Size Dependence of Validation in Linear Modeling

Péter Király
2nd year PhD report
Supervisor: Gergely Tóth

May 21, 2020

Outline

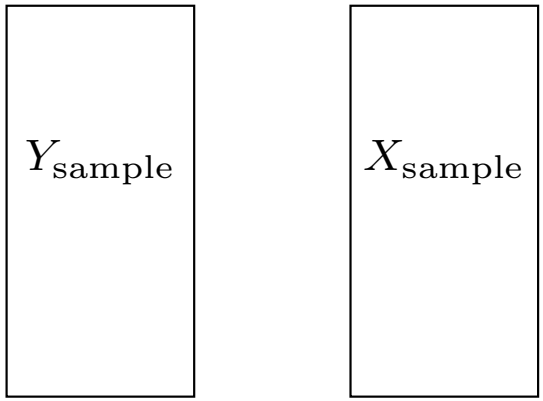
- Sample size dependence of validation parameters (VPs)
  - studied on multiple linear regression (MLR) and multi-target partial least squares regression (PLS2) models
  - Scramble or randomize?
  - Leave-one-out (LOO) or leave-many-out (LMO) CV?
  - R^2 is superior to CCC (concordance correlation coefficient)
  - Assessment of predictivity on 2D Roy-Ojha-Kovács diagrams
  - Rank correlation of VPs

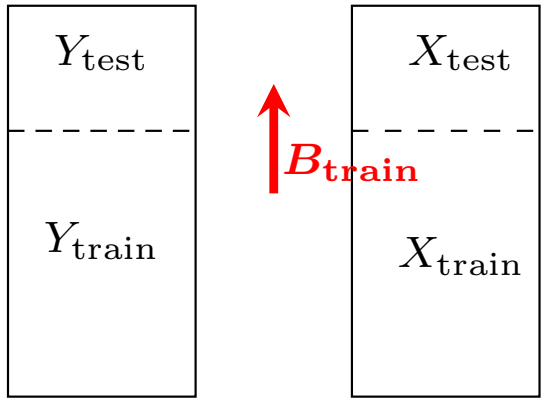
Validation schemes

Internal validation

External validation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$$





goodness of fit, robustness

predictivity

Validation parameters (VPs)

internal VPs

$$R^2 = \frac{1}{d}\sum_{i=1}^{d} 1 - \frac{\mathrm{RSS_i}}{\mathrm{TSS_i}} \qquad R^2_{\mathrm{inc}} = 1 - \frac{\sum_{i=1}^{d} \mathrm{RSS}_i}{\sum_{i=1}^{d} \mathrm{TSS}_i}$$

$$Q^2_{\mathrm{LOO}} = \frac{1}{d}\sum_{i=1}^{d} 1 - \frac{\mathrm{PRESS}_i}{\mathrm{TSS}_i}$$

$$\mathrm{CCC} = \frac{1}{d}\sum_{i=1}^{d} \frac{2\sigma(y_i, \hat{y}_i)}{\mathrm{RSS}_i + \mathrm{MSS}_i + n_i\left(\bar{y}_i - \bar{\hat{y}}_i\right)^2}$$

$$\mathrm{RMSE} = \frac{1}{d}\sum_{i=1}^{d} \sqrt{\frac{\mathrm{RSS}_i}{n_i}} \quad \ldots$$

external VPs

$$Q^2_{F1} = \frac{1}{d}\sum_{i=1}^{d} 1 - \frac{\mathrm{PRESS}_{\mathrm{test},i}}{\mathrm{TSS}_{\mathrm{test,i}}\left(\bar{y}_{\mathrm{train},i}\right)}$$

$$Q^2_{F2} = \frac{1}{d}\sum_{i=1}^{d} 1 - \frac{\mathrm{PRESS}_{\mathrm{test},i}}{\mathrm{TSS}_{\mathrm{test},i}}$$

$$Q^2_{F3} = \frac{1}{d}\sum_{i=1}^{d} 1 - \frac{\mathrm{PRESS}_{\mathrm{test},i}/n_{\mathrm{test},i}}{\mathrm{TSS}_{\mathrm{train},i}/n_{\mathrm{train},i}}$$

$\vdots$

Datasets

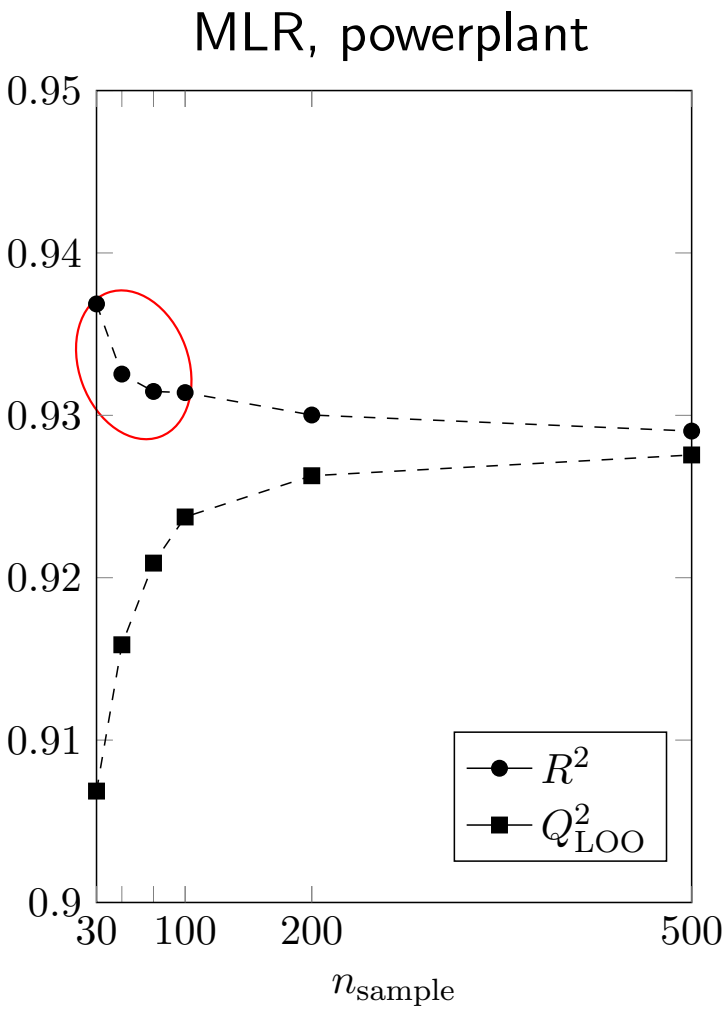
dataset	targets	features	instances
powerplant	1	4	9568
concrete	1	8	1030
gramatica1-3	1-1-1	4	369
temperature	2	26(1)	400
slump	3	7(7)	103
air pollution	6	38(19)	363

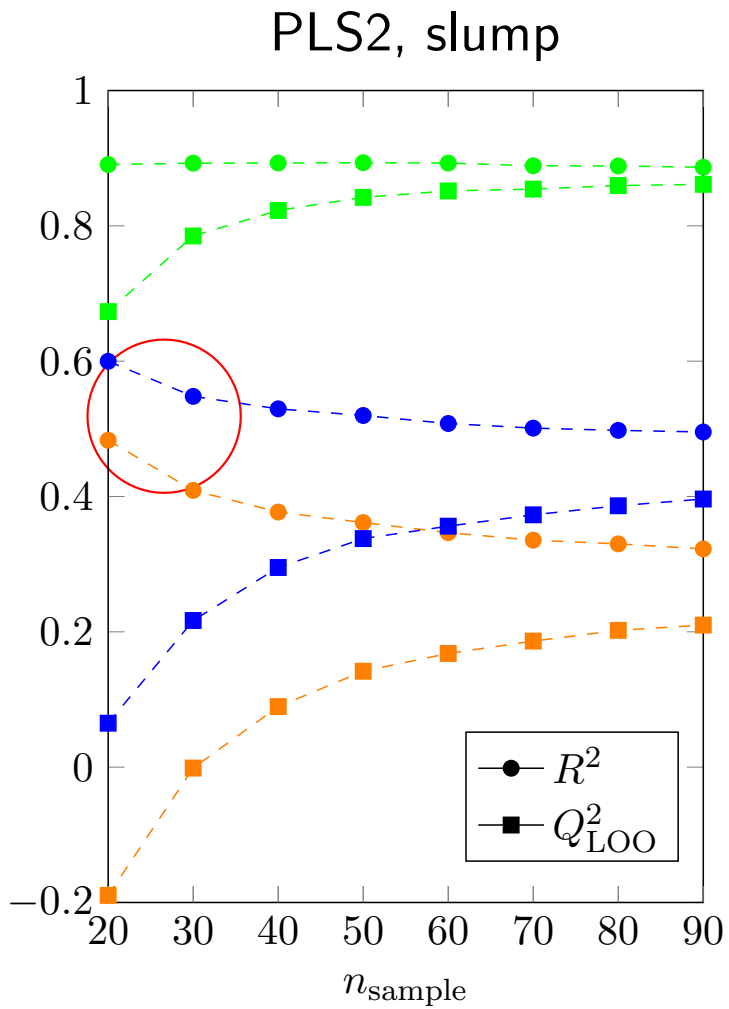
Specialities for multi-target sets

- sets are modeled with *multi-target partial least squares regression* (PLS2)
- predictor and response matrices standardized
- number of latent variables is set by *repeated double cross validation* (rdCV), R^2

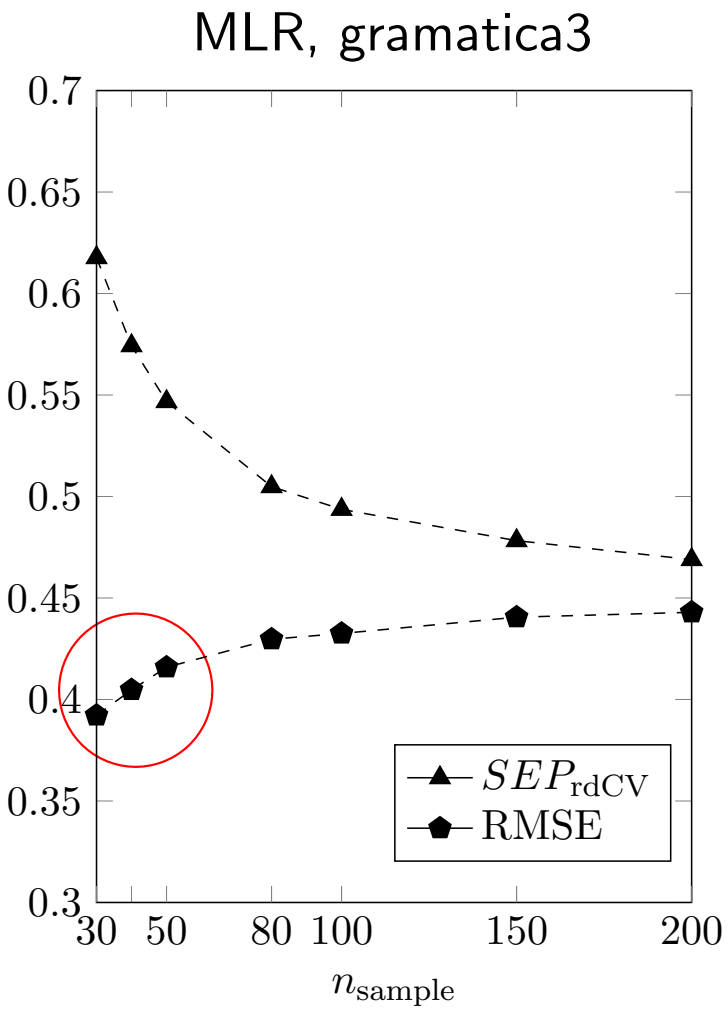
Tasks for size dependence study

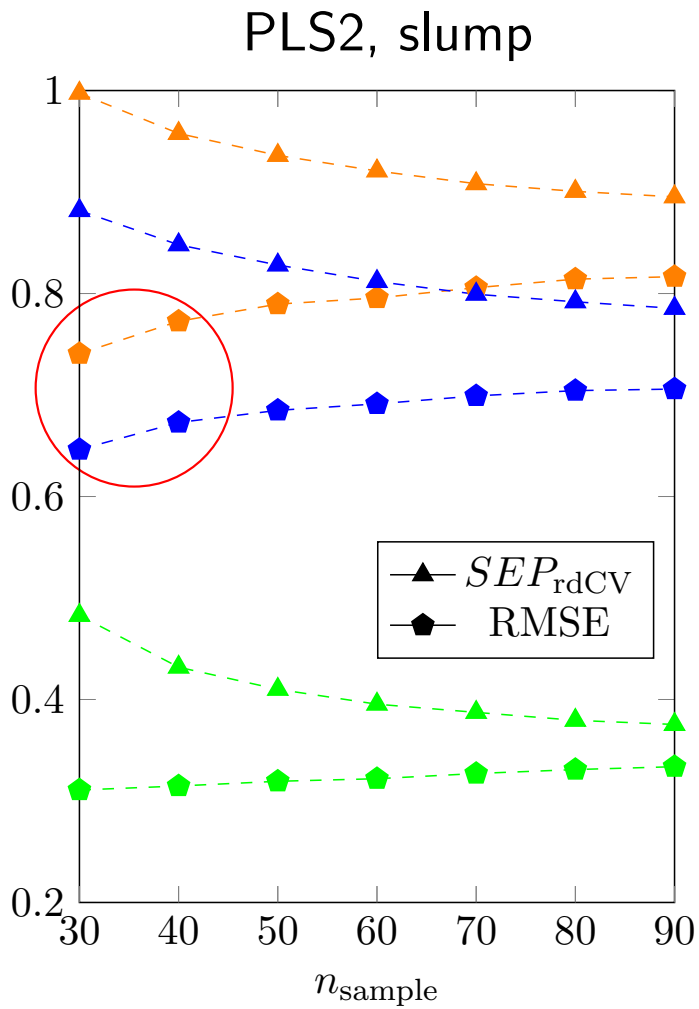
- random sampling from population
- *training/test* split of sample for external validation (80/20 for training)
- linear modeling and computation of *validation parameters* (VPs)
- tasks above repeated $1000\times$ for given sample size, VPs averaged





Is the smaller model the better one?

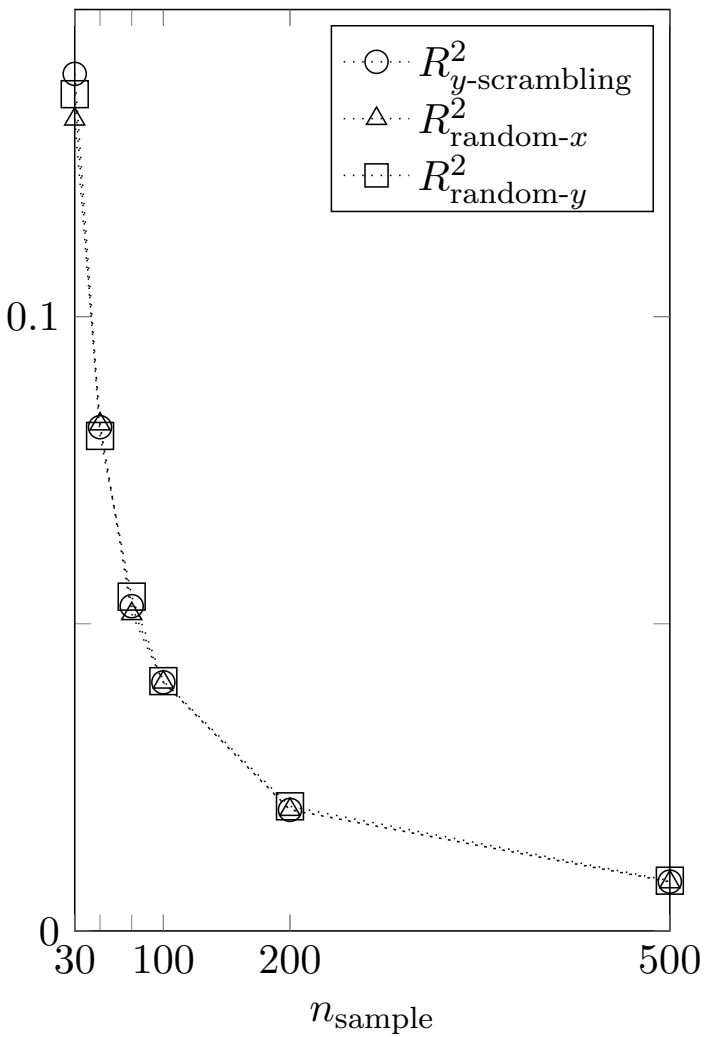




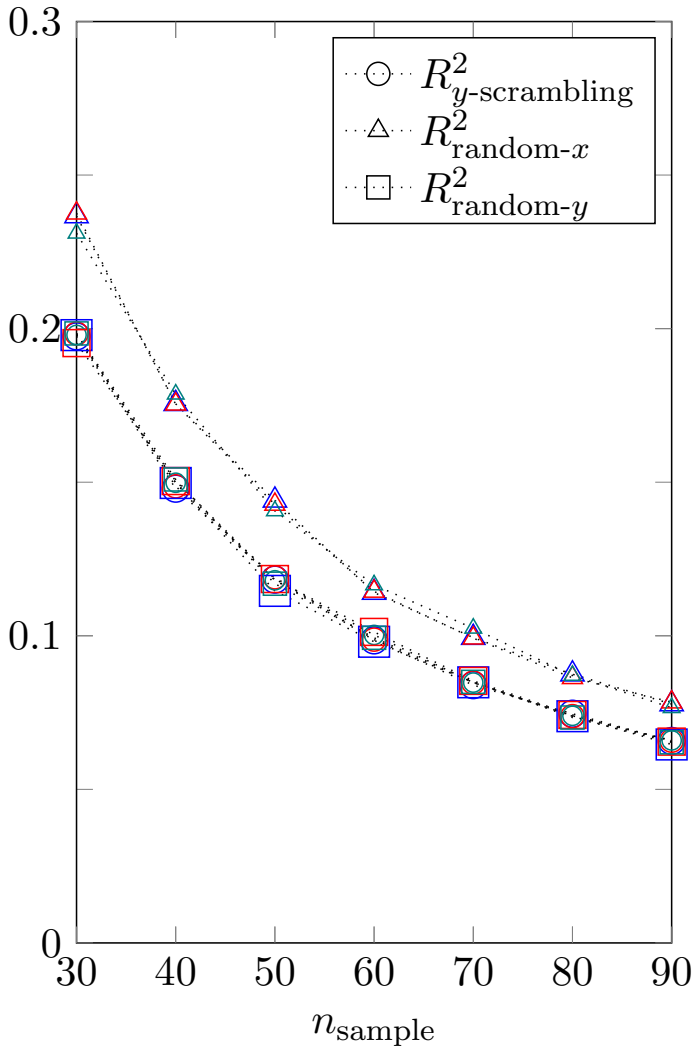
Is the smaller model the better one?

Internal VPs - scrambling, randomization

MLR, powerplant



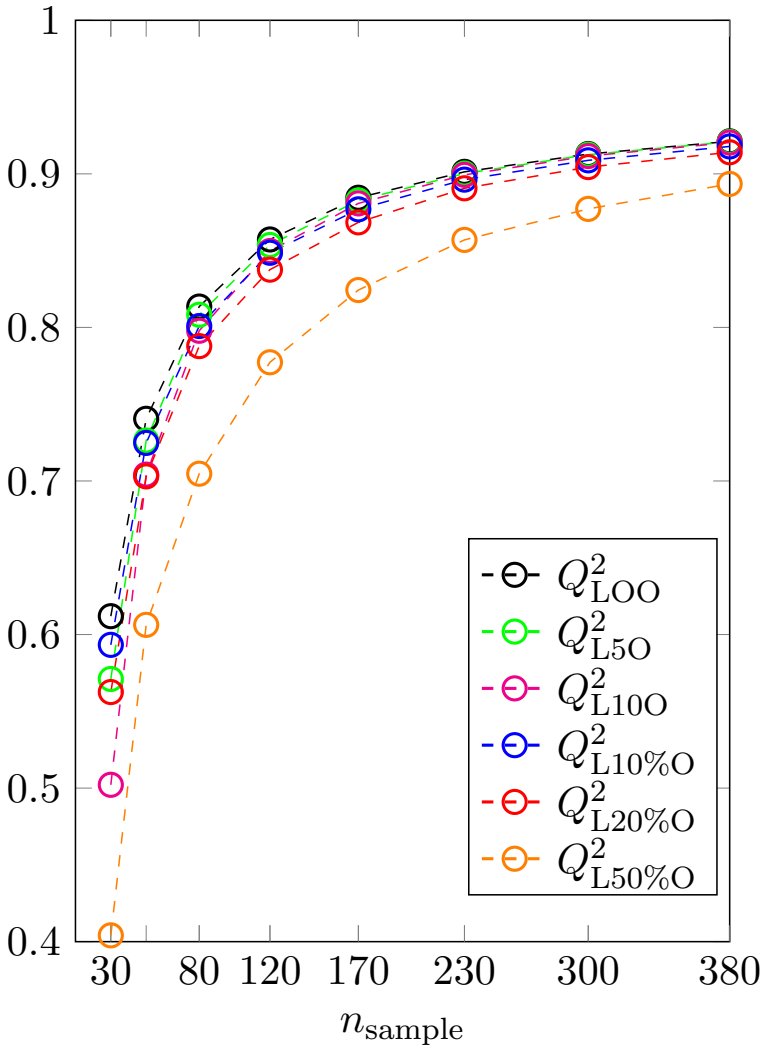
PLS2, slump



y-scrambling easiest to calculate

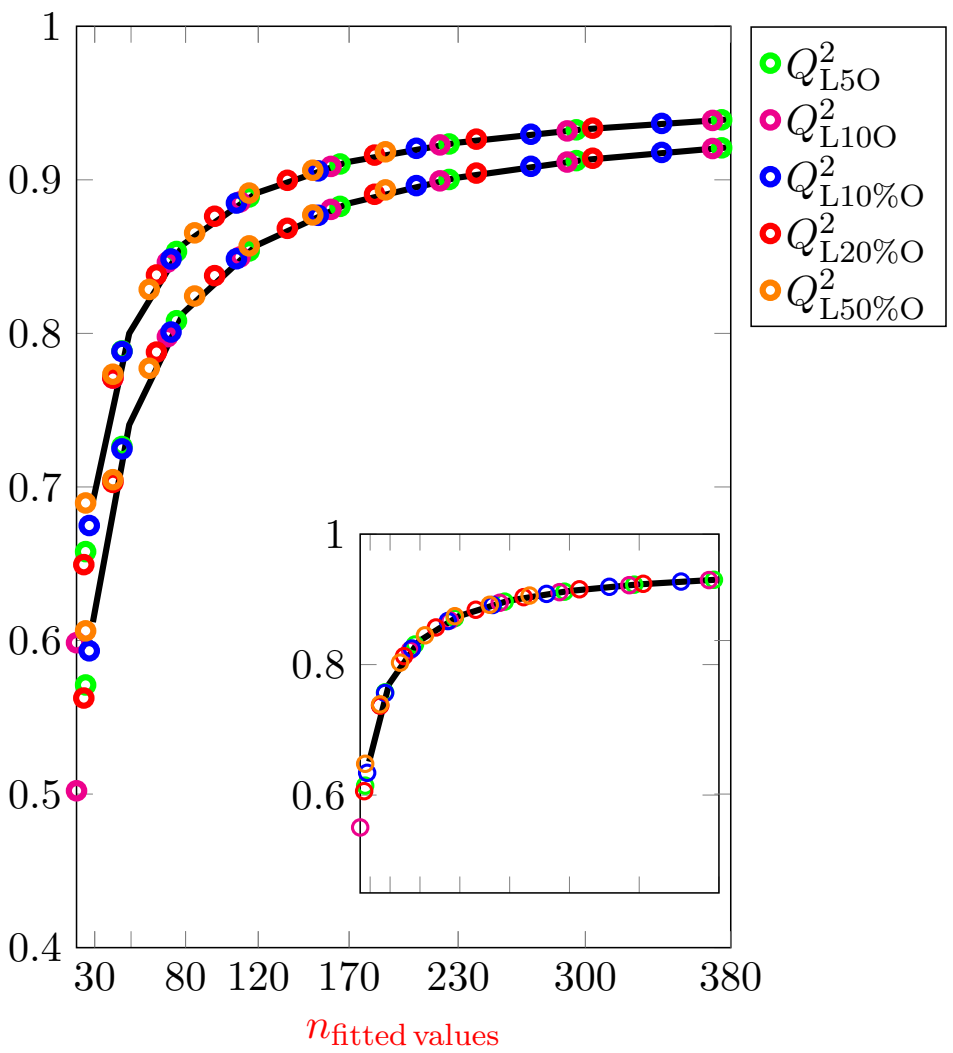
Internal VPs - leave many out

PLS2, temperature y_1

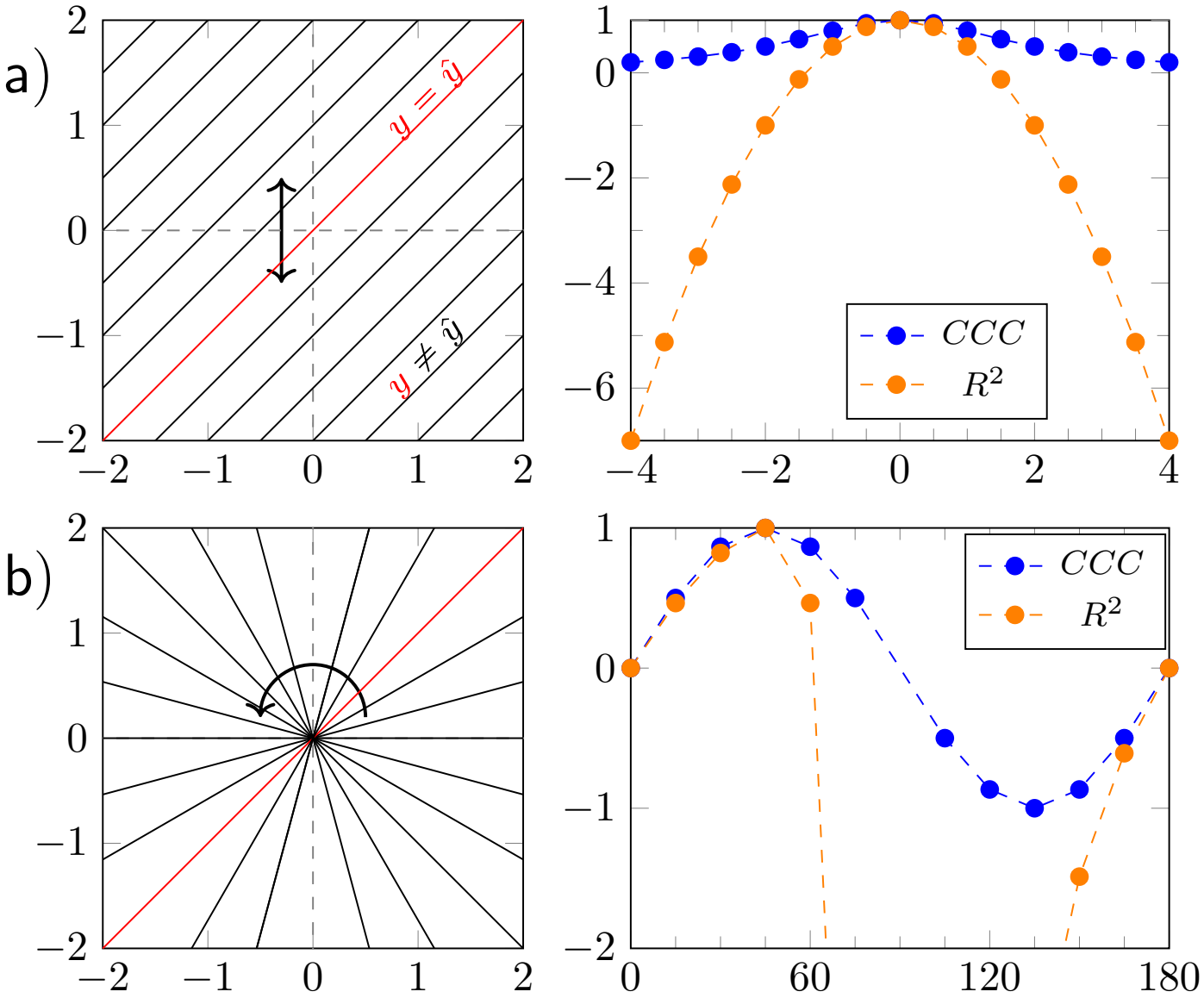


$\Rightarrow$

PLS2, temperature



Calculation of Q^2_{LOO} is straightforward

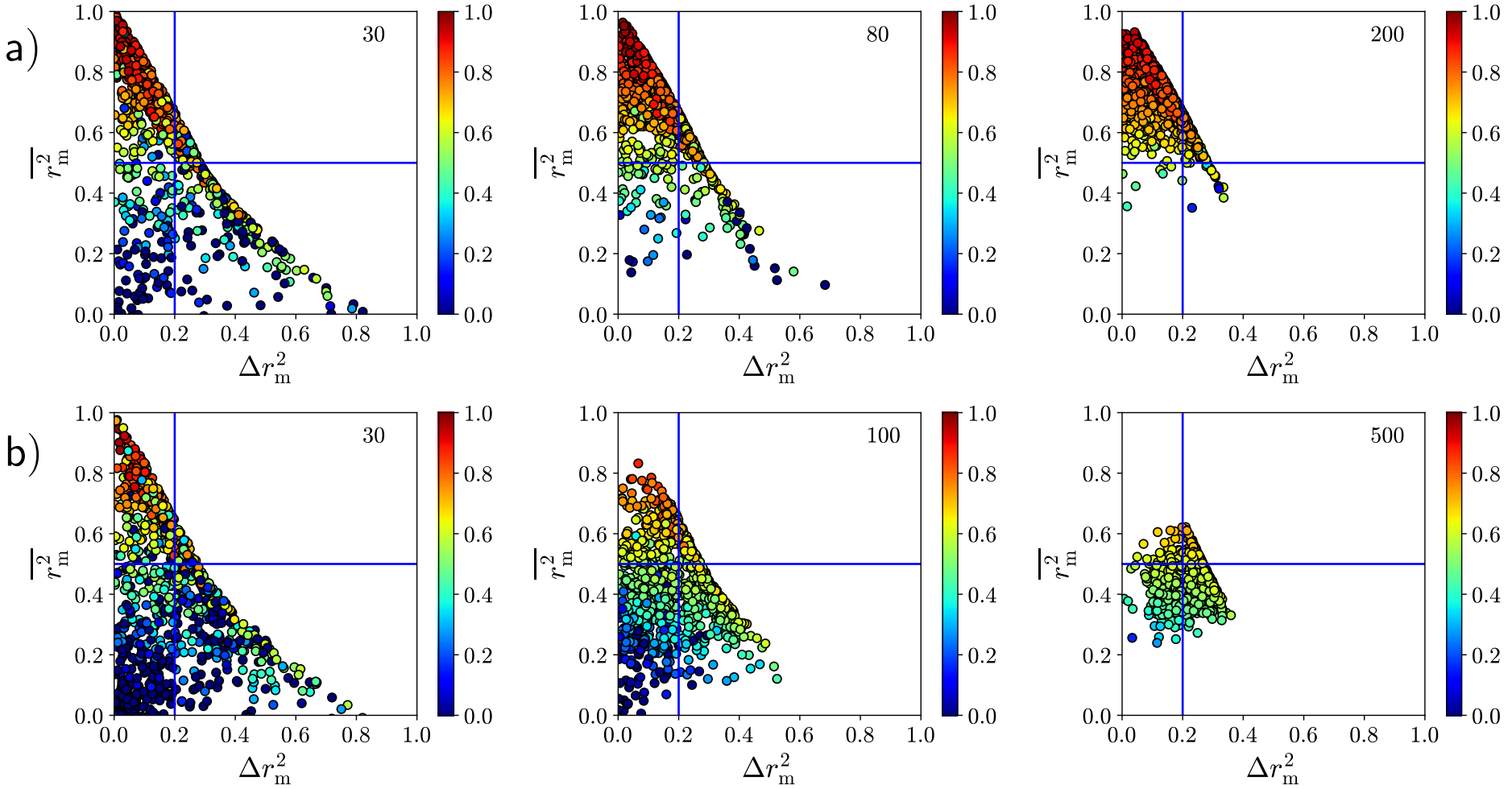

a)
$y = \hat{y}$
$y \neq \hat{y}$
CCC
R^2
b)
CCC
R^2

Roy-Ojha-Kovács diagram (test set)

a) MLR, gramatica3

b) MLR, concrete

} models colored according to $Q^2_{F2}(R^2_{\text{test}})$



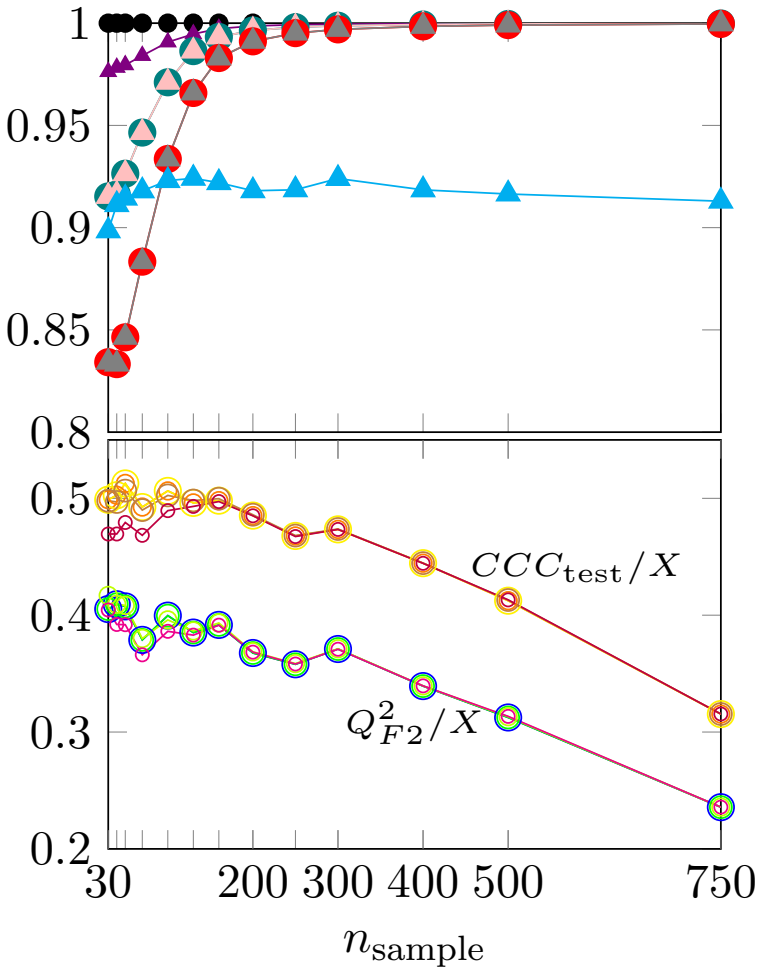
Rank correlations

$\underbrace{R^2, \mathrm{CCC}}$ internal (goodness-of-fit)

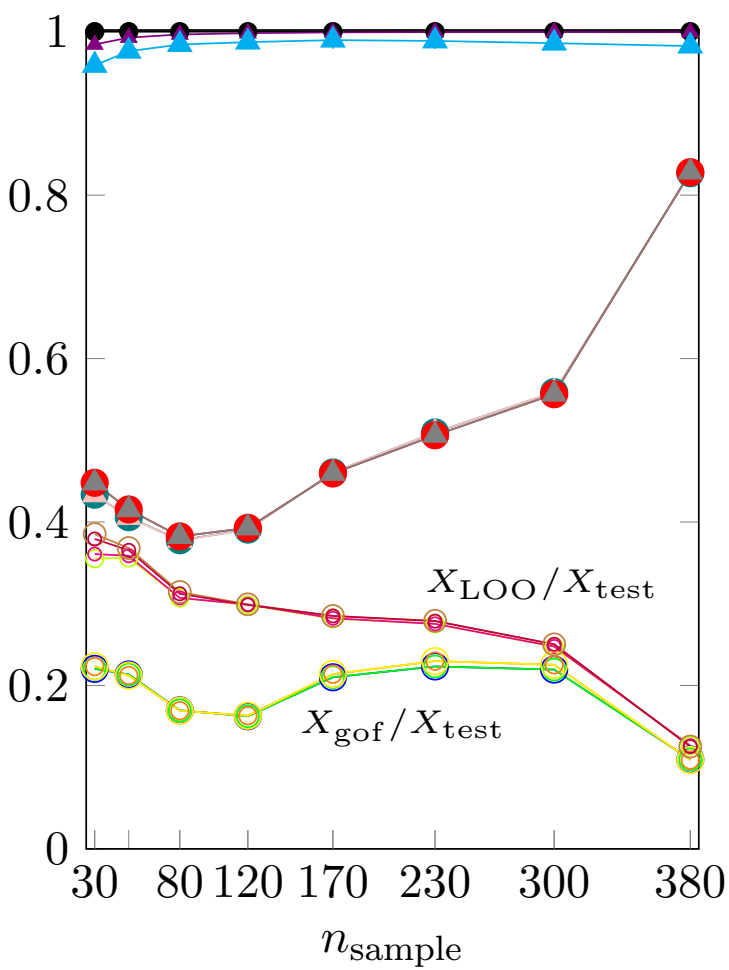
$\underbrace{Q^2_{\mathrm{LOO}}, \mathrm{CCC}_{\mathrm{LOO}}}$ internal (robustness)

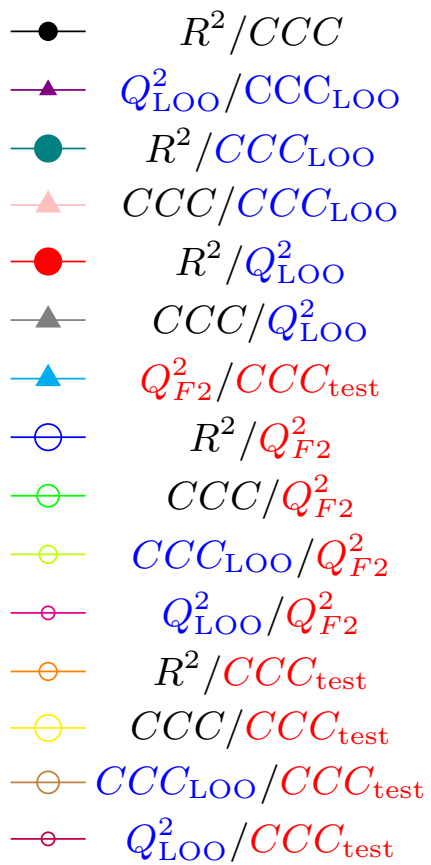
$\underbrace{Q^2_{F2}, \mathrm{CCC}_{\mathrm{test}}}$ external (predictivity)

MLR, concrete



PLS2, temperature mean





Summary

- model validation: goodness of fit (R^2), robustness (Q^2_{LOO}), predictivity (Q^2_{F2})
- sample size dependence of R^2, RMSE show anomaly
- scrambling can be used instead of randomization
- Q^2_{LOO} is sufficient to use instead of LMO variants
- model predictivity is categorized on Roy-Ojha-Kovács diagrams
- internal and external schemes have different information content for all sample sizes
- Q^2_{F1-3} for external validation show different sensitivity for different sample size
- rdCV redesigned for PLS2

We thank Prof. Imre Salma for providing us the air pollution dataset.

Co-workers: Ramóna Kiss, Dániel Kovács, Gergely Tóth