Remarks on some validation parameters

Péter Király, Dániel Kovács and Gergely Tóth

Institute of Chemistry, Eötvös University, Budapest

Abstract

In our spin-off poster of the size dependence study of validation parameters we take a closer look at emerged issues. It is known that QSAR validation conventions divide internal validation parameters into two groups in terms of assessing *goodness-of-fit* or *robustness*, while the external ones are applied to check the *predictivity* [1].

Rank correlation of validation parameters

We calculated the rank correlation of validation parameters for given sample sizes. The parameters were chosen according to the three aspects of validation:

$$\underbrace{R^2, \mathrm{CCC}}_{\text{internal (goodness-of-fit)}} \quad \underbrace{Q^2_{\mathrm{LOO}}, \mathrm{CCC}_{\mathrm{LOO}}}_{\text{internal (robustness)}} \quad \underbrace{Q^2_{F2}, \mathrm{CCC}_{\mathrm{test}}}_{\text{external (predictivity)}}$$

We found that the internal/internal, external/external, internal/external pairings behave differently.

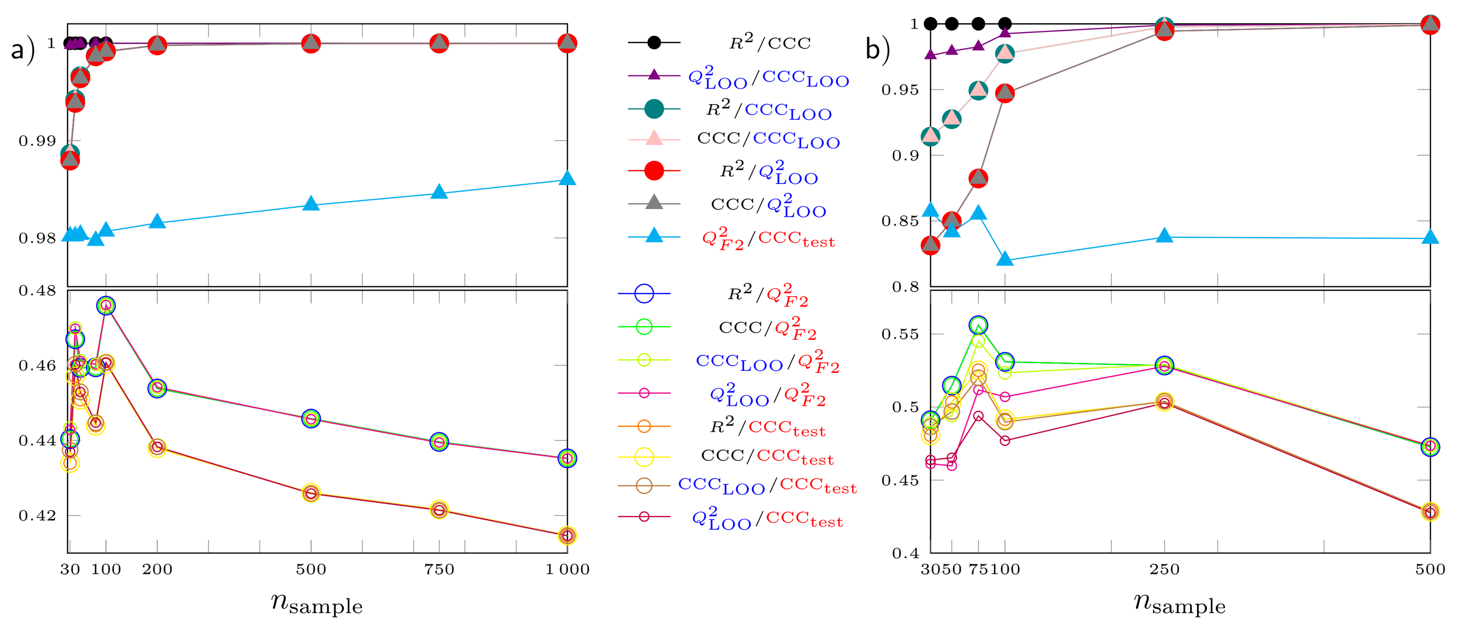

Figure 1: Rank correlations as function of sample size. **(a)** Powerplant dataset ($R^2 = 0.92$). **(b)** Concrete dataset ($R^2 = 0.62$). The Powerplant set is more adequate for linear modeling.

- largest correlations for internal/internal pairings
- mixed internal pairings have slightly weaker rank correlation
- external/external correlation is still significant
- external/internal pairings are weakly correlated

Up to a given sample size the internal validation aspects goodness-of-fit and robustness are distinguishable, while above it they become *redundant*. The internal and external schemes of validation have different information content for all sample sizes, therefore they are *not redundant*. This novel insight contributes to the long-standing debate of advocates of internal and external validation.

Comparison of internal parameters in constrained/unconstrained models

Following table surveys the attributes that can be deducted from definitions of the basic validation parameters of goodness-of-fit (R^2,R^2_{adj}) and robustness (Q^2_{LOO}).

	linear regression with arbitrary intercept	all other models
R^2	$[0,1]$	$(-\infty,1]$
R^2_{adj}	$\left[-\frac{p}{n-p-1},1\right]$	$(-\infty,1]$
Q^2_{LOO}	$(-\infty,1]$	$(-\infty,1]$
$R^2 \equiv \rho^2\left(\underline{y},\underline{\hat{y}}\right)$	true	false
$R^2_{\mathrm{adj}} \leq R^2$	true (=, if $R^2=1$)	true (=, if $R^2=1$)
$Q^2_{\mathrm{LOO}} \leq R^2$	true	true
$\mathrm{TSS} = \mathrm{MSS} + \mathrm{RSS}$	true	not true
meaning of $R^2 = 0$	best model is the average	?
$\overline{y} = \overline{\hat{y}}$	true	false
R^2 meaning	model explained part of TSS	?
Definitions: $R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$	$R^2_{\mathrm{adj}} = 1 + (R^2-1)\frac{n-1}{n-p-1}$	$Q^2_{\mathrm{LOO}} = 1 - \frac{\mathrm{PRESS}}{\mathrm{TSS}}$

Proposal for Q^2_{F3} correction

We propose to correct the biased estimator of variance in the denominator of Q^2_{F3} [2]. Speciality of Q^2_{F3}: independent of test allocation.

$$Q^2_{F3} = 1 - \frac{\mathrm{RSS}_{\mathrm{test}}/n_{\mathrm{test}}}{\mathrm{TSS}_{\mathrm{train}}/\left(n_{\mathrm{train}}-1\right)}$$

Roy-Ojha diagrams

Roy and Ojha suggested to assess the predictivity of the models with their $\overline{r^2_m}$ and Δr^2_m parameters in addition to the Q^2_{F1-3} family [3]. Models with feasible Q^2_{F1-3} values are acceptable if $\overline{r^2_m} > 0.5$ and $\Delta r^2_m < 0.2$. We propose to visualize these in a 2D diagram. Therefore the domain of acceptance is the upper left part in the diagrams partitioned by blue lines in Fig. 2. Additionally the models were colored according to their Q^2_{F2} value.

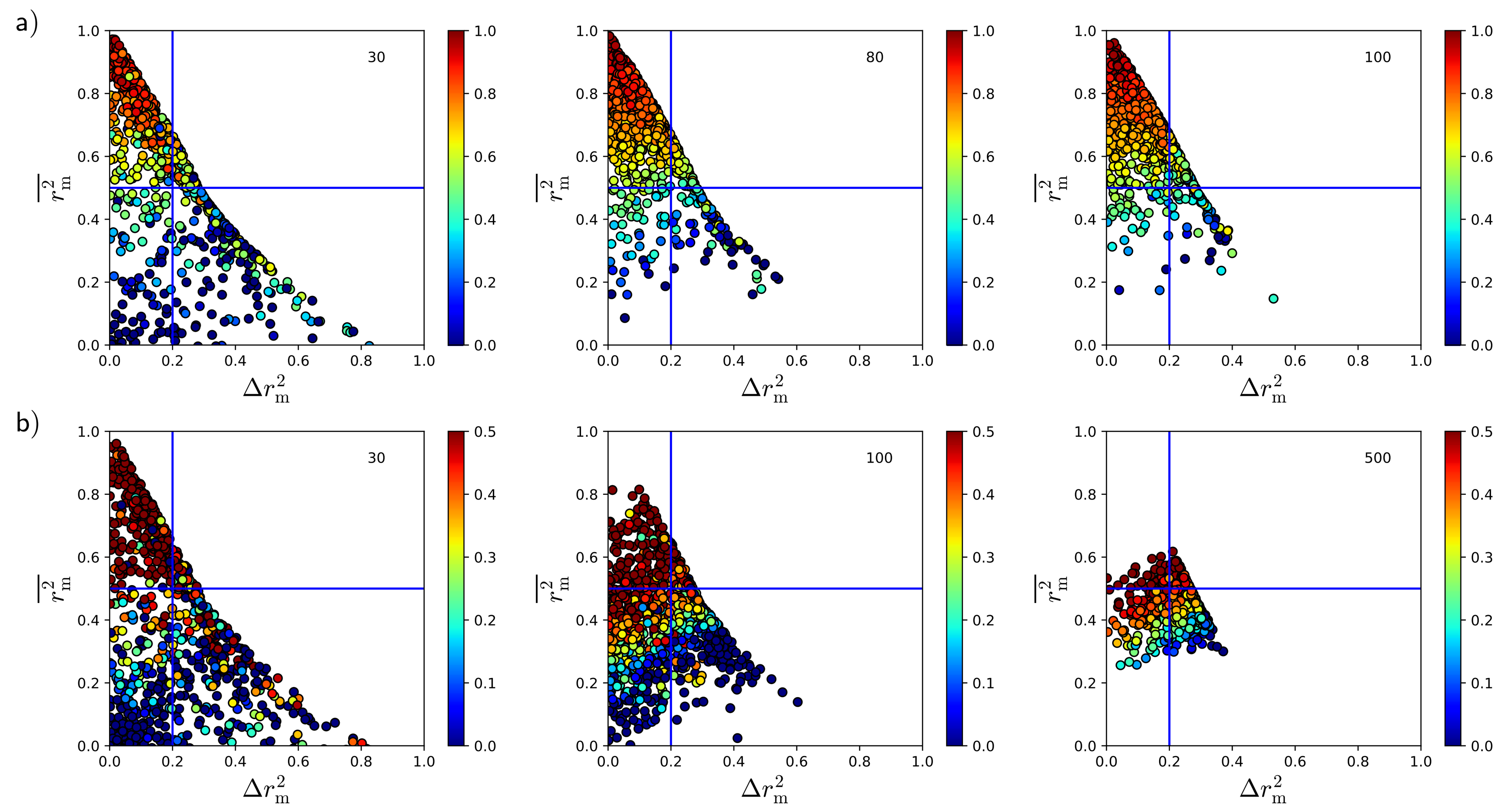

Figure 2: **(a)** Roy-Ojha diagrams for the Gramatica QSAR dataset with increasing sample size, $n_{\mathrm{sample}} = (30, 80, 200)$. **(b)** Roy-Ojha diagrams for the Concrete benchmark dataset with increasing sample size, $n_{\mathrm{sample}} = (30, 100, 500)$. The models were colored according to their Q^2_{F2} value. Domain of acceptable predictivity is set by $\overline{r^2_m} > 0.5$ and $\Delta r^2_m < 0.2$.

In Fig. 2 the Roy-Ojha diagrams are shown for two datasets for sample sizes 30, 80, 200 and 30, 100, 500. The Gramatica dataset in Fig. 2(a) is better modelable, already at small sample sizes almost all models are located in the domain of acceptance. The Concrete set studied in Fig. 2(b) is less suitable for linear modeling, the tendency is similar, however the domain of acceptance is emptied with increasing sample size and the models are concentrated around the R^2 value of the full set. The stratification of Q^2_{F1-3} values through the partitions confirms correlation of Q^2_{F1-3} and the Roy-Ojha parameters. The internal parameters R^2, R^2_{adj} and Q^2_{LOO} did not show significant stratification as a function of Roy-Ojha parameters. The Roy-Ojha diagrams visualized according to the idea of Dániel Kovács serve as a visual asset in monitoring the predictivity of the models.

R^2 superior to CCC

Lin introduced the Concordance Correlation Coefficient (CCC) as an index to measure departure of measurement from the concordance line (straight line through origin with slope $= 1$) [4]. CCC detects vertical shift of the measurement relative to the concordance line (location shift) and rotation of the concordance line about the ordinate axis (scale shift). CCC has clear benefits compared to Pearson's ρ. Departure from the concordance line is an unconstrained problem ($R^2 \neq \rho^2$) and certainly can be addressed with R^2 too. Fig. 3 shows that in case of vertical shift and rotation R^2 has superior sensitivity and warning capacity compared to CCC.

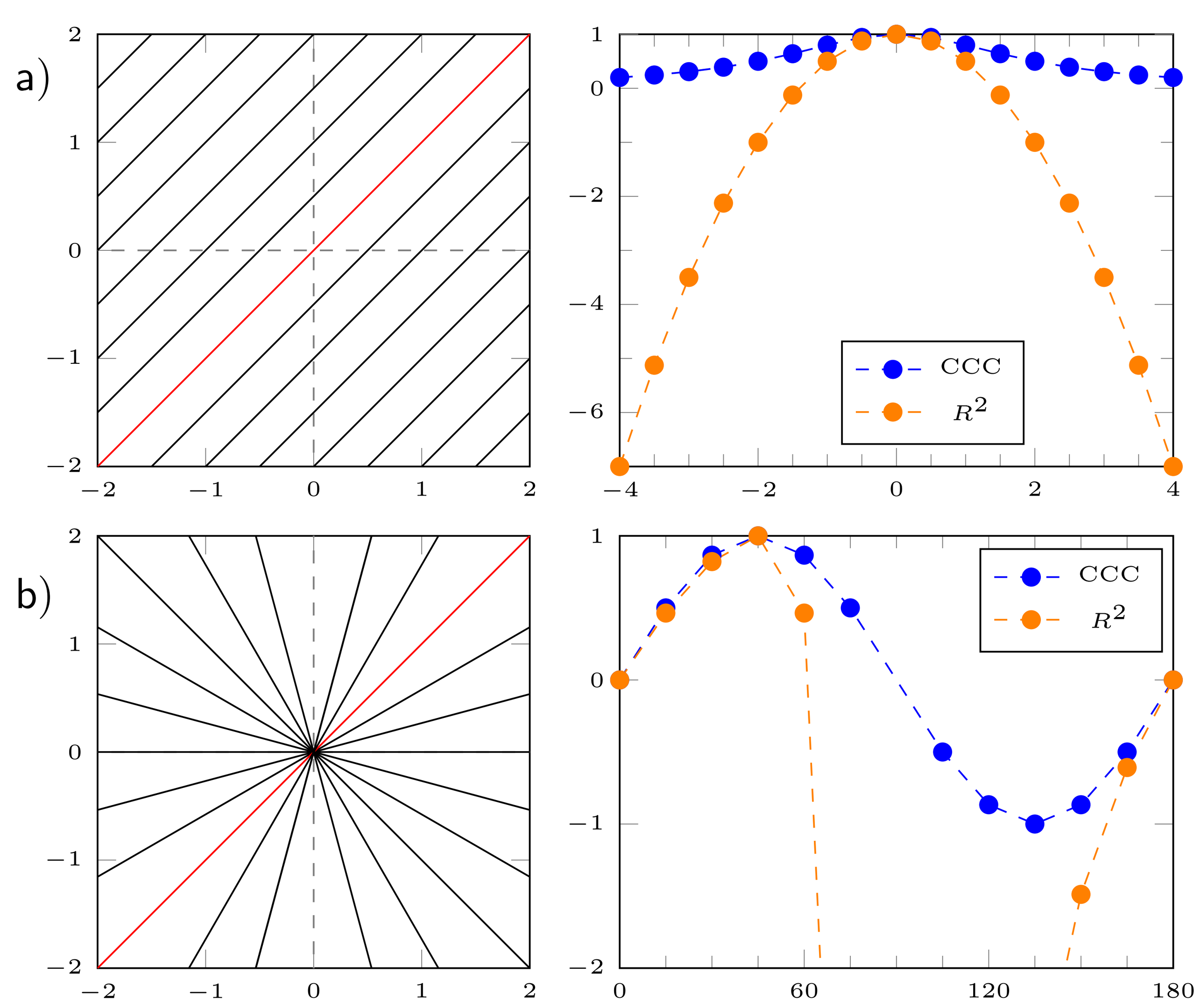

Figure 3: R^2 and CCC values of shifted and rotated measurements. Black lines representing the departed measurements, the red line is the concordance line, representing full agreement. **(a)** The measurement is vertically shifted relative to the concordance line. **(b)** The measurement is rotated by 15° about the ordinate axis.

References

(1) OECD Principles for the Validation, for Regulatory Purposes, of QSAR Models., 2004.

(2) Consonni, V.; Ballabio, D.; Todeschini, R. *J. Chem. Inf. Model.* **2009**, *49*, 1669–1678.

(3) Ojha, P. K.; Mitra, I.; Das, R. N.; Roy, K. *Chemom. Intell. Lab. Syst.* **2011**, *107*, 194–205.

(4) Lin, L. I.-K. *Biometrics* **1989**, *45*, 255–268.