A Novel PLS2-Based Algorithm for Imputing Missing Values in Foodomics/Metabolomics Studies with multiple response variables

Tashk, Ashkan; Engelsen, Søren Balling; Khakimov, Bekzod; Steenstrup Pedersen, Kim; Sørensen, Klavs Martin; kwr854, kwr854

Publication date: 2023

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Title:
A Novel PLS2-Based Algorithm for Imputing Missing Values in Foodomics/Metabolomics Studies with multiple response variables

Author: Ashkan Tashk (1)

Co-authors: Søren Balling Engelsen (1), Bekzod Khakimov (1), Kim Steenstrup Pedersen (2,3), Klavs Martin Sørensen (4), Ole-Christian Galbo Engstrøm (1,2,4)

Affiliations: (1) Department of Food Science, University of Copenhagen, (2) Department of Computer Science, University of Copenhagen, (3) Natural History Museum of Denmark, University of Copenhagen, and (4) FOSS Analytical A/S.

Abstract (400 words):
Missing values are frequent problems in data analytics studies, especially when a calibration model should be designed [1]. This study introduces a novel method, utilizing the PLS2 algorithm [2, 3] for imputing missing values within Ultra Centrifugation (UC) measurements of lipoprotein (LP) subfractions in human plasma. LP subfractions are essential biomarkers of food-related diseases such as obesity and cardiovascular diseases. They are categorized into four main fractions based on density and size: very low-density (VLDL), intermediate density (IDL), low-density (LDL), and high-density (HDL) LPs. Our proposed algorithm leverages proton ($^1$H) nuclear magnetic resonance (NMR) spectroscopy of LP in human blood plasma as a promising analytical method for rapid quantification of LP subfractions [5]. The robust and reliable NMR spectral data (p=1500) serves as the input X variable, while the UC variables (p=65) are used as the response variables. The UC variables are prone to measurement errors, occasionally considered as missing variables.

The proposed imputation method is iterative and consists of several stages to impute the missing values effectively. First, the samples are stratified based on the number of missing values in each sample. Subsequently, bootstrapping cross-validation is applied using PLS2 modeling. PLS2 models and their associated number of Latent Variables (LVs) with root mean squared error of cross-validation (RMSECVs) falling within a confidential interval of the RMSECV distribution, i.e., $[\mu-\sigma, \mu]$ are extracted. Next, the weighted mean of the predicted values for the extracted PLS2 models is calculated. The imputation process then progresses to the next stratification level until all samples are included and all missing values are imputed. Finally, the whole procedure is repeated until all the imputed variables converge.

Comparative analysis reveals that the proposed PLS2-based imputation method outperforms other commonly used imputation strategies such as iterative PCA [6] and missMDA [7]. The algorithm demonstrates superior performance in accurately imputing missing values within UC measurements of LP subfractions, enabling researchers to obtain more reliable and accurate results.

In conclusion, this study highlights the significance of addressing missing values in chemometrics studies. The proposed algorithm, which combines PLS2 modeling and NMR spectroscopy, offers a robust approach for imputing missing values within UC measurements of LP subfractions. Initial results demonstrate that the method's efficacy surpasses traditional
imputation strategies and may contribute to enriching and improving lipoprotein prediction models based on NMR spectroscopy.

References: