Supporting Information for:

Numerical Representations of Metabolic Systems

Age K. Smilde*,† and Thomas Hankemeier‡

†Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands

‡Analytical Biosciences, LACDR, Leiden University, Leiden, The Netherlands

*E-mail: a.k.smilde@uva.nl

Short explanation of the methods cited

PCA: Unsupervised method that reduces large data sets in scores (representing the samples) and loadings (representing the variables). Used for data exploration and visualization.

OPLS-DA: Supervised method that finds differences between groups of samples. Used for biomarker discovery and investigating treatment effects.

PARAFAC: Extension of PCA for more complex data structures where the data can be arranged in a three-way table (e.g multiple metabolites measured at different time points for different samples).

SCA: Class of unsupervised methods to investigate multiple sets of related groups of data.

Formal treatment of measurement scales

One of the dominant theories of measurement is representational theory and this will be explained briefly in the following[1]. There are two notions important in this theory: a representation of a system and uniqueness properties of the numerical representation of that system. A small example will be used to explain these ideas.

Suppose we consider all sticks in the world; these are shown as an Empirical Relational System (ERS) in the upper left of Figure S1. Although the sticks may have different colors, we are only interested in their lengths. The relationships between the sticks can be represented numerically with the numbers in the upper right panel of Figure S1. An equally valid representation of the lengths of the sticks is given in the below right panel of Figure S1. Hence, we have two numerical representational systems (NRS1 and NRS2) that can both represent the ERS. Although the two NRSs are different the ratios between the numbers within both systems is the same: that property of the NRS is unique. Such systems (and associated

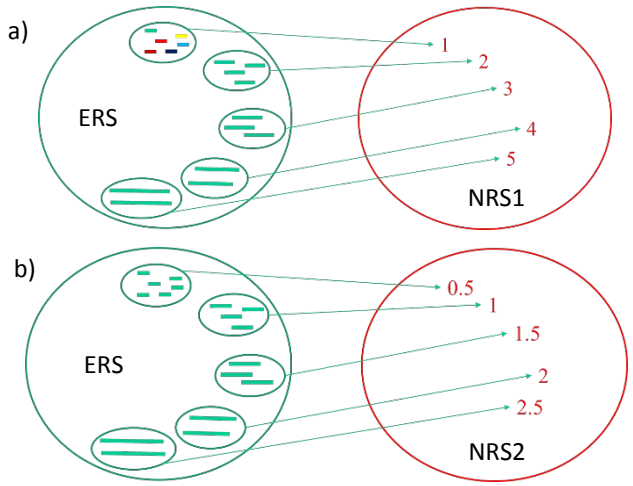

Figure S1: Numerical representations of the lengths of sticks: a) left: the empirical relational system (ERS) of which only the length is studied1 right: a numerical representation (NRS1)1 b) an alternative numerical representation (NRS2) of the same ERS carrying essentially the same information.

measurements) are therefore called ratio-scaled measurements (the unit is arbitrary). Likewise, it is possible to define interval-scaled measurements, ordinal-scaled measurements and nominal-scaled measurements. In all these cases, the permissible transformations allow for different NRSs that keep the properties of the represented ERSs intact.

Calibration models

Calibration models (or curves) are used to make build a relation between a measured intensity (or relative intensity) and the concentration of an analyte (i.e. metabolite). In general, a calibration model consists of four regions: i) a below limit of quantification (LOQ region, ii) a region of linearity and iii) a non-linear region and a saturation region (iv). These are indicated in Figure S2. The calibration model does not necessarily pass through the origin. It may be that the analyte is absorbed somewhere in the measurement equipment thereby generating a negative intercept or, alternatively, the analyte is present in one of the solvents as contaminant resulting in a positive intercept. Calibration models can also be estimated using weighted least squares to account for non-ideal situations[2].

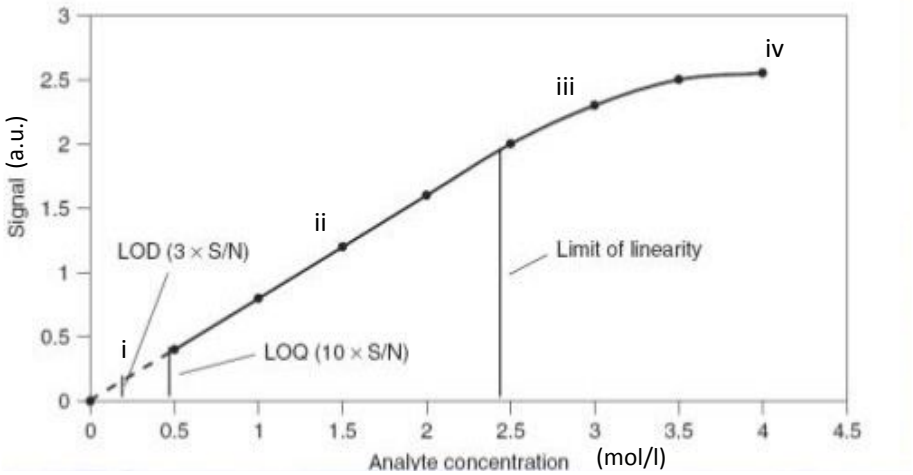

Figure S2: General shape of a calibration model. Abbreviations: LOD-limit is detection; LOQ is limit of quantification.

If only relative intensities are measured then the ratios of such intensities do not necessarily translate to the same ratio in concentrations. This holds only for measurements in region ii. In region i), the intensities cannot be used to infer something about concentrations and in region iii) the calibration model is nonlinear. Note that in this region, the intensities are still ordinal scaled. Actually, a bit more since the permissible transformation of ordinal-scaled variables is a monotonic increasing function, whereas in the case of calibration models, this function should also be concave.

Internal standards

Internal standards are used for different purposes: i) compensate for shifts in retention time or mass calibration to support alignment of features or identification of metabolites: ii) compensate for variations in sample preparation (e.g. sample volume) or injection volume, iii) compensate for variation in response factors due to e.g. variation in MS sensitivity and iv) support establishment of calibration models for metabolites.

The first type of correction is important for alignment of features (level 1) and for identification (prerequisite of level 2/3); the second and third type of corrections are important to allow for comparing abundances of features or metabolites between samples (level 1) or when creating semi-quantitative data (level 2) or determining concentrations (level 3). For determining concentrations often more internal standards are used, in its most extensive form one isotopically labelled internal standard per metabolite (so called isotope dilution analysis). It should be noted that internal standards can also be used to monitor the performance of a method rather than to correct numbers to (ultimately) obtain data.

References

(1) Hand, D. J. Statistics and the theory of measurement. Journal of the Royal Statistical Society Series A-statistics in Society 1996, 159, 445-473.

(2) Gu, H.; Liu, G.; Wang, J.; Aubry, A.; Arnold, M. Selecting the correct weighting factors for linear and quadratic calibration curves with least-squares regression algorithm in bioanalytical LC-MS/MS assays and impacts of using incorrect weighting factors on curve stability, data quality, and assay performance. Analytical Chemistry 2014, 86, 8959-8966.