Accepted Manuscript

Probabilistic partial least squares model: Identifiability, estimation and application

Said el Bouhaddani, Hae-Won Uh, Caroline Hayward, Geurt Jongbloed, Jeanine Houwing-Duistermaat

PII: S0047-259X(17)30276-2
DOI: https://doi.org/10.1016/j.jmva.2018.05.009
Reference: YJMVA 4368

To appear in: *Journal of Multivariate Analysis*

Received date : 6 May 2017

Please cite this article as: S. el Bouhaddani, H.-W. Uh, C. Hayward, G. Jongbloed, J. Houwing-Duistermaat, Probabilistic partial least squares model: Identifiability, estimation and application, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.05.009

Probabilistic partial least squares model: Identifiability, estimation and application

Said el Bouhaddani[a,*], Hae-Won Uh[b,a], Caroline Hayward[e], Geurt Jongbloed[c], Jeanine Houwing-Duistermaat[d,a]

[a] *Department of Medical statistics and bioinformatics, Leiden University Medical Center, The Netherlands*
[b] *Department of Biostatistics and Research Support UMC Utrecht, div. Julius Centrum, University Medical Center Utrecht, The Netherlands*
[c] *Department of Applied Mathematics, Delft University of Technology, The Netherlands*
[d] *Department of Statistics, University of Leeds, United Kingdom*
[e] *MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Scotland*

Abstract

With a rapid increase in volume and complexity of data sets, there is a need for methods that can extract useful information, for example the relationship between two data sets measured for the same persons. The Partial Least Squares (PLS) method can be used for this dimension reduction task. Within life sciences, results across studies are compared and combined. Therefore, parameters need to be identifiable, which is not the case for PLS. In addition, PLS is an algorithm, while epidemiological study designs are often outcome-dependent and methods to analyze such data require a probabilistic formulation. Moreover, a probabilistic model provides a statistical framework for inference. To address these issues, we develop Probabilistic PLS (PPLS). We derive maximum likelihood estimators that satisfy the identifiability conditions by using an EM algorithm with a constrained optimization in the M step. We show that the PPLS parameters are identifiable up to sign. A simulation study is conducted to study the performance of PPLS compared to existing methods. The PPLS estimates performed well in various scenarios, even in high dimensions. Most notably, the estimates seem to be robust against departures from normality. To illustrate our method, we applied it to IgG glycan data from two cohorts. Our PPLS model provided insight as well as interpretable results across the two cohorts.

Keywords: Dimension reduction, EM algorithm, Identifiability, Inference, Probabilistic partial least squares

1. Introduction

With the exponentially growing volume of data sets, multivariate methods for reducing dimensionality are an important research area in statistics. For combining two data sets, Partial Least Squares (PLS) regression [28] is a popular dimension reduction method [1]. PLS decomposes variation in each data set in a joint part and a residual part. The joint part is a linear projection of one data set on the other that best explains the covariance between the two data sets. These projections are obtained by iterative algorithms, such as NIPALS [28]. Partial Least Squares is popular in chemometrics [3]. In this field, the focus is on development of algorithms with good prediction performance, while the underlying model is less important. For applications in life sciences, interpretation of parameter estimates is necessary to gain understanding of the underlying molecular mechanisms.

For interpretation, a model needs to be identifiable. A model is said to be unidentifiable if the model corresponds to more than one set of parameter values. For PLS, rotation of the parameters does not change the model [26]. Hence, PLS does not provide an identifiable model. By constraining the parameter space, identifiability can be obtained. This involves solving a challenging optimization problem, since PLS requires estimating a structured covariance matrix [19].

For many problems in life sciences the study design needs to be accounted for, and algorithmic approaches such as PLS cannot be applied. Hence, a probabilistic formulation is necessary. Since likelihood method provides asymptotic standard errors of parameter estimates, computer-intensive resampling procedures can be avoided.

[*]Corresponding author

Email addresses: S.el_Bouhaddani@lumc.nl (Said el Bouhaddani), H.W.Uh@umcutrecht.nl (Hae-Won Uh), Caroline.Hayward@igmm.ed.ac.uk (Caroline Hayward), G.Jongbloed@tudelft.nl (Geurt Jongbloed), J.Duistermaat@leeds.ac.uk (Jeanine Houwing-Duistermaat)