Chemometrics and Intelligent Laboratory Systems 149 (2015) 127–139

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

# Novel unified framework for latent modeling and its interpretation

Thanh N. Tran [a,e,*], Lionel Blanchet [b,c], Nelson Lee Afanador [d,e], Lutgarde M.C. Buydens [e]

[a] Center for Mathematical Sciences, Merck, Sharp & Dohme, Oss, The Netherlands
[b] Department of Toxicology, Maastricht University Medical Center, The Netherlands
[c] Top Institute Food and Nutrition (TIFN), Wageningen, The Netherlands
[d] Center for Mathematical Sciences, Merck, Sharp & Dohme, West Point, PA, USA
[e] Radboud University Nijmegen, Institute for Molecules and Materials, Analytical Chemistry & Chemometrics, The Netherlands

ARTICLE INFO

*Article history:*
Received 25 April 2015
Received in revised form 2 September 2015
Accepted 2 September 2015
Available online 8 September 2015

*Keywords:*
Chemometrics theory
Partial least squares
Principal component regression
Multivariate analysis
Interpretation

ABSTRACT

An important characteristic of chemometrics has been its need to manage the tradeoff between computational, mathematical and statistical performance against data interpretability. Additionally, being mostly seen as a conglomeration of data analytic methods that target the solution to real-world problems, the development of chemometrics as an independent and well-defined field has been hampered by its applied nature. Consequently, the broad range and diversity of application of chemometric tools has hindered the development of a unified theory able to propel it beyond its current use in analytical and industrial chemistry to larger and more complex data problems.

In this paper, we provide a mathematical vehicle for the understanding and improvement of current methods popular in chemometrics. Starting from a historical solution to matrix factorization we develop a novel unified framework for the fundamentals of latent variable modeling methods, elucidate major properties and clarify controversies between major PLS implementations and interpretations. The concepts presented in this work aims at contributing to a deeper understanding of the underlying theory of chemometrics methods, and strengthen their use in practice. Furthermore, this effort attempts to bridge the gap between chemometrics and big data problems and contribute to the development and acceptance of chemometrics as a mature and independent scientific field by the broader data analytic community.



## 1. Introduction

Chemometrics followed the evolution of analytical chemistry and resulted in many successful applications in chemical, petrochemical [1], pharmaceutical [2–4], and food processing [5,6]. This progress came with a number of challenges when dealing with more complex datasets. This is the case when advanced analytical instruments are used such as chromatography [7–12], metabolomics [13–15], mass spectrometry [16–18], hyperspectral imaging [2,19], and industrial processes [20,21].

In most of its applications, a central dogma in chemometrics is the need for interpretable models. This helps explain the importance given to visualization of the data in loading, residual, and score graphs where group identification, outlier detection, and ultimately data analysis become a visual and intuitive task. In this context, prior technical or scientific knowledge can be used to enhance any analysis. This interest in understanding the data, rather than simply use it for prediction, explains the focus of chemometrics on methods offering a balanced tradeoff between statistical performance and interpretability, as shown in Fig. 1.

The enormous increase of its applications creates an imbalance with basic research [22]. As stated in [23], the lack of theoretical developments seriously challenges the sustainability of chemometrics as a data science of the future. It has led to some misuse of algorithms [8,22,24] resulting in unfortunate false discoveries, and a loss in confidence to use the methods in other scientific fields [25]. The most classical examples are highly over-fitted models, prediction performance evaluated on the training set (implying that no validation was performed), and non-robust sampling methods for uncertainty estimation. Such errors easily lead to an incorrect assessment of important variables and interpretation of model parameters such as weights and loadings. As the data becomes more complex the confidence given to prediction and interpretation results progressively diminishes. This is the case in genomics and other areas of biology with highly multidimensional data consisting of an underdetermined stochastic nature inherent in the biological networks themselves. In most cases, the results obtained from latent variable modeling are not stable and hardly theoretically explainable. The topic becomes even more complex when one realizes that there are multiple PLS implementations [26] and that a controversy exists on which implementation is the most adequate. The NIPALS algorithm (also known as Wold-PLS) [27] is probably the most popular, yet SIMPLS [28], Martens' PLS [29] and Bidiag2 PLS [30] have all been successfully applied to regression and classification problems. Pell et al.

* Corresponding author. Tel.: +31 41 2662690, +31 24 3653180.
*E-mail addresses:* thanh.tran@merck.com, chemomtrics@science.ru.nl (T.N. Tran).

http://dx.doi.org/10.1016/j.chemolab.2015.09.002
0169-7439/

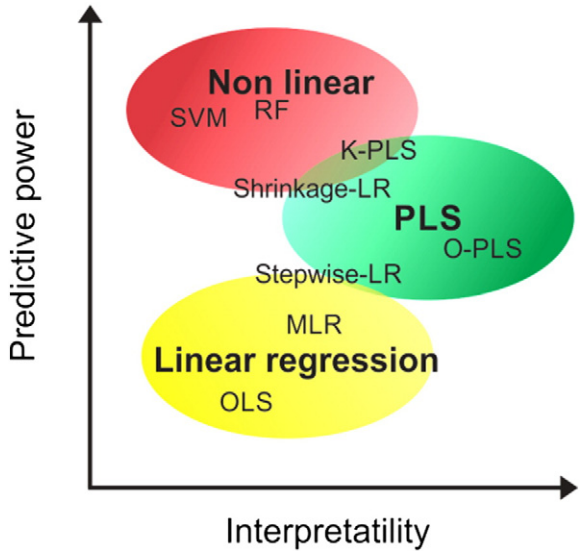


**Fig. 1.** Tradeoff between interpretation and predictive power.

pointed toward an internal inconsistency in PLS [31,32] which led to a debate within the chemometrics community [33,34]. The NIPALS inconsistency was later confirmed and its consequences on error estimation illustrated on industrial process monitoring [35,36]. Recently, Indahl [37] stated that the inconsistency does not exist and that the two models are rotated version of each other.

In this paper we propose a unified framework to better understand and discuss different latent variable modeling methods. For this we introduce the basic sequence (BS) as a special case of the Krylov sequence. The Krylov sequence has been used to interpret PLS regression coefficients [38] and to provide a geometric interpretation of PLS [39]. We will take the BS as a building block in order to benefit from its properties and develop a novel unified framework for latent variable modeling allowing us to further study latent variable modeling methods. This unified framework will firstly reduce confusion because different implementations can be positioned in one way or another in the unified framework. Secondly, the framework helps to better understand strengths and weakness of different implementations, guiding towards a better interpretation of latent variable models, and providing an understanding as to why it does not necessarily work for complex data. Finally, the framework encourages better development of latent variable models for more complex, and big, data.

In Section 2, we will first review the Krylov sequence and its historical perspective and uses, and from there we move on to explain the BS as a special case of the Krylov sequence and as a building block of the unified framework for as well single latent variable as multiple latent variable models for a dataset (Section 3). Here, some important, commonly used implementations for PCA, PCR, and PLS will be positioned in this framework to explore their strength and weakness. In Sections 4 and 5, we will focus on the implication of different PLS implementations on their predictive model performance and the interpretation of their latent variable model space and parameters, respectively. In-depth discussions for target projection and O-PLS methods will be provided. Additionally, we will provide illustrative examples of using the framework property to design a new technique for specific purposes; in this case for the selection of important variables.

## 2. Unified framework for single latent modeling

In this section, we will introduce a unified framework to fully explain the most classical chemometric tools: PCA and PLS. The framework uses basic sequence as a special case of the Krylov sequence in constructing a single latent variable. In Section 3, it will be extended to multiple layers with different data deflation types to complete the decomposition of $\boldsymbol{X}$ into its latent variables by means of a set of multiple sequences, leading to the unified framework of latent variable modeling methods.

### 2.1. Krylov sequence and its historical perspective

The current state of chemometrics can be better understood by examining the historical development of its analysis methods. The decomposition of a matrix into its corresponding left and right singular vectors and singular values is the cornerstone of most popular chemometrics methods. Historically, one of the first methods proposed to perform this task was the Krylov (also known as power) sequence [40]. The original article can be traced to a Soviet publication in 1931 [41]. Yet the concept of power sequences should be attributed to Chaim Müntz whose work had been reported to the *Académie des Sciences*, in Paris in 1913 by Emile Picard [42]. In the corresponding proceedings, one can find the definition of this sequence extended for any *kernel* matrix (e.g. for a covariance matrix). Simply stated, this work demonstrates that using a set of $n$ random directions, the principal components can be obtained by successive orthogonalization and normalization steps. For PLS, Krylov has been used to illustrate regression coefficients and loading vectors as belonging to a spanned space of a *special* Krylov sequence [38][39].

### 2.2. United framework for single latent variable models

We introduce here the basic sequence (BS) as a special case of the Krylov sequence and we will use this as the building block in order to benefit from its properties in constructing a unified framework for a single latent variable modeling. Given a mean-centered dataset $\boldsymbol{X}$, the basic sequence (BS) of length $k$ is defined as a sequence of $k$ basic vectors $\boldsymbol{Z} = \{\boldsymbol{z}^{[0]}, \boldsymbol{z}^{[1]}, \ldots, \boldsymbol{z}^{[k]}\}$ obtained by an iterative procedure to update the initial vector $\boldsymbol{z}^{[0]}$ with the cross-product matrix $\boldsymbol{X}'\boldsymbol{X}$ and normalization (2)–(3). This procedure converges on the dominant eigenvector of $\mathbf{X}'\mathbf{X}$. The scalar $k$ represents the number of iterations required satisfactory convergence. As such, the basic sequence is a special case of a Krylov sequence (or the power method) [40,41] in which we replace a general symmetric matrix with the rotation or cross-product matrix $\boldsymbol{X}'\boldsymbol{X}$. To avoid confusion with other Krylov cases, we call it the basic sequence. Throughout this paper, the sequence $\boldsymbol{Z}$ is also denoted as $\boldsymbol{B}^{[k]}(\boldsymbol{z}^{[0]}, \boldsymbol{X})$.

$$\boldsymbol{z}^{[i]} = \left(\boldsymbol{X}'\boldsymbol{X}\right)\boldsymbol{z}^{[i-1]} \quad (\textit{Basic rotation}) \tag{2}$$

$$\boldsymbol{z}^{[i]} = \frac{\boldsymbol{z}^{[i]}}{\|\boldsymbol{z}^{[i]}\|} \quad (\textit{Normalization}) \tag{3}$$

As an inherited property of the Krylov sequence, within the sequence $Z$ the basic vector $\boldsymbol{z}^{[i]}$ is rotated progressively towards the dominant eigenvector of the cross-product matrix $\boldsymbol{X}'\boldsymbol{X}$ [42]. The normalization in (3) is required to consider the operation as a rotation. In this procedure, moving from $\boldsymbol{z}^{[i-1]}$ to $\boldsymbol{z}^{[i]}$ represents a basic rotation and, as illustrated in our earlier work [50], the rotation speed depends on the ratio of the dominant eigenvalue to the 2nd largest eigenvalue, $\frac{|\lambda_1|}{|\lambda_2|}$. The procedure of the basic sequence is illustrated in Fig. 2. The vector $\boldsymbol{z}^{[0]}$ can be initialized in many different ways; randomly, predefined by user, or one of the rows from the data matrix $\boldsymbol{X}$. The sequence length (k) is also varied depending on the initialization.

To formulate a unified framework for a single latent vector we define the score sequence, denoted as $\{\boldsymbol{t}^{[1]}, \boldsymbol{t}^{[2]}, \ldots, \boldsymbol{t}^{[m]}\}$, and the loading sequence, denoted as $\{\boldsymbol{p}^{[2]}, \ldots, \boldsymbol{p}^{[m]}\}$, from the basic sequence via projection and regression steps, respectively. In particular, the score sequence is defined by the orthogonal projection of the rows of $X$ onto the basic sequence $\{\boldsymbol{z}^{[1]}, \ldots, \boldsymbol{z}^{[m]}\}$ and a loading sequence is obtained by the ordinary least square (OLS) solution of the columns of $X$ onto