Comparison of Multivariate Estimation Methods

Raju Rimal[a,*], Trygve Almøy[a], Solve Sæbø[b]

[a]*Faculty of Chemistry and Bioinformatics, Norwegian University of Life Sciences, Ås, Norway*
[b]*Prorector, Norwegian University of Life Sciences, Ås, Norway*

Abstract

Prediction performance often does not reflect the estimation behaviour of a method. High error in estimation not necessarily results in high prediction error but can leads to an unreliable prediction when test data are in a different direction than the training data. In addition, the effect of a variable becomes unstable and can not be interpreted in such situations. >> Fix from here << Here we have extend the previous study on prediction comparison to compare the estimation of the methods used in the study.

While Data science is battling to extract information from the enormous explosion of data, many estimators and algorithms are being developed for better prediction. Researchers and data scientists often introduce new methods and evaluate them based on various aspects of data. However, studies on the impact of/on model with multiple response model is limited. This study compares some newly-developed (envelope) and well-established (PLS, PCR) prediction methods based on simulated data specifically designed by varying properties such as multicollinearity, correlation between multiple responses and amount of information content in predictor variables. This study aims to give some insight on these methods and help researcher to understand and use them for further study.

Keywords: model-comparison,multi-response,simrel

*Corresponding Author
Email addresses: raju.rimal@nmbu.no (Raju Rimal), trygve.almoy@nmbu.no (Trygve Almøy), solve.sabo@nmbu.no (Solve Sæbø)

1. Introduction

Estimation of parameters in a regression model is an integral part in many research study. Research fields such as social science, econometric, psychology and medical study are more interested in measuring the impact of certain indicator or variable rather than performing prediction. Such studies has large influence in people's perception and also help in policy making and decisions.

This study extends the [Refer to my current paper] and compares some well established estimators such as Principal Components Analysis (PCA), Partial Least Squares (PLS) together with two new methods based on envelope estimation: Envelope estimation in predictor space (Xenv) and simultaneous estimation of envelope (Senv). The comparison tests the estimation performance of these methods using simulated data with controlled properties. The properties includes the number of predictors, level of multicollinearity, correlation between different response variables and the position of relevant predictor components.

Prediction has been an essential components of modern data science, weather it is statistical analysis or machine learning. Modern technology has facilitated a massive explosion of data, however, such data often contain irrelevant information consequently making prediction difficult. Researchers are devising new methods and algorithms in order to extract information to create robust predictive models. Mostly such models contain predictor variables that are directly or indirectly correlated with other predictor variables. In addition studies often constitute of many response variables correlated with each other. These interlinked relationships influence any study, whether it is predictive modeling or inference.

Modern inter-disciplinary research fields such as chemometrics, econometrics and bioinformatics are handling multi-response models extensively. This paper attempts to compare some multivariate prediction methods based on their prediction performance on linear model data with specific properties. The properties includes correlation between response variables, correlation between predictor variables, number of predictor vari-

ables and the position of relevant predictor components. These properties are discussed more in the [Experimental Design] section. Sæbø et al. (2015) and Almøy (1996) have made a similar comparison in the single response setting. In addition, Rimal et al. (2018) has also made a basic comparison on some prediction methods and their interaction with the data properties of a multi-response model. The main aim of this paper is to present a comprehensive comparison of contemporary prediction methods such as simultaneous envelope estimation (Senv) (Cook and Zhang, 2015) and envelope estimation in predictor space (Xenv) (Cook et al., 2010) with customary prediction methods such as Principal Component Regression (PCR), Partial Least Squares Regression (PLS) using simulated dataset with controlled properties. An experimental design and the methods under comparison are discussed further, followed by a brief discussion of the strategy behind the data simulation.

References

Almøy, T., 1996. A simulation study on comparison of prediction methods when only a few components are relevant. Computational Statistics & Data Analysis 21, 87–107. doi:doi:10.1016/0167-9473(95)00006-2.

Cook, R.D., Li, B., Chiaromonte, F., 2010. Envelope Models for Parsimonious and Efficient Multivariate Linear Regression. Statistica Sinica 20, 927–1010.

Cook, R.D., Zhang, X., 2015. Simultaneous envelopes for multivariate linear regression. Technometrics 57, 11–25. doi:doi:10.1080/00401706.2013.872700.

Rimal, R., Almøy, T., Sæbø, S., 2018. A tool for simulating multi-response linear model data. Chemometrics and Intelligent Laboratory Systems 176, 1–10. doi:doi:10.1016/j.chemolab.2018.02.009.

Sæbø, S., Almøy, T., Helland, I.S., 2015. Simrel - A versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. Chemometrics and Intelligent Laboratory Systems 146, 128–135. doi:doi:10.1016/j.chemolab.2015.05.012.