Summary of

Fast Partial Quantile Regression

Álvaro Méndez Civieta M. Carmen Aguilera-Morillo Rosa E. Lillo

Work published in Chemometrics and Intelligent Laboratory Systems, one of the top journals in the statistical field of chemometry. See Mendez-Civieta et al. [2022].

1 State of the art

Chemometry is a research field that studies the chemical composition of different objects from a statistical perspective. A very common problem in this field is dealing with colinear datasets in which predictive variables are highly correlated or with data that is high dimensional. An additional problem faced in this area is the fact that the response is typically not univariate, as is common in regression problems, but multivariate. An example of this type of datasets is that of the usage of near infrared measurements (which are high dimensional and colinear) for the prediction of the chemical composition of a product (which is multivariate), but this framework of colinear, possibly high dimensional data and multivariate responses is observed in other research areas like econometrics.

As a solution to this problem, [Wold, 1973] proposed the partial least squares (PLS) algorithm, a dimensionality reduction technique commonly applied to two data blocks X and Y that projects the independent data matrix X into a subspace of uncorrelated new variables that maximize the covariance with the response matrix Y. This projection is obtained as a result of an iterative process based on least squares, which are known to behave nicely when the errors are normally distributed. However there is no guarantee that the normality will be satisfied in many experimental data problems, where heavy tailed distributions, outliers or heteroscedasticity are expected to be found. This makes PLS extremely sensitive to the presence of outliers or non normal data. The solution to this problem has traditionally been centered in robustifying the least squares estimator in which PLS is based, see for example [Hubert and Branden, 2003] where they consider a robust covariance matrix estimator, [Serneels et al., 2005] where they make use of a robust M-regression estimator or [Acitas et al., 2020], where a partial robust adaptive modified maximum likelihood estimator is proposed among others.

An alternative solution is based on quantile regression [Koenker and Bassett, 1978], a statistical methodology that provides estimates of the conditional quantiles of a response given a set of covariates, as opposed to the mean based estimates provided by least squares. Being based on the quantiles, quantile regression is resistant to outliers, and can deal with heavy

tailed distributions and heteroscedasticity. Regarding partial least squares, [Dodge and Whittaker, 2009] extended a specific version of the PLS algorithm for univariate response problems to the quantile regression framework. They proposed a quantile covariance metric and used this metric to modify the univariate PLS, a modification that they called partial quantile regression (PQR). [Dodge and Whittaker, 2009] lays the foundation for an extension of PLS to the quantile regression framework, however we find some shortcomings in this methodology that should be addressed.

- It is an algorithmic modification with no background on the optimization problem that their PQR algorithm is solving.
- It is centered in univariate response problems, providing no solution for multivariate response problems commonly found in fields such as chemometrics.
- The computation time of their quantile coviariance grows linearly with the number of variables, making solving high dimensional problems computationally and time expensive.

2 Main contributions

In this paper we introduce the fast partial quantile regression (fPQR), a dimensionality reduction technique that addresses the above mentioned problems and extends the multivariate PLS to quantile regression. In particular, the main contributions of this work are summarized here:

Methodological contribution: fPQR as a quantile covariance maximization problem

Partial least squares is an iterative process based on least squares, but at each iteration, the algorithm can be posed as a covariance maximization problem. This fact opens the door to the possibility of using alternative covariance definitions. In this work, we pose the fPQR algorithm as a quantile covariance maximization problem.

Theoretical contribution: quantile covariance definitions

There is a clear definition on what the covariance is when dealing with mean based estimates, however, there is not a clear definition on what a quantile covariance metric should be. For this reason in this work we study three possible definitions that are: based on the traditional covariance between a random variable and an indicator function [Li et al., 2015], based on the quantile regression slope [Dodge and Whittaker, 2009] and based on the pearson correlation coefficient [Choi and Shin, 2018].

Computational contribution: algorithmic implementation

We provide an efficient implementation of the fPQR algorithm based on a singular value decomposition and include the definitions of the three quantile covariances under study. All

these methods are studied through a series of synthetic datasets and a chemometrics real dataset.

The resulting methodology, fPQR, is an algorithm that shares many of the PLS nice properties. First, it is a dimension reduction technique suitable for multicollinear or high dimensional data. Second, the new scores obtained by the algorithm are orthogonal. Third, it maximizes the quantile covariance between predictor and response. But it also has some additional properties, as it is a robust methodology, suitable for dealing with outliers or heteroscedastic data, and can provide an estimation of any quantile of interest of the response matrix, conditional to the predictors, obtaining a complete view of the distribution of the response.

3 Future work

We identify here three main lines of future work consisting in further methodological research, open source development, and real data applications.

Sparse formulation of fPQR

The fPQR algorithm provides an estimation of a new set of variables built as linear combinations of all the original variables. In high dimensional problems this may affect the interpretability of the results. The fact that the objective function that the algorithm solves has been clearly specified allows us to consider extensions of the methodology. This way, we can include penalizations that perform variable selection by producing sparse linear combinations of the original predictors, enhancing the interpretability.

Open source development

As a way to make high dimensional and quantile regression methodologies available for researchers, we launched some months ago an open source python package for variable selection called asgl that has been very well received in the research community, achieving more than 11000 downloads. We will work in the improvement of this package, developing and including the fPQR methodology, making it available for any interested user.

Real data applications

The work presented here has shown the main benefits of using the fPQR methodology in a series of simulations and a chemometrics real example, and has been mainly centered in the study of the central behaviour of the data, so it could be compared against PLS methodologies. However, in our opinion it would be very interesting to see further applications of fPQR benefiting from the study of other quantiles. Examples from fields like climate data, genetics or econometrics, where the study of the behaviour of the response variables at the tails of the distributions can be very helpful.

References

S. Acitas, P. Filzmoser, and B. Senoglu. A new partial robust adaptive modified maximum likelihood estimator. Chemometrics and Intelligent Laboratory Systems, 204:104068, 2020. ISSN 18733239. doi: 10.1016/j.chemolab.2020.104068. URL https://doi.org/10.1016/j.chemolab.2020.104068.

J.-E. Choi and D. W. Shin. Quantile correlation coefficient: a new tail dependence measure. 2018. ISBN 8223277360. URL http://arxiv.org/abs/1803.06200.

Y. Dodge and J. Whittaker. Partial quantile regression. Metrika, 70:35–57, 2009. ISSN 00261335. doi: 10.1007/s00184-008-0177-4.

M. Hubert and K. V. Branden. Robust methods for partial least squares regression. Journal of Chemometrics, 17(10):537–549, 10 2003. ISSN 0886-9383. doi: 10.1002/cem.822. URL http://doi.wiley.com/10.1002/cem.822.

R. Koenker and G. Bassett. Regression Quantiles. Econometrica, 46(1):33–50, 1 1978. ISSN 00129682. doi: 10.2307/1913643.

G. Li, Y. Li, and C. L. Tsai. Quantile Correlations and Quantile Autoregressive Modeling. Journal of the American Statistical Association, 110(509):246–261, 2015. ISSN 1537274X. doi: 10.1080/01621459.2014.892007.

A. Mendez-Civieta, M. C. Aguilera-Morillo, and R. E. Lillo. Fast partial quantile regression. Chemometrics and Intelligent Laboratory Systems, 223(March), 2022. ISSN 01697439. doi: 10.1016/j.chemolab.2022.104533. URL https://linkinghub.elsevier.com/retrieve/pii/S0169743922000442.

S. Serneels, C. Croux, P. Filzmoser, and P. J. Van Espen. Partial robust M-regression. Chemometrics and Intelligent Laboratory Systems, 79(1-2):55–64, 2005. ISSN 01697439. doi: 10.1016/j.chemolab.2005.04.007.

H. Wold. Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments. In P. R. Krishnaiah, editor, Multivariate Analysis?III, pages 383–407. Academic Press, 1973. ISBN 978-0-12-426653-7.