Multivariate linear L_1 regression for cluster-correlated data

Klaus Nordhausen[1], **Jaakko Nevalainen**[2], **and Hannu Oja**[1]

[1] *University of Tampere, Finland*
[2] *University of Turku, Finland*

Abstract

We consider the multivariate linear regression model $\mathbf{y}_i = \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\epsilon}_i$ where $\mathbf{y}_i$ is a p-variate response variable, $\mathbf{x}_i$ is a q-variate vector of explanatory variables, and $\boldsymbol{\epsilon}_i$ is a p-variate random error, $i = 1, ..., n$. We wish to make inference on the unknown $q \times p$ regression coefficient matrix $\boldsymbol{\beta}$. In the case of independent and identically distributed random errors, the estimates and related tests are commonly based on the use of the L_2 criterion $\sum_{i=1}^n ||\mathbf{y}_i - \boldsymbol{\beta}'\mathbf{x}_i||^2$ where $||\cdot||$ is the Euclidean norm. In an alternative approach, Bai et al. (1990) and Arcones (1998) considered the L_1 objective function $\sum_{i=1}^n ||\mathbf{y}_i - \boldsymbol{\beta}'\mathbf{x}_i||$. This approach then yields spatial median-type estimates and sign-type tests. Oja (2010) provides further references.

The assumption that $\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_n$ are independent is not true, however, if the data are clustered. The typical situation then is that instead of sampling independent and identically distributed random variables, the researcher samples observations in clusters. Observations within a cluster tend to be similar (correlated) but the clusters are independent. The clusters may, for example, be clinics with patients, schools with students, litters of rats, and so on. Also repeated measures on the same individual serves as an example.

In this talk we first shortly review the multivariate L_1 regression theory in the case of iid error variables. Then we show how the theory can be extended to the cluster-correlated case. The theory is illustrated with several examples. A small simulation study is conducted to compare the properties of different estimates.

Keywords

Clustered data, L_1 criterion, Repeated measures, Spatial median, Spatial sign.

References

Arcones, M.A. (1998). Asymptotic theory for M-estimators over a convex kernel. *Econometric Theory 14*, 387–422.

Bai, Z.D., Chen, R., Miao, B.Q., and Rao, C.R. (1990). Asymptotic theory of least distances estimate in multivariate linear models. *Statistics 4*, 503–519.

Oja, H. (2010). *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*. New York: Springer.