Adjusted Pareto scaling for optimum prediction performance of chemometric multivariate models

Kurt Varmuza [1,2] and ***Peter Filzmoser*** [1]

TU Wien - Vienna University of Technology

[1] Institute of Statistics and Mathematical Methods in Economics | Computational Statistics

[2] Institute of Chemical, Environmental and Bioscience Engineering | Sustainable Technologies and Process Simulation

Multivariate models are routinely applied for the determination of a property y of samples using m measurements made on the samples.

Widely used are linear regression models $\hat{y} = \hat{\boldsymbol{b}}^T \boldsymbol{x}$, with $\hat{y}$ for the predicted property, $\hat{\boldsymbol{b}}$ the vector with estimated regression coefficients, and $\boldsymbol{x}$ the centered vector with m measurements (variables).

The performance of such chemometric models often depends on the applied method of scaling the x-variables.

Autoscaling and Pareto scaling are widely used [1].

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

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

x_C Centered variable; based on arithmetic mean or median.

s Spread measure of variable; classical standard deviation (sd) or from IQR (interquartile range) or another robust measure.

P **Pareto exponent** 0 … 1, step for instance 0.1

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

Calibration and evaluation

- PLS regression is used together with the strategy **r**epeated **d**ouble **C**ross **V**alidation (rdCV) [2].
- **Performance** of models is characterized by the prediction errors for test set samples, obtained separately from model optimization.
- **Performance measure** used is **SEP** (standard error of prediction), the standard deviation of prediction errors, estimated from several repetitions of CV (typ. 50 values give a boxplot). The range ±2 SEP defines a ~95% confidence interval for predictions of y [3].

Experimental

P Pareto exponent varied 0, 0.1, 0.2, …, 1

SEP 50 values estimated by rdCV (50 repetitions with different random splits), presented as boxplot.

A Number of PLS components (model complexity) estimated by rdCV (*well* fitted, mid plot), under-fitted (top plot), and over-fitted (bottom plot).

rdCV: 3 segments for test/calibration, 7 segments for optimization of A (double cross validation)

Scaling: median for center; 0.7413*IQR for spread

Software: R programming environment [4,5]

Data set **HEAT**

n = 122 biomass samples [6]

m = 13 variables (**C,H,N contents** and derived thereof)

y, **heating value**, 15719 – 25948 kJ/kg, sd = 1415 kJ/kg

Data set **GLU-IR**

n = 166 cereal fermentation samples [7]

m = 197 variables (**NIR** absorb., 1100-2300 nm, 1st derivative)

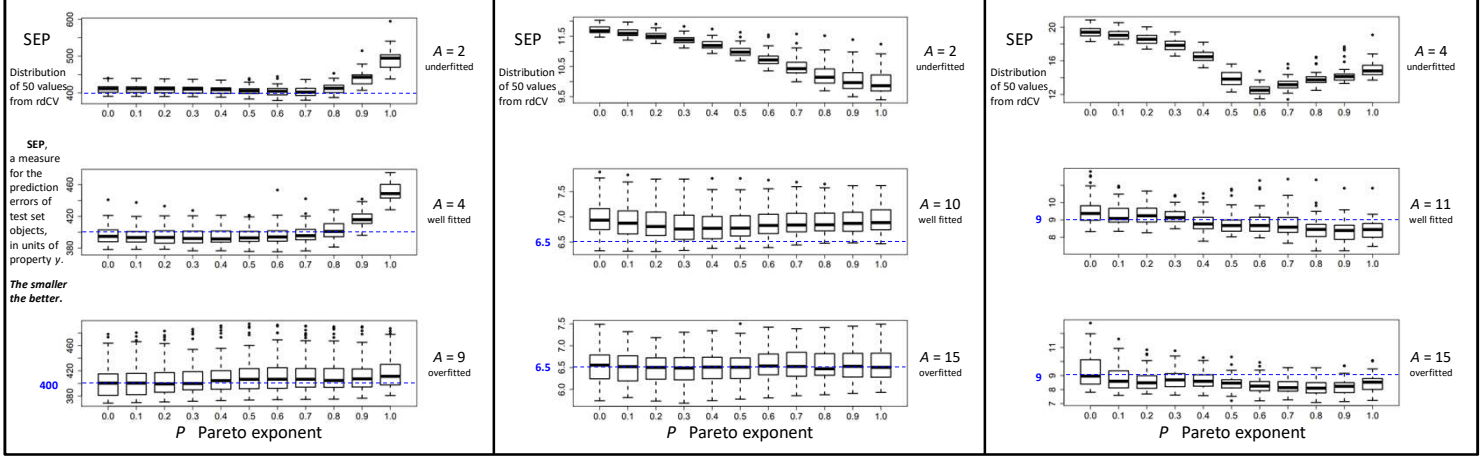
y, **glucose content**, 0.3 – 54,4 g/L, sd = 14.2 g/L

Data set **PAC-RI**

n = 209 polycyclic aromatic compound structures [8,9]

m = 2234 variables (**molecular descriptors,** *Dragon* [10])

y, **gas-chromatographic retention index**, 197 – 504, sd = 80.8



Conclusions *based on the results in these three examples – and an outlook*

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 simple 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 clear 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 (work in progress).

[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] Friedl A., et al.: *Anal. Chim. Acta*, **544**, 191 (2005)

[7] Liebmann B., et al.: *Biochem. Eng. J.*, **52**, 187 (2010)

[8] Lee M.L., et al.: *Anal. Chem.*, **51**, 768 (1979)

[9] Varmuza K., et al.: *Comput. Struct. Biotechn. J.*, **5** [6], e201302007, 1 (2013)

[10] Todeschini R., et al.: Molecular descriptors for chemoinformatics, Wiley-VCH, Weinheim, 2009

[230329a]

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

Poster at conference ***ANAKON 2023*, 11-14 April 2023, Vienna, Austria**