Adjusted Pareto scaling for optimum prediction performance of chemometric multivariate models

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

Email: kurt.varmuza@tuwien.ac.at

[1] Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria
[2] Institute of Chemical, Environmental and Bioscience Engineering, Vienna University of Technology, Vienna, Austria

---

Introduction

Multivariate models are routinely applied in analytical chemistry for the determination of a property *y* of samples from a set of *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 measurements (variables). A data set from typical n = 30–200 samples with known y and typical m = 10–500 variables is required for the development of a model [1]. Aim is an optimum prediction of y for samples not used in model building. The performance of the model (measured by the prediction errors) 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. Examples from chemistry are presented.

Methods

(1) *Autoscaling* of a variable is performed by x_c/s with x_c the centered original variable, and s the standard deviation of the variable. Autoscaling eliminates the units of the variables and makes them equal for the model building. Drawback is a blow-up of variables with small values perhaps originating from noise. (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. It is popular in biomarker identification and for metabolomics data. (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. The optimum P is selected by an evaluation of the prediction errors for test set samples resulting from repeated double cross validation [2] and PLS regression.

Results

Adjusted Pareto scaling has been tested together with PLS regression for various data sets: (A) Heating value of biomass modelled by elemental content data. (B) GC retention indices modelled by molecular descriptors. (C) Glucose content of fermentation samples modelled by NIR absorbances. Results indicate that classical Pareto scaling often can be improved by searching for an optimum exponent P.

Innovative aspects

- Improvement of the prediction performance of multivariate calibration models.
- Performance is cautiously estimated for test set samples.
- Generalized, adjustable approach for scaling methods based on the variable spread.

References

[1] Varmuza, K., Filzmoser P., Introduction to multivariate statistical analysis in chemometrics, CRC Press, Boca Raton, FL, USA, **2009**

[2] Filzmoser, P. et al., Repeated double cross validation. Journal of Chemometrics **2009**, 23, 160-171