Adjusted Pareto scaling

K. Varmuza, P. Filzmoser

Vienna University of Technology, Institute of Statistics and Mathematical Methods in Economics, Computational Statistics, Vienna, Austria
Email: kurt.varmuza@tuwien.ac.at

The performance of multivariate models in chemometrics often depends on the applied method of scaling the x-variables. Autoscaling and Pareto scaling are widely used; here an *adjusted Pareto scaling* is suggested – covering the range from no scaling via classical Pareto scaling to autoscaling, and is compared with range scaling and vast scaling [1].

(1) *Autoscaling* of a variable is performed by x_c / s with x_c for the centred original variable, and s the standard deviation of the variable. (2) *Pareto scaling* is performed by $x_c / s^{0.5}$. The scaling effect is weaker than with autoscaling, noise is less amplified, and variables with a high original variance retain part of their importance for the model. (3) *Adjusted Pareto scaling* is performed by x_c / s^P with P varying between 0 (no scaling) and 1 (autoscaling), typically in steps of 0.1, thus including classical Pareto scaling with $P = 0.5$. (4) *Range scaling* is defined by $x_c/(x_{HIGH} - x_{LOW})$ with the variable spread given for instance by the quantiles 0.98 and 0.02. (5) *Vast scaling* considers the relative standard deviation $s_{RSD} = s/c$ (*c* is the mean or median of the variable) possessing rather small values for stable variables and scaling by $x_c/(s.s_{RSD})$.

These scaling methods are compared for PLS regression models and applying the strategy repeated double cross validation [2]. In one example GC retention indices are modelled by molecular descriptors, in the other the glucose content of fermentation samples is modelled by NIR absorbances. Parameters for scaling are varied and robust versions are considered.

Results show that appropriate scaling of x-variables by methods based on the spread may improve the performance of multivariate regression models. However, the effect has to be tested and optimized. Standard Pareto scaling with $P = 0.5$ may be not optimal and varying P between 0 and 1 is recommended. For some data sets, (robust) scaling methods based on the variable spread, like range scaling or vast scaling, may outperform (adjusted) Pareto scaling. No general rules, related to data properties, seem to be evident.

References

[1] J. Walach, et al.: In J. Jaumot, et al.: Comprehensive analytical chemistry. Data analysis for omics sciences: Methods and applications, Elsevier, Amsterdam, 165-196 (2018).

[2] P. Filzmoser, B. Liebmann, K. Varmuza, *J. Chemometr.,* **23** (2009) 160-171. Repeated double cross validation.