Adjusted Pareto Scaling

Kurt Varmuza and Peter Filzmoser

TU Wien - Vienna University of Technology, Austria
Institute of Statistics and Mathematical Methods in Economics | Computational Statistics

Adjusted Pareto scaling $x_{PARETO} = x_C / s^P$

x_C Centered variable; based on arithmetic mean or median

s Spread measure of variable; classic standard deviation (sd) or robust from IQR (interquartile range, s = 0.7413 IQR)

P **Pareto exponent** 0 … 1, varied for instance in steps of 0.1

$P = 0$	no scaling
$P = 0.5$	standard Pareto scaling
$P = 1$	autoscaling

Further scaling methods based on variable spread [1]

Range scaling $x_{RANGE} = x_C / (x_{HIGH} - x_{LOW})$

x_{LOW}, x_{HIGH} Range of the variable with borders given by quantile 0, 0.01, 0.02, …, and 1, 0.99, 0.98, …

Vast scaling $x_{VAST} = x_C / (s \cdot s_{RSD})$

$s_{RSD} = s / c$ **R**elative **s**tandard **d**eviation (coefficient of variation); c, center (mean or median)

Variable stability scaling: stable variables have small s_{RSD}

The performance of multivariate calibration or classification models often depends on the applied scaling of the x-variables.

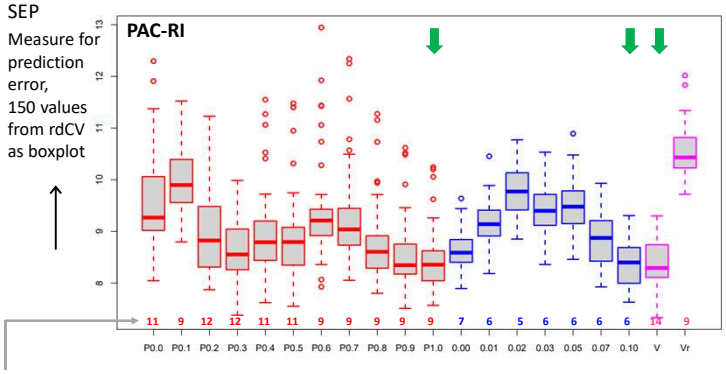
Autoscaling and Pareto scaling are widely used [1]; however, rarely strictly evaluated or compared.

Here an **adjusted Pareto scaling** is presented – covering the range from no scaling via standard Pareto scaling to autoscaling.

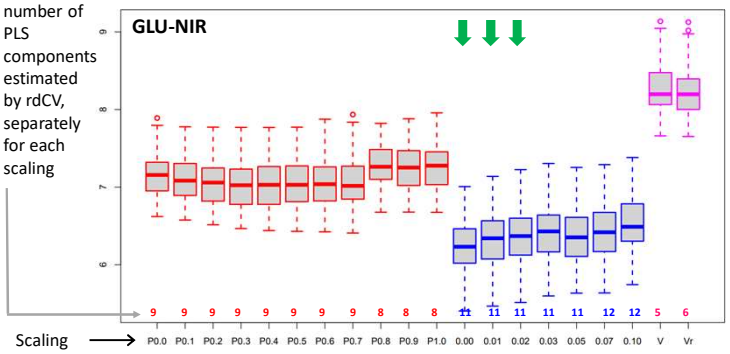
Aim: Model performance (prediction of y in new cases) may be improved by optimizing a parameter P between 0 and 1. Linear regression models $\hat{y} = \hat{\boldsymbol{b}}^T \boldsymbol{x}$, made by PLS, are considered.

- **r**epeated **d**ouble **C**ross **V**alidation (**rdCV**) [2] with PLS is used; 3 segments for test set split (outer CV); 7 segments for optimization of A, number of PLS components (inner CV); 50 repetitions; [3-5].
- Model performance is measured by **SEP** (standard error of prediction), the standard deviation of prediction errors for test set objects, estimated from CV and the repetitions in rdCV. Variations of SEP are presented by boxplots. Range ±2 SEP is a ~95% confidence interval for predictions of y [3].

SEP
Measure for prediction error, 150 values from rdCV as boxplot



A, optimum number of PLS components estimated by rdCV, separately for each scaling



Scaling parameter →

Adjusted Pareto scaling: Pareto exponent P = 0, 0.1, …, 1

Range scaling: Quantiles for low border 0, 0.01, …, 0.1

Vast scaling: V, classic Vr, robust

Examples

PAC-RI

n = 209 polycyclic aromatic compound structures [6,7]
m = 2234 variables (**molecular descriptors,** *Dragon* [8])
y, **gas-chromatographic retention index**, 197 – 504, sd = 80.8

Best scaling: "Pareto" scaling with P = 1 (= autoscaling)
Range scaling with range 0.1 – 0.9 quantiles
Vast scaling, classic (with mean and standard deviation)

GLU-NIR

n = 166 cereal fermentation samples [9]
m = 197 variables (**NIR** absorb., 1100-2300 nm, 1st derivative)
y, **glucose content**, 0.3 – 54,4 g/L, sd = 14.2 g/L

Best scaling: Range scaling with the ranges minimum to maximum, quantiles 0.01 to 0.99 or 0.02 to 0.98

Tentative conclusions

Scaling of x-variables by methods based on the spread (s) may improve the performance of multivariate regression models. However, the **effect has to be tested**, and **depends on the data**. No general rules, related to data properties, seem to be evident.

Standard Pareto scaling $x_{PARETO} = x_C / s^P$ *with* P = 0.5 may be not optimal. **Varying the Pareto exponent *P* between 0 (no scaling) and 1 (autoscaling) is recommended.**

A dependence of the model performance from P is found for under-fitted models, but may be low for well-fitted or over-fitted models.

Other (robust) scaling methods, based on the variable spread, like **range scaling** or **vast scaling,** may give better model performances for some data sets than adjusted Pareto scaling.

[1] Walach J., et al.: In Jaumot J., et al.: Comprehensive analytical chemistry. Data analysis for omics sciences: Methods and applications, Elsevier, Amsterdam, p. 165-196, 2018
[2] Filzmoser P., et al.: *J. Chemometrics*, **23**, 160 (2009)
[3] Varmuza K., Filzmoser P.: Introduction to multivariate statistical analysis in chemometrics, CRC Press, Boca Raton, FL, USA, 2009
[4] R, A language and environment for statistical computing, R Development Core Team, Foundation for Statistical Computing, http://www.r-project.org, Vienna, Austria, 2023
[5] Filzmoser P., Varmuza K.: R package chemometrics, Vienna, Austria, 2010, http://cran.at.r-project.org/web/packages/chemometrics/index.html
[6] Lee M.L., et al.: *Anal. Chem.*, **51**, 768 (1979)
[7] Varmuza K., et al.: *Comput. Struct. Biotechn. J.*, **5** [6], e201302007, 1 (2013)
[8] Todeschini R., et al.: Molecular descriptors for chemoinformatics, Wiley-VCH, Weinheim, 2009
[9] Liebmann B., et al.: *Biochem. Eng. J.*, **52**, 187 (2010)

230619

kurt.varmuza@tuwien.ac.at | www.lcm.tuwien.ac.at/vk/

Poster at *Conferentia Chemometrica 2023, Sopron (Hungary)*, 10-13 Sep 2023