Identifying Possible Socio-Demographic Factors in Infant Mortality Rate Classification With Orthogonal Projections to Latent Structures Discriminant Analysis

Noviana Pratiwi[1,2], *Dedi* Rosadi[1,*], and *Abdurakhman*[1]

[1]Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia
[2]Department of Statistics, Institut Sains & Teknologi Akprind Yogyakarta, Indonesia

Abstract. The measurement of infant mortality rate (IMR) stands as a crucial and complex indicator within global public health. Particularly for countries with constrained resources necessitating easily calculable and accurate gauges of population health, IMR retains its relevance. Several studies investigating infant mortality rates adopt a multilevel perspective and connect underlying factors with proximate variables. Conversely, interdisciplinary viewpoints, such as socio-demographic, offer a comprehensive overview of the intricate connections between socio-demographic disparities and mortality rates. This research focuses on the analysis of infant mortality rates, aiming to identify the socio-demographic factors influencing them to facilitate the formulation of more targeted intervention strategies. Employing Orthogonal Projection to Latent Structures - Discriminant Analysis (OPLS-DA) as the analytical method is motivated by its capability to navigate complex multivariable structures and elucidate the relationships between socio-demographic variables and infant mortality rates. The OPLS-DA model is used to extract significant patterns and relationships in the data so that it is possible for socio-demographic factors to have a significant impact on infant mortality rates. It is hoped that the results of the classification will provide insight and become the basis for developing policies by considering socio-demographic aspects in efforts to prevent infant mortality.
keywords : OPLS-DA, centering, scaling, pareto scaling, IMR

1 Introduction

Infant Mortality Rate (IMR), defined as the number of infant deaths per 1000 live births, stands as a critical indicator of the overall health and well-being of a population [1]. Reducing infant mortality is a global health priority, and understanding the factors influencing this rate is crucial for effective public health intervention. The geographical differences in infant mortality rates across different regions can be partially attributed to disparities in environmental exposures, including environmental factors in the workplace [2], [3], and [4]. The variation in infant health may also appear as an effect of differences in socio-demograpgic, behavioral, and medical aspects, as well as disparities in the accessibility of healthcare service [5], [6], and [7].

*e-mail: dedirosadi@ugm.ac.id

Among the myriad determinants, socio-demographic factors play a significant role in shaping the health outcomes of infants. Understanding which socio-demographic factors are associated with higher IMR allows policymakers and public health officials to target interventions more effectively. By addressing specific socio-demographic factors, such as maternal age, education, Population, Dependency Ratios, population growth rate, urban pop, migration, birth rate, doctor, nursing, death rate, GDP, maternal mortality, literacy, obesity, children underweight, child marriage, air pollutants, poor, etc, interventions can be tailored to the communities most in need. Findings on socio-demographic variables associated with IMR can inform the development of policies aimed at reducing infant mortality. This may include policies related to maternal and child healthcare, education, poverty alleviation, access to clean water and sanitation, and social support services. Identifying socio-demographic variables associated with IMR can guide further research and surveillance efforts. Researchers can delve deeper into the mechanisms through which socio-demographic factors influence infant mortality, leading to a better understanding of the root causes and potential interventions. Several researchers have discussed the influence of socio-demographics on IMR, such as: [8] who conducted research in Assam, [9] in Australia, and most recently research by [10] regarding the impact of socio-demographics and others on IMR in Oman.

To understand the socio-demographic factors that contribute to IMR, we can carry out advanced machine learning. The comparison between the amount of data and the variables that influence IMR makes the dimensions in data analysis complex and should be simplified. Undesirable systematic variance needs to be simplified by creating new components that capture more data. Recently, orthogonal projection to latent structure (OPLS) methods such as OPLS [11], OPLS-DA [12], K-OPLS [13], and O2OPLS [14] were discovered as alternative modeling techniques capable of simplifying and facilitating the separation and interpretation of various structures. The method used in this research is Orthogonal Projection to Latent Structures - Discriminant Analysis (OPLS-DA) method [12]. Socio-demographic data often have high levels of complexity with many interrelated variables. OPLS-DA can address this challenge by extracting hidden patterns that may be difficult to discern directly using simpler analysis methods.

OPLS-DA allows us to handle a large number of predictor variables (socio-demographics) simultaneously, which is important when we want to understand the relative contributions of various socio-demographic factors to IMR. OPLS-DA can help identify the variables that most influence or contribute to differences between groups in a specific context. OPLS-DA only requires that some variation in the measured data correlates with group membership, regardless of whether that variation is signal or noise [15], [16], [17] and relative stability in the R^2 and Q^2 value despite the deterioration in the model reliability [18]. OPLS-DA can provide insights into which variables contribute most significantly to the classification of IMR. This allows us to prioritize interventions or policies targeted at the most significant factors.

This study focuses on investigating the potential influence of socio-demographic factors on IMR classification.

Utilizing OPLS-DA approach, we aim to discern patterns and relationships socio-demographic variables that contribute to the classification of IMR. The investigation of various socio-demographic dimensions, which including but not limited to maternal age, education, Population, Dependency Ratios, population growth rate, urban pop, migration, birth rate, doctor, nursing, death rate, GDP, maternal mortality, literacy, obesity, children underweight, child marriage, air pollution, poor, etc mostly taken from the world factbook downloadable via cia.gov. By integrating advanced machine learning techniques, we seek to unravel intricate patterns within this multi-dimensional space, shedding light on the nuanced relationship that contributes to the classification of IMR.

The result of this study could be used to inform policy intervention and public health strategies, particularly in addressing the challenges associated with infant mortality in socio-demographic contexts and understanding the various socio-demographic factors that contribute to IMR classification. This study seeks to utilize machine learning techniques to examine high-dimensional data and recognize the socio-demographic variables that impact IMR classification. The objective is to perspective that can guide more efficient public health interventions and enhance infant health.

2 Orthogonal Projection to Latent Structures - Discriminant Analysis (OPLS-DA)

2.1 A comparative analysis of Partial Least Square (PLS) and Orthogonal Projection to Latent Structures (OPLS)

PLS and OPLS are both multivariate statistical techniques commonly used for modeling and analyzing relationships between sets of variables. While PLS [20] aims to maximize covariance between independent and dependent variables, reducing dimensionality, OPLS [21] shares this objective with an additional emphasis on separating systematics variation (predictive components) from orthogonal (unrelated) variation. Let X and Y be $n \times p$ and $n \times m$ matrics, respectively, where p and m represent the number of variables in each collection and n represent the number of observation. We assume that Y can be accurately approximated by

$$\hat{\boldsymbol{Y}} = \boldsymbol{X\beta} \tag{1}$$

$\boldsymbol{\beta}$ represents the matrix of regression coefficient with dimensions $p \times m$. While both $\hat{Y}$ and β are recognized as sample estimates. When the number of observation, $n > p$ and X has full rank, β_{OLS} is found :

$$\beta_{OLS} = (X^T X)^{-1} X^T Y \tag{2}$$

When $n < p$, the matrix $X^T X$ may become singular, and therefore, there may not be a unique solution for β_{OLS} that satisfies (1).

PLS assumes that the dataset contains meaningful patterns that can be represented in a low-dimensional space. Therefore, only a few predictor variables are needed to predict the outcome variable, Y accurately. The subsets are commonly referred to as *latent variables* due to their inherent nature of being unmeasurable independently [19]. The PLS model can be expressed concisely as

$$\begin{aligned} X &= TP^T + E \\ \hat{Y} &= TC^T + F \end{aligned} \tag{3}$$

The matrices T, P, and C have a relatively low rank. In a similar manner to principle components analysis (PCA), the columns of matrix T are referred to as *scores*, the columns of P are known as the loadings (or $X - loadings$), the columns of C are the $Y - loadings$, E and F represents the $residual$ matrix.

O-PLS can be thought of as PLS combined with a preprocessing step that filters out systematic variation from X that is orthogonal or statistically uncorrelated with the Y variables. The O-PLS [21] model is structured in the following :

$$\begin{aligned} X &= TP^T + T_{orth}P_{orth}^T + E \\ \hat{Y} &= TC^T + F \end{aligned} \tag{4}$$

The subscript $(\cdot)_{orth}$ indicates orthogonal components T, P, and C are generally differ from those of PLS.

The new predictions are acquired using the conventional method through Equation (1), following the initial filtration of the new samples with orthogonal variation. The OPLS steps can be seen as follows:

1. Data Preprocessing

   Partial Least Square [20]

2. Calculate weight $W^T = \frac{Y^T X}{Y^T Y}$
3. Normalize $W = \frac{W}{\|W\|}$
4. Calculate input scores $T = \frac{XW}{W^T W}$
5. Calculate output loading $C^T = \frac{T^T Y}{T^T T}$
6. Calculate output scores $u = \frac{YC}{C^T C}$
7. Calculate input loadings $P^T = \frac{T^T X}{T^T T}$

   Orthogonal Projection to Latent Structures [21]

8. calculated W_{orth}, the most significant loading of PCA($T^T E_{XY}$) with $E_{XY} = X - TW^T$
9. sequentially remove structure noise from X

$$t_{orth} = XW_{orth}$$
$$p_{orth}^T = t_{orth}^T X / \|t_{orth}\|^2$$
$$x \leftarrow X - t_{orth} p_{orth}^T$$

10. Repeat steps 8 and 9 for additional orthogonal components. Otherwise, repeat step 1 using the filtered X to update the matrix T and go to step 11 to construct a PLS model from Y and filtered X
11. To predict new sample, $x^{(n-1)}$, first filter any orthogonal variation, then $\hat{Y}^{(n+1)} = x^{(n+1)} \beta_{OPLS}$

2.2 Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA)

Partial Least Square (PLS) is a multivariate statistical technique that simultaneously handles many response variables (Y) and X variables. The PLS algorithm is used to predict continuous or discrete/categorical variables. The PLS used for classification is called PLS-DA with DA Discriminant Analysis. The PLS-DA algorithm knows how to handle collinearity and the ability of the model to rank the capacity of predicting variables in a multivariate context. OPLS-DA is an extension of PLS-DA used to maximize the variance between groups in one dimension or latent variable. In OPLS-DA, an orthogonal Y is considered to represent part of the variance in a class, which is sometimes useful for discrimination [12]. This approach has a similar concept to Soft independent modelling (SIMCA) classification, which uses residual observations to determine class membership based on local models of each class. OPLS-DA is a classification algorithm that combines PLS-DA (maximum class separation) with SIMCA

classification (handling of within-class variance) and retains interpretative superiority. Variables and coefficient weights from a validated OPLS-DA model can be used to improve all variables concerning their performance in discriminating between groups, and these can be used as part of a dimensionality reduction or feature selection task to predict the top predictor for a given model. OPLS-DA uses the information in the categorical response matrix Y to decompose the matrix X into three distinct parts, as in Equation (4).

Classification with OPLS-DA estimation is performed in two steps, firstly removing orthogonal variables from the X-matrix.

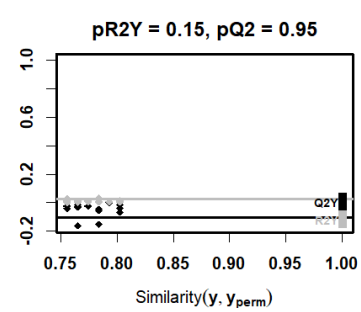
$$X_p = X - T_{orth} P_{orth}^T \tag{5}$$

Second, Y_{hat} is estimated using the most recent X_p, and the components of the estimated OPLS-DA model are predicted using the training data. Discriminant analysis concentrates on the boundaries that separate the different classes in the multidimensional space [22]. The class separation process is performed according to the patterns extracted from the X-matrix and the information about the class is entered in the binary Y-matrix, where one is used for indicating membership to a class, and the value zero is used for non-membership.

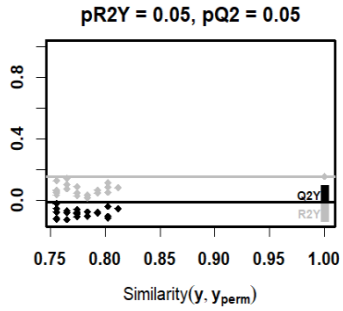
3 Result

In this chapter we will learn about multivariate analysis in the form of IMR classification in the world, mostly the data taken from the World Factbook downloadable via cia.gov. X variables used are Socio-Demographic Possible Factors in Infant Mortality Rate including maternal age, education, Population, Dependency Ratios, population growth rate, urban pop, migration, birth rate, doctor, nursing, death rate, GDP, maternal mortality, literacy, obesity, children underweight, child marriage, air pollutants, poor, etc taken from countries in the world that are included in the list of the world factbook. In data analysis with OPLS-DA, different scale approaches will be applied. The main objective was to evaluate the influence of the scale on the ability of OPLS-DA to differentiate and classify IMR based on socio-demographic variables.

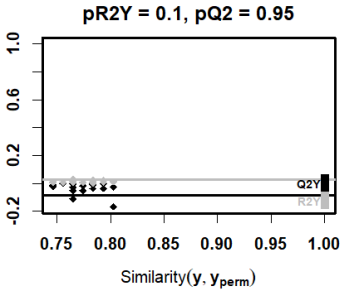
The first step in OPLS-DA is a scaling process which aims to make all variables equally important. The scaling processes used are mean centering and unit variance scaling, mean centering and Pareto scaling and mean centering only. Before looking at the model evaluation, a checking diagnosis will be carried out. The first checking diagnosis is permutation. Random permutation in diagnostic analysis aims to test the statistical significance of the model being built. This provides a basis for assessing the extent to which the observed model could have been generated by chance or random. The results of the checking diagnosis can be seen in Figure 1. The three approaches in Figure 1 show that the lowest pR2Y and pQ2 values are found in the scaling approach. This offers better model significance. The gray dots represent the points resulting from the permutation, while the black dots represent the points of the original Y matrix. The scaling approach clearly shows the points of separation, indicating that the resulting model cannot be obtained randomly. In contrast to the center and Pareto scaling approaches, there is an overlap separation between the gray and black dots. Therefore, the separation between dots is unclear, which suggests that the model was generated randomly. The next diagnostic checking is outlier detection. Outlier detection results can be seen in the Figure 2



(a) checking diagnosis with center

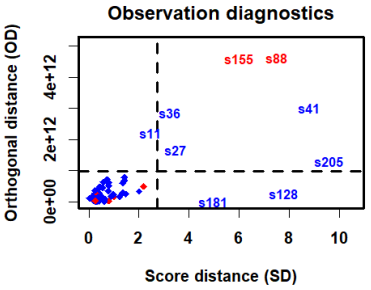


(b) checking diagnosis with scaling

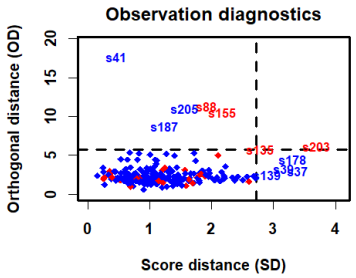


(c) checking diagnosis with Pareto scaling

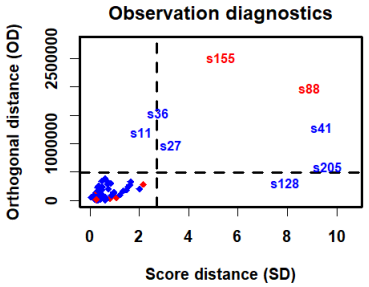
Figure 1: checking diagnosis



(a) outlier diagnosis with center



(b) outlier diagnosis with scaling



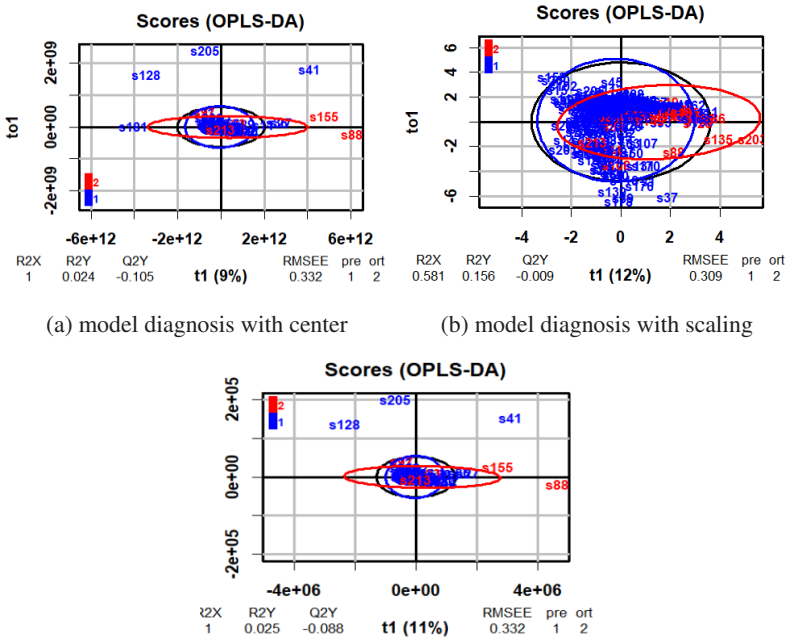
(c) outlier diagnosis with Pareto scaling

Figure 2: outlier and checking diagnosis

Figure 2 shows insight into whether there are outliers in the data and how much they impact the model. Score distance measures how far each observation is from the model produced by the independent variable (X), so an increase in score distance can indicate that an observation tends to be an outlier to the model. From the three approaches in figure 2, it can be concluded that the standard scaling approach has the closest distance compared to centering and Pareto scaling, so it can be said that the model is better than the other approaches.

The following diagnosis is about the quality and structure of the model. This diagnosis is seen with an ellipse plot of the first component between the predicted and orthogonal components. To better see the differences between the three approaches, see the figure 3.

The figure 3 shows that $t1$ with the scaling approach has the most considerable (%) value compared to other scaling approaches. This shows that observations can be explained by the first component better compared to other scale approaches. The resulting ellipse plot also shows significant variance for the standard scaling approach and can separate classes more clearly than others.



(a) model diagnosis with center (b) model diagnosis with scaling

(c) model diagnosis with Pareto scaling

Figure 3: model diagnosis

Classification evaluation with OPLS-DA involves commonly used classification evaluation metrics. To show the comparison between OPLS-DA and its predecessor method, PLS-DA, an analysis will be conducted with PLS-DA, but it turns out that PLS-DA only produces one component, so the model is not optimal in capturing data.

Table 1 shows the evaluation values that the three approaches of accuracy, sensitivity, and specificity have similar values for OPLS-DA, but the RMSEE looks different. The accuracy value measures how well the model is able to distinguish samples into different classes. Whereas the RMSEE measures how well the model fits the training data, the RMSEE measures the prediction error against the data used to build the model. The smaller the RMSEE

Table 1: Table Diagnostic Checking

	OPLS-DA (center)	OPLS-DA (scaling)	OPLS-DA (Pareto)	PLS-DA
R2Y	0.024	0.156	0.025	0.077
Q2Y	-0.105	-0.009	-0.088	0.020
RMSEE	0.332	0.309	0.332	0.321
Accuracy	0.878	0.878	0.878	0.873
sensitivity	0.877	0.877	0.877	1.000
Specificity	1.000	1.000	1.000	0.000

and the greater the accuracy, the better the model. The RMSEE of PLS-DA is neither the smallest nor the largest, but PLS-DA cannot be said to be a good model because it shows overfitting. PLS-DA classification results show the same class for all observations (sensitivity value 1 and specificity 0), henceforth PLS-DA is not included in further analysis. The difference in RMSEE values can also be seen from different diagnostic examinations. In the diagnosis check, the largest R2Y and Q2Y values are found in the scaling approach. This means that OPLS-DA, with the scaling approach, can measure the model's ability to predict the Y response in the test data better than other approaches.

Table 2: Table Variable Importance in Projection

Variable	contribution
Death rate	2.66970003
Medical doctors per 10,000	1.82221405
Med, Age	1.74541913
Population growth rate	1.26770885
Birth rate	1.23366598
Nursing and midwifery personnel per 10,000	1.06960373
Urban Pop %	0.88913252
Net Migration rate	0.67395048
Literacy	0.52445936
Children underweight (under the age of 5 years)	0.49329628
Martel mortality ratio	0.42652563
Education expenditures	0.30959973
Land Area (Km²)	0.29186661
Population	0.18629713
Air Pollutans From Methane Emissions (Megatons)	0.17403819
GDP	0.16632796
Obesity rate	0.070473
Dependency Ratios	0.06279546
Poor population	0.05268462
Child marriage by age 15 (%)	0.04697775

OPLS-DA can provide insights into which variables contribute most significantly to the classification of IMR. Below are the results of most significantly Variable Importance in Projection (VIP) from the best method, namely OPLS-DA with scaling.

Table 2 is the VIP value of OPLS-DA model in classifying IMR in the world. Higher values indicate more important variables in distinguishing between different classes in the classification analysis. The highest VIP value is Death rate which indicates that death rate has the most significant contribution in distinguishing between classes or groups in IMR analysis. The death rate is significantly different from the second place which is medical doctors and followed not so far by Med, Age and birth rate. Variables with lower VIP values such as Air Pollutans, GDP, Obesity rate, dependency ratio, poor population and child marriage still have an influence in distinguishing between classes, but the contribution tends to be lower compared to variables that have higher VIP values. It is important to remember that the interpretation of VIP should be done relative to other variables in the model. Variables with higher VIP values tend to be the main focus in analysis and decision-making.

4 Conclusion

In conclusion, the scaling approach in OPLS-DA for Infant Mortality Rate (IMR) classification with socio-demographic variables, as assessed through cross-validation and permutation, provides superior model diagnostics compared to centering and Pareto scaling. Although the accuracy (0.8779), sensitivity (0.87736), and specificity (1.000) metrics yield identical values, it is noteworthy that the RMSEE measures the model's fitness to the training data, while the RMSEE also quantifies the prediction error against the data used for model construction. Meanwhile, the accuracy value gauges the model's efficacy in distinguishing samples across different classes.

It is important to remember that the interpretation of VIP should be done relative to other variables in the model. Variables with higher VIP values tend to be the main focus in analysis and decision-making. Based on the model analysis and level of accuracy, it is evident that OPLS-DA effectively classifies IMR using socio-demographic variables. This underscores the potential role of socio-demographic factors in infant mortality rate classification. These findings are anticipated to inform policymakers in addressing the underlying determinants of infant mortality. Furthermore, OPLS-DA offers insights into the variables that contribute most significantly to IMR classification, enabling the prioritization of interventions or policies targeting the most impactful factors. Table 2 presents the sequential order of variables, based on socio-demographic considerations, that could be addressed to mitigate high IMR.

5 Acknowledgments

The author would like to thank IST AKPRIND Yogyakarta, Beasiswa Pendidikan Indonesia (BPI), and the Ministry of Finance, Republic of Indonesia, for their contributions to financing doctoral studies.

References

[1] DD. Reidpath, P.Allotey, JECH. May;**57(5)**:344-6. 2003 doi:10.1136/jech.57.5.344. PMID: 12700217; PMCID: PMC1732453.

[2] DR. Mattison, COP, **vol. 22** (pg. 208-218),2010

[3] DT. Wigle, TE. Arbuckle, MC. Turner, A. Berube, Q. Yang , S. Liu, D. Krewski, JTEH, **vol. 11** (pg. 373-517), 2008

[4] A. Burdorf, T. Brand, VW. Jaddoe, A. Hofman, JP. Mackenbach, EA Steegers, OEM, **vol. 68** (pg. 197-204), 2011

[5] D. Kim, A. Saada, JERPH, **10**, 2296-2335. 2013 https://doi.org/10.3390/ijerph10062296

[6] GC. Smith, J.Gutovich, C. Smyser, R. Pineda, C. Newnham, TH. Tjoeng, C. Vavasseur, M. Wallendorf, J. Neil, T. Inder, **70**: 541-549. , (2011), https://doi.org/10.1002/ana.22545

[7] ACJ. Ravelli, M. Tromp, MV. Huis, et al, JoECH;**63**:761-765, 2009

[8] Barman, Nityananda, and T. Dipul, IjoSET **3.5** : 1893-1900.(2014)

[9] G.L. Dasvarma, "Differential Mortality in Australia 1970-72", Department of Demography. Australian National University, Canberra,1980.

[10] Eltayib, RA Abuelgassim, M. Al-Azri, and MF. Chan, EJoIH, Psychology and Education **13.6**: 986-999. (2023)

[11] J. Trygg, and S. Wold, Journal of Chemometrics, **16.3** : 119-128.(2002)

[12] M. Bylesjö, et al.,Journal of Chemometrics, **20.8-10** : 341-351.(2006)

[13] M. Rantalainen, et al., Journal of Chemometrics, **21.7-9**: 376-385. (2007)

[14] J. Trygg, and S. Wold, Journal of chemometrics **17.1** : 53-64.(2003)

[15] S. Wold, S. Michael, and L. Eriksson, Chemometrics and intelligent laboratory systems **58.2** : 109-130.(2001)

[16] Brereton, Richard G., and GR. Lloyd, Journal of Chemometrics **28.4** : 213-225.(2014)

[17] S. Wiklund, et al, Analytical chemistry **80.1**: 115-122. (2008)

[18] Worley, Bradley, and R. Powers, Current Metabolomics **4.2**: 97-103. (2016)

[19] H. Martens, T. Naes, Multivariate Calibration., Wiley: Chichester (1989).

[20] S. Wold, I Ruhe, H. Wold, & WJ.Dunn Iii, In SIAM J. ScI. STAT. COMPUT **(Vol. 5, Issue 3)**.(1984)

[21] J. Trygg, & S. Wold, Journal of Chemometrics, **16(3)**, 119–128. (2002). https://doi.org/10.1002/cem.695

[22] M. Bylesjö, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, J.Trygg, Journal of Chemometrics, **20** (2006) 341-351.