Further experience with a novel randomization test for PLS component selection

Klaas Faber

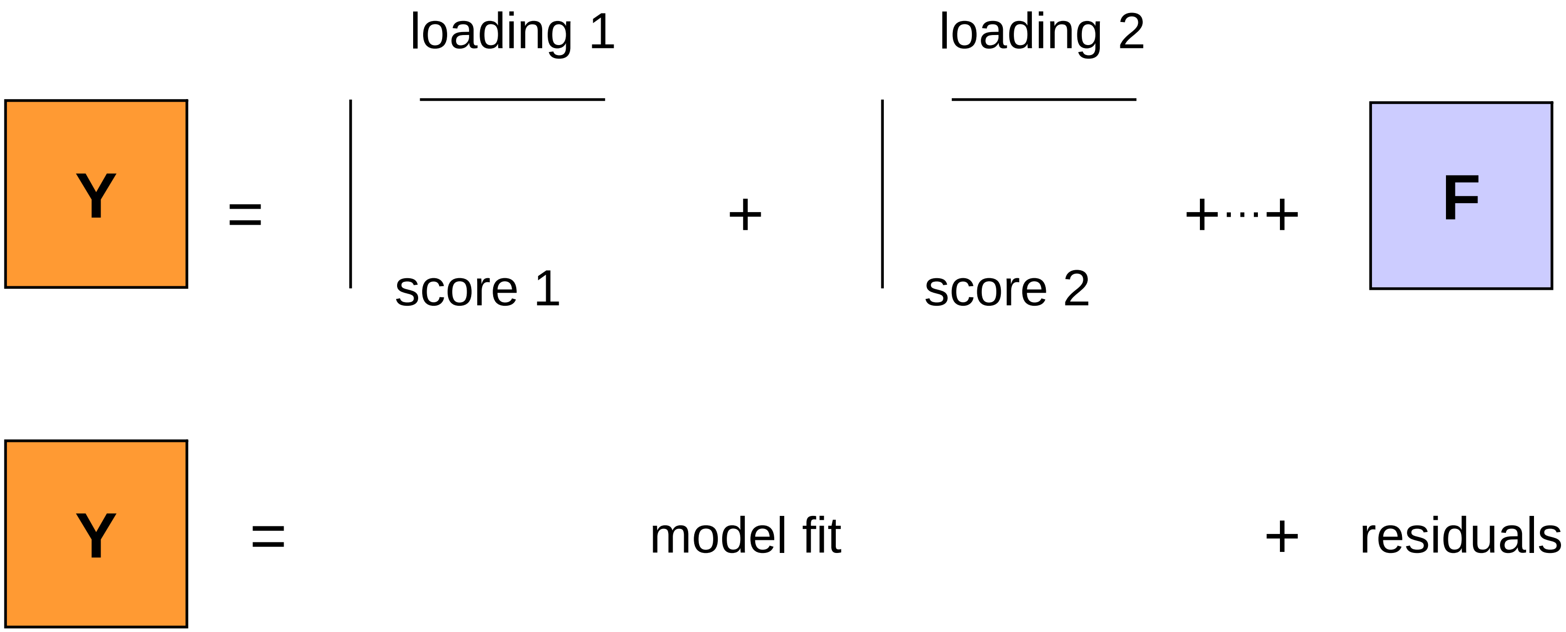
Chemometry Consultancy

www.chemometry.com

Outline

1. An evergreen problem 2
2. The proposed solution 15
3. Example data set 18
4. Concluding remarks 21
5. Acknowledgments 22

1. An evergreen problem



Y = loading 1 score 1 + loading 2 score 2 +···+ F

Y = model fit + residuals

property of interest

When to stop adding PLS components???

State of the art of commercial software

Tony Davies, Analytical computing survey
Spectroscopy Europe, **16** (2004) 26-27

“Back in 1998 more advanced chemometric tools were being made available as standard in spectrometer control packages. This had, however, raised fears that the **inherent dangers of over-fitting data were not being sufficiently addressed** in order to help inexperienced spectroscopists handle the additional computing power that was becoming available. I must admit that the work of my co-column Editor in pushing for “Good Chemometrics Practice” has hopefully raised awareness in the community of the potential pitfalls in using these packages without due consideration, but I personally have not been aware of **clear unambiguous automated warnings** starting to appear when data was being overfitted.”

Current approach to component selection: validation

Comparison of model predictions with known reference values of validation objects for increasing number of PLS components:

$$\mathrm{RMSEP}=\sqrt{\frac{1}{I_{\mathrm{val}}}\sum_{i}(\hat{Y}_{i}-Y_{i,\mathrm{ref}})^{2}}$$

# validation objects

prediction residual

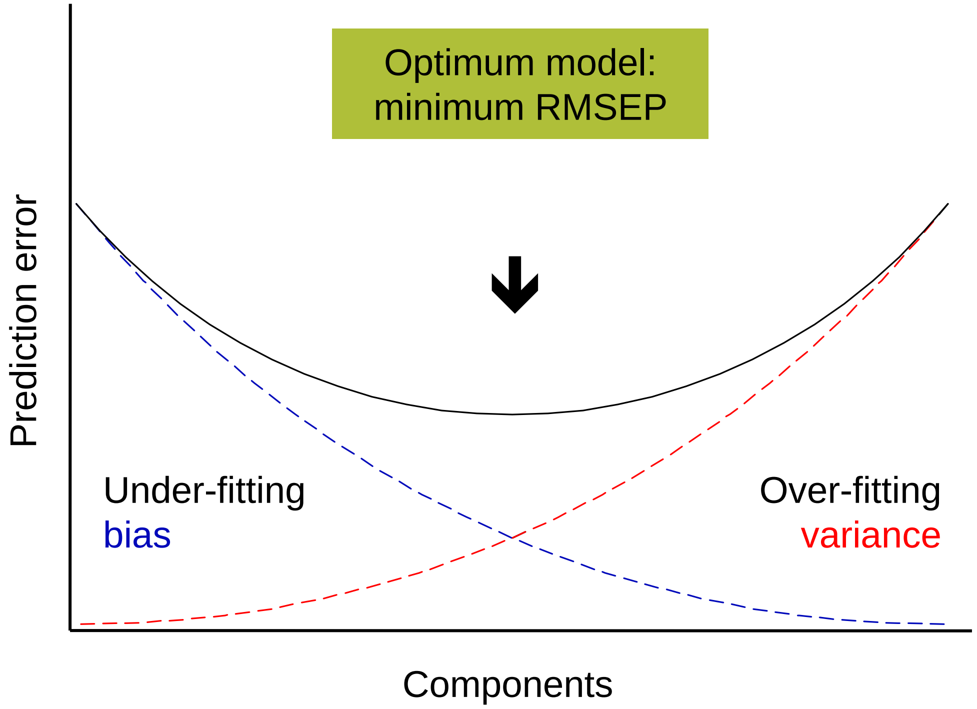
Ideally, the optimum number of components minimizes RMSEP (root mean squared error of prediction).

Schematic view of the variance-bias trade-off

With too few components, one misses predictive features

⇩

systematic error


Optimum model: minimum RMSEP

With too many components, one also fits noise in the data

⇩

random error

Common validation approaches for PLS

- External validation:
  - independent validation set is best (“test = best”), but it requires a lot of data and is therefore rather “wasteful”.

- Internal validation:
  - cross-validation is “economic”, but
    - it tends to select too many components (over-fit);
    - it can fail for small designed data sets, e.g. in sensory or quantitative structure activity relationship (QSAR) modeling;
    - it can fail when a model requires updating for new sources of variation and one needs to decide about additional PLS components (not further considered).
  - leverage correction is a quick-and-dirty alternative to cross-validation that is even more likely to over-fit the data.

Over-fitting tendency of cross-validation

- Well-documented in the statistics literature.

- Example of a chemometrics paper:
  - Q.-S. Xu and Y.-Z. Liang

    Monte Carlo cross validation

    Chemometrics and Intelligent Laboratory Systems, **56** (2001) 1-11

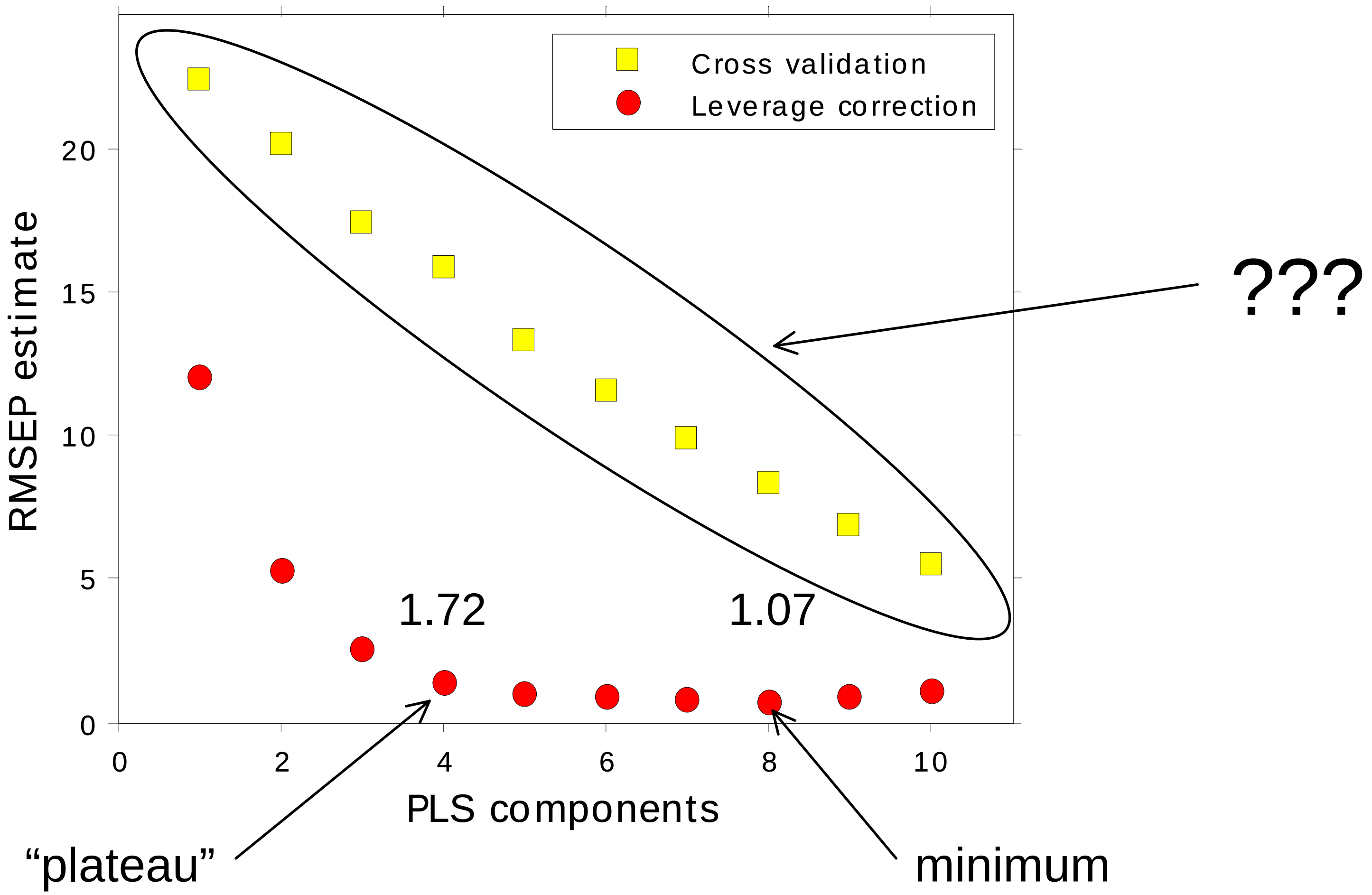
Cross-validation and designed data

- Cross-validation is a re-sampling technique like the jack-knife and the bootstrap.

- An underlying assumption for correct use is therefore that the calibration data at hand are sampled from a population, i.e. not designed.

- The consequences can be particularly grave for relatively small designed data sets.

QSAR application: hexapeptides synthesized according to a molecular design (16 objects and 18 X-variables)



???

“plateau”

minimum

Discussion

- Cross-validation does not give a hint about the optimum number of PLS components.
- Leverage correction yields a global minimum RMSEP for 8 components. However, it is doubtful whether 16 objects can span an 8-dimensional space. This model is likely to over-fit the data.
- A “soft” decision rule (“plateau”) suggests 4 components. However, the associated RMSEP estimate is considerably higher than the one obtained for the global minimum (1.72 *vs*. 1.07).
- How to decide???

Conclusion

Each validation approach has serious drawbacks…

Obvious research question

… is it possible to select PLS components without relying on validation?

Requirement

An approach is required that makes minimum assumptions about the data: it must be data-driven.

2. The proposed solution

A so-called randomization test is suitable here. It has been successfully applied to solve related problems in chemometrics.

The main result

One obtains a p-value for each component that is added to the PLS model.

More details and applications

- S. Wiklund, D. Nilsson, L. Eriksson, M. Sjöström, S. Wold and N.M. Faber

  A randomization test for PLS component selection

  Journal of Chemometrics, submitted.

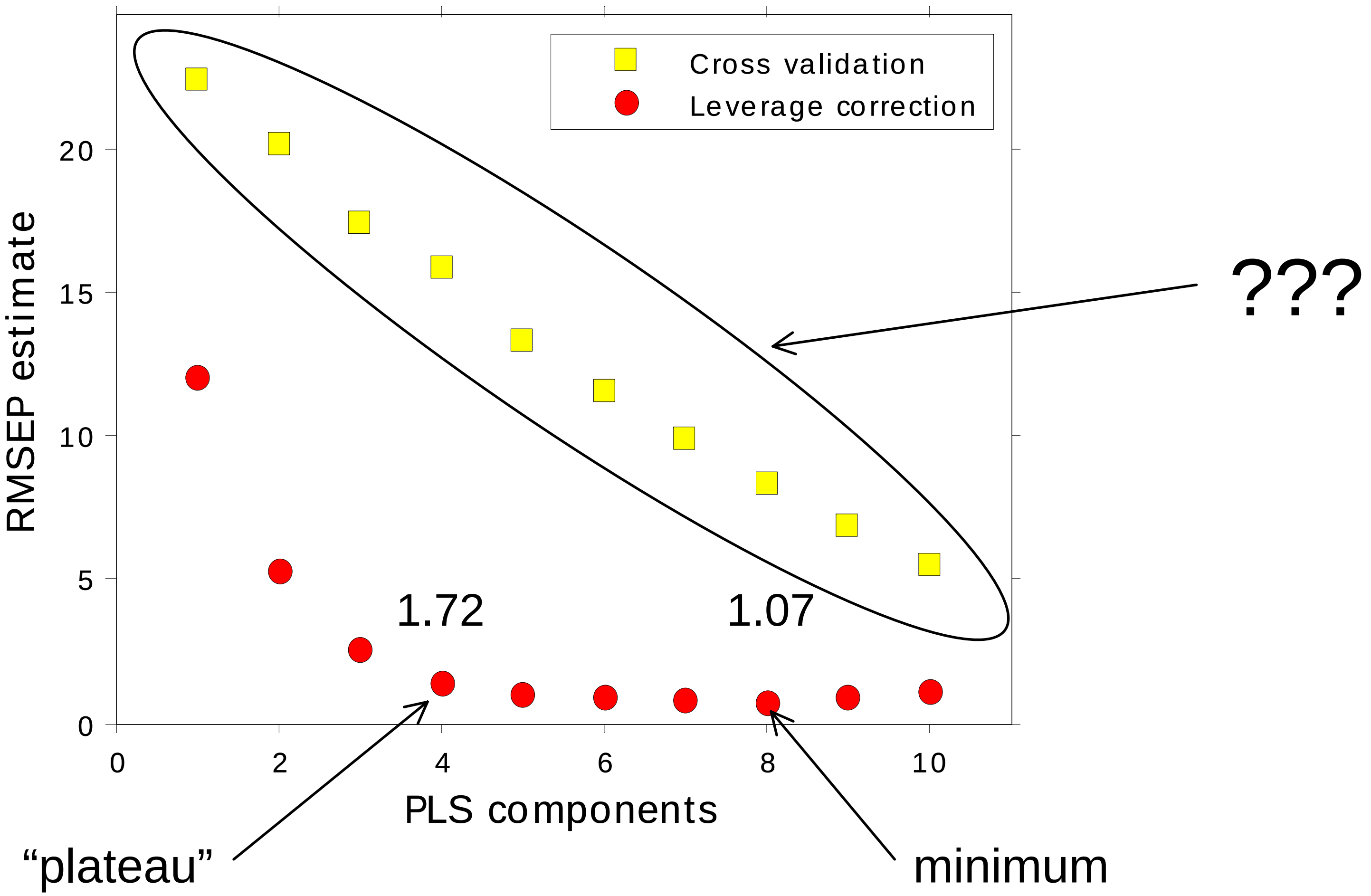
- N.M. Faber and R. Rajkó

  How to avoid over-fitting in multivariate calibration – the conventional validation approach and an alternative

  Spectroscopy Europe, submitted.

Request: nmf@chemometry.com

3. Example data set



Results for randomization test

Comp	p-value (%)	Comp	p-value (%)
1	32	4	2.2
2	0.36	5	58
3	5.1	6	82

Discussion

- Support for (at most) 4 PLS components. This coincides with the beginning of the “plateau” for leverage correction. The agreement with leverage correction was poor, however, for other data sets (not shown).

- Component 1 is not significant while higher-numbered components are. The natural behavior is that significant components are followed by the non-significant ones. This phenomenon has been observed for spectral data sets that require pre-treatment to remove irrelevant X-variation (not further investigated; historical data set).

4. Concluding remarks

- One has to accept as a fact that correct and/or adequate validation is not always feasible. In the context of PLS component selection, the proposed method intends to fill a gap.

- The randomization test can be used in stand-alone fashion, as shown, or in combination with e.g. cross-validation if the RMSEP curve does not exhibit a clear minimum and one has to resort to “soft” decision rules like “first local minimum” or “plateau”.

5. Acknowledgments

- Testing:
  - Susanne Wiklund, David Nilsson, Lennart Eriksson, Michael Sjöström and Svante Wold (Umeå University and Umetrics)
  - Jose Andrade (University of a Corunna)
  - Douglas Rutledge (Institut National Agronomique)
  - Randy Pell (Dow Chemical)
  - Lin Zhang (Pfizer)
  - Scott Ramos (Infometrix)
- Implementation:
  - Chris Brown (InLight Solutions)
  - Alejandro Olivieri (Universidad Nacional de Rosario)