

# A PCA-based Data Prediction Method

Peteris DAUGULIS<sup>1</sup>, Vija VAGALE<sup>1</sup>, Emiliano MANCINI<sup>2,3</sup>, Filippo CASTIGLIONE<sup>4</sup>

<sup>1</sup> Daugavpils University, Daugavpils, Latvia

<sup>2</sup> Data Science Institute, Hasselt University, Diepenbeek, Belgium

<sup>3</sup> Department of Global Health, Amsterdam UMC, Amsterdam, The Netherlands

<sup>4</sup> Institute for Computing Applications, Rome, Italy

`peteris.daugulis@du.lv`, `vija.vagale@du.lv`, `emiliano.mancini@uhasselt.be`,  
`filippo.castiglione@cnr.it`

0000-0003-3866-514X, 0000-0002-5428-6441, 0000-0002-5613-234X, 0000-0002-1442-3552

**Abstract.** The problem of choosing appropriate values for missing data is often encountered in the data science. We describe a novel method containing both traditional mathematics and machine learning elements for prediction (imputation) of missing data. This method is based on the notion of distance between shifted linear subspaces representing the existing data and candidate sets. The existing data set is represented by the subspace spanned by its first principal components. Solutions for the case of the Euclidean metric are given.

## 1 Introduction

### 1.1 Outline

In this article we describe a method of predicting unknown values of variables which is based on PCA and metric (Euclidean or other) in the ambient real linear space of variables - *the PCA-distance method*. Our motivation and goal is the problem of recovering missing data assuming that the set of complete data samples is approximated by the hyperplane spanned by the first principal components and the set of candidate points form another shifted subspace. Regression is not used in this method. In the case of the Euclidean metric we give exact solutions which use orthogonal projections and extrema of quadratic functions. We prove all the included mathematical statements. The computation algorithm and the activity diagram of the PCA-distance method are given. It is assumed that arithmetical operations of various data entries are justified, i.e. all data units are dimensionless. Steps of the method can be interpreted in terms of machine learning.

## 1.2 Background and previous work

*Prediction as a mathematical modelling problem.* In most sciences and research areas it is necessary to generate (predict, impute, estimate) missing information (data, relations, rules etc.) if partial initial information is given. In this article we use the term *prediction* to denote such methods. Such methods usually are based on minimizing errors and finding extremal values of functions and discrete objects. Most prediction methods start with building a suitable mathematical object - a *model*, a discrete or continuous subset of an ambient space, to represent the most important properties of the given body of information (e.g. a discrete data set). Models may mean representation in at least two senses - the evolution of a system or a process, or the simplified description of an existing system or a collection of data. Apart from a model of existing data we must also have a set of candidate values for predictions.

A range of general purpose methods which can also be used for prediction purposes, such as approximation, interpolation, extrapolation and others, have been developed (Meijering, 2002), (Mittal, 2016). Machine learning approaches are used (Bengio, Courville et al., 2013), (Bzdok, Altman et al., 2018).

The simplest prediction methods involve using mean values of specified components of suitable data points, (Little and Rubin, 1987). There are prediction methods using least-square (linear regression) ideas, (Bu, Dysvik et al., 2004). There are methods assuming that the completely defined data samples belong to a multivariate distribution, (van Buuren, 2007). In such methods (Expectation-Maximization methods) parameters of the distribution corresponding to the completely defined samples and missing values are computed to maximize the likelihood function. A popular direction is based on the K-nearest neighbour (K-NN) idea which uses a metric or a similarity measure in certain subspaces of the ambient space, (Jonsson and Wohlin, 2004). In this approach missing values are defined as means of corresponding values of nearest completely defined data points. Nearness is defined using Euclidean-like metrics in the subspace having dimensions where values are defined for all data points. See (Bertsimas, Pawlowski et al., 2018).

*Principal Components Analysis.* An effective and widely used tool of data modelling and analysis is the Principal Component Analysis (PCA), (Pearson, 1901), (Eckart and Young, 1936), (Hestenes, 1958), (Hotelling, 1933), (Hotelling, 1936). It is used to represent a discrete set of data points as a shifted linear subspace which shows the most important variables and their linear combinations. PCA is used for linearization of data, dimensionality reduction, filtering out noise and finding the most important linear combinations of data variables (Meglen, 1991), (Gorban, Kegl et al., 2007).

PCA is a mathematical procedure that transforms the basis of the space (i.e., a change of variable) which includes the set of data we are interested in. It is mainly used to reduce the dimensionality of a large data set. In fact, after the transformation, the new coordinates of the basis (i.e., the new variables) are ranked in terms of the ability to embed most of the variability of the data set. Thus, focusing on few variables and neglecting the others, one can keep most of the information contained in the data set and focus on that.

In other words, to preserve as much variability as possible one finds new variables that are linear functions of those in the original dataset, that have the property of successively maximizing the variance and, at the same time, are uncorrelated with each other. The mathematical operations to perform a PCA are based on the eigendecomposition of the data covariance matrix (hence the principal components are eigenvectors of the data's covariance matrix) also known as singular value decomposition of the data matrix.

PCA is mainly a statistical tool developed by statisticians that has found a large number of applications in many fields of science and technology and that is currently used as a step to perform predictions with machine learning methodologies. There is a data prediction method - Bayesian Principal Component analysis, which uses PCA (Oba, Sato et al., 2003).

## 2 Main results

In this section we describe the minimal distance idea and prove the mathematical results for the Euclidean case.

### 2.1 The minimal distance idea

*Notations and basic facts.* Given two subsets  $\mathcal{A}, \mathcal{B}$  of a metric space  $(\mathbb{M}, d(\cdot, \cdot))$  we denote by  $d(\mathcal{A}, \mathcal{B})$  the distance between  $\mathcal{A}$  and  $\mathcal{B}$ :  $d(\mathcal{A}, \mathcal{B}) = \inf_{a \in \mathcal{A}, b \in \mathcal{B}} d(a, b)$ . If  $\mathcal{U}, \mathcal{V}$  are shifted linear subspaces in a real Euclidean (inner-product) space  $E$  then  $d(\mathcal{U}, \mathcal{V}) = \min_{u \in \mathcal{U}, v \in \mathcal{V}} d(u, v)$ , a nonnegative real number. We consider  $\mathbb{R}^n$  as the Euclidean space with the norm  $\|x\| = \sqrt{x^T x}$  and the metric  $d(x, y) = \|x - y\|$ , for consistency its elements are defined as columns. We denote by  $proj_{\mathcal{V}}$  the orthogonal projection onto  $\mathcal{V} \leq \mathbb{R}^m$ . We denote the subspace spanned by the columns of a matrix  $V$  by  $\mathcal{V}$  (using *mathcal* letters) or  $\langle V \rangle$ .

*Predicting one variable.* First we describe the problem we are trying to solve in the case of predicting one variable. Suppose we have a system described by  $m - 1$  independent variables (indicators)  $x_1, x_2, \dots, x_{m-1}$  and one dependent variable  $y$ . We consider  $\mathbb{R}^m$  with some metric, for example, the Euclidean metric.

Suppose we have a set of complete measurements  $\mathcal{S} = \{(x_{11}, \dots, x_{1,m-1}, y_1), \dots, (x_{s1}, \dots, x_{s,m-1}, y_s)\}$ ,  $|\mathcal{S}| = s$ . We also have an incomplete measurement - a sequence of values  $X_0 = (x_{01}, \dots, x_{0,m-1})$  for which we want to find a  $y$ -value which would be most appropriate in a rigorous sense. It means finding a point on the line  $\mathcal{L} = \{x_1 = x_{01}, \dots, x_{m-1} = x_{0,m-1}\}$  (*the prediction line*) which is special with respect to  $\mathcal{S}$ .

In order to extract the most important property of  $\mathcal{S}$  we choose "the first term of approximation" - the linear approximation, which is sufficient for most meaningful predictions dealing with missing data coming from various sources with possibly different

standards and formats. For this purpose we can use PCA. Arrange the coordinates of  $S$ -elements as a matrix  $S = \begin{bmatrix} x_{11} & \dots & x_{1,m-1} & y_1 \\ \dots & \dots & \dots & \dots \\ x_{s1} & \dots & x_{s,m-1} & y_m \end{bmatrix}$ . Choose  $n$  (first) principal components (PC) of  $S$ . Construct the subspace  $\mathcal{P}$  (*shifted principal subspace*) spanned by these PC. It is a shifted linear subspace of dimension  $n$  in  $\mathbb{R}^m$ . By construction  $\mathcal{P}$  is spanned by a basis (the principal components) which diagonalizes the covariance matrix of data with maximal diagonal elements, in every dimension there is the shift by the column average.

We have that  $\dim(\mathcal{P}) = r$ ,  $\dim(\mathcal{L}) = 1$ . Suppose that  $\mathcal{L} \not\subseteq \mathcal{P}$ . If  $\mathcal{P}$  and  $\mathcal{L}$  intersect (in one point  $(x_{01}, \dots, x_{0,m-1}, y_0)$ ) then take the intersection point as the prediction - the predicted  $y$ -value of the sequence  $X_0$  is  $y_0$ . This See Fig.1 for the case  $n = 3, r = 2$ . We note that this case is essentially the least square prediction.

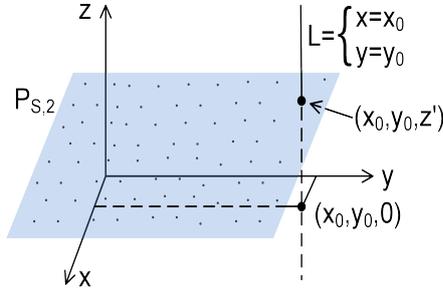


Fig.1. The case  $r = 2$  in  $\mathbb{R}^3$ .

Consider the case when  $\mathcal{P}$  and  $\mathcal{L}$  do not intersect. Our proposal is to choose a point in  $\mathcal{L}$  minimizing the distance to  $\mathcal{P}$  as our prediction. At least one such point exists. We want to find  $l_0 \in \mathcal{L}$  for which there is  $p_0 \in \mathcal{P}$  such that  $d(l_0, p_0) = \min_{l \in \mathcal{L}, p \in \mathcal{P}} d(l, p) = d(\mathcal{L}, \mathcal{P})$ ,  $l_0$  gives the desired "predicted"  $y$ -value.

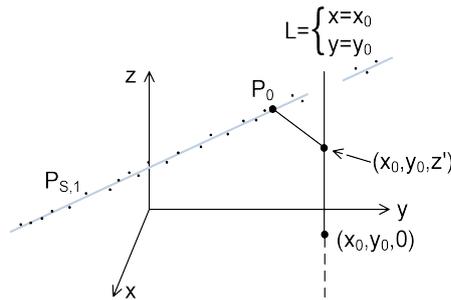


Fig.2. The case  $r = 1$  in  $\mathbb{R}^3$ .

If we use the Euclidean distance then  $p_0 = proj_{\mathcal{P}}(l_0)$ .

*Predicting more than one variable.* We can have a situation where more than one entry of data points are missing - we need to predict more than one variable for each measured data sample. We work in  $\mathbb{R}^n$  and have a sequence  $X_0 = (x_{01}, \dots, x_{0k})$ ,  $k < n - 1$ , for which we want to find the missing  $n - k$  values. In this case the shifted linear subspace  $\mathcal{L} = \{x_1 = x_{01}, \dots, x_k = x_{0k}\}$  (the prediction space) has dimension  $k$ . Again we can use the minimal distance idea: find a  $l_0 \in \mathcal{L}$  for which the minimal distance to  $\mathcal{P}$  is achieved.

We note that in all cases  $\mathcal{L} \cap \mathcal{P} \neq \emptyset$  is equivalent to  $d(\mathcal{L}, \mathcal{P}) = 0$ .

Prediction by minimizing distances can also be generalized for the cases when data or candidate data models are nonlinear varieties in ambient spaces.

## 2.2 Solutions for Euclidean spaces

*One dimensional prediction space - prediction line.* In this section we describe exact solutions for a prediction line  $\mathcal{L}$ ,  $\dim(\mathcal{L}) = 1$ , and an arbitrary principal subspace  $\mathcal{P}$  in case of the Euclidean metric. These solutions are based on orthogonal projections and extrema of quadratic functions.

The first proposition deals with the case of a linearly independent generating set of the subspace  $\mathcal{P}$  with respect to which we find the special point on the prediction line  $\mathcal{L}$ .

**Proposition 2.1.** *Let  $p_1, \dots, p_n$  be linearly independent elements in  $\mathbb{R}^m$ ,  $P = [p_1 | \dots | p_n]$  is the  $m \times n$  matrix obtained by joining  $p_1, \dots, p_n$ . Denote*

$$W = P(P^T P)^{-1} P^T - E_m = [w_1 | W'], \quad (1)$$

where  $w_1$  is the first column of  $W$ . Let  $L = \left\{ \begin{bmatrix} t \\ l' \end{bmatrix} \mid t \in \mathbb{R} \right\}$ ,  $l' \in \mathbb{R}^{m-1}$  fixed, an affine line in  $\mathbb{R}^m$ . Let  $\mathcal{P} = \langle p_1, \dots, p_n \rangle \leq \mathbb{R}^m$ .

1. If  $w_1 = 0$  then for any  $l \in \mathcal{L}$  there is a point  $p \in \mathcal{P}$  such that  $d(l, p) = d(\mathcal{L}, \mathcal{P})$ .
2. If  $w_1 \neq 0$  then  $d(l_{pred}, p_0) = d(\mathcal{L}, \mathcal{P}) = \min_{l \in \mathcal{L}, p \in \mathcal{P}} d(l, p)$  for  $l_{pred} \in \mathcal{L}$  and  $p_0 \in \mathcal{P}$

if and only if

$$l_{pred} = \begin{bmatrix} t_{pred} \\ l' \end{bmatrix} \in \mathbb{R}^n \text{ where}$$

$$t_{pred} = -\frac{1}{\|w_1\|^2} w_1^T W' l'. \quad (2)$$

*Proof.* Let  $l = \begin{bmatrix} t \\ l' \end{bmatrix} \in \mathcal{L}$ .  $P^T P$  is invertible since columns of  $P$  are linearly independent. It is known that  $proj_{\mathcal{P}} = P(P^T P)^{-1} P^T$ , (Meyer, 2010). Furthermore,  $d(l, \mathcal{P}) = \|proj_{\mathcal{P}}(l) - l\| = \|(P(P^T P)^{-1} P^T - E_m)l\| = \|Wl\|$ .

We express  $Wl$  as the linear combination of  $W$ -columns:

$$Wl = tw_1 + W'l'. \quad (3)$$

1. Let  $w_1 = 0$ . Then for any  $l \in \mathcal{L}$   $d(l, \mathcal{P}) = \|W'l'\|$ , it does not depend on  $l$ ,  $d(l, proj_{\mathcal{P}}(l)) = d(\mathcal{L}, \mathcal{P})$ .
2. Let  $w_1 \neq 0$ . We use the fact  $\|x\|^2 = \|proj_V(x)\|^2 + \|proj_{V^\perp}(x)\|^2$  (the generalized Pythagorean theorem). Taking  $x = Wl$  and  $V = \langle w_1 \rangle$  we get

$$\begin{aligned} \|Wl\|^2 &= \|proj_{\langle w_1 \rangle}(Wl)\|^2 + \|proj_{\langle w_1 \rangle^\perp}(Wl)\|^2 = \\ &= \|tw_1 + proj_{\langle w_1 \rangle}(W'l')\|^2 + \|proj_{\langle w_1 \rangle^\perp}(W'l')\|^2. \end{aligned} \quad (4)$$

$l$  such that  $\|Wl\|$  is minimal will be achieved for the unique  $t$  satisfying  $tw_1 = -proj_{\langle w_1 \rangle}(W'l')$ , hence

$$t_{pred} = -\frac{1}{\|w_1\|^2} w_1^T W'l'. \quad (5)$$

*Remark 2.2.* Note that in any case  $d(l, p) = d(l, \mathcal{P})$ ,  $l \in \mathcal{L}$ ,  $p \in \mathcal{P}$ , if and only if  $p = proj_{\mathcal{P}}(l)$ .

*Remark 2.3.* If  $(p_1, \dots, p_n)$  is an (ordered) orthonormal basis of  $\mathcal{P}$  then  $P^T P = E_n$  therefore  $proj_{\mathcal{P}} = W = PP^T$ . Orthonormality of  $(p_1, \dots, p_n)$  takes place if  $p_1, \dots, p_n$  are principal components.

*Remark 2.4.* If  $(p_1, \dots, p_n)$  is a (not necessarily orthonormal, ordered) basis of  $\mathcal{P}$  then it may be computationally more efficient to compute  $P(P^T P)^{-1} P$  via the  $QR$  factorization of  $P$ . See 2010. The  $QR$  factorization is suitable for matrices with large condition number, it can be made computationally stable using Householder or Givens reductions. If  $P = QR$  where columns of  $Q$  are orthonormal and  $R$  is a triangular matrix with positive diagonal entries then  $P(P^T P)^{-1} P^T = QQ^T$ .

The next proposition deals with the case when the generators of the subspace  $\mathcal{P}$  are not linearly independent.

**Proposition 2.5.** *Let  $p_1, \dots, p_n$  be elements of  $\mathbb{R}^m$ ,  $P = [p_1 | \dots | p_n]$  is the  $m \times n$  matrix obtained by joining  $p_1, \dots, p_n$ . Let  $PC = [P_e | O_{m, n-r}]$  be such that  $rank(P_e) = rank(P)$  (for example, a column echelon form of  $P$ ),  $C$  is a  $n \times n$  matrix of elementary column operations.*

Denote

$$W = P_e(P_e^T P_e)^{-1} P_e^T - E_m = [w_1 | W'], \quad (6)$$

where  $w_1$  is the first column of  $W$ . In these notations the statements of Proposition 2.2 are true.

*Proof.* Columns of  $P_e$  form a basis for  $\mathcal{P}$ . Denote the columns of  $P_e$  by  $p'_1, \dots, p'_r$ ,  $\mathcal{P}_e = \langle p'_1, \dots, p'_r \rangle$ . Then  $proj_{\mathcal{P}} = proj_{\mathcal{P}_e} = P_e(P_e^T P_e)^{-1} P_e^T$ . Repeat the proof of Proposition 2.2 substituting  $P$  by  $P_e$ .

The final proposition of this section gives another interpretation and the same solution of the problem using the fact that the square of the distance between a point on a line and a subspace is quadratic function of a line parameter.

**Proposition 2.6.** *Let  $p_1, \dots, p_n$  be elements in  $\mathbb{R}^m$ ,  $\mathcal{P} = \langle p_1, \dots, p_n \rangle \leq \mathbb{R}^m$ . Let  $\mathcal{L} = \{t \in \mathbb{R} \mid \begin{bmatrix} t \\ l' \end{bmatrix} \in \mathbb{R}^m\}$ ,  $l' \in \mathbb{R}^{m-1}$  fixed, an affine line in  $\mathbb{R}^m$ . Let  $l_i = \begin{bmatrix} t_i \\ l' \end{bmatrix} \in \mathcal{L}$ ,  $i \in \{1, 2, 3\}$ ,  $t_i$  distinct. Let  $proj_{\mathcal{P}}(x) = Wx$ ,  $x \in \mathbb{R}^m$ , with the first column of  $W$  being nonzero. Let  $\|proj_{\mathcal{P}}(l_i) - l_i\|^2 = d_i$ . Then  $d(l_{pred}, p_0) = d(\mathcal{L}, \mathcal{P})$  for  $l_{pred} \in \mathcal{L}$  and  $p_0 \in \mathcal{P}$  iff  $l_{pred} = \begin{bmatrix} t_{pred} \\ l' \end{bmatrix}$  where  $t_{pred} = -\frac{a_1}{2a_2}$  and  $[a_0, a_1, a_2]^T$  is the solution of the linear system*

$$\begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} \quad (7)$$

*Proof.* If  $l = \begin{bmatrix} t \\ l' \end{bmatrix}$  then  $\|proj_{\mathcal{P}}(l) - l\|^2$  is a nonconstant nonnegative quadratic function  $a_0 + a_1 t + a_2 t^2$  of  $t$ . Its coefficients can be determined by considering its values at 3 values of  $t$ , say,  $t_1, t_2, t_3$ , coefficients are solutions of (7). The minimum of  $\|proj_{\mathcal{P}}(l) - l\|^2$  is achieved when  $t = -\frac{a_1}{2a_2}$ .

Alternatively,  $Wl = tw_1 + W'l'$ , therefore

$$\|Wl\|^2 = (Wl)^T Wl = \|w_1\|^2 t^2 + 2w_1^T W'l' \cdot t + \|W'l'\|^2. \quad (8)$$

The minimum of  $\|Wl\|^2$  as a function of  $t$  is achieved when  $t = -\frac{1}{\|w_1\|^2} w_1^T W'l'$ .

*Remark 2.7.* If we use a general inner product  $(x, y) = x^T M y$  inducing the norm  $\|x\|_M = \sqrt{x^T M x}$  where  $M$  is a symmetric positive definite matrix then (2) has to be substituted by another formula

$$t_{pred} = -\frac{1}{\|w_1\|_M^2} w_1^T M W'l'. \quad (9)$$

*Remark 2.8.* Other norms such as  $\|\cdot\|_p$ ,  $1 \leq p$ , or  $\|\cdot\|_\infty$  can be considered in a similar way. The distance from  $l \in \mathcal{L}$  to  $\mathcal{P}$  can be found using the unit circle of the norm.

*Multidimensional prediction space.* This section contains results when  $\dim(\mathcal{L}) > 1$ . The first proposition gives a general solution if  $\mathcal{P}$  is spanned by linearly independent generators.

**Proposition 2.9.** Let  $p_1, \dots, p_n$  be linearly independent elements in  $\mathbb{R}^m$ ,  $P = [p_1 | \dots | p_n]$  the  $m \times n$  matrix obtained by joining  $p_1, \dots, p_n$ . Let  $\mathcal{P} = \langle p_1, \dots, p_n \rangle \leq \mathbb{R}^m$ . Let

$$W = P(P^T P)^{-1} P^T - E_m = [W_k | W'], \quad (10)$$

where  $W_k$  is the block of the first  $k$  columns of  $W$ . Let  $\mathcal{L} = \left\{ \begin{bmatrix} t \\ l' \end{bmatrix} \in \mathbb{R}^m \right\}$ ,  $t = \begin{bmatrix} t_1 \\ \dots \\ t_k \end{bmatrix} \in \mathbb{R}^k$ ,  $l' \in \mathbb{R}^{m-k}$  fixed, an affine  $k$ -dimensional subspace in  $\mathbb{R}^m$ .

1. If  $W_k = 0$  then for any  $l \in \mathcal{L}$  there is a point  $p \in \mathcal{P}$  such that  $d(l, p) = d(\mathcal{L}, \mathcal{P})$ .
2. If  $W_k \neq 0$  then  $d(l_{pred}, p_0) = \min_{l \in \mathcal{L}, p \in \mathcal{P}} d(l, p)$  for  $l_{pred} \in \mathcal{L}$  and  $p_0 \in \mathcal{P}$  if and only

if  $l_{pred} = \begin{bmatrix} t_{pred} \\ l' \end{bmatrix}$  where  $t_{pred} \in \mathbb{R}^k$  is such that

$$W_k t_{pred} = -proj_{\langle W_r \rangle}(W' l'). \quad (11)$$

If  $rank(W_k) = k$  then

$$t_{pred} = -W_{k,L} W_k (W_k^T W_k)^{-1} W_k^T W' l' \quad (12)$$

where  $W_{k,L}$  is a left-inverse of  $W_k$  ( $W_{k,L} W_k = E_k$ ).

*Proof.* Let  $l = \begin{bmatrix} t \\ l' \end{bmatrix} \in \mathcal{L}$ . Again we have that  $P^T P$  is invertible and  $proj_{\mathcal{P}} = P(P^T P)^{-1} P^T$ ,  $d(l, \mathcal{P}) = \|proj_{\mathcal{P}}(l) - l\| = \|(P(P^T P)^{-1} P^T - E_m)l\| = \|Wl\|$ . Again we express the product  $Wl$  as a linear combinations of columns:

$$Wl = W_k t + W' l' = W_k t + proj_{\langle W_k \rangle}(W' l') + proj_{\langle W_k \rangle^\perp}(W' l'). \quad (13)$$

Using the orthogonality we have

$$\|Wl\|^2 = \|W_k t + proj_{\langle W_k \rangle}(W' l')\|^2 + \|proj_{\langle W_k \rangle^\perp}(W' l')\|^2. \quad (14)$$

1. Let  $W_k = 0$ . Then for any  $l \in \mathcal{L}$   $d(l, \mathcal{P}) = \|W' l'\|$ , it does not depend on  $l$ .
2. Let  $W_r \neq 0$ .  $\|Wl\|$  is minimal if and only if

$$proj_{\langle W_k \rangle}(Wl) = W_k t + proj_{\langle W_k \rangle}(W' l') = 0.$$

$t$  can be chosen such that  $proj_{\langle W_k \rangle}(Wl) = 0$  since  $proj_{\langle W_k \rangle}(W' l')$  is an element generated by the columns of  $W_k$  and it can be expressed in the form  $W_k t$ . If  $rank(W_k) = k$  then  $proj_{\langle W_k \rangle} = W_k (W_k^T W_k)^{-1} W_k^T$  and  $W_{k,L}$  - the left inverse of  $W_k$ , exists, thus  $W_k t = -proj_{\langle W_k \rangle}(W' l')$  implies  $t$  is given by (12).

*Remark 2.10.* The condition

$$W_k t = -proj_{\langle W_k \rangle}(W' l') \quad (15)$$

can be interpreted that  $t$  is the coordinate column of  $-proj_{\langle W_k \rangle}(W' l')$  with respect to the sequence of  $W_k$ -columns, a generating set of  $\langle W_k \rangle$ .

**Proposition 2.11.** Let  $p_1, \dots, p_n$  be elements in  $\mathbb{R}^m$ ,  $P = [p_1 | \dots | p_n]$ . Let  $PC = [P_e | O_{m, n-r}]$  be such that  $\text{rank}(P_e) = \text{rank}(P)$  (for example, a column echelon form of  $P$ ),  $C$  is a  $n \times n$  matrix of elementary column operations.

Denote

$$W = P_e(P_e^T P_e)^{-1} P_e^T - E_m = [W_k | W'], \quad (16)$$

where  $W_k$  is the block of the first  $k$  columns of  $W$ . In these notations the statements of Proposition 2.9 are true.

*Proof.* See proof of Proposition 2.5.

The  $\mathcal{L}$ -subset with minimal distance to  $\mathcal{P}$  can also be found similarly to Proposition 2.6. Consider the nonnegative quadratic surface in independent variables  $t_1, \dots, t_k$  and the dependent variable  $d$ :  $\|\text{proj}_{\mathcal{P}}(l) - l\|^2 = d$ , find its coefficients using scalar products or  $\frac{(k+1)(k+2)}{2}$  points in general position, use theory of quadratic forms or find partial derivatives and solve the corresponding linear system. Details are given in the proposition below. For other metrics the solution must be modified accordingly.

**Proposition 2.12.** Let all notations be as in Proposition 2.9. Let  $w_j$  be the  $j$ th column of  $W$ . Then  $t_{\text{pred}} \in \mathbb{R}^k$  in the case 2. of Proposition 2.9 is a solution of the  $k \times k$  linear system

$$At = b, \text{ where } [A]_{ij} = w_i^T w_j, \quad b_i = -w_i^T W' l' \quad (17)$$

*Proof.* Let  $l = \begin{bmatrix} t \\ l' \end{bmatrix} \in \mathcal{L}$ ,  $t = \begin{bmatrix} t_1 \\ \dots \\ t_k \end{bmatrix} \in \mathbb{R}^k$ . Again we interpret the product  $Wl$  a linear combination of  $W$ -columns:

$$Wl = \sum_{i=1}^k t_i w_i + W' l'. \quad (18)$$

Using orthogonality we get  $\|Wl\|^2 = \sum_{i,j} t_i t_j w_i^T w_j + 2 \sum_{i=1}^k t_i w_i^T W' l' + \|W' l'\|^2$ . We have that

$$\frac{\partial \|Wl\|^2}{\partial t_i} = 2 \sum_{j=1}^k t_j w_i^T w_j + 2 w_i^T W' l'. \quad (19)$$

Equating it to 0 for each  $i$  we get the  $k \times k$  linear system of equations given in the statement.

*Removing outliers (the most influential points) - an extension of the Cook's distance idea to the PCA setting.* A desirable step in the process of finding hidden data features is detection of outliers, (Zimek and Schubert, 2017). Outliers can be removed using the ideas of the Cook's distance, (Cook, 1979), (Kim, 2017). The idea is for each data point  $x$  to compare projections of data points onto two principal component hyperplanes - the

PC hyperplane constructed with the whole data set  $\mathcal{S}$  and the PC hyperplane of the same dimension constructed with the data set  $\mathcal{S} \setminus x$ . If the difference between these projections is relatively large, then  $x$  is considered an outlier (or an unduly influential point) with respect to the construction of the PCA hyperplane. It is related to the leave-one-out cross-validation.

In general, if we are given the coordinate column of a data point  $y$  and two projection matrices  $H$  and  $H'$  then  $Hy - H'y = (H - H')y$  or its norm represents the difference of the two projections. To estimate the absolute difference between  $H$  and  $H'$  on the whole data set, the sum over all data points of norms (squared) of such projection differences  $\sum_{y \in \mathcal{S}} \|(H - H')y\|^2$  is computed. For the relative difference we divide it by the sum of norms of distances from data points to their projections onto a chosen hyperplane, say,  $\sum_{y \in \mathcal{S}} \|(H - E_m)y\|^2$ .

We explain it in more detail for the PCA setting. We use the notations of Section 2.2. Let  $P = [p_1 | \dots | p_n]$  be the  $m \times n$  matrix where  $p_j$  is the  $j$ th principal component for the data matrix  $S$ . Let  $P_i = [p_{i1} | \dots | p_{in}]$  be the  $m \times n$  matrix where  $p_{ij}$  is the  $j$ th principal component for the data matrix  $S \setminus S_{i*}$  ( $i$ th row removed). Construct the projection matrices  $H = P(P^T P)^{-1} P^T$ ,  $H_i = P_i(P_i^T P_i)^{-1} P_i^T$ , note that  $H, H_i$  are  $m \times m$  matrices. Columns of  $S^T = [x_1 | \dots | x_m]$  are vectors of data points, columns of  $(H - H_i)S^T$  are differences of projections of data points. We use the Frobenius norm of matrices. We have that

$$\sum_{x \in \mathcal{S}} \|(H - H_i)x\|^2 = \|(H - H_i)S^T\|^2. \quad (20)$$

We can assume that

$$C_i = \|(H - H_i)S^T\| \quad (21)$$

measures the total (absolute) influence of the removal of the  $i$ th data point. Columns of  $(H - E_m)S^T$  are vectors orthogonal to the PC hyperplane having data points and their projection as endpoints,  $\|(H - E_m)S^T\|^2$  is total sum of distance squares from data points to their projections onto the initial PC hyperplane.  $RC_i = \frac{\|(H - H_i)S^T\|}{\|(H - E_m)S^T\|}$  can be chosen as relative influence or outlying measure of the data point  $i$ .

Clusters of data points with large  $RC$ -value can be removed after computing all  $RC_i$  or iteratively.

*Cross-validation issues.* Cross-validation of the model can be done estimating in-sample and out-of-sample mean square error (MSE) of predictions. Leave- $p$ -out and  $k$ -fold cross-validation can be used to split the whole data set into the training and test subsets.

We explain in some detail the leave-1-out cross-validation for our method. We use the notations of the previous section. For each  $i, 1 \leq i \leq |\mathcal{S}|$  we construct  $P_i$  as described above (the projection to the PCA hyperplane of dimension  $n$  which is computed removing the  $i$ -th data point), define  $W_i = P_i(P_i^T P_i)^{-1} P_i - E_m$ . We proceed according to Proposition 2.2, use the formula (2) and get the prediction  $t_{pred} = y'_i$  for the  $i$ 'th

data point using the other data points as a training set. We can now compare vectors  $[y_1, \dots, y_s]$  and  $[y'_1, \dots, y'_s]$ , find MSE etc.

*Confidence interval estimation.* Confidence intervals for this method can be estimated using the natural jackknife or bootstrap methods. Leave- $p$ -out jackknife method can be used to generate a prediction distribution for a given initial data vector. Sufficient number of leave- $p$ -out subsets of full data points and corresponding PCA hyperplanes are generated, predictions are computed for the given incomplete data vector. Another way to generate a prediction distribution is using bootstrapping by randomly choosing with replacement sufficiently large subsets of data points. Additionally, sufficient number of fictitious incomplete data vectors with a given distribution can be generated to estimate confidence intervals with a fixed full data matrix  $S$ .

### 3 Implementation of the PCA-distance method

*Scaling.* The data matrix  $S$  can be scaled columnwise - each  $s_{ij} \in S$  is substituted by  $\frac{s_{ij} - \bar{s}_{*j}}{\sigma_j}$  where  $\bar{s}_{*j}$  and  $\sigma_j$  are the mean and the standard deviation, respectively, of the  $j$ th column. If  $\sigma_j = 0$  then the function  $s_{ij} \mapsto s_{ij} - \bar{s}_{*j}$  is applied. After the prediction is found the inverse transformation is computed. We usually assume that the data is scaled.

#### 3.1 An algorithm

We describe the main steps for an algorithm implementing the PCA-prediction method developed in Propositions 2.2, 2.5, 2.9, 2.11, 2.12. In particular,  $p_i \in \mathbb{R}^m$ . Notations of these propositions are used. See also Fig.3.

- Step 1 Identify vectors  $p_1, \dots, p_n$ , form the  $m \times n$  matrix  $P = [p_1 | \dots | p_n]$ . Go to Step 2.
- Step 2 Determine  $\text{rank}(P)$ . If  $\text{rank}(P) = n$  (i.e.  $p_1, \dots, p_n$  are linearly independent) then go to Step 3.1, otherwise go to Step 3.2.
- Step 3.1 Compute the  $m \times m$  matrix  $W = P(P^T P)^{-1} P^T - E_m$ . Go to Step 4.
- Step 3.2 Find a column-echelon form of  $P - [P_e | O_{m, n-r}]$ . Compute  $W = P_e(P_e^T P_e)^{-1} P_e^T - E_M$ . Go to Step 4.
- Step 4 Identify the subspace  $\mathcal{L}$ . If  $\dim(\mathcal{L}) = 1$  then go to Step 5.1, otherwise go to Step 5.2.
- Step 5.1 Subdivide  $W = [w_1 | W']$ . Identify  $l'$ . Compute  $t_{pred} = -\frac{1}{\|w_1\|^2} w_1^T W' l'$ . Go to Step 6.

Step 5.2 Subdivide  $W = [W_k|W'] = [w_1|\dots|w_k|W']$ . Identify  $l'$ . Compute the matrix elements  $a_{ij} = w_i^T w_j$ ,  $b_i = -w_i^T W' l'$ . Solve the linear system (17). If there are free unknowns then use additional arguments to find a unique prediction  $t_{pred} = \begin{bmatrix} t_1 \\ \dots \\ t_k \end{bmatrix} \in \mathbb{R}^k$ . Go to Step 6.

Step 6 Complete the algorithm by returning  $l_{pred} = \begin{bmatrix} t_{pred} \\ l' \end{bmatrix}$ .

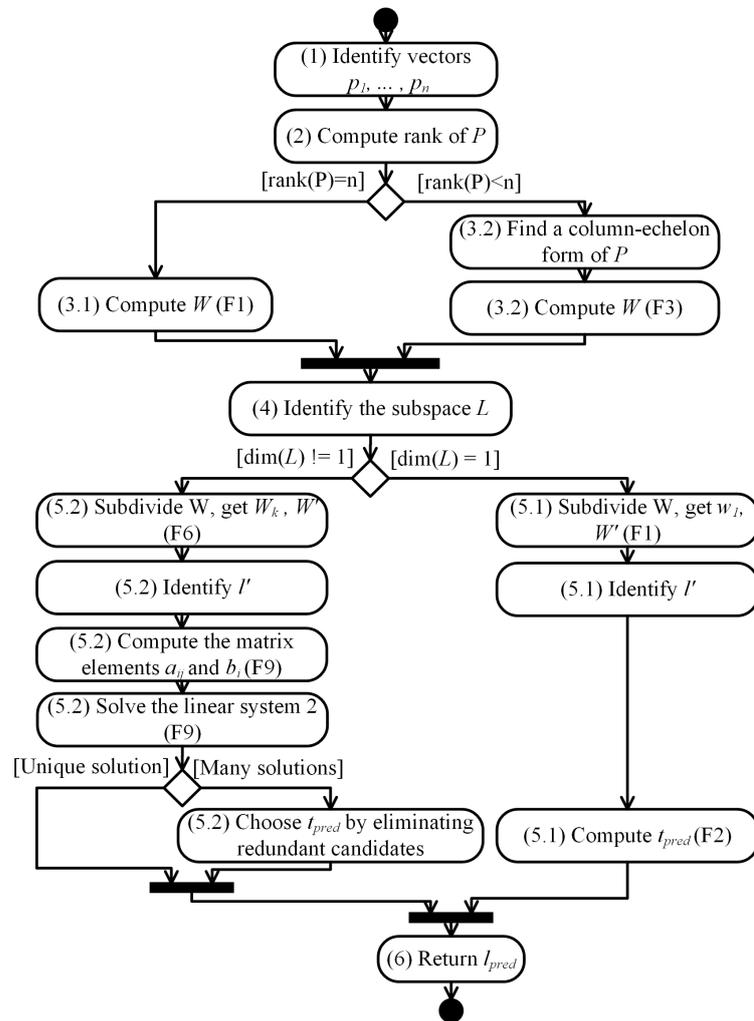


Fig.3. The activity diagram of the PCA-distance algorithm

### 3.2 Implementation example

The PCA-distance method have been implemented as a part of the research project dealing with predictions of drug-resistant pathogen strains for medical and pharmacological purposes. In our case the independent variables  $x_{ij}$  are certain socio-economic indicators and  $y_i$ 's measure antimicrobial resistance of pathogens, see Acknowledgements. Fig.4 shows one of the outcomes of this implementation - the world map coloured according to the PCA-distance predictions of the antimicrobial resistance. Typical size of data matrices (the matrix  $S$ ) was about  $300 \times 7000$ . Top 5% of outliers were removed using the approach given in subsection 2.2.

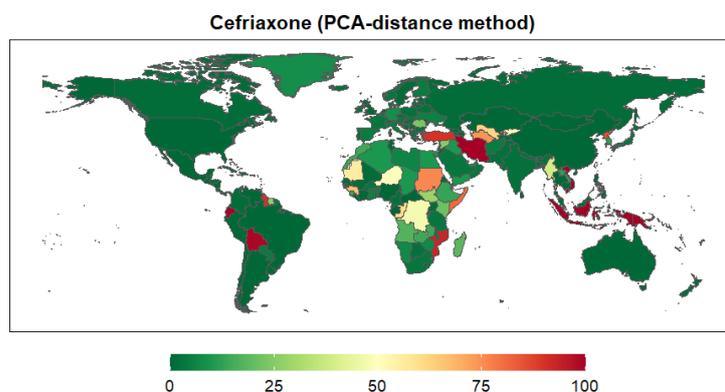


Fig.4. Predictions of ceftriaxone-resistant pathogen percentage rates by the PCA-distance method

This approach has been compared with the one used in (Oldenkamp, Schultsz et al, 2021) - a betabinomial vector generalized linear model with a logit link function. For the considered cases prediction errors of this method are close to those obtained by the Oldenkamp-Schultsz-Mancini-Cappuccio method, see Fig.5. It must be noted that confidence intervals provided by the OSMC method contain predictions by the PCA-method in 20% – 30% cases.

Antimicrobial	N	OSMC prediction mean absolute error	OSMC prediction mean relative error	PCA-distance prediction mean absolute error	PCA-distance prediction mean relative error	Percentage of OSMC confidence intervals containing PCA-distance predictions
Azithromycin	303	0.007	0.042	0.006	0.038	19.4
Cefepime	359	0.005	0.049	0.004	0.047	33.5
Cefepime/Ceftriaxone	389	0.005	0.043	0.005	0.041	20.7
Ceftriaxone	220	0.005	0.055	0.006	0.052	19.9
Ciprofloxacin	405	0.012	0.019	0.011	0.019	33.3

Fig.5. Comparing OSMC and PCA-distance methods

## 4 Discussion

*Machine learning features.* The described prediction method may be interpreted as a technique having features of unsupervised and semi-supervised machine learning. The PCA-based dimensionality reduction which results in the approximation of the initial data set by a low-dimensional hyperplane, is a case of an unsupervised representation (feature) learning which identifies the most important data indicators for prediction purposes (Bengio, Courville et al., 2013). Unsupervised PCA-based anomaly detection is used to find outliers of the data set. In case the data matrix has undefined elements, it can be filled by appropriate methods which can be interpreted as cases of semisupervised learning.

Our prediction method is subject to typical machine learning limitations and failures being caused by badly chosen assumptions, conjectures, by data overfitting or underfitting.

*Comparison of the PCA-distance method with other prediction methods.* The PCA-distance method seems to be more advanced and sensitive compared to the naive mean value methods which do not use linearization. One possible advantage of the mean value method is that it always returns values within the interval specified by existing measurements, e.g. it can not return a negative value if all existing values are positive.

The PCA-distance method does not assume that data is distributed in a special and uniform way therefore it seems more suitable to process data having high dimensionality and data distributed in different ways. Thus is it markedly different from methods in the expectation-maximization family, (van Buuren, 2007).

The PCA-distance method takes into account the whole data set of complete samples. In this sense it is different from the K-NN methods which take into account only a few complete samples, (Bertsimas, Pawlowski et al., 2018). In our method there is no need to arbitrarily specify the integer K. Taking into account the linearization of the whole data set seems more justified for large data sets with high dimensionality. We note that the metric idea is used in the space of all variables, including the dependent variables, not just the subspace of variables which are defined for all samples.

Existing prediction methods using PCA seem to recover missing values as coordinates of points on the subspace of principal components. The PCA-distance method is different from such methods since it finds extremal points on the candidate subspace with respect to the shifted subspace of principal components.

## 5 Conclusion

We offer a novel method for prediction of missing data which uses only the linearization - the most important approximation idea, and the notion of metric. The data set of full samples is linearized using the PCA. The point or points on the candidate subspace having the minimal distance to the linearized data subspace is chosen as the prediction. Closed formulas are obtained for the Euclidean (canonical or generalized) metric case. Our method may be suitable for data sets having high dimensionality and different data distribution patterns for different variables since it does not explicitly assume any data distribution.

## Acknowledgements

The authors acknowledge partial funding from the following national funding agencies participating in the project MAGICIAN JPI-AMR (<https://www.magician-amr.eu/>): State Education Development Agency (VIAA, Latvia) and Italian Ministry of Education and Research (MIUR, Italy).

## References

- Bengio, Y.; Courville A., Vincent, P. (2013). Representation Learning: A Review and New Perspectives, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8): 1798-1828. arXiv:1206.5538. doi:10.1109/tpami.2013.50. PMID 23787338. S2CID 393948.
- Bertsimas, D., Pawlowski, C. Zhuo, Y.D. (2018). From predictive methods to missing data imputation: An optimization approach, *Journal of Machine Learning Research*, 18, pp.1-39.
- Bu, T.H., Dysvik, B., Jonassen, I. (2004). LSImpute: accurate estimation of missing values in microarray data with least squares methods, *Nucleic Acids Res.*, 32(3):e34.
- van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification, *Stat Methods Med Res.*, 16(3):219-42.
- Bzdok, D., Altman, N., Krzywinski, M. (2018). Statistics versus Machine Learning, *Nature Methods*, 15 (4): 233-234. doi:10.1038/nmeth.4642. PMC 6082636. PMID 30100822.
- Cook, R. D. (1979). Influential Observations in Linear Regression, *Journal of the American Statistical Association. American Statistical Association*. 74 (365): 169?174.
- Eckart, C., Young, G. (1936). The approximation of one matrix by another of lower rank, *Psychometrika*, 1 (3): 211?8.
- Gorban, A.N, Kegl, B., Wunsch, D.C, Zinovyev, A. (eds) (2007). *Principal Manifolds for Data Visualisation and Dimension Reduction*, LNCSE 58, Springer, Berlin ? Heidelberg ? New York, 2007.
- Hestenes, M. R. (1958). Inversion of Matrices by Biorthogonalization and Related Results, *Journal of the Society for Industrial and Applied Mathematics*, 1, pp.51?90.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, 24, pp.417?441, and 498?520.
- Hotelling, H. (1936). Relations between two sets of variates, *Biometrika*, 28 (3/4), pp.321?377.
- Jonsson, P., Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using Likert data, *10th International Symposium on Software Metrics, 2004. Proceedings*, pp.108-118.
- Little, R.J., Rubin, D.B. (1987). *Statistical analysis and missing data*, John Wiley & Sons.
- Kim, M.G. (2017). A cautionary note on the use of Cook's distance. *Communications for Statistical Applications and Methods.*, 24 (3): 317?324.
- Meijering, E. (2002). A chronology of interpolation: from ancient astronomy to modern signal and image processing, *Proceedings of the IEEE*, 90 (3), pp.319?342.
- Meglen, R.R. (1991). Examining Large Databases: A Chemometric Approach Using Principal Component Analysis, *Journal of Chemometrics*, 5 (3), pp.163?179.
- Meyer, C.D. (2010). *Matrix analysis and applied linear algebra*, SIAM.
- Mittal, S. (2016). A Survey of Techniques for Approximate Computing, *ACM Comput. Surv. ACM.*, 48 (4), pp.62:1?62:33.
- Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K., Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data, *Bioinformatics*, Nov 1;19(16), pp.2088-96.

- Oldenkamp R., Schultsz C., Mancini E., Cappuccio A. (2021) Filling the gaps in the global prevalence map of clinical antimicrobial resistance, *Proc Natl Acad Sci U S A*, 2021 Jan 5;118(1):e2013515118. doi: 10.1073/pnas.2013515118. Erratum in: *Proc Natl Acad Sci U S A*. 2021 Oct 19;118(42): PMID: 33372157; PMCID: PMC7817194.
- Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space, *Philosophical Magazine*, 2 (11), pp.559-572.
- Zimek, A., Schubert, E. (2017), Outlier Detection, *Encyclopedia of Database Systems*, Springer New York, pp. 1-5.

Received January 19, 2021 , revised February 11, 2022, accepted February 16, 2022