayes models using score-based structure learning". In: *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*. IEEE, pp. 122–127.

Bae, Harold, Stefano Monti, Monty Montano, Martin H Steinberg, Thomas T Perls, and Paola Sebastiani (2016). "Learning Bayesian networks from correlated data". In: *Scientific reports* 6, p. 25156.

Balakrishnan, Sivaraman, Martin J Wainwright, Bin Yu, et al. (2017). "Statistical guarantees for the EM algorithm: From population to sample-based analysis". In: *The Annals of Statistics* 45.1, pp. 77–120.

Cai, Jian-Feng, Emmanuel J Candès, and Zuowei Shen (2010). "A singular value thresholding algorithm for matrix completion". In: *SIAM Journal on optimization* 20.4, pp. 1956–1982.

Celeux, Gilles, Stéphane Chrétien, Florence Forbes, and Abdallah Mkhadri (2001). "A component-wise EM algorithm for mixtures". In: *Journal of Computational and Graphical Statistics* 10.4, pp. 697–712.

Combettes, Patrick L and Jean-Christophe Pesquet (2011). "Proximal splitting methods in signal processing". In: *Fixed-point algorithms for inverse problems in science and engineering*. Springer, pp. 185–212.

Demirtas, Hakan (2004). "Modeling incomplete longitudinal data". In: *Journal of Modern Applied Statistical Methods* 3.2, p. 5.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Dong, Yiran and Chao-Ying Joanne Peng (2013). "Principled missing data methods for researchers". In: *SpringerPlus* 2.1, p. 222.

Emmanuel, Candes and Benjamin Recht (2009). "Exact matrix completion via convex optimization". In:

Fekade, Berihun, Taras Maksymyuk, Maryan Kyryk, and Minho Jo (2017). "Probabilistic recovery of incomplete sensed data in IoT". In: *IEEE Internet of Things Journal* 5.4, pp. 2282–2292.

Friedman, Nir (2013). "The Bayesian structural EM algorithm". In: *arXiv preprint arXiv:1301.7373*.

Genes, Cristian, Inaki Esnaola, Samir M Perlaza, Luis F Ochoa, and Daniel Coca (2016). "Recovering missing data via matrix completion in electricity distribution systems". In: *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, pp. 1–6.

Geng, Xin, Kate Smith-Miles, Zhi-Hua Zhou, and Liang Wang (2010). "Face image modeling by multilinear subspace analysis with missing values". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41.3, pp. 881–892.

Griffiths, T and Alan Yuille (2008). "A primer on probabilistic inference". In: *The probabilistic mind: Prospects for Bayesian cognitive science*, pp. 33–57.

Heijden, Geert JMG Van der, A Rogier T Donders, Theo Stijnen, and Karel GM Moons (2006). "Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example". In: *Journal of clinical epidemiology* 59.10, pp. 1102–1109.

Hershey, John R and Peder A Olsen (2007). "Approximating the Kullback Leibler divergence between Gaussian mixture models". In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. Vol. 4. IEEE, pp. IV–317.

Jerez, José M, Ignacio Molina, Pedro J García-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, and Leonardo Franco (2010). "Missing data imputation using statistical and machine learning methods in a real breast cancer problem". In: *Artificial intelligence in medicine* 50.2, pp. 105–115.

Juhola, Martti and Jorma Laurikkala (2013). "Missing values: how many can they be to preserve classification reliability?" In: *Artificial Intelligence Review* 40.3, pp. 231–245.

Kaggle. URL: https://www.kaggle.com/mishra5001/credit-card?select=application_data.csv (visited on 2019).

Keshavan, Raghunandan H, Andrea Montanari, and Sewoong Oh (2010). "Matrix completion from a few entries". In: *IEEE transactions on information theory* 56.6, pp. 2980–2998.

Liu, Yuzhe and Vanathi Gopalakrishnan (2017). "An overview and evaluation of recent machine learning imputation methods using cardiac imaging data". In: *Data* 2.1, p. 8.

McLachlan, Geoffrey J and Thriyambakam Krishnan (2007). *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons.

Nelwamondo, Fulufhelo V, Dan Golding, and Tshilidzi Marwala (2013). "A dynamic programming approach to missing data estimation using neural networks". In: *Information Sciences* 237, pp. 49–58.

Nigam, Kamal, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell (2000). "Text classification from labeled and unlabeled documents using EM". In: *Machine learning* 39.2-3, pp. 103–134.

Purwar, Archana and Sandeep Kumar Singh (2015). "Hybrid prediction model with missing value imputation for medical data". In: *Expert Systems with Applications* 42.13, pp. 5621–5631.

Schafer, Joseph L (1997). *Analysis of incomplete multivariate data*. Chapman and Hall/CRC.

Schafer, Joseph L and John W Graham (2002). "Missing data: our view of the state of the art." In: *Psychological methods* 7.2, p. 147.

Schlomer, Gabriel L, Sheri Bauman, and Noel A Card (2010). "Best practices for missing data management in counseling psychology." In: *Journal of Counseling psychology* 57.1, p. 1.

Schmitt, P, J Mandel, and M Guedj (2015). "A Comparison of Six Methods for Missing Data Imputation. J Biomet Biostat 6: 224. doi: 10.4172/2155-6180.100022 4 J Biomet Biostat ISSN: 2155-6180 JBMBS, an open access journal Page 2 of 6 Volume 6• Issue 1• 1000224 The Breast cancer 2 dataset provides a 70 genes signature for prediction of metastasis-free survival, measured on 89 tumor samples [17]". PhD thesis. These 70 genes highlight three grades of tumors:"poorly," intermediate".

Sehgal, Muhammad Shoaib B, Iqbal Gondal, and Laurence S Dooley (2005). "Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data". In: *Bioinformatics* 21.10, pp. 2417–2423.

Solanki, Kaushal, Kenneth Sullivan, Upamanyu Madhow, BS Manjunath, and Shivkumar Chandrasekaran (2006). "Provably secure steganography: Achieving zero KL divergence using statistical restoration". In: *Image Processing, 2006 IEEE International Conference on*. IEEE, pp. 125–128.

Stefanakos, Ch N and GA Athanassoulis (2001). "A unified methodology for the analysis, completion and simulation of nonstationary time series with missing values, with application to wave data". In: *Applied ocean research* 23.4, pp. 207–220.

Stephens, Matthew and Paul Scheet (2005). "Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation". In: *The American Journal of Human Genetics* 76.3, pp. 449–462.

Stuart, Elizabeth A, Melissa Azur, Constantine Frangakis, and Philip Leaf (2009). "Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative". In: *American journal of epidemiology* 169.9, pp. 1133–1139.

Takács, Gábor, István Pilászy, Bottyán Németh, and Domonkos Tikk (2008). "Matrix factorization and neighbor based algorithms for the netflix prize problem". In: *Proceedings of the 2008 ACM conference on Recommender systems*, pp. 267–274.

Wang, Hai and Shouhong Wang (2010). "Mining incomplete survey data through classification". In: *Knowledge and information systems* 24.2, pp. 221–233.

White, Ian R and John B Carlin (2010). "Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values". In: *Statistics in medicine* 29.28, pp. 2920–2931.

Yu, Qi, Yoan Miche, Emil Eirola, Mark Van Heeswijk, Eric SéVerin, and Amaury Lendasse (2013). "Regularized extreme learning machine for regression with missing data". In: *Neurocomputing* 102, pp. 45–51.

Zhang, Wen, Ye Yang, and Qing Wang (2011). "Handling missing data in software effort prediction with naive Bayes and EM algorithm". In: *Proceedings of the 7th International Conference on Predictive Models in Software Engineering*, pp. 1–10.