Data imputation for visualisation

Contents

[1 Introduction 2](#_Toc519674270)

[2 Data with gaps 2](#_Toc519674271)

[2.1 Representation of incomplete points 2](#_Toc519674272)

[2.2 Distance between complete and incomplete points 3](#_Toc519674273)

[2.3 Distance between two incomplete points 3](#_Toc519674274)

[2.4 kNN data repair 4](#_Toc519674275)

[2.5 SVD with gaps 4](#_Toc519674276)

[3 Results 6](#_Toc519674277)

[4 Conclusion 11](#_Toc519674278)

[References 12](#_Toc519674279)

# Introduction

Let us consider problem of visualisation of incomplete data (data with missing values).

There are two main approaches of data visualisation: multidimensional scaling and principal objects. Principal object approaches can be presented as partial case of multidimensional scaling [1]. The main difference for us is creation of close form procedure of finding of coordinates of any point in low dimensional space by principal objects approach. Alternative multidimensional scaling approach works with similarity matrix only. This means that there is no way to find coordinates of low dimensional projection of any new point. There are many different principal objects [2, 3, 4, 5, 6, 7, 8, 9] from Principal Components (PCs) and principal curves to elastic maps and principal graphs.

All data visualisation methods are based on the distance between two points or distance between data point and principal object. Really we always can transfer any particular task to calculation distances between two points.

Distance between two incomplete points can be easily defined (see “Distance between two incomplete points”). The main drawback of such approach is long calculation. This does not mean usage of twice more time but adding one or two orders for time consumption.

As a result it is preferable to complete data (impute all missing values) before visualisation. There are many different data imputation approaches [10]. Widely used multiple imputation is not appropriate for our purpose because these methods create cloud of points instead of one point and confuse user by showing, for example, wrong density.

We consider four methods of data imputation: probabilistic PCA [11], kNN [12], unrestricted and restricted singular value decomposition proposed in this work.

There are very important question for method comparison: what does it mean good data imputation?

# Data with gaps

In this section we consider calculation of distance for data with gaps.

## Representation of incomplete points

There are two representation of points with gaps (missed coordinates): unrestricted (we have linear manifold with missed coordinates as basis) or restricted (for each coordinate we have interval of acceptable values).

Restricted representation can have individual restriction (intervals) for each data point or common restriction (the same intervals for all points).

Let us consider set of intervals where is low border and us upper border of interval . Let incomplete point contains missed values in coordinates where Let us denote

|  |  |
| --- | --- |
|  | (1) |

point with zeros on the place of missed coordinates. Then point can be represented as multidimensional interval

|  |  |
| --- | --- |
|  | (2) |

It is necessary to note that representation (2) is valid for both restricted and unrestricted representation with finite intervals in case of restricted representation and infinite intervals otherwise.

## Distance between complete and incomplete points

There are two possible tasks: calculation of distance between complete and incomplete data points and calculation of distance between two incomplete data points. In this section se consider calculation of distance between complete point and multidimensional interval (2) for data point with set of missed coordinates . Set of known coordinates is . In accordance with definition of distance between point and set [13] we have

|  |  |
| --- | --- |
|  | (3) |

For Euclidean distance we can write

All summands in the sum under infimum operator are independent and we can rewrite this formula as

|  |  |
| --- | --- |
|  | (4) |

For unrestricted representation we have the zero second summand in (4). For restricted intervals we have

|  |  |
| --- | --- |
|  | (5) |

Formula (5) can be used for unrestricted case too (with ).

## Distance between two incomplete points

Let us consider two incomplete points and with set of missed attributes for the first point and for the second one. We will use infimum of distances between points of different sets as distance between sets. In this case we can write

|  |  |
| --- | --- |
|  | (6) |

It can be seen that the closest points in intervals and are points and defined as:

|  |  |
| --- | --- |
|  | (7) |

where

|  |  |
| --- | --- |
|  | (8) |

is the mean point of intervals and intersection. For infinite intervals we select .

## kNN data repair

The main idea of this method is to use mean (weighted mean) value of required attribute of k nearest neighbours of target point. For this method we have to calculate distance between two incomplete points. Currently implemented version required space for complete distance matrix.

## SVD with gaps

The problem of SVD with missing data was considered in many works (see, for example, [14, 15, 16]). The main difference of our approach is consideration of restricted problem.

We have data matrix with elements where is number of object and is number of attribute. Attributes can have missed values. Let us denote missed values as ‘@’. For each attribute of each data point we define interval of acceptable values .

We want to use found principal components to restore data:

|  |  |
| --- | --- |
|  | (9) |

where is number of principal component, are vectors of principal component (PC) and is length of projection of data point to principal component. We need to find such components that for known values the difference between known value and reconstruction is as small as possible and for unknown values a reconstruction has to belong the interval of accepted values. This means that we want to find principal components such that minimise function

|  |  |
| --- | --- |
| subject to | (10) |

This problem can be solved by greedy algorithm. Let us define Then we find one PC for data , subtract reconstruction from this data and will search the second principal component as the first principal component for data matrix .

As a result we can consider problem of one principal component search only. For this we have to find vectors which minimise

|  |  |
| --- | --- |
| subject to | (11) |

Let us denote number of known values in attribute of data matrix as . We now will minimise function (11). Let us initialise vector by following values:

|  |  |
| --- | --- |
|  | (12) |

To initialise vector we randomly generate all coordinates of this vector and then normalise vector to unit length.

Now we can use vectors and to calculate vector . Then use fixed vectors and to calculate , and then use vectors and to calculate .

Let us consider all three steps of algorithm.

Search of for fixed and . First of all, function (11) has minimum in the point of zero derivative or in the border of intervals . Let us find the derivative with respect to :

After some algebra we have

|  |  |
| --- | --- |
|  | (13) |

where .

Now we can find projection point as . If for all missed values of data point we have then found vector is solution of problem (11) with fixed and . Otherwise we have to provide hold of restriction. We can reformulate problem (11) with fixed and for one point d as finding of projection of point with unknown coordinates onto unit vector with restriction that . It means that we have to minimise distance between points and :

|  |  |
| --- | --- |
|  | (14) |

Solution of this problem for fixed is

|  |  |
| --- | --- |
|  | (15) |

Now we can solve problem (14) with following algorithm:

1. Solve unconstrained problem (14) by formula (13).
2. Correct projection by formula (15).
3. Solve problem (14) for point

|  |  |
| --- | --- |
|  | (16) |

by formula

|  |  |
| --- | --- |
|  | (17) |

1. Repeat steps 2 and 3 until convergence.

Really we firstly search projection by usage of known coordinates only. Then we project found point into multidimensional interval . Then we change the length of projection to minimise target function and repeat this process. Unfortunately it is possible to have final projection which is not satisfied restrictions. In this case we have contradiction in restrictions for specified vectors and y.

The next step is search of vector for known and . For this purpose we calculate derivative

After some algebra we have

|  |  |
| --- | --- |
|  | (18) |

It is evident that during this step function (11) can decrease only.

The last step is to find vector for known and .

After some algebra we have

|  |  |
| --- | --- |
|  | (19) |

And the last operation is normalisation of vector to unit length.

It is very important to stress that the last step of algorithm must be calculation of vector because this step provides hold of conditions.

# Results

Test of work with gaps is designed as following.

1. We take a complete database
2. Create specified fraction of randomly selected gaps.
3. Calculate projection of original data onto the first three PCs
4. Repair data by one of considered method.
5. Calculate projection of repaired data onto the first three principal components.
6. Compare results and spent time.

Results comparison assumed comparison of

1. Sum of angles of the first three PCs calculated for original () and repaired () data (ideal value 0):
2. Sum of length of projections of PCs calculated for repaired data onto three first PCs calculated for repaired data (ideal value 3):
3. Fraction of the same neighbours (FSNN) from the first 10 NN. Calculated mean value for all points. Ideal value is 1.
4. Fraction of coordinates with statistically significant mean (variance) in the original and repaired databases. This two indicators are calculated for multidimensional space and for projections onto three first PCs.

Time comparison was presented for the usual laptop (CPU Intel® Core™ i5-6300U, 2.40GHz, 16 GB RAM). For PPCA on breast cancer [15] database required time exceed 16 hours on high performance computer. For bladder cancer database [16] time of PPCA is more than hundred time greater than for RSDV for 20% gaps and more than 10,000 times greater for 10% gaps. Unexpectedly for 5% gaps 16 hours were not enough.

Table . Time spent (seconds)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Database | Fraction of gaps | Time kNN (10) | Time USVD | Time RSVD | PPCA |
| Breast cancer | 20% | 129.273 | 2,589.258 | 7,877.319 | >57,600(16h) |
| Breast cancer | 10% | 107.370 | 2,704.260 | 5,880.690 | >57,600(16h) |
| Breast cancer | 5% | 79.458 | 2,673.139 | 5,618.206 | >57,600(16h) |
| Breast cancer | 20% | 126.794 | 2,611.911 | 7,970.122 | >57,600(16h) |
| Breast cancer | 10% | 93.727 | 2,547.161 | 5,955.101 | >57,600(16h) |
| Breast cancer | 5% | 81.858 | 2,578.382 | 5,310.829 | >57,600(16h) |
| Bladder cancer | 20% | 0.901 | 2.523 | 6.616 | 701.190 |
| Bladder cancer | 10% | 0.579 | 2.219 | 5.798 | 52,085.284 |
| Bladder cancer | 5% | 0.407 | 2.666 | 6.861 | >57,600(16h) |

Table . Comparison with original database, green background highlights the best method for each database

| Database | Projection onto 3 PCS | | High dimensional space | |
| --- | --- | --- | --- | --- |
| 10FSNN | 20FSNN | 10FSNN | 20FSNN |
| Breast cancer 3 PCs of original | 0.2605 | 0.3668 | N/A | N/A |
| Breast cancer 20% gaps 10NN | 0.2573 | 0.3629 | 0.7780 | 0.8161 |
| Breast cancer 20% gaps USVD | 0.2577 | 0.3647 | 0.8773 | 0.9080 |
| Breast cancer 20% gaps RSVD | 0.2584 | 0.3643 | 0.8773 | 0.9077 |
| Breast cancer 10% gaps 10NN | 0.2545 | 0.3650 | 0.8829 | 0.9089 |
| Breast cancer 10% gaps USVD | 0.2587 | 0.3661 | 0.9210 | 0.9455 |
| Breast cancer 10% gaps RSVD | 0.2587 | 0.3654 | 0.9199 | 0.9442 |
| Breast cancer 5% gaps 10NN | 0.2563 | 0.3652 | 0.9308 | 0.9533 |
| Breast cancer 5% gaps USVD | 0.2580 | 0.3668 | 0.9556 | 0.9626 |
| Breast cancer 5% gaps RSVD | 0.2587 | 0.3666 | 0.9556 | 0.9633 |
| Breast cancer 20% gaps 10NN | 0.2535 | 0.3615 | 0.7818 | 0.8208 |
| Breast cancer 20% gaps USVD | 0.2535 | 0.3652 | 0.8769 | 0.9051 |
| Breast cancer 20% gaps RSVD | 0.2524 | 0.3649 | 0.8794 | 0.9072 |
| Breast cancer 10% gaps 10NN | 0.2510 | 0.3649 | 0.8839 | 0.9128 |
| Breast cancer 10% gaps USVD | 0.2587 | 0.3654 | 0.9241 | 0.9395 |
| Breast cancer 10% gaps RSVD | 0.2587 | 0.3656 | 0.9227 | 0.9386 |
| Breast cancer 5% gaps 10NN | 0.2580 | 0.3650 | 0.9308 | 0.9537 |
| Breast cancer 5% gaps USVD | 0.2608 | 0.3664 | 0.9535 | 0.9629 |
| Breast cancer 5% gaps RSVD | 0.2608 | 0.3659 | 0.9531 | 0.9631 |
| Bladder cancer 3 PCs of original | 0.7775 | 0.8450 | N/A | N/A |
| Bladder cancer 20% gaps 10NN | 0.7775 | 0.8450 | 0.8750 | 0.9050 |
| Bladder cancer 20% gaps USVD | 0.7725 | 0.8450 | 0.9375 | 0.9563 |
| Bladder cancer 20% gaps RSVD | 0.7725 | 0.8450 | 0.9375 | 0.9563 |
| Bladder cancer 20% gaps PPCA | 0.2575 | 0.5475 | 0.2175 | 0.5113 |
| Bladder cancer 10% gaps 10NN | 0.7775 | 0.8425 | 0.9325 | 0.9475 |
| Bladder cancer 10% gaps USVD | 0.7775 | 0.8500 | 0.9625 | 0.9800 |
| Bladder cancer 10% gaps RSVD | 0.7775 | 0.8500 | 0.9625 | 0.9800 |
| Bladder cancer 10% gaps PPCA | 0.2400 | 0.5163 | 0.6825 | 0.7238 |
| Bladder cancer 5% gaps 10NN | 0.7775 | 0.8438 | 0.9625 | 0.9825 |
| Bladder cancer 5% gaps USVD | 0.7800 | 0.8438 | 0.9675 | 0.9863 |
| Bladder cancer 5% gaps RSVD | 0.7800 | 0.8438 | 0.9675 | 0.9863 |

Table . Comparison with projection to the first three PCs with original data projection

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 3 PCs for | Angle (degree) | Len | 10FSNN | 20FSNN |
| Original | 0.0000 | 3.0000 | 1.0000 | 1.0000 |
| Breast cancer 20% gaps 10NN | 24.6418 | 2.9722 | 0.8685 | 0.9047 |
| Breast cancer 20% gaps USVD | 17.9434 | 2.9836 | 0.9538 | 0.9668 |
| Breast cancer 20% gaps RSVD | 17.9522 | 2.9836 | 0.9542 | 0.9664 |
| Breast cancer 10% gaps 10NN | 15.5647 | 2.9876 | 0.9224 | 0.9441 |
| Breast cancer 10% gaps USVD | 11.8301 | 2.9928 | 0.9647 | 0.9757 |
| Breast cancer 10% gaps RSVD | 11.8263 | 2.9928 | 0.9647 | 0.9757 |
| Breast cancer 5% gaps 10NN | 10.6468 | 2.9942 | 0.9549 | 0.9689 |
| Breast cancer 5% gaps USVD | 8.0580 | 2.9967 | 0.9829 | 0.9846 |
| Breast cancer 5% gaps RSVD | 8.0568 | 2.9967 | 0.9829 | 0.9850 |
| Breast cancer 20% gaps 10NN | 24.0917 | 2.9708 | 0.8615 | 0.8930 |
| Breast cancer 20% gaps USVD | 17.9314 | 2.9836 | 0.9524 | 0.9638 |
| Breast cancer 20% gaps RSVD | 17.9283 | 2.9836 | 0.9528 | 0.9640 |
| Breast cancer 10% gaps 10NN | 16.2129 | 2.9879 | 0.9283 | 0.9530 |
| Breast cancer 10% gaps USVD | 11.7238 | 2.9930 | 0.9682 | 0.9811 |
| Breast cancer 10% gaps RSVD | 97.2065 | 2.9929 | 0.9678 | 0.9801 |
| Breast cancer 5% gaps 10NN | 10.5382 | 2.9944 | 0.9538 | 0.9692 |
| Breast cancer 5% gaps USVD | 8.0236 | 2.9967 | 0.9780 | 0.9825 |
| Breast cancer 5% gaps RSVD | 8.0299 | 2.9967 | 0.9769 | 0.9823 |
| Bladder cancer 20% gaps 10NN | 59.0673 | 2.8518 | 0.9200 | 0.9313 |
| Bladder cancer 20% gaps USVD | 37.9001 | 2.9204 | 0.9700 | 0.9825 |
| Bladder cancer 20% gaps RSVD | 37.8821 | 2.9205 | 0.9700 | 0.9825 |
| Bladder cancer 20% gaps PPCA | 266.726 | 0.0710 | 0.2675 | 0.5350 |
| Bladder cancer 10% gaps 10NN | 37.3143 | 2.9328 | 0.9375 | 0.9613 |
| Bladder cancer 10% gaps USVD | 24.7549 | 2.9655 | 0.9750 | 0.9800 |
| Bladder cancer 10% gaps RSVD | 24.7492 | 2.9655 | 0.9750 | 0.9800 |
| Bladder cancer 10% gaps PPCA | 266.338 | 0.0777 | 0.2725 | 0.5138 |
| Bladder cancer 5% gaps 10NN | 20.2705 | 2.9767 | 0.9775 | 0.9838 |
| Bladder cancer 5% gaps USVD | 17.4027 | 2.9825 | 0.9825 | 0.9850 |
| Bladder cancer 5% gaps RSVD | 17.4022 | 2.9825 | 0.9825 | 0.9850 |

Table . Comparison of means (variances): fraction of attributes which have mean (variance) statistically significantly (with significance level 95%) different from mean (variance) of original dataset

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Database | Mean | | | | Variance | | | |
| 10NN | USVD | RSVD | PPCA | 10NN | USVD | RSVD | PPCA |
| Breast cancer 20% gaps | 0.0035 | 0.0001 | 0.0001 | N/A | 0.2868 | 0.0102 | 0.0100 | N/A |
| Breast cancer 10% gaps | 0.0000 | 0.0000 | 0.0000 | N/A | 0.0075 | 0.0015 | 0.0015 | N/A |
| Breast cancer 5% gaps | 0.0000 | 0.0000 | 0.0000 | N/A | 0.0007 | 0.0002 | 0.0002 | N/A |
| Breast cancer 20% gaps | 0.0038 | 0.0000 | 0.0000 | N/A | 0.2956 | 0.0103 | 0.0101 | N/A |
| Breast cancer 10% gaps | 0.0000 | 0.0000 | 0.0000 | N/A | 0.0097 | 0.0015 | 0.0016 | N/A |
| Breast cancer 5% gaps | 0.0000 | 0.0000 | 0.0000 | N/A | 0.0005 | 0.0003 | 0.0003 | N/A |
| Bladder cancer 20% gaps | 0.0000 | 0.0000 | 0.0000 | 0.0339 | 0.0313 | 0.0158 | 0.0158 | 1.0000 |
| Bladder cancer 10% gaps | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0066 | 0.0040 | 0.0040 | 0.3343 |
| Bladder cancer 5% gaps | 0.0000 | 0.0000 | 0.0000 | N/A | 0.0026 | 0.0023 | 0.0023 | N/A |

Table . Comparison of means (variances) in projection onto the first three PCs: fraction of attributes which have mean (variance) statistically significantly (with significance level 95%) different from mean (variance) of original dataset

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Database | Mean | | | | Variance | | | |
| 10NN | USVD | RSVD | PPCA | 10NN | USVD | RSVD | PPCA |
| Breast cancer 20% gaps | 0.3333 | 0.0000 | 0.0000 | N/A | 1.0000 | 0.0000 | 0.0000 | N/A |
| Breast cancer 10% gaps | 0.0000 | 0.0000 | 0.0000 | N/A | 0.0000 | 0.0000 | 0.0000 | N/A |
| Breast cancer 5% gaps | 0.0000 | 0.0000 | 0.0000 | N/A | 0.0000 | 0.0000 | 0.0000 | N/A |
| Breast cancer 20% gaps | 0.0000 | 0.0000 | 0.0000 | N/A | 1.0000 | 0.0000 | 0.0000 | N/A |
| Breast cancer 10% gaps | 0.0000 | 0.0000 | 0.0000 | N/A | 0.0000 | 0.0000 | 0.0000 | N/A |
| Breast cancer 5% gaps | 0.0000 | 0.0000 | 0.0000 | N/A | 0.0000 | 0.0000 | 0.0000 | N/A |
| Bladder cancer 20% gaps | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 |
| Bladder cancer 10% gaps | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 |
| Bladder cancer 5% gaps | 0.0000 | 0.0000 | 0.0000 | N/A | 0.0000 | 0.0000 | 0.0000 | N/A |

Table 4 and Table 5 show that PPCA saves mean value of attributes but significantly distorts all variances. USVD and RSVD perfectly save mean of attributes and mostly save variance (less than 2% of attributes have statistically significant difference between variances of original and repaired database). 10NN usually saves mean but sometimes essentially distorts variance.

Table 2 and Table 3 show that almost always USVD and RSVD are better than 10NN (there is only one case when 10NN is slightly better that SVD based methods). Differences between USVD and RSVD are almost always negligible. This means that it is preferable to use RSVD because this method is protected against problem with very small number of known attributes. The main drawback of RSVD is approximately twice more calculation time compared to USVD.

Comparison of results of PPCA with all other methods (see Table 2, Table 3, Table 4, and Table 5) shows that PPCA is very poor for database under consideration. For example, Table 3 shows that all three fist PPCA PCs are almost orthogonal to the first three PCs which are calculated for original database. To check calculation of used indicators we directly compare outputs of all four compared methods. Fragments of data for original database, and for database fixed by four methods after creation of 20% of gaps are presented in Table 6. As we can see for all methods exclude PPCA imputed values are different from original but not so far. Unfortunately for PPCA we observe huge differences in the imputed values: these values cannot be observed for such attributes.

Table . Comparison of fragment of original database and four repaired databases after removing of 20% of data, yellow background highlights locations of removed and repaired values

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Row | Original | | | | | |
| 1 | 0.07 | 0.14 | 0.01 | 0.10 | 0.06 | -0.12 |
| 2 | 0.07 | 0.25 | 0.32 | 0.30 | 0.43 | 0.20 |
| 3 | 0.39 | 0.08 | 0.58 | 0.30 | 0.10 | -0.28 |
| 4 | -0.06 | 0.18 | 0.02 | 0.22 | 0.23 | 0.12 |
| 6 | -0.19 | -0.12 | 0.31 | 0.00 | 0.05 | 0.02 |
| 8 | 0.85 | 0.93 | 0.91 | 0.87 | 0.96 | 1.05 |
| 9 | 1.13 | 1.13 | 1.21 | 1.00 | 1.05 | 1.21 |
| 10 | 0.21 | -0.20 | 0.02 | -0.02 | 0.20 | -0.02 |
|  | 10NN | | | | | |
| 1 | 0.07 | 0.14 | 0.01 | -0.27 | -0.24 | -0.18 |
| 2 | 0.07 | -0.20 | 0.32 | -0.27 | 0.43 | 0.20 |
| 3 | 0.39 | 0.08 | 0.58 | 0.30 | 0.10 | -0.18 |
| 4 | -0.06 | 0.18 | -0.36 | 0.22 | 0.23 | 0.12 |
| 6 | -0.19 | -0.12 | 0.31 | -0.35 | 0.05 | 0.02 |
| 8 | 0.85 | 0.93 | -0.56 | 0.87 | 0.96 | 1.05 |
| 9 | 1.13 | 1.13 | 1.21 | 1.00 | 1.05 | -0.18 |
| 10 | -0.34 | -0.20 | -0.56 | -0.02 | -0.35 | -0.26 |
|  | USVD | | | | | |
| 1 | 0.07 | 0.14 | 0.01 | 0.00 | 0.10 | -0.26 |
| 2 | 0.07 | 0.05 | 0.32 | 0.25 | 0.43 | 0.20 |
| 3 | 0.39 | 0.08 | 0.58 | 0.30 | 0.10 | 0.32 |
| 4 | -0.06 | 0.18 | 0.23 | 0.22 | 0.23 | 0.12 |
| 6 | -0.19 | -0.12 | 0.31 | 0.08 | 0.05 | 0.02 |
| 8 | 0.85 | 0.93 | 1.39 | 0.87 | 0.96 | 1.05 |
| 9 | 1.13 | 1.13 | 1.21 | 1.00 | 1.05 | 0.65 |
| 10 | 0.20 | -0.20 | -0.17 | -0.02 | 0.05 | -0.31 |
|  | RSVD | | | | | |
| 1 | 0.07 | 0.14 | 0.01 | 0.00 | 0.10 | -0.26 |
| 2 | 0.07 | 0.05 | 0.32 | 0.25 | 0.43 | 0.20 |
| 3 | 0.39 | 0.08 | 0.58 | 0.30 | 0.10 | 0.32 |
| 4 | -0.06 | 0.18 | 0.23 | 0.22 | 0.23 | 0.12 |
| 6 | -0.19 | -0.12 | 0.31 | 0.08 | 0.05 | 0.02 |
| 8 | 0.85 | 0.93 | 1.39 | 0.87 | 0.96 | 1.05 |
| 9 | 1.13 | 1.13 | 1.21 | 1.00 | 1.05 | 0.65 |
| 10 | 0.20 | -0.20 | -0.17 | -0.02 | 0.05 | -0.31 |
|  | PPCA | | | | | |
| 1 | -0.44 | 0.14 | 0.28 | 159944.44 | -13139393.54 | -8266682.62 |
| 2 | -5.86 | -374187.69 | -0.50 | -486186.49 | 0.39 | 0.18 |
| 3 | -32.88 | 0.10 | 3.39 | 0.30 | 0.04 | -3481867.47 |
| 4 | -8.63 | 0.18 | -649187989.52 | 0.22 | 0.08 | 0.18 |
| 6 | 3.15 | -0.12 | 1.13 | 27995.06 | -0.06 | 0.02 |
| 8 | 2.65 | 0.93 | -136715600.94 | 0.87 | 0.73 | 1.01 |
| 9 | 7.26 | 1.13 | 0.88 | 1.00 | 1.02 | 5931685.01 |
| 10 | -3117235880.80 | -0.20 | -268413800.25 | -0.02 | -16463071.64 | -16397386.70 |

We have to stress that for 10% of gaps data of PPCA are significantly more similar to original data (see Table 7).

Table . Comparison of fragment of original database and four repaired databases after removing of 10% of data, yellow background highlights locations of removed and repaired values

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Row | Original | | | | | | | | | |
| 1 | 0.0681 | 0.1402 | 0.0081 | 0.1049 | 0.0609 | -0.1163 | 0.1276 | 0.1362 | 0.1455 | -0.7093 |
| 2 | 0.0720 | 0.2517 | 0.3179 | 0.3031 | 0.4262 | 0.2005 | 0.4596 | 0.4215 | 0.3676 | 0.6045 |
| 3 | 0.3944 | 0.0843 | 0.5802 | 0.3048 | 0.1040 | -0.2816 | 0.3751 | 0.0645 | 0.1463 | -0.6372 |
| 4 | -0.0606 | 0.1794 | 0.0162 | 0.2214 | 0.2270 | 0.1192 | 0.2379 | 0.3017 | 0.2997 | -0.6594 |
| 5 | 0.0607 | -0.0739 | 0.1239 | 0.3020 | 0.2524 | 0.5181 | 0.5711 | 0.3580 | 0.2754 | 0.1586 |
| 6 | -0.1874 | -0.1199 | 0.3133 | 0.0028 | 0.0458 | 0.0202 | -0.0208 | 0.1198 | 0.3549 | -0.6372 |
| 7 | -0.0851 | -0.3928 | -0.2041 | -0.0751 | -0.2997 | -0.3704 | 0.0279 | -0.0218 | 0.1040 | 0.6368 |
| 8 | 0.8541 | 0.9316 | 0.9105 | 0.8688 | 0.9610 | 1.0530 | 0.8081 | 0.9261 | 0.9495 | 1.6978 |
| 10 | 0.2127 | -0.1975 | 0.0162 | -0.0239 | 0.2008 | -0.0155 | 0.2018 | 0.1597 | 0.2098 | 0.7500 |
|  | 10NN | | | | | | | | | |
| 1 | 0.0681 | 0.1402 | -0.4276 | 0.1049 | 0.0609 | -0.1163 | 0.1276 | 0.1362 | 0.1455 | -0.7093 |
| 2 | 0.0720 | 0.2517 | -0.4306 | -0.3617 | 0.4262 | 0.2005 | 0.4596 | 0.4215 | 0.3676 | 0.1178 |
| 3 | 0.3944 | 0.0843 | 0.5802 | 0.3048 | 0.1040 | -0.2816 | 0.3751 | 0.0645 | 0.1463 | 0.1672 |
| 4 | -0.3893 | 0.1794 | 0.0162 | 0.2214 | 0.2270 | 0.1192 | 0.2379 | 0.3017 | 0.2997 | -0.6594 |
| 5 | 0.0607 | -0.3606 | 0.1239 | 0.3020 | 0.2524 | 0.5181 | 0.5711 | -0.3065 | 0.2754 | 0.1586 |
| 6 | -0.1874 | -0.4128 | 0.3133 | 0.0028 | 0.0458 | 0.0202 | -0.0208 | 0.1198 | 0.3549 | -0.6372 |
| 7 | -0.3893 | -0.3928 | -0.2041 | -0.0751 | -0.2997 | -0.3704 | 0.0279 | -0.0218 | 0.1040 | 0.6368 |
| 8 | -0.4435 | 0.9316 | 0.9105 | -0.3117 | 0.9610 | 1.0530 | 0.8081 | 0.9261 | -0.4862 | 1.6978 |
| 10 | 0.2127 | -0.1975 | 0.0162 | -0.0239 | -0.3808 | -0.2437 | 0.2018 | 0.1597 | 0.2098 | 0.7500 |
|  | USVD | | | | | | | | | |
| 1 | 0.0681 | 0.1402 | 0.0932 | 0.1049 | 0.0609 | -0.1163 | 0.1276 | 0.1362 | 0.1455 | -0.7093 |
| 2 | 0.0720 | 0.2517 | 0.4380 | 0.2402 | 0.4262 | 0.2005 | 0.4596 | 0.4215 | 0.3676 | 0.5126 |
| 3 | 0.3944 | 0.0843 | 0.5802 | 0.3048 | 0.1040 | -0.2816 | 0.3751 | 0.0645 | 0.1463 | -0.1574 |
| 4 | 0.0508 | 0.1794 | 0.0162 | 0.2214 | 0.2270 | 0.1192 | 0.2379 | 0.3017 | 0.2997 | -0.6594 |
| 5 | 0.0607 | -0.1206 | 0.1239 | 0.3020 | 0.2524 | 0.5181 | 0.5711 | -0.0715 | 0.2754 | 0.1586 |
| 6 | -0.1874 | 0.0812 | 0.3133 | 0.0028 | 0.0458 | 0.0202 | -0.0208 | 0.1198 | 0.3549 | -0.6372 |
| 7 | 0.3030 | -0.3928 | -0.2041 | -0.0751 | -0.2997 | -0.3704 | 0.0279 | -0.0218 | 0.1040 | 0.6368 |
| 8 | 1.4867 | 0.9316 | 0.9105 | 1.0399 | 0.9610 | 1.0530 | 0.8081 | 0.9261 | 0.9962 | 1.6978 |
| 10 | 0.2127 | -0.1975 | 0.0162 | -0.0239 | 0.2428 | -0.0415 | 0.2018 | 0.1597 | 0.2098 | 0.7500 |
|  | RSVD | | | | | | | | | |
| 1 | 0.0681 | 0.1402 | 0.0931 | 0.1049 | 0.0609 | -0.1163 | 0.1276 | 0.1362 | 0.1455 | -0.7093 |
| 2 | 0.0720 | 0.2517 | 0.4381 | 0.2403 | 0.4262 | 0.2005 | 0.4596 | 0.4215 | 0.3676 | 0.5124 |
| 3 | 0.3944 | 0.0843 | 0.5802 | 0.3048 | 0.1040 | -0.2816 | 0.3751 | 0.0645 | 0.1463 | -0.1577 |
| 4 | 0.0508 | 0.1794 | 0.0162 | 0.2214 | 0.2270 | 0.1192 | 0.2379 | 0.3017 | 0.2997 | -0.6594 |
| 5 | 0.0607 | -0.1206 | 0.1239 | 0.3020 | 0.2524 | 0.5181 | 0.5711 | -0.0714 | 0.2754 | 0.1586 |
| 6 | -0.1874 | 0.0813 | 0.3133 | 0.0028 | 0.0458 | 0.0202 | -0.0208 | 0.1198 | 0.3549 | -0.6372 |
| 7 | 0.3029 | -0.3928 | -0.2041 | -0.0751 | -0.2997 | -0.3704 | 0.0279 | -0.0218 | 0.1040 | 0.6368 |
| 8 | 1.4866 | 0.9316 | 0.9105 | 1.0392 | 0.9610 | 1.0530 | 0.8081 | 0.9261 | 0.9961 | 1.6978 |
| 10 | 0.2127 | -0.1975 | 0.0162 | -0.0239 | 0.2425 | -0.0421 | 0.2018 | 0.1597 | 0.2098 | 0.7500 |
|  | PPCA | | | | | | | | | |
| 1 | 0.0685 | 0.1407 | 0.5269 | 0.1057 | 0.0613 | -0.1155 | 0.1284 | 0.1378 | 0.1459 | -0.7093 |
| 2 | 0.0722 | 0.2537 | 0.6256 | 0.2471 | 0.4263 | 0.1996 | 0.4588 | 0.4220 | 0.3675 | -0.8839 |
| 3 | 0.3945 | 0.0854 | 0.5811 | 0.3059 | 0.1047 | -0.2802 | 0.3759 | 0.0664 | 0.1467 | -1.2626 |
| 4 | -2.2673 | 0.1800 | 0.0158 | 0.2203 | 0.2260 | 0.1174 | 0.2369 | 0.3013 | 0.2993 | -0.6602 |
| 5 | 0.0610 | 0.1604 | 0.1238 | 0.3024 | 0.2523 | 0.5193 | 0.5720 | 0.0451 | 0.2755 | 0.1584 |
| 6 | -0.1882 | 0.9722 | 0.3131 | 0.0018 | 0.0449 | 0.0187 | -0.0217 | 0.1194 | 0.3544 | -0.6389 |
| 7 | -6.1328 | -0.3928 | -0.2045 | -0.0756 | -0.2998 | -0.3712 | 0.0273 | -0.0224 | 0.1038 | 0.6372 |
| 8 | 0.9151 | 0.9315 | 0.9101 | 0.9107 | 0.9612 | 1.0530 | 0.8083 | 0.9277 | 0.9930 | 1.6978 |
| 10 | 0.2131 | -0.1968 | 0.0166 | -0.0234 | 0.2464 | 0.0988 | 0.2021 | 0.1606 | 0.2100 | 0.7502 |

# Conclusion

As we can see from Table 1 kNN is the fastest method and for very big dimensions it looks like only usable method. Probabilistic PCA is very slow method and usage of this method for big datasets is impossible. Moreover for database with large number of missed values PPCA sometimes impute inappropriate values (see Table 6).

# References

1. Borg,, I.; [Groenen, P.](https://en.wikipedia.org/wiki/Patrick_Groenen" \o "Patrick Groenen) (2005). Modern Multidimensional Scaling: theory and applications (2nd ed.). New York: Springer-Verlag. pp. 207–212. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [0-387-94845-7](https://en.wikipedia.org/wiki/Special:BookSources/0-387-94845-7).
2. Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. 559–572 (1901).
3. Smola, A. J., Williamson, R. C., Mika, S. & Sch, B. Regularized Principal Manifolds. 214–229 (1999).
4. Gorban, A. & Zinovyev, A. Elastic principal graphs and manifolds and their practical applications. Computing (Vienna/New York) 75, 359–379 (2005).
5. Gorban, A. N. & Zinovyev, A. Principal manifolds and graphs in practice: from molecular biology to dynamical systems. Int. J. Neural Syst. 20, 219–232 (2010).
6. McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv (2018).
7. Gorban, A., Kégl, B., Wunch, D. & Zinovyev, A. Principal Manifolds for Data Visualisation and Dimension Reduction. Lect. notes Comput. Sci. Eng. 340 (2008). doi:10.1007/978-3-540-73750-6
8. Carlsson, G. & de Silva, V. Topological approximation by small simplicial complexes. Preprint 1–36 (2003).
9. Gorban, A. N. & Zinovyev, A. Y. in Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques (2008). doi:10.4018/978-1-60566-766-9
10. Enders, C.K. (2010). Applied missing data analysis. New York: Guilford Press.
11. Tipping, M. E. and Bishop, C. M., Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1999.
12. K.L. ClarksonNearest-neighbor searching and metric space dimensions T. Darrell T, et al. (Eds.), Nearest-Neighbor Methods in Learning and Vision, The MIT Press (2005), pp. 15-59
13. Deza, E.; [Deza, M.](https://en.wikipedia.org/wiki/Michel_Deza" \o "Michel Deza) (2006), Dictionary of Distances, Elsevier, [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [0-444-52087-2](https://en.wikipedia.org/wiki/Special:BookSources/0-444-52087-2).
14. Gorban & A Rossiev
15. Brand M. (2002) Incremental Singular Value Decomposition of Uncertain Data with Missing Values. In: Heyden A., Sparr G., Nielsen M., Johansen P. (eds) Computer Vision — ECCV 2002. ECCV 2002. Lecture Notes in Computer Science, vol 2350. Springer, Berlin, Heidelberg
16. Kurucz, M., Benczúr, A.A. and Csalogány, K., 2007, August. Methods for large scale SVD with missing values. In Proceedings of KDD cup and workshop (Vol. 12, pp. 31-38).
17. Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J. et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet, 365, 671-679.
18. L.Dyrskjot et al (2003) Identifying distinct classes of bladder carcinoma using microarrays. Nat Genetics 33(1):90-6.