Covariance penalty criteria for selecting dimensions of PLSR models

Illustration on NIRS data

matthieu.lesnoff@cirad.fr

ChemHouse Montpellier France, 17 November 2020

https://github.com/mlesnoff/rnirs

- Introduction
- DoFs
- llustration on a NIR dataset
- Appendix: some theoretical details

Introduction

Selection of the number of components (LV) in a PLSR model

- $a = 0, 1, \ldots, A$ components?

= Model selection

Many indicators

- **Scores, loadings, b-coefficients**
- **Prediction errors**
  - Cross-validation, Bootstrap
  - Permutations (Not presented here)
  - Covariance penalty criteria
- **Etc.**

Covariance penalty criteria

- $\in$ "**Information criteria**"

- Estimate of **prediction error (future)** from the training data set

- **Bias-variance** compromises

- Depend on the **model complexity** df

For OLS regression

- **a well-known covariance penalty** criteria is the Mallows' Cp

$$\mathrm{Cp} = \frac{\mathrm{SSR}}{n} + \frac{2}{n}\, p\, \hat{\sigma}^2$$ (p = number of variables)

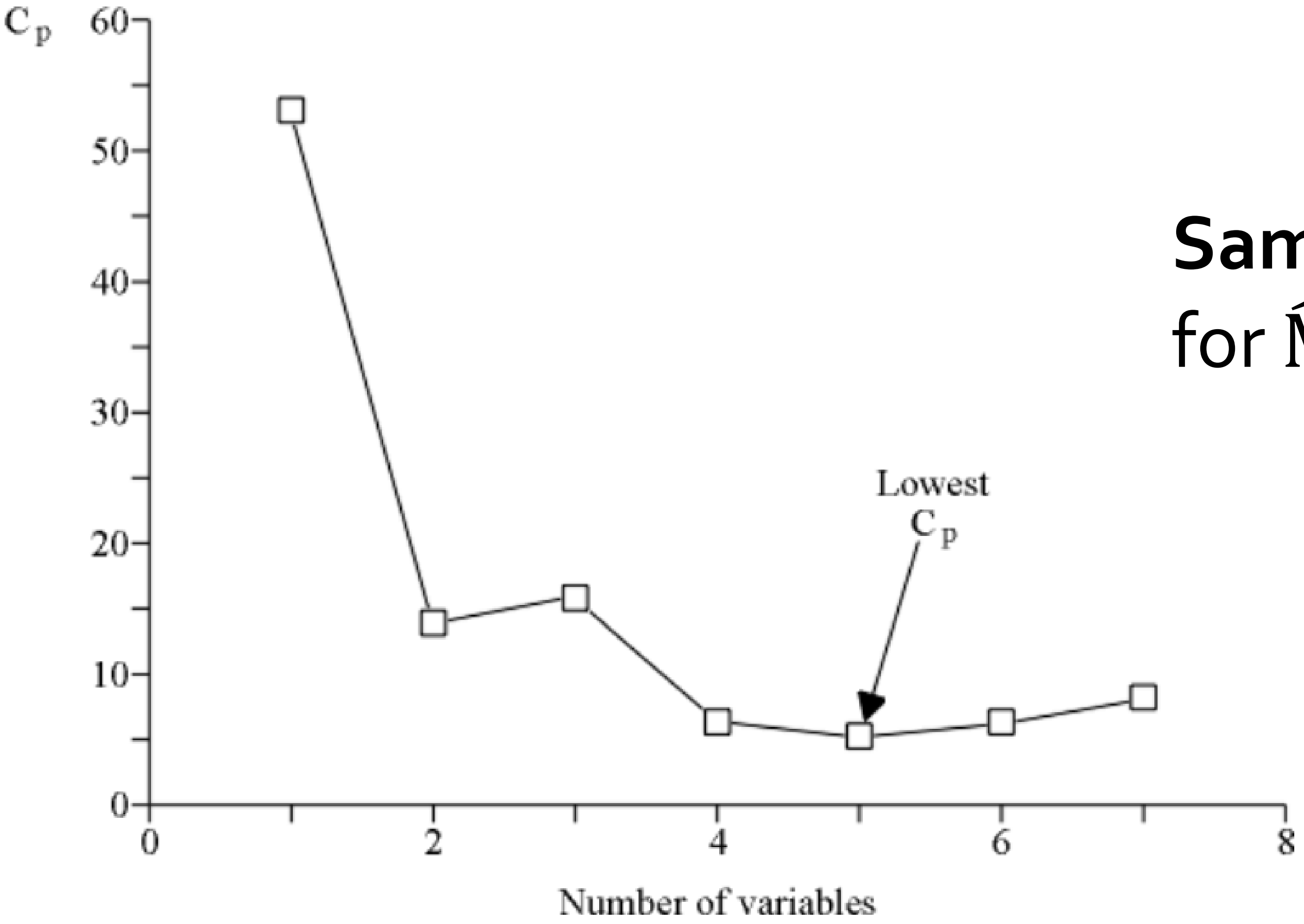
Mallows, C.L., 1973. Some Comments on Cp. *Technometrics* 15, 661–675.
https://doi.org/10.1080/00401706.1973.10489103

- $$\mathrm{Cp} = \frac{\mathrm{SSR}}{n} + \frac{2}{n}\,\underset{\uparrow}{p}\,\hat{\sigma}^2$$

Model complexity *df*

- $$\mathrm{Cp} = \frac{\mathrm{SSR}}{n} + \frac{2}{n}\boxed{p\ \hat{\sigma}^2}$$

Covariance penalty



Same interpretation as for $\widehat{\mathrm{MSEP}}_{\mathrm{CV,\ BOOT}}$

Zuccaro, C., 1992. Mallows’ Cp Statistic and Model Selection in Multiple Linear Regression. *International Journal of Market Research*. 34, 1–10. https://doi.org/10.1177/147078539203400204

For OLS (and some other models)

Cp is calculated from the training set without any simulation

$\Rightarrow$ very fast

Cp

- Estimate of a type of **(future) prediction error**, referred to as $\mathrm{Err}_{\mathrm{in}}$ in Hastie et al 2009
- Not exactly the same error estimated by CV or Bootstrap (Err) but both approaches often return close models

See Appendix for $\mathrm{Err}_{\mathrm{in}}$ and Err

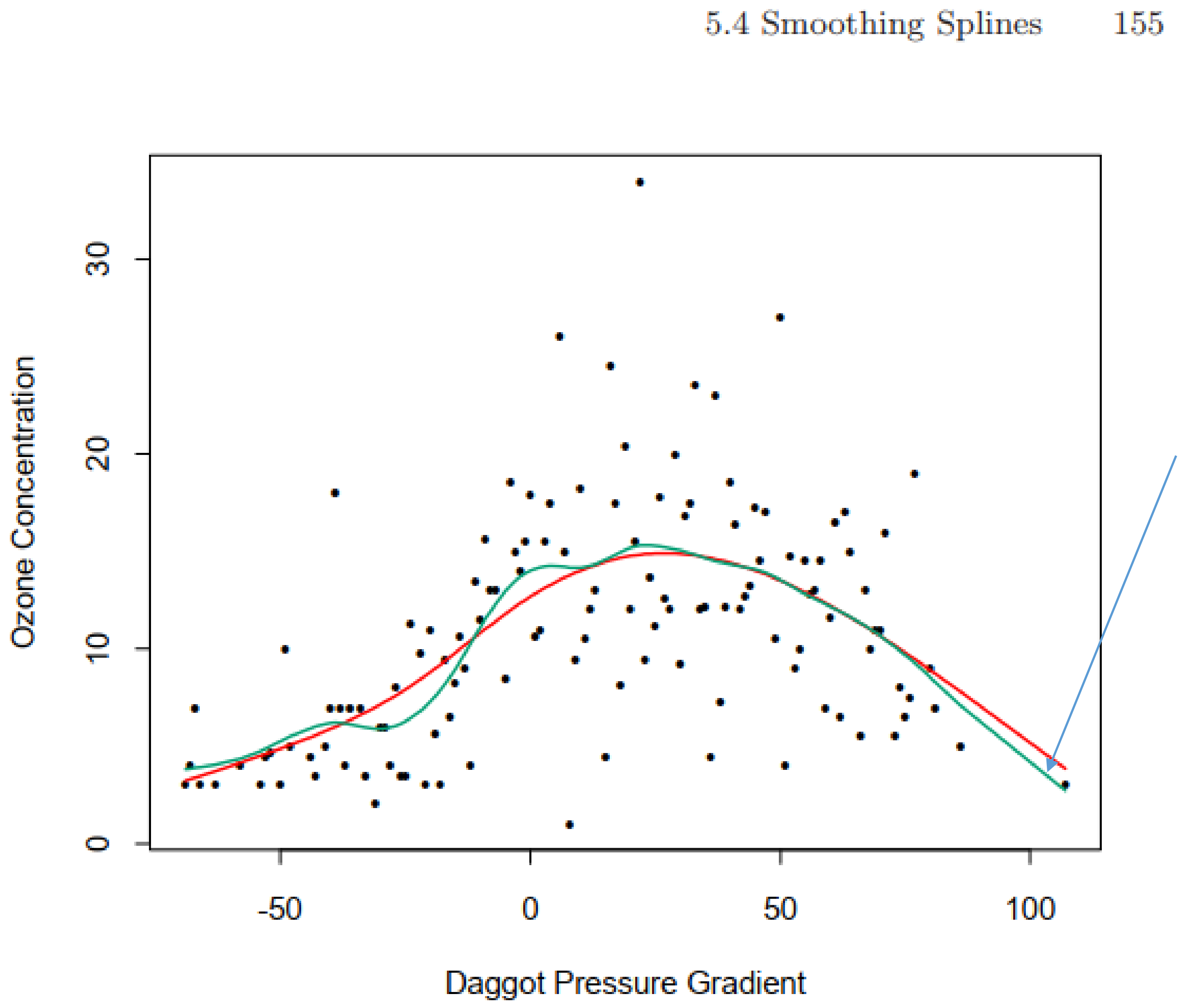
- Prediction errors = Bias-variance compromises

$$= \mathrm{Variance}_{\tau}\,(\hat{\boldsymbol{y}}) + \mathrm{Bias}_{\tau}\,(\hat{\boldsymbol{y}})^2 + \ldots$$

training

τ: training data set of size n

Hastie *et al* 2009 Fig 5.7

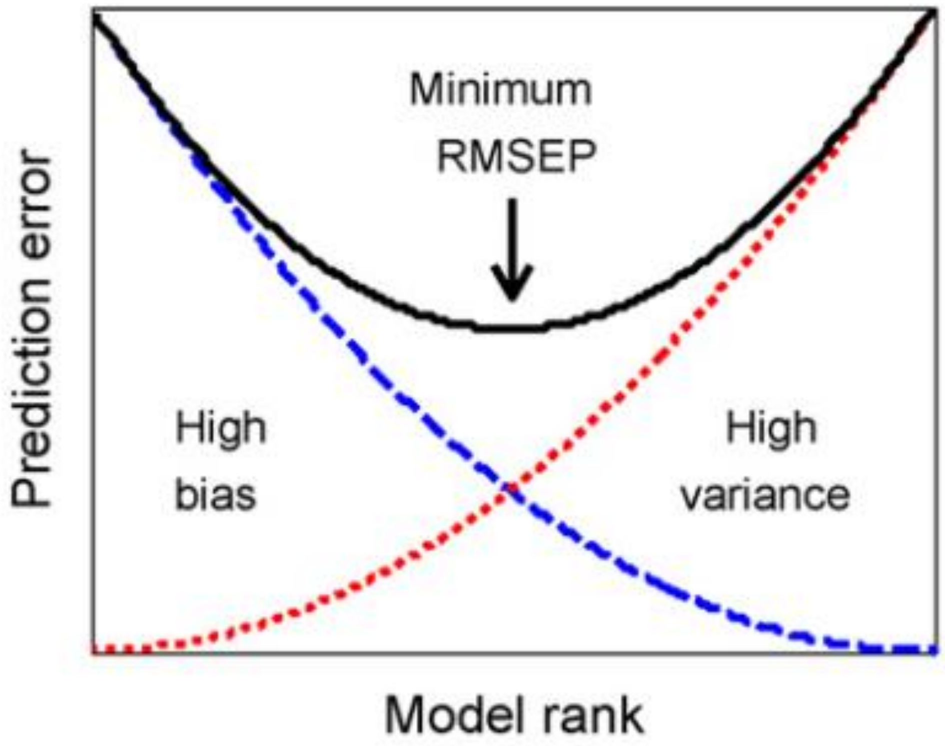


5.4 Smoothing Splines 155

More complex model

Less biased but more variable with τ

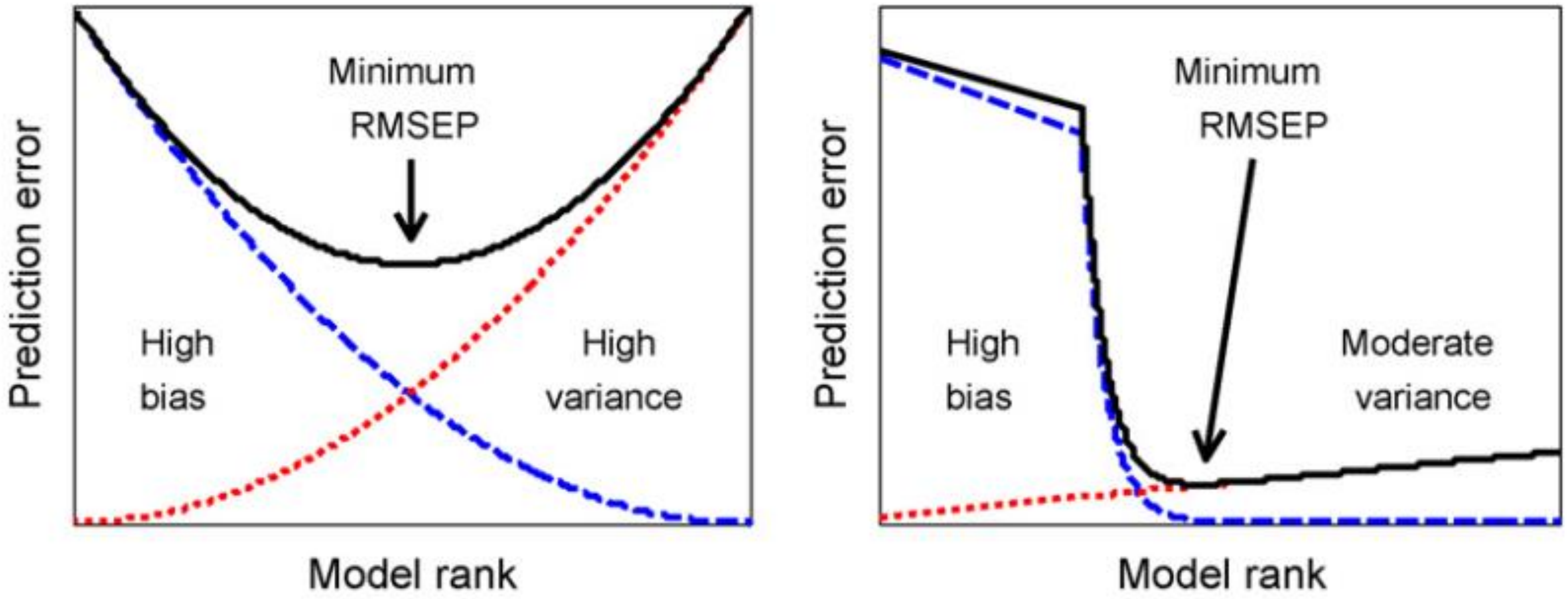
N.M. Faber, R. Rajkó / Analytica Chimica Acta 595 (2007) 98–106



Faber, N. (KLAAS) M., 1999. A closer look at the bias–variance trade-off in multivariate calibration. Journal of Chemometrics 13, 185–192. https://doi.org/10.1002/(SICI)1099-128X(199903/04)13:2<185::AID-CEM538>3.0.CO;2-N

Faber, N.M., Rajkó, R., 2007. How to avoid over-fitting in multivariate calibration—The conventional validation approach and an alternative. Analytica Chimica Acta, Papers presented at the 10th International Conference on Chemometrics in Analytical Chemistry 595, 98–106. https://doi.org/10.1016/j.aca.2007.05.030

N.M. Faber, R. Rajkó / Analytica Chimica Acta 595 (2007) 98–106



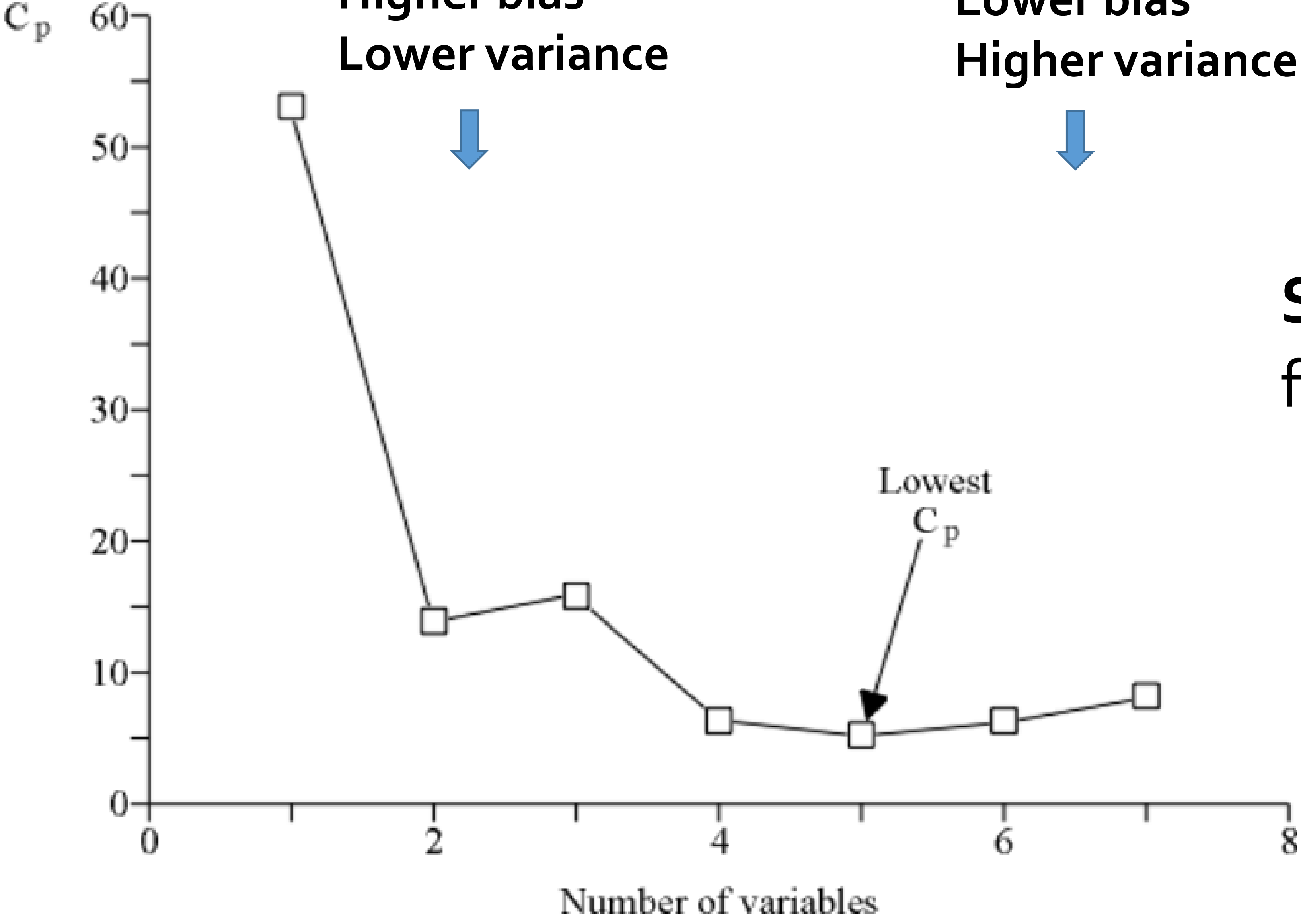
Faber, N. (KLAAS) M., 1999. A closer look at the bias–variance trade-off in multivariate calibration. Journal of Chemometrics 13, 185–192. https://doi.org/10.1002/(SICI)1099-128X(199903/04)13:2<185::AID-CEM538>3.0.CO;2-N

Faber, N.M., Rajkó, R., 2007. How to avoid over-fitting in multivariate calibration—The conventional validation approach and an alternative. Analytica Chimica Acta, Papers presented at the 10th International Conference on Chemometrics in Analytical Chemistry 595, 98–106. https://doi.org/10.1016/j.aca.2007.05.030

$$Var(\overline{\boldsymbol{x}}) = Var(\hat{\mu}) = \frac{\sigma^2}{n}$$

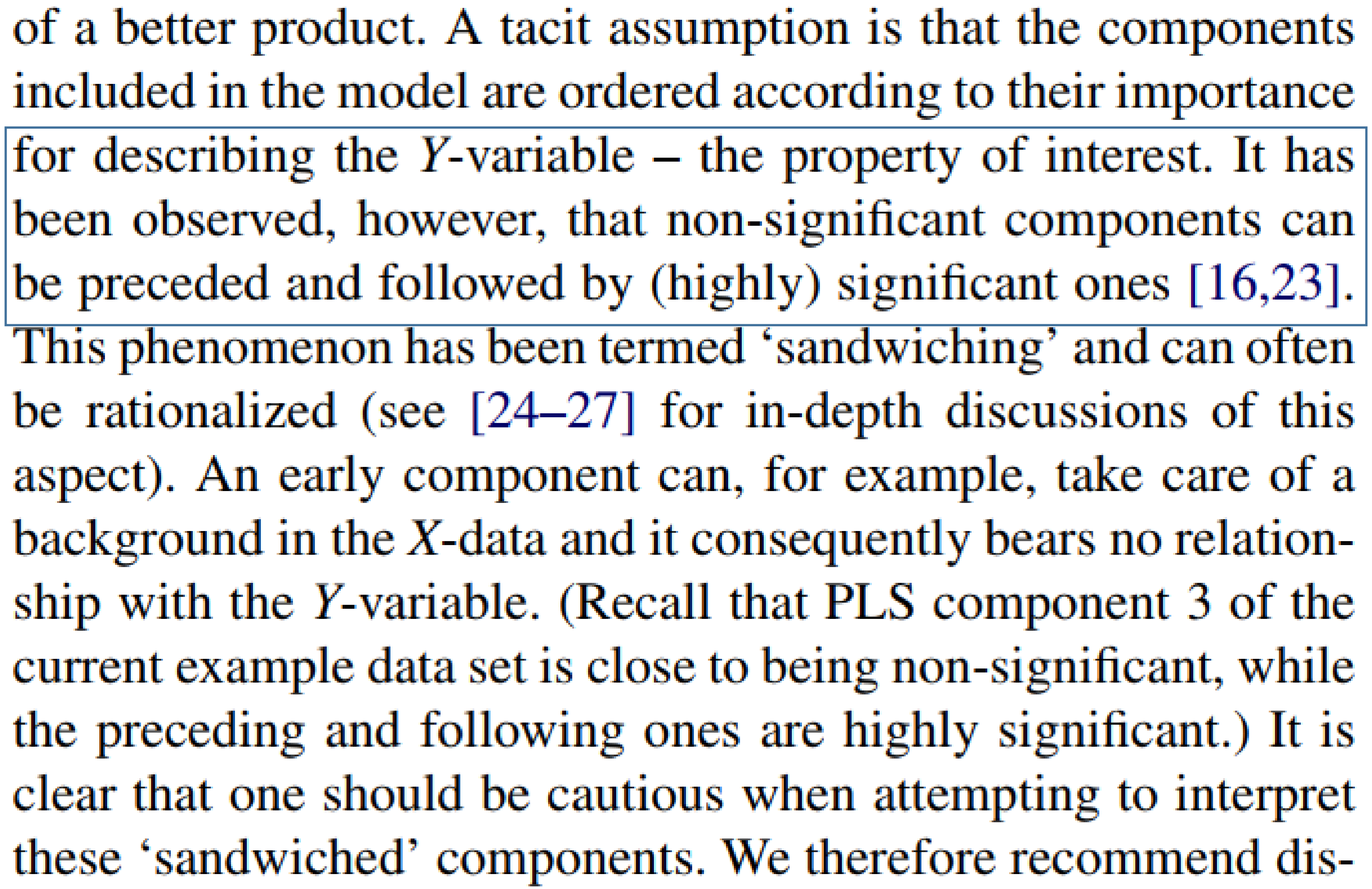
Higher bias
Lower variance

Lower bias
Higher variance



Same as for $\widehat{\mathrm{MSEP}}_{\mathrm{CV,\ BOOT}}$

Faber & Rajkó 2007



of a better product. A tacit assumption is that the components included in the model are ordered according to their importance for describing the *Y*-variable – the property of interest. It has been observed, however, that non-significant components can be preceded and followed by (highly) significant ones [16,23]. This phenomenon has been termed 'sandwiching' and can often be rationalized (see [24–27] for in-depth discussions of this aspect). An early component can, for example, take care of a background in the *X*-data and it consequently bears no relationship with the *Y*-variable. (Recall that PLS component 3 of the current example data set is close to being non-significant, while the preceding and following ones are highly significant.) It is clear that one should be cautious when attempting to interpret these 'sandwiched' components. We therefore recommend dis-

Remark
"Sandwiching"
effect in PLSR

For models with Gaussian errors

- Cp is equivalent to the **Akaike Criterion** (AIC)

$$\text{AIC} = n\log(\text{SSR}) + 2\,(p+1)$$ (after removing non-useful constants)

About AIC

Hurvich, C.M., Tsai, C.-L., 1989. Regression and Time Series Model Selection in Small Samples. *Biometrika* 76, 297. https://doi.org/10.2307/2336663

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed. Springer, New York, NY, USA.

Variants of Cp consist in increasing the penalty coefficient “2” for preventing overfitting → Ex: BIC penalty

- $\mathrm{Cp} = \frac{\mathrm{SSR}}{n} \quad + \frac{2}{n} p\, \hat{\sigma}^2$

- $\mathrm{Cp(BIC)} \ = \frac{\mathrm{SSR}}{n} \quad + \quad \frac{\log(n)}{n} p\, \hat{\sigma}^2$

$\log(100) = 4.6$

$\log(1000) = 6.9$

See discussion in:

Eubank, R.L., 1999. Nonparametric Regression and Spline Smoothing, 2nd ed, Statistics: Textbooks and Monographs. *Marcel Dekker, Inc.*, New York, USA.

Bias correction for Cp $\rightarrow$ For “small” n

- $\mathrm{Cp(AICc)} = \frac{\mathrm{SSR}}{n} + \frac{2}{n} p\, \hat{\sigma}^2 \frac{n}{n-p-1}$

Hurvich, C.M., Tsai, C.-L., 1989. Regression and Time Series Model Selection in Small Samples. *Biometrika* 76, 297. https://doi.org/10.2307/2336663

Generalization to other models than OLS regression

- General form $$\mathrm{Cp} = \frac{\mathrm{SSR}}{n} + \frac{2}{n}\, df\, \hat{\sigma}^2$$

Need to be calculated

- df is easier to calculate for **linear smoothers** (OLS, Splines, Ridge, etc.)
  - $\hat{\boldsymbol{y}} = \boldsymbol{S}\,\boldsymbol{y}$ $\boldsymbol{y}$ not involved in $\boldsymbol{S}$
  - Ex: OLS regression $\hat{\boldsymbol{y}} = \boldsymbol{S}\,\boldsymbol{y}$ ($\boldsymbol{S}$: hat matrix)

$$= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\boldsymbol{y}$$

$$df = tr(2\boldsymbol{S} - \boldsymbol{SS}') = tr(\boldsymbol{S}) = p$$

- PLSR is a non linear smoother …

- PLSR $=$ OLS regression on a scores
  - But $\hat{\boldsymbol{y}} = \boldsymbol{S}_{\boldsymbol{a,y}}\,\boldsymbol{y}$ $\boldsymbol{y}$ is involved in $\boldsymbol{S}$
  - Consequence: $df > a + 1$

("+ 1" is for the intercept)

- Same difficulty in PCA $\mathrm{vec}(\widehat{\boldsymbol{X}}) = \boldsymbol{S}_{\boldsymbol{a},\boldsymbol{X}}\,\mathrm{vec}(\boldsymbol{X})$

- In PCR $df = a + 1$

The idea $df_{\mathrm{PLSR}} > a + 1$ is not new in chemometrics, for instance …

- **Martens, H. and Naes, T. (1989)**. Multivariate Calibration. *Wiley*, New York.

- **Frank, Ildiko E., Friedman, J.H., 1993**. A Statistical View of Some Chemometrics Regression Tools. *Technometrics* 35, 109–135. https://doi.org/10.1080/00401706.1993.10485033

- **Voet, H. van der, 1999**. Pseudo-degrees of freedom for complex predictive models: the example of partial least squares. *Journal of Chemometrics* 13, 195–208. https://doi.org/10.1002/(SICI)1099-128X(199905/08)13:3/4<195::AID-CEM540>3.0.CO;2-L

- **Denham, M.C., 2000**. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. *Journal of Chemometrics* 14, 351–361. https://doi.org/10.1002/1099-128X(200007/08)14:4<351::AID-CEM598>3.0.CO;2-Q

- **Krämer, N., Braun, M.L., 2007**. Kernelizing PLS, degrees of freedom, and efficient model selection, in: Proceedings of the 24th *International Conference on Machine Learning, ICML* '07. Association for Computing Machinery, New York, NY, USA, pp. 441–448. https://doi.org/10.1145/1273496.1273552

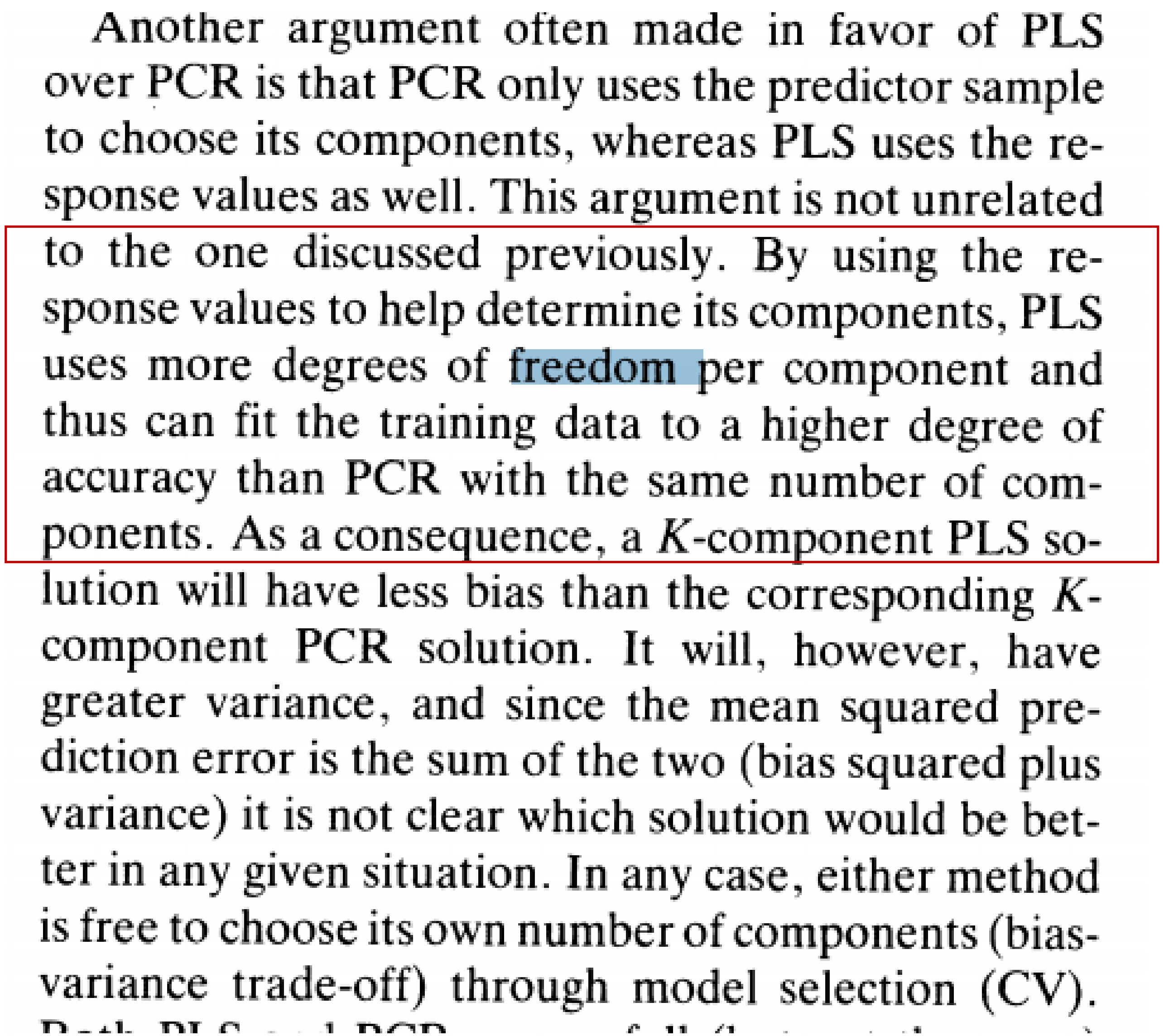
- **Krämer, N., Sugiyama, M., 2011**. The Degrees of Freedom of Partial Least Squares Regression. *Journal of the American Statistical Association* 106, 697–705. https://doi.org/10.1198/jasa.2011.tm10107

… But often forgotten

Example

- **Li, B., Morris, J., Martin, E.B., 2002**. Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* 64, 79–89. https://doi.org/10.1016/S0169-7439(02)00051-5

Authors used the naïve $df = a + 1$ in the AIC criterion



Another argument often made in favor of PLS over PCR is that PCR only uses the predictor sample to choose its components, whereas PLS uses the response values as well. This argument is not unrelated to the one discussed previously. By using the response values to help determine its components, PLS uses more degrees of freedom per component and thus can fit the training data to a higher degree of accuracy than PCR with the same number of components. As a consequence, a K-component PLS solution will have less bias than the corresponding K-component PCR solution. It will, however, have greater variance, and since the mean squared prediction error is the sum of the two (bias squared plus variance) it is not clear which solution would be better in any given situation. In any case, either method is free to choose its own number of components (bias-variance trade-off) through model selection (CV).

Franck & Friedman Technometrics 1993

Krämer & Sugiyama Jasa 2011

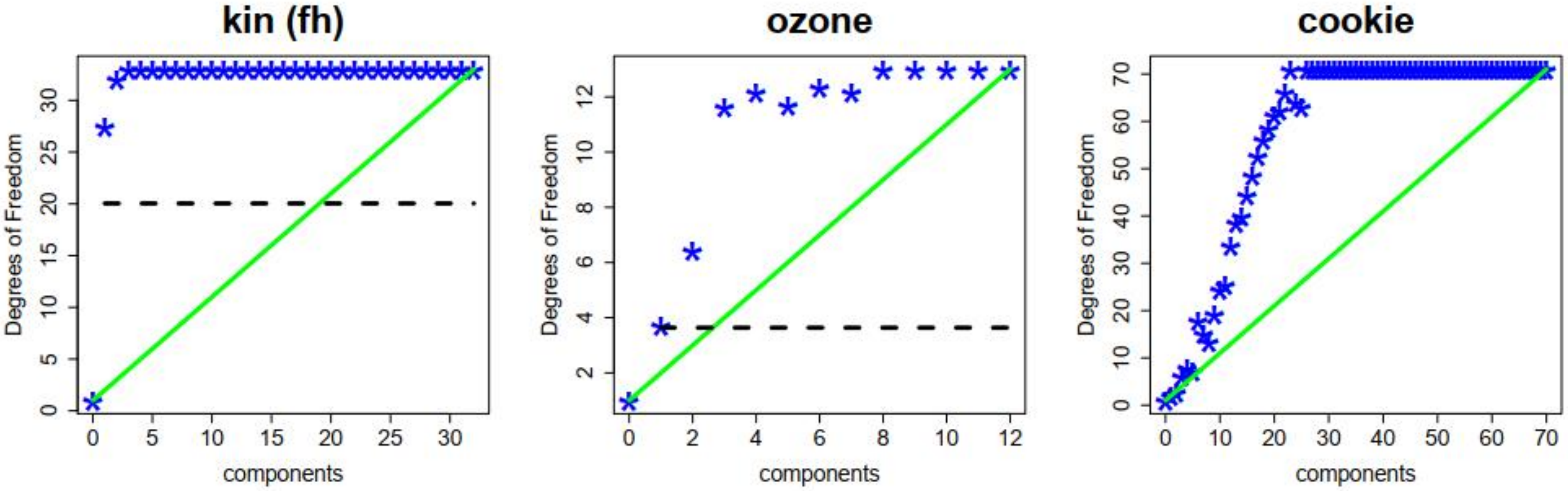


Figure 1: Estimated Degrees of Freedom (stars) for the three benchmark data sets. The solid line displays the naive estimate $\text{DoF}(m) = m + 1$. If the assumption of theorem 3 is fulfilled, we also display the lower bound on the Degrees of Freedom for 1 component (dashed horizontal line).

Krämer & Sugiyama Jasa 2011

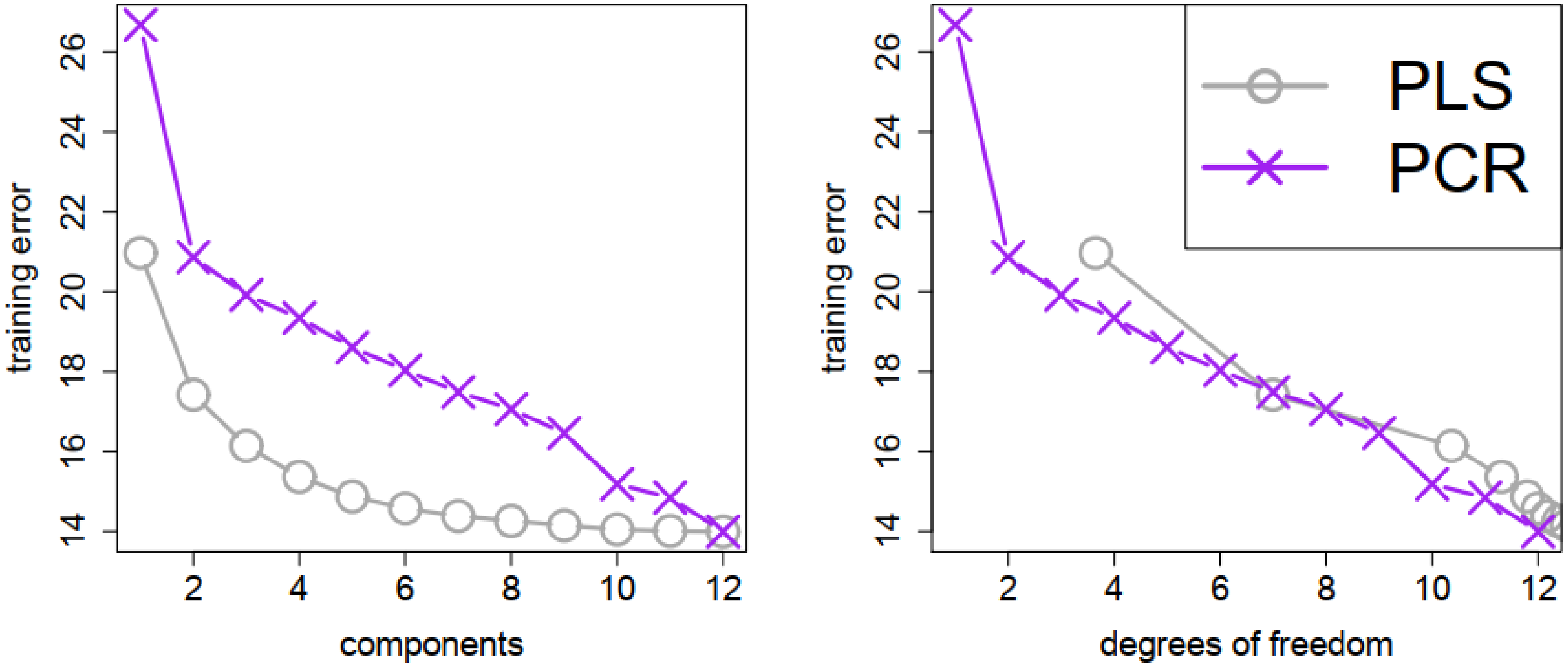


Figure 3: Training error of PLSR and PCR. Left: Training error as a function of the number of components. Right: Training error as a function of the Degrees of Freedom.

Can we use Cp (AIC, etc.) for PLSR?

- yes if one uses the **relevant model complexity** df

Model’s degree of freedom

- **Hastie, T., Tibshirani, R.J., 1990**. Generalized Additive Models, Monographs on statistics and applied probablity. *Chapman and Hall/CRC*, New York, USA.

- **Ye, J., 1998**. On Measuring and Correcting the Effects of Data Mining and Model Selection. *Journal of the American Statistical Association* 93, 120–131. https://doi.org/10.1080/01621459.1998.10474094

- **Eubank, R.L., 1999**. Nonparametric Regression and Spline Smoothing, 2nd ed, Statistics: Textbooks and Monographs. *Marcel Dekker, Inc.*, New York, USA.

- **Efron, B., 2004**. The Estimation of Prediction Error. J*ournal of the American Statistical Association* 99, 619–632. https://doi.org/10.1198/016214504000000692

- **Zou, H., Hastie, T., Tibshirani, R., 2007**. On the “degrees of freedom” of the lasso. *The Annals of Statistics* 35, 2173–2192. https://doi.org/10.1214/009053607000000127

- **Hastie, T., Tibshirani, R., Friedman, J., 2009**. The elements of statistical learning: data mining, inference, and prediction, 2nd ed. *Springer*, New York.

- **Hastie, T., Tibshirani, R., Wainwright, M., 2015**. Statistical Learning with Sparsity: The Lasso and Generalizations. *CRC Press*.

df

- Model complexity
- **Effective number of parameters** in the model

For any model (linear or not) with additive homoscedastic error

the **consensus** is to consider that

$$df \;= \sum_{i=1}^{n} Cov_{\boldsymbol{\varepsilon}}(y_i\,, \hat{y}_i)/\sigma^2$$

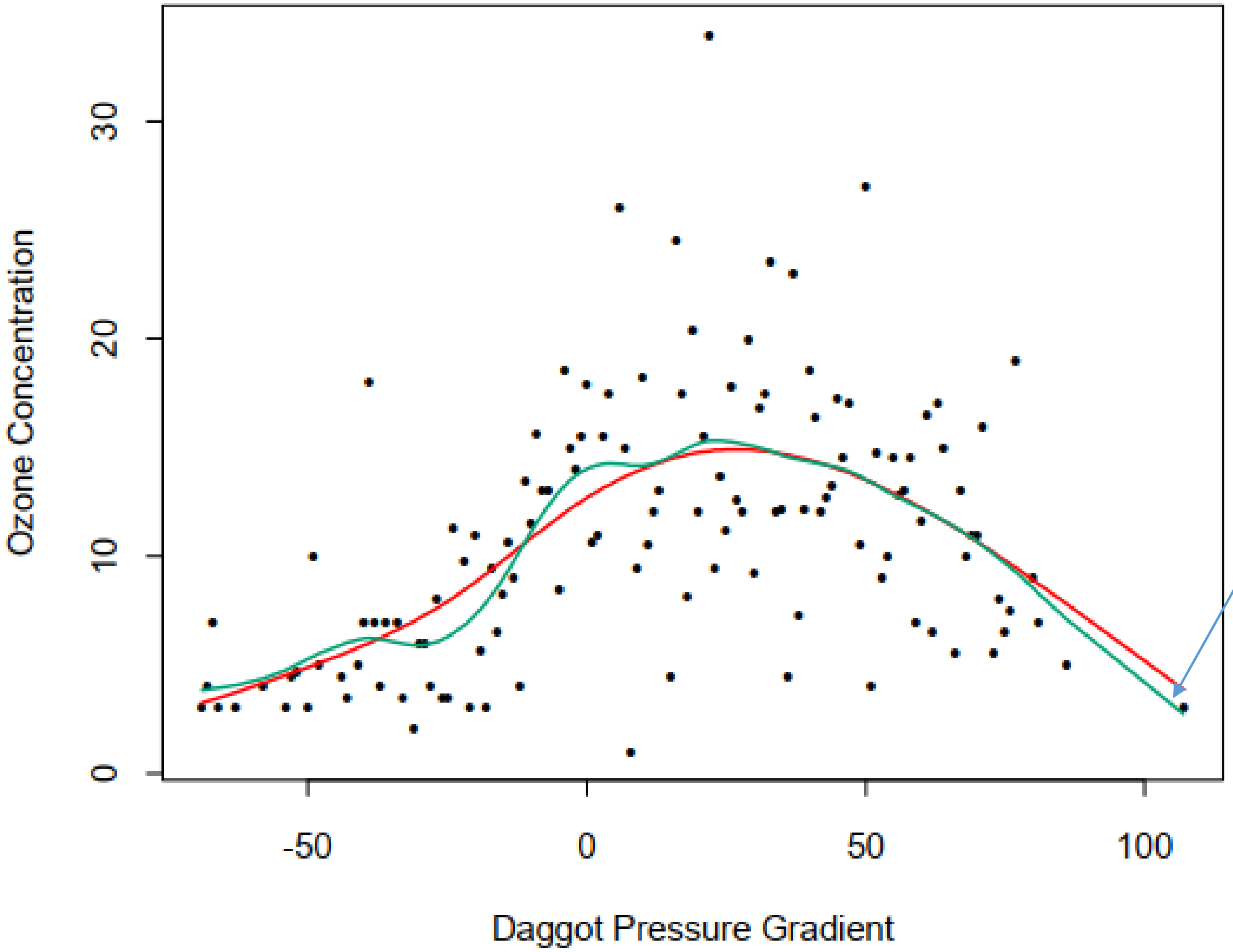
Variation of the training $\boldsymbol{y}$ ($\boldsymbol{X}$ fixed)

Hastie et al 2015 p.18

Self-influence that each response measurement has on its prediction

5.4 Smoothing Splines 155

Hastie *et al* 2009 Fig 5.7



More complex model

A change in y_i will have more impact

For OLS regression

- $Cov_{\boldsymbol{\varepsilon}}(yi, \hat{y}_i) = p\ \sigma^2/n$

$\Rightarrow \quad df\ = \sum_{i=1}^{n} Cov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i)/\sigma^2 = p$

$\Rightarrow$ This covariance definition of df is consistent with the historical definition for OLS regressions $df_{\text{OLS}}\ = p$

For Gaussian errors $\boldsymbol{\varepsilon}$ (Stein 1981, Efron 2004)

$$Cov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i) = \sigma^2 E_{\boldsymbol{\varepsilon}}\left(\frac{\partial \hat{y}_i}{\partial y_i}\right)$$

Stein, C.M., 1981. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics* 9, 1135–1151.

Leading to two usual forms for df

1) $df = \sum_{i=1}^{n} Cov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i)/\sigma^2$

2) $df = \sum_{i=1}^{n} E_{\boldsymbol{\varepsilon}}\left(\frac{\partial \hat{y}_i}{\partial y_i}\right)$

Gaussian error (Stein)

Generalized df
(**Ye Jasa 1998**)

Divergence div$_i$

Estimation of df

- Formal derivation (exact or approximation)
  - Either $Cov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i)$ or $\frac{\partial \hat{y}_i}{\partial y_i}$

- Monte Carlo simulations

Monte Carlo

1) $\widehat{df} = \sum_{i=1}^{n} \hat{C}ov_{\varepsilon}(y_i, \hat{y}_i)/\hat{\sigma}^2$

Parametric bootstrap
(e.g. Efron 2004)

2) $\widehat{df} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i}$ (SURE)

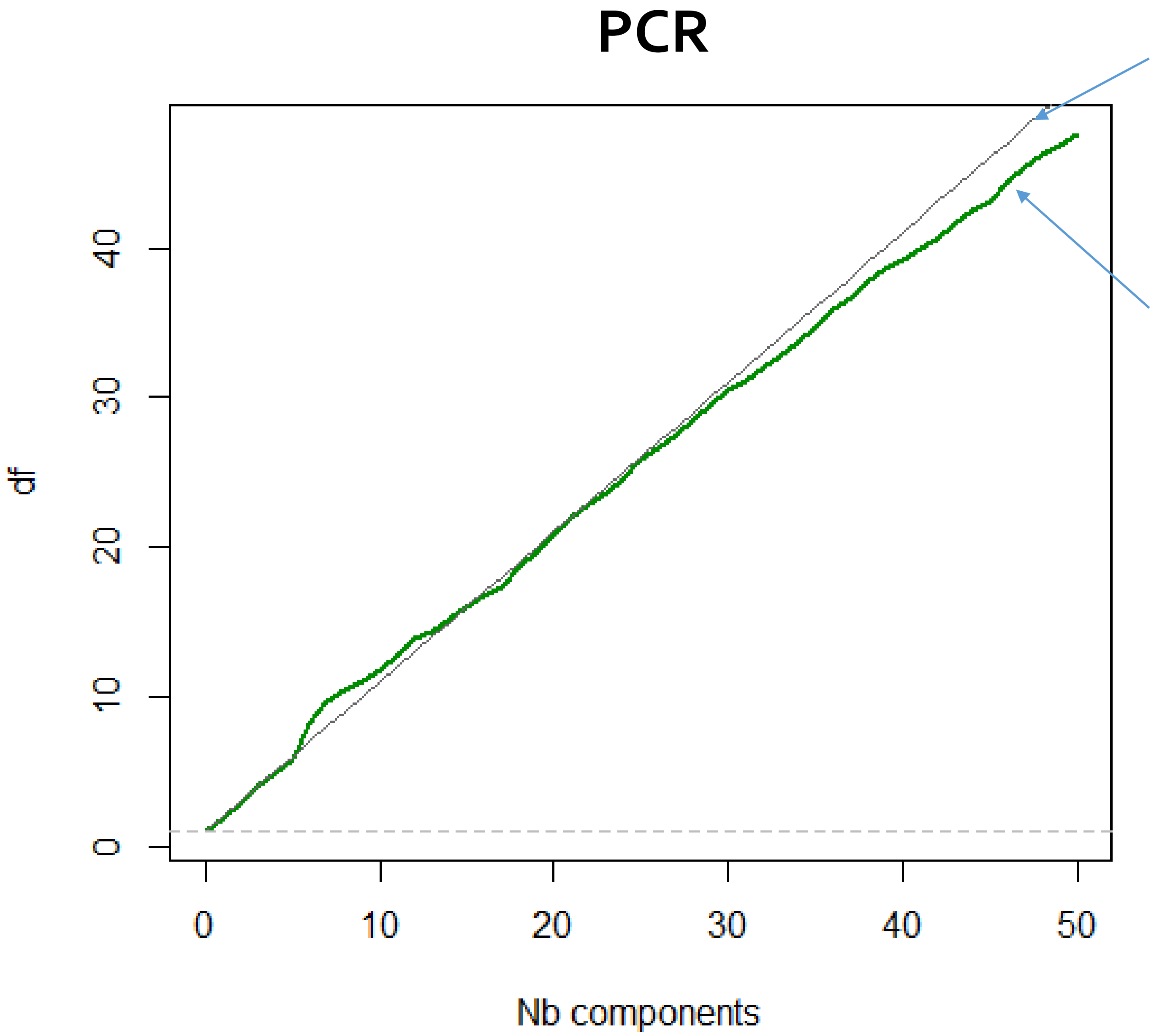
Stein Unbiased
Risk Estimation

Sensitivity analysis
(perturbations
e.g. Ye 1998)

Sorghum NIR data

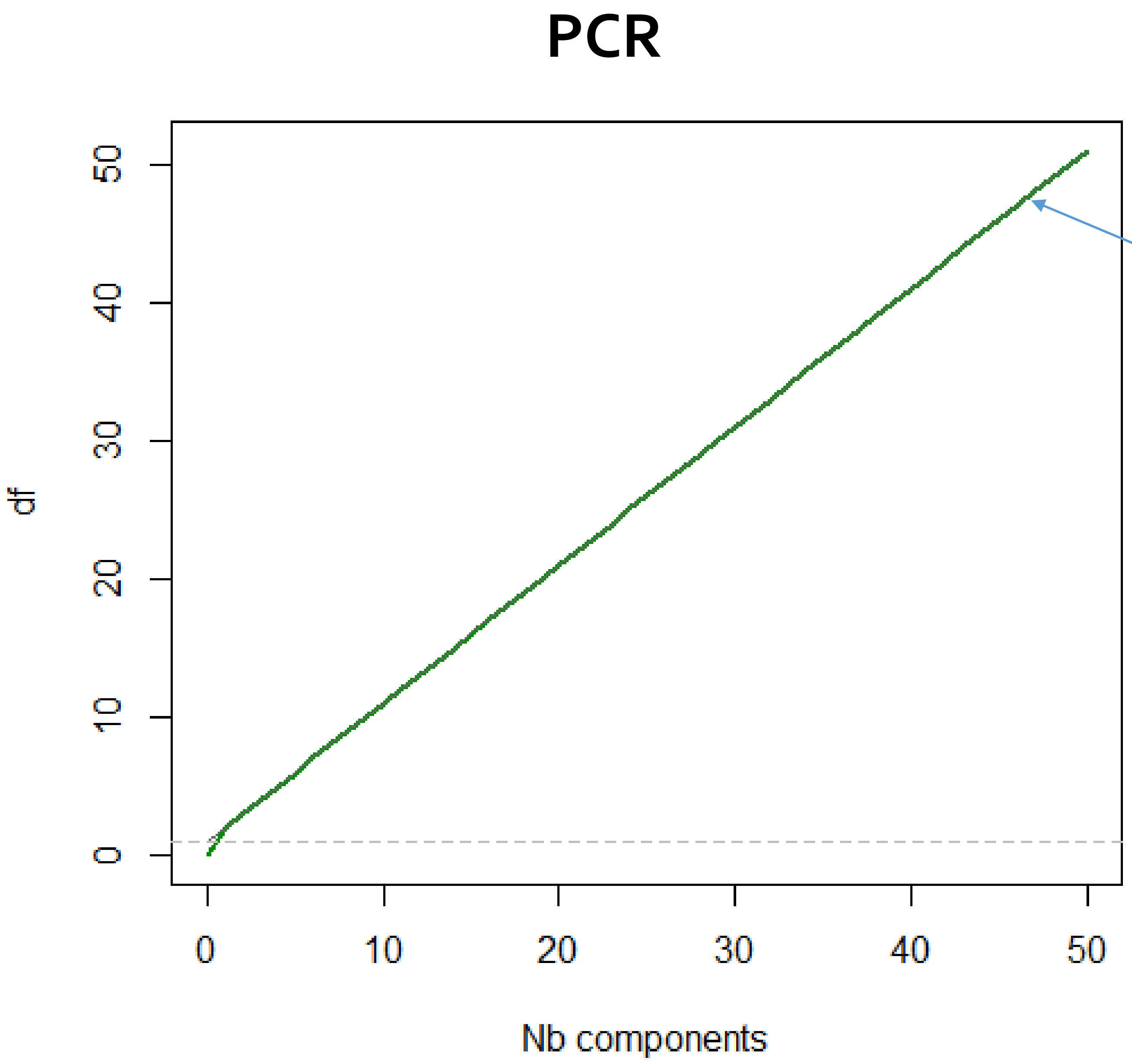
n = 1006 p = 700

% Crude fibers prediction



$df_{\text{naive}} = a + 1$

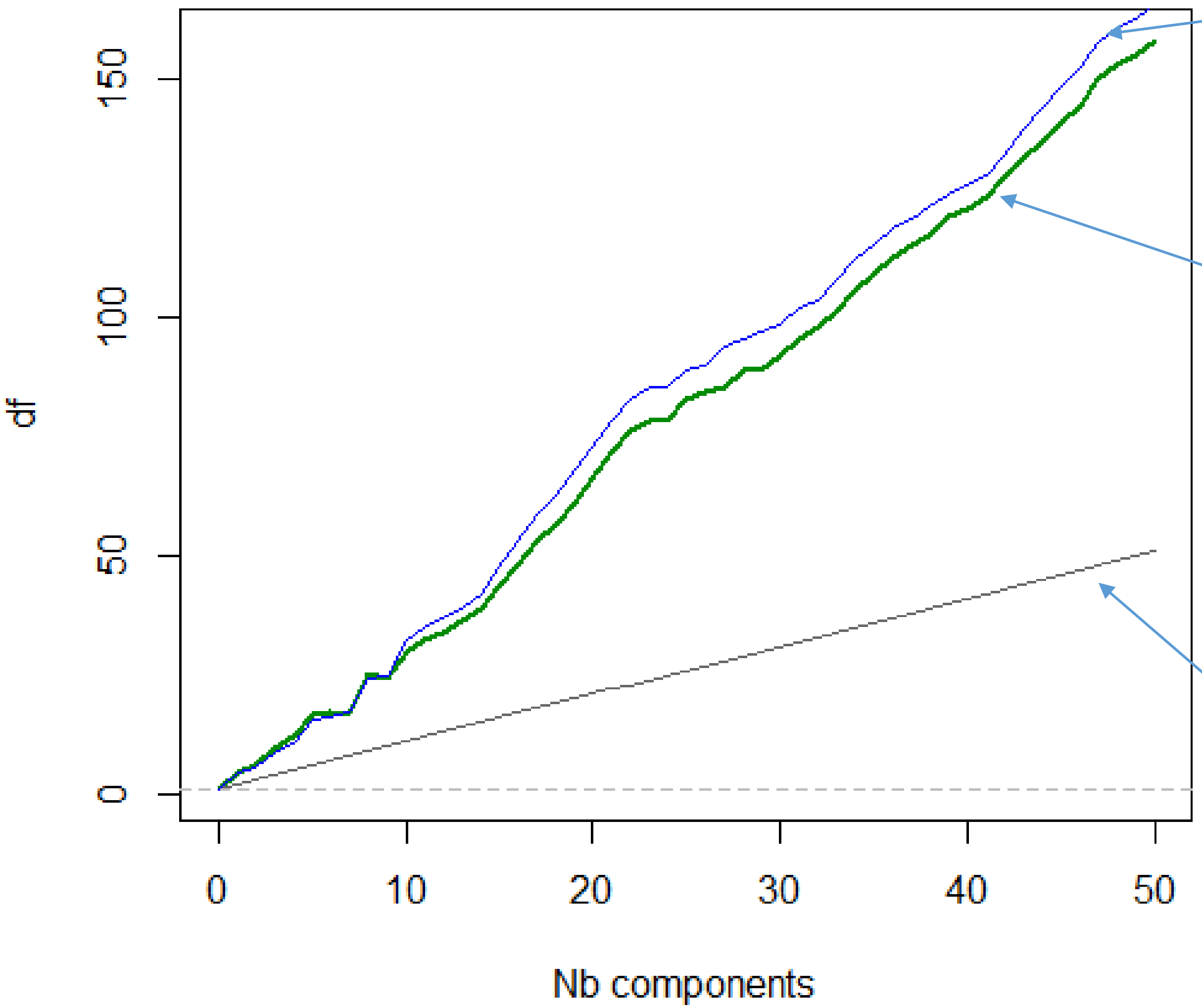
$\widehat{df}_{\text{SURE}}$
with m = 50 perturbation points



$\widehat{df}_{\text{SURE}}$

with m **= 1000 perturbation points**

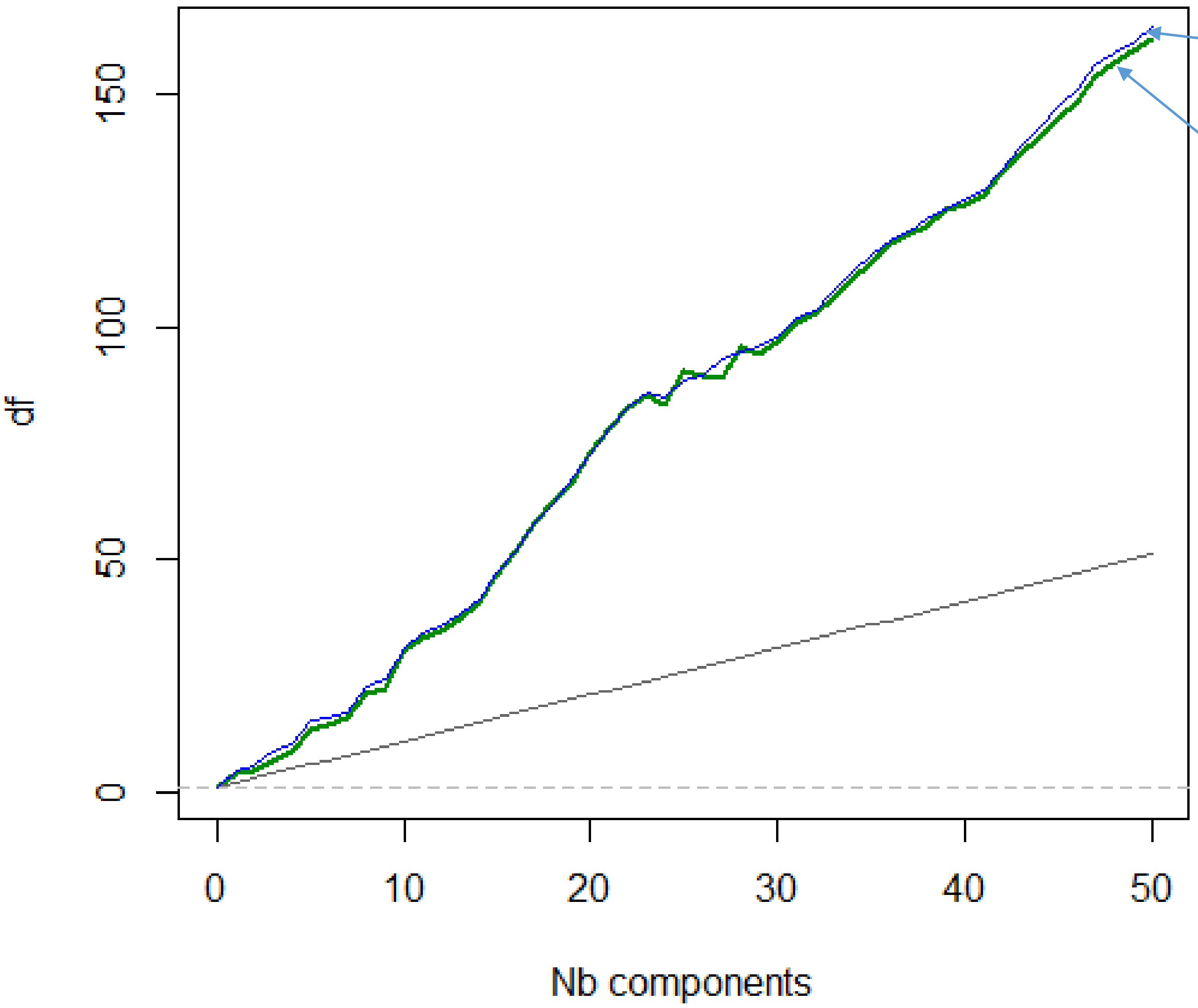
PLSR



$\widehat{df}_{\text{Cov}}$
with B = 50 bootstrap replications

$\widehat{df}_{\text{SURE}}$
with m = 50 perturbation points

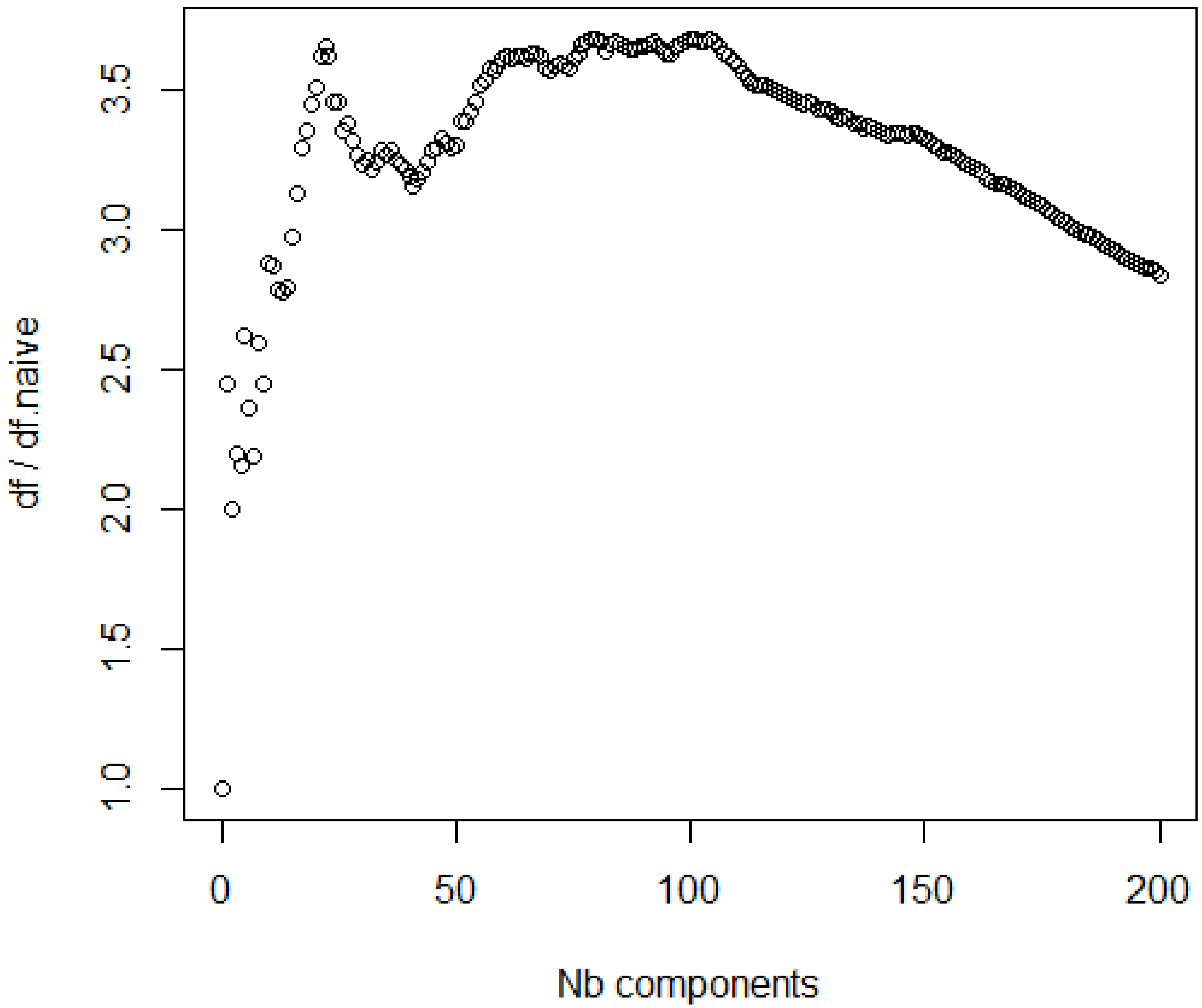
$df_{\text{naive}} = a + 1$



$\widehat{df}_{\text{Cov}}$ with B = **100 bootstrap replications**

$\widehat{df}_{\text{SURE}}$ with m = **100 perturbation points**

Ratio to the naive *df*



Krämer & Sugiyama Jasa 2011

Proposed formulas for calculating $\widehat{df}_{\text{SURE}}$

$$\hat{\boldsymbol{y}} = \boldsymbol{S}_{\boldsymbol{a},\boldsymbol{y}}\ \boldsymbol{y} \qquad \text{with} \qquad \boldsymbol{S}_{\boldsymbol{a},\boldsymbol{y}} = \boldsymbol{T}_{\boldsymbol{a},\boldsymbol{y}}\ (\boldsymbol{T}_{\boldsymbol{a},\boldsymbol{y}}'\ \boldsymbol{T}_{\boldsymbol{a},\boldsymbol{y}})^{-1}\boldsymbol{T}_{\boldsymbol{a},\boldsymbol{y}}'$$

$$\operatorname{tr}\left(\frac{\partial\hat{\boldsymbol{y}}}{\partial\boldsymbol{y}}\right) = \operatorname{tr}\left(\frac{\partial(\boldsymbol{S}_{\boldsymbol{a},\boldsymbol{y}}\ \boldsymbol{y})}{\partial\boldsymbol{y}}\right)$$

R package **`plsdof`** (CRAN)

Algorithm 1

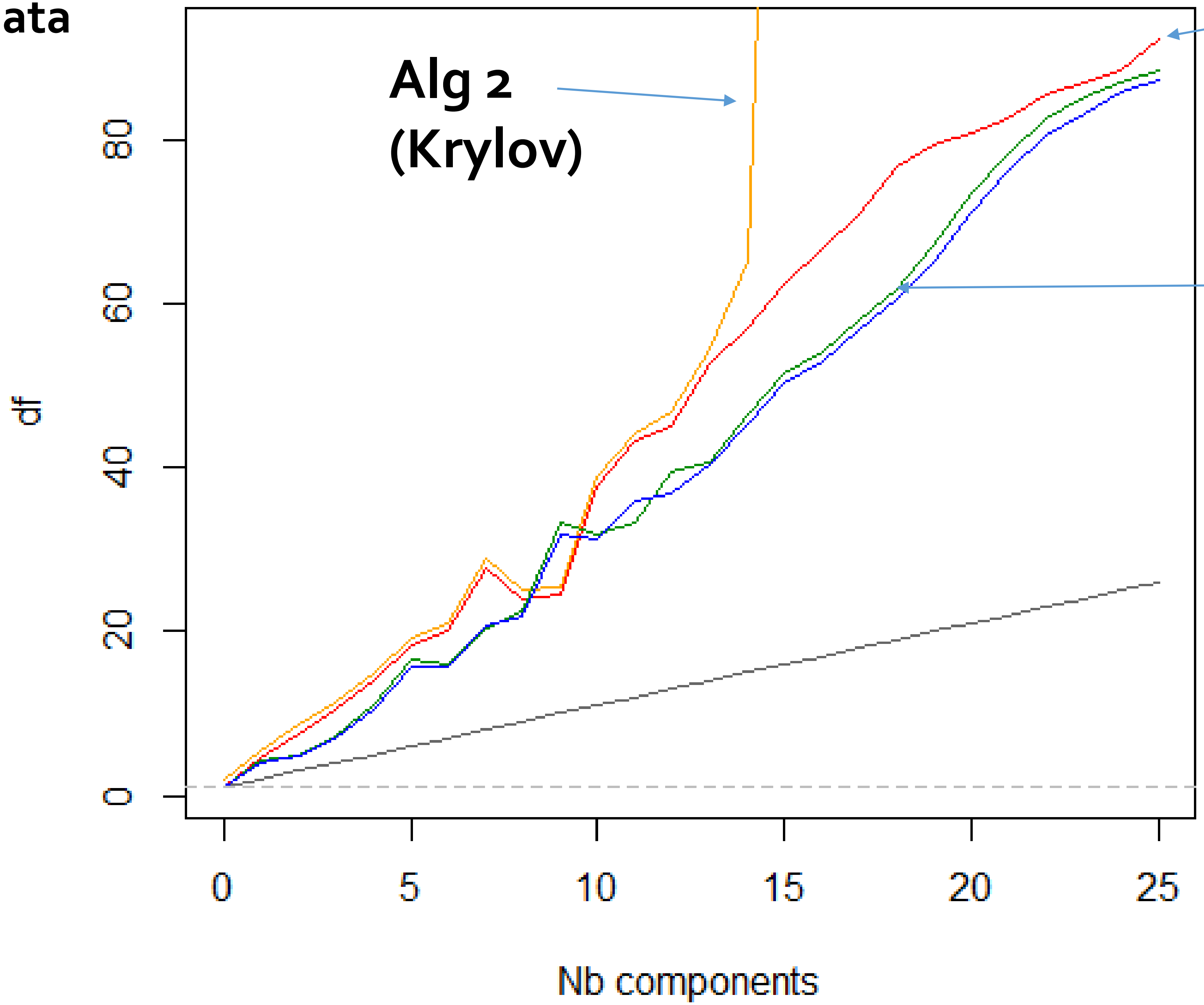
- Lanczos algorithm for PLSR
- Step of derivation for each component (Kramer & Braun 2007)
- **High calculation times**

Algorithm 2

- $\hat{\boldsymbol{y}}$ is expressed using Krylov subspaces
- This simplifies the calculation of the trace of the derivative
- **Much faster calculation times**

Sorghum NIR data

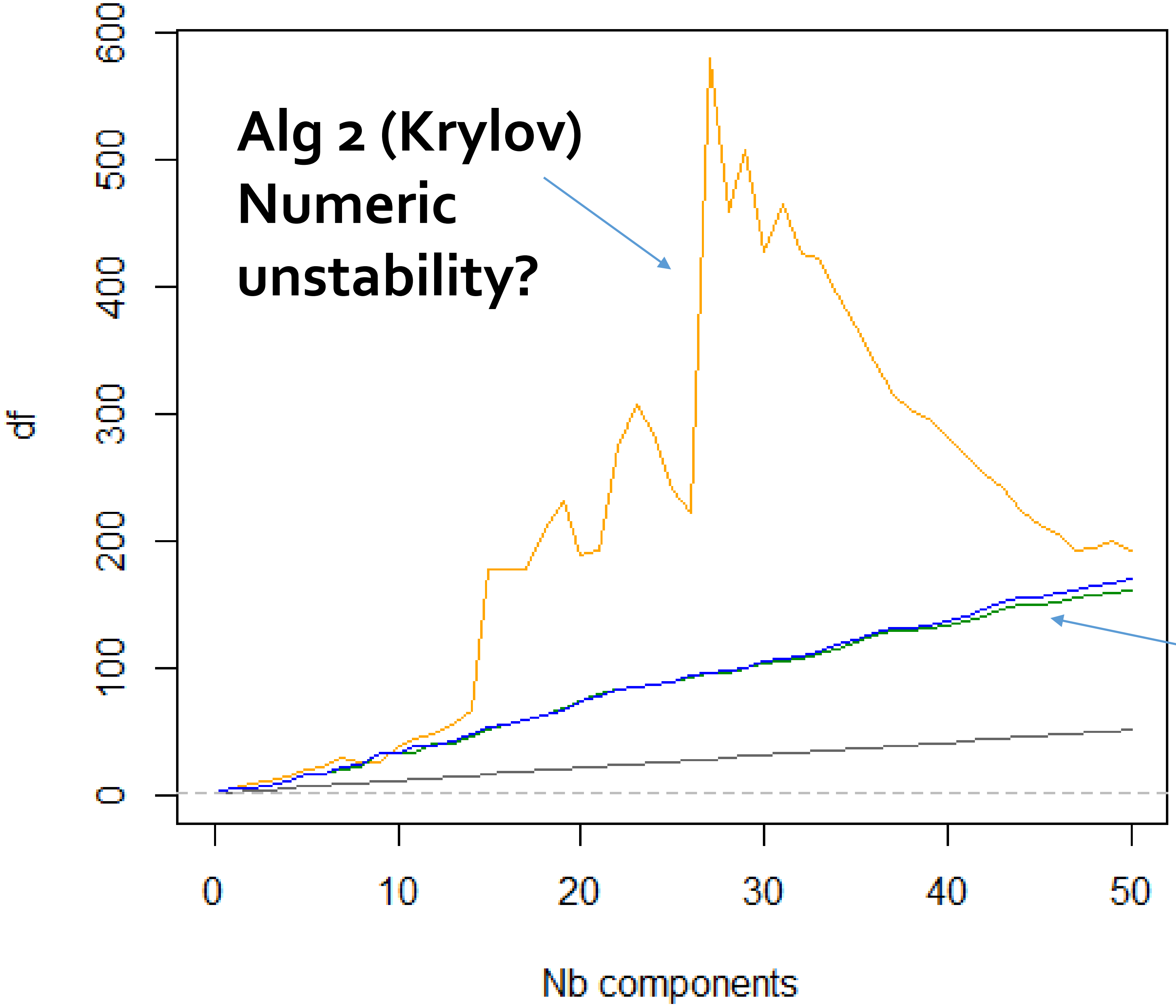
$n = 1006$



Alg 2 (Krylov)

Alg 1 (Lanczos)

Monte Carlo Cov & SURE



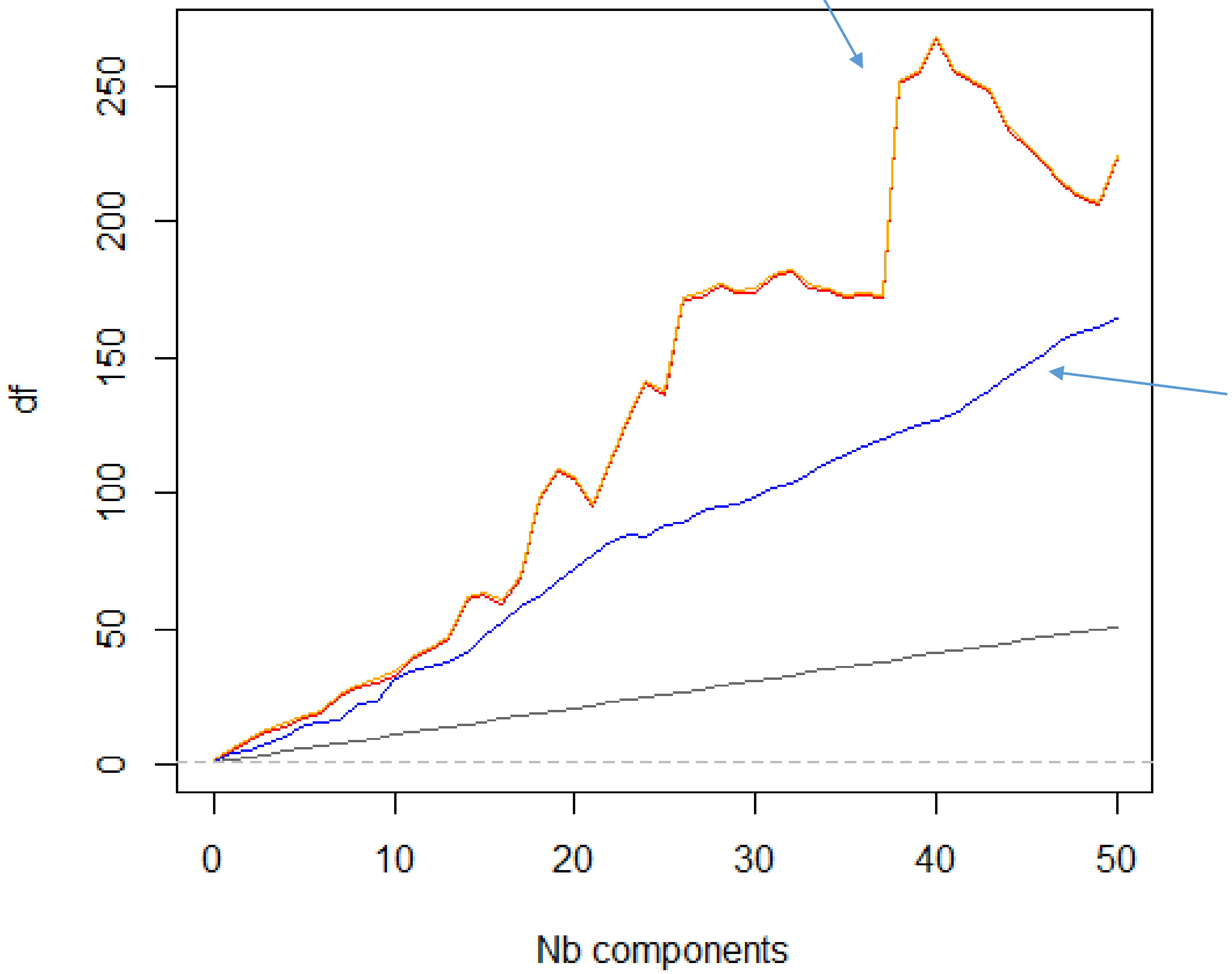
Alg 2 (Krylov)
Numeric
unstability?

(Alg 1
crashed)

Monte Carlo
Cov & SURE

Same problem with another dataset

Cassava $n = 200$ (carotenoids)



Alg 2 (Krylov)

Monte Carlo Cov

Coming back to covariance penalty criteria

- $\mathrm{Cp} = \quad \frac{\mathrm{SSR}}{n} \quad + \quad \frac{2}{n}\,\widehat{df}\,\hat{\sigma}^2$

- $\widehat{df} = \sum_{i=1}^{n} \hat{C}ov_{\varepsilon}(y_i, \hat{y}_i)/\widehat{\sigma}^2 \quad = \quad (\mathrm{SURE}) \;\; \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i}$

Two estimates

1) $$\mathrm{Cp} = \frac{\mathrm{SSR}}{n} + \frac{2}{n}\sum_{i=1}^{n} \hat{C}ov_{\varepsilon}(y_i, \hat{y}_i)$$

2) $$\mathrm{Cp} = \frac{\mathrm{SSR}}{n} + \frac{2}{n}\hat{\sigma}^2 \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i}$$

SURE estimate

llustration of PLSR model selection
on a NIR dataset

- Presentation of the data

- Separation between training vs test sets

- CV results

- Examination of the PLS scores

- Examination of the loadings and b-coefficients

- Cp results

- Sensitivity to the test set

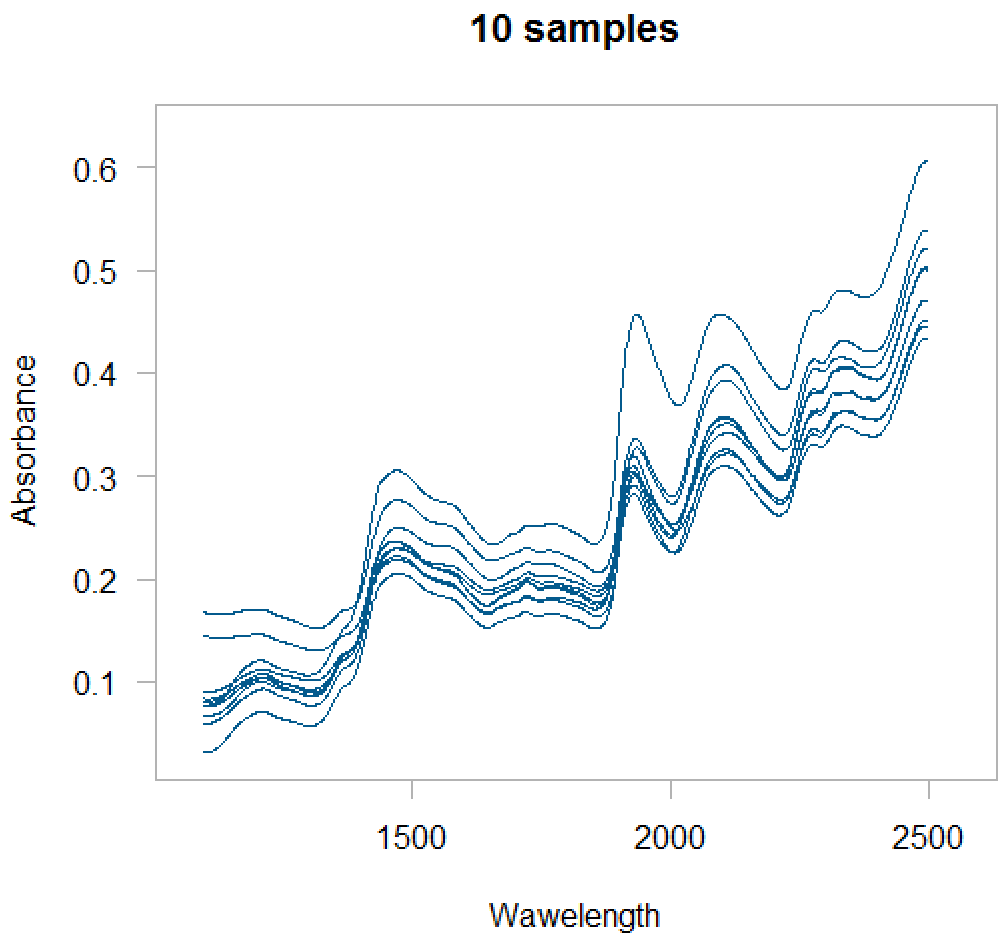
Data

Sorghum (stem, leafs etc.)

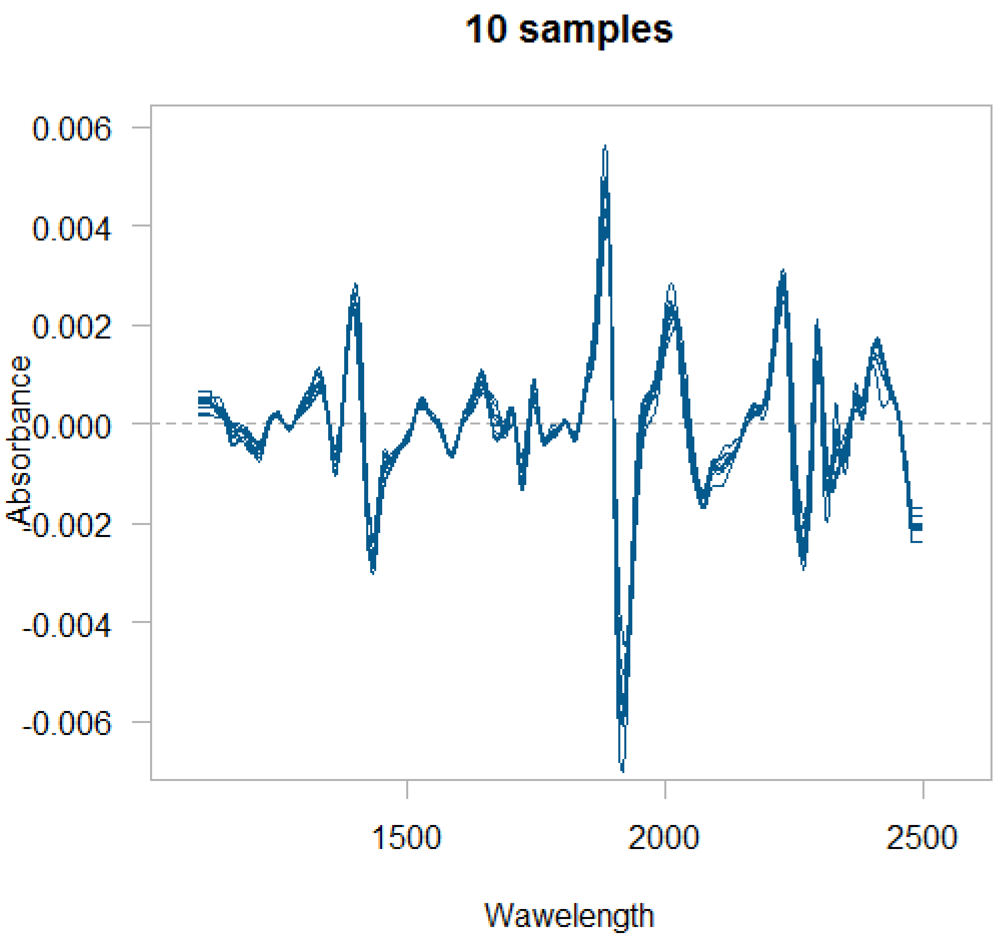
Dried and grounded

Spectra FOSS 1100-2498 nm (step = 2 nm)

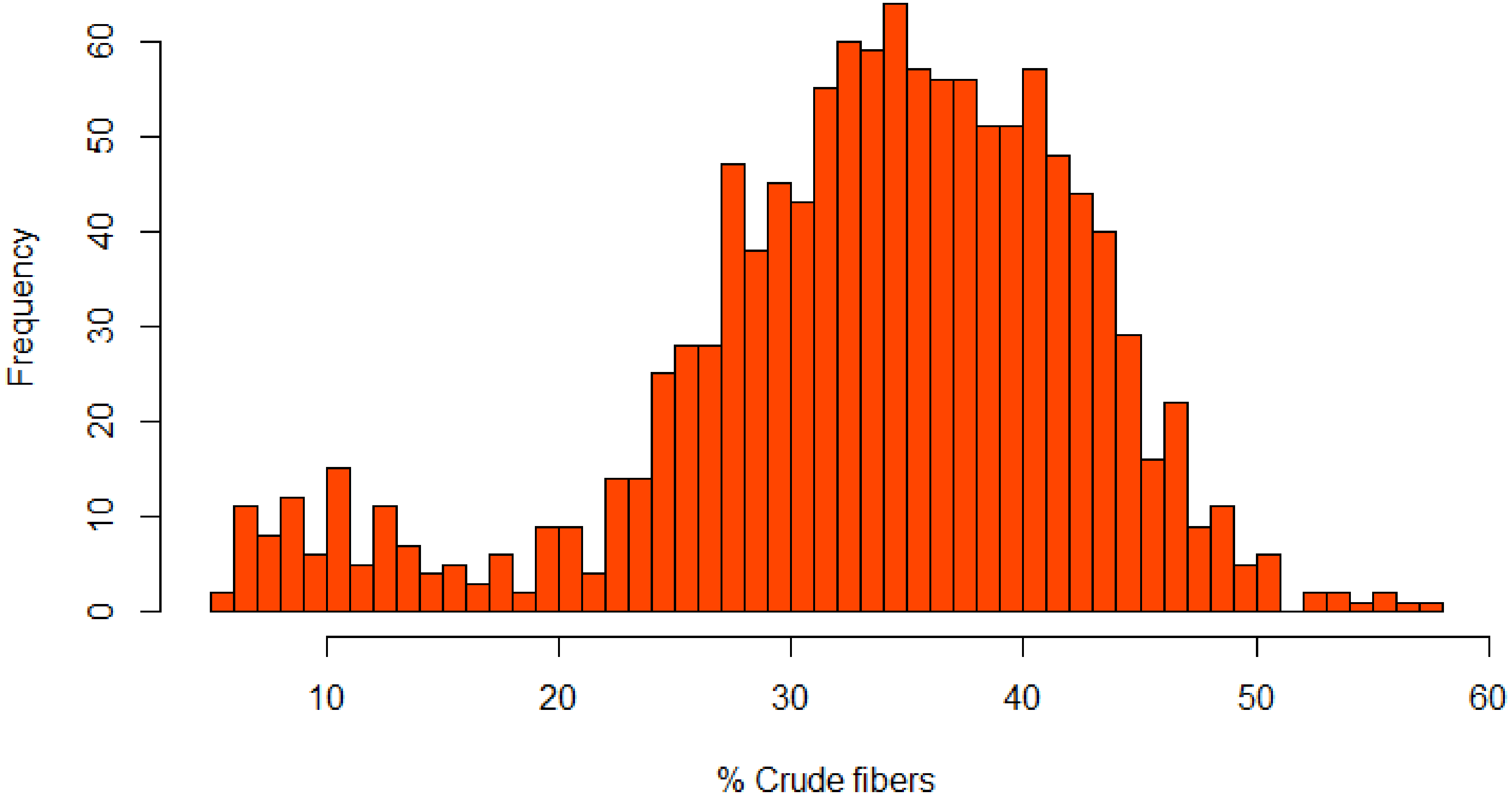
n = 1206 p = 700

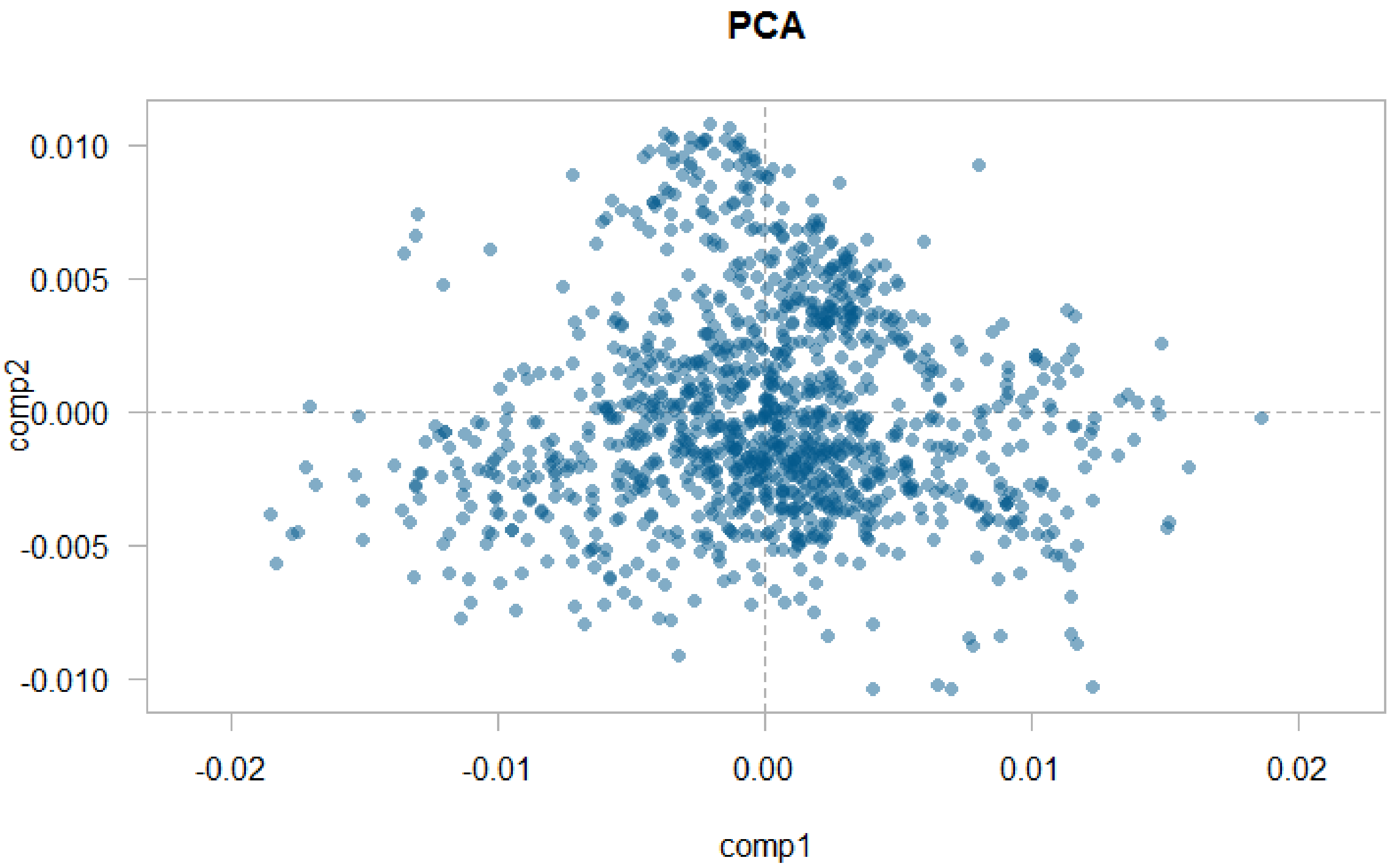


savgol(snv(X), m = 2, n = 21, p = 2)



Prediction of % crude fibers




PCA
0.010
0.005
0.000
-0.005
-0.010
comp2
-0.02
-0.01
0.00
0.01
0.02
comp1

Splitting the data

Training *vs.* Test sets

Hypothesis Future = same mechanism as the training (probability distribution F)

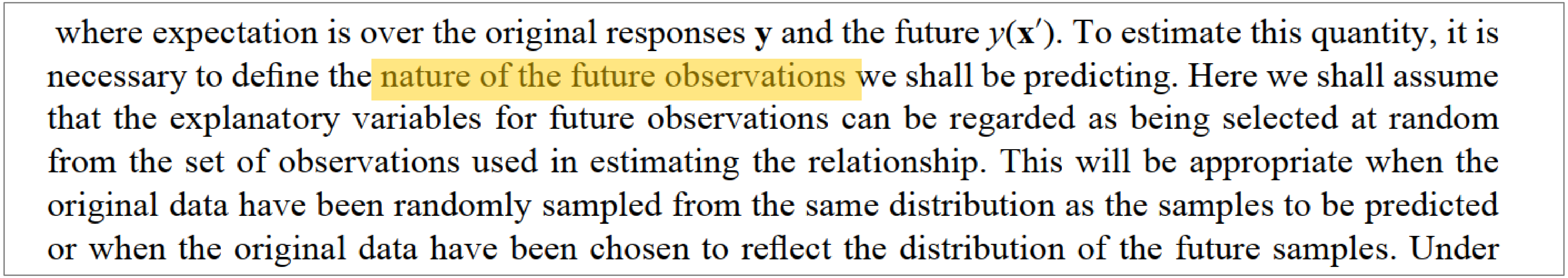
- **Training set** $F \rightarrow \boldsymbol{\tau} = \{ (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \}$

- **New observation** $F \rightarrow (\boldsymbol{x}^*, y^*)$

(Will be relaxed later)

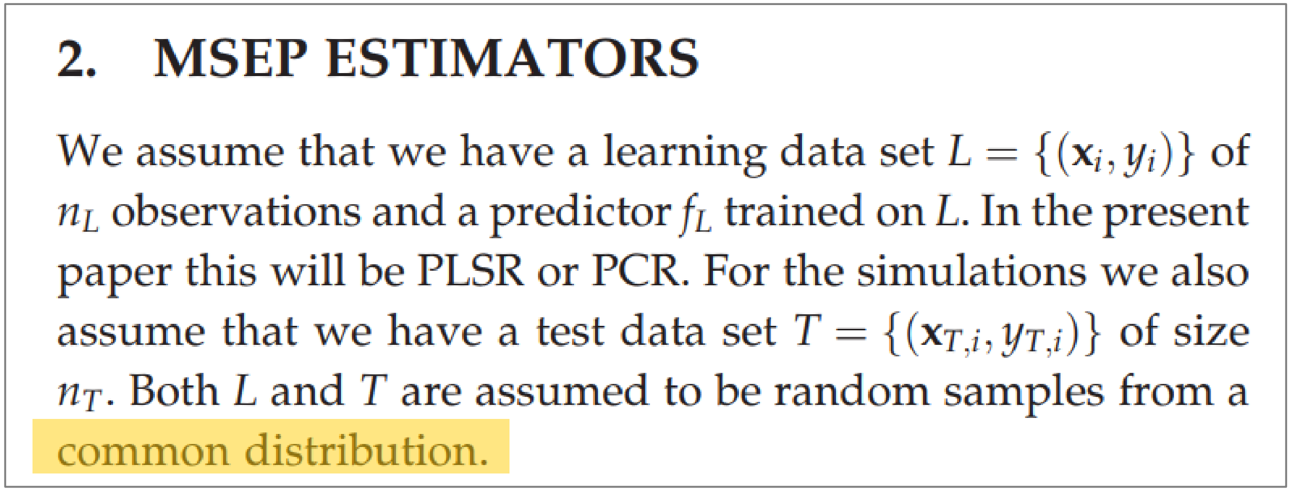
Examples in chemometric journals

- **Denham, M.C., 2000**. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. *Journal of Chemometrics* 14, 351–361. https://doi.org/10.1002/1099-128X(200007/08)14:4<351::AID-CEM598>3.0.CO;2-Q



where expectation is over the original responses $\mathbf{y}$ and the future $y(\mathbf{x}')$. To estimate this quantity, it is necessary to define the nature of the future observations we shall be predicting. Here we shall assume that the explanatory variables for future observations can be regarded as being selected at random from the set of observations used in estimating the relationship. This will be appropriate when the original data have been randomly sampled from the same distribution as the samples to be predicted or when the original data have been chosen to reflect the distribution of the future samples. Under

- **Mevik, B.-H., Cederkvist, H.R., 2004**. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *Journal of Chemometrics* 18, 422–429. https://doi.org/10.1002/cem.887



2. MSEP ESTIMATORS

We assume that we have a learning data set $L = \{(\mathbf{x}_i, y_i)\}$ of n_L observations and a predictor f_L trained on L. In the present paper this will be PLSR or PCR. For the simulations we also assume that we have a test data set $T = \{(\mathbf{x}_{T,i}, y_{T,i})\}$ of size n_T. Both L and T are assumed to be random samples from a common distribution.

- See also **Faber, N. (Klaas) M., 1999**. Estimating the uncertainty in estimates of root mean square error of prediction: application to determining the size of an adequate test set in multivariate calibration. *Chemometrics and Intelligent Laboratory Systems* 49, 79–89. https://doi.org/10.1016/S0169-7439(99)00027-1 **Etc. !!!**

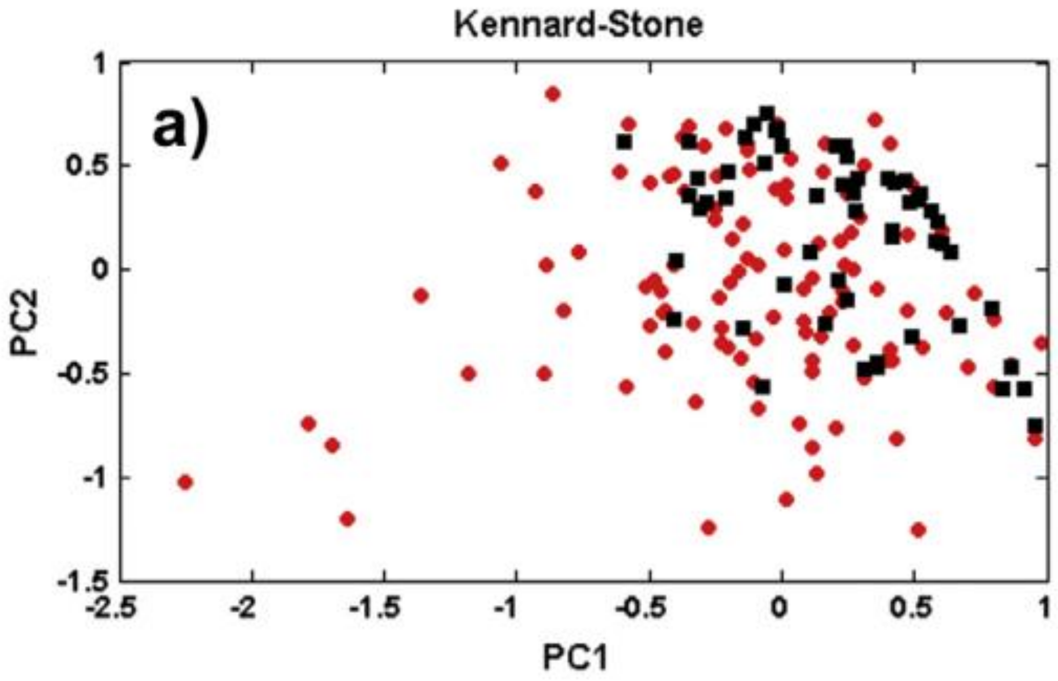
Remark 1) True generating distribution F

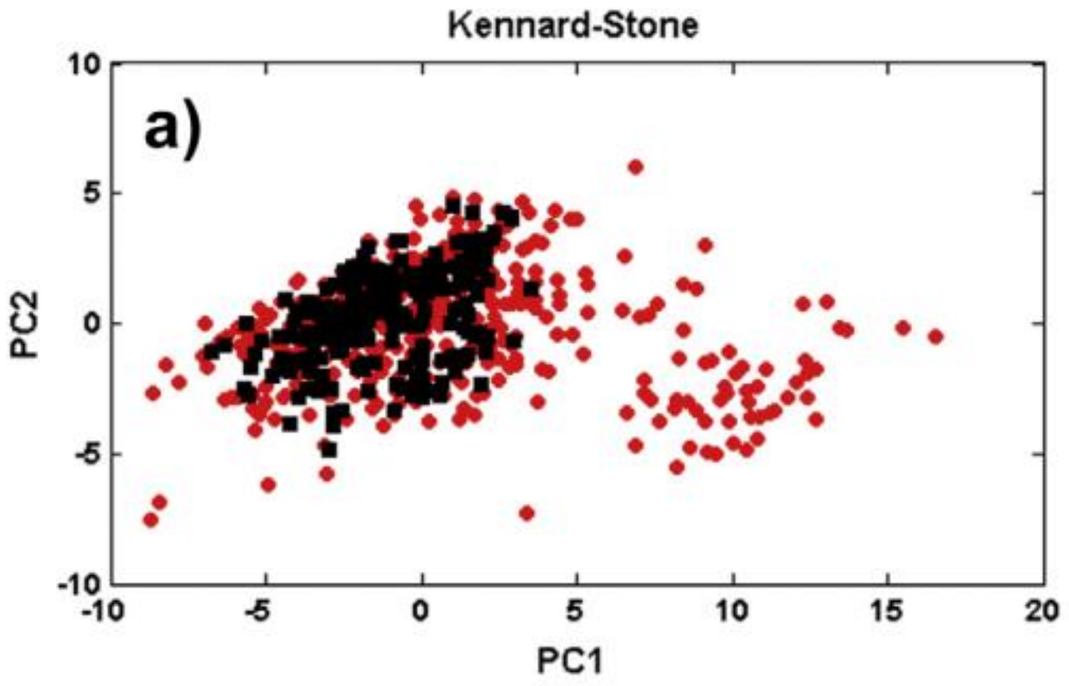
- **Traditional chemometrics**
  - Hidden F is often considered having low dimension
  - Bringing out these dimensions

- **Statistical ecology, epidemiology etc.**
  - F often considered having high dimension (even “infinite”)
  - Model with relevant bias-variance compromise for a training set of size n

Remark 2)

The Kennard-Stone algorithm does not follow the previous hypothesis, e.g.





Westad, F., Marini, F., 2015. Validation of chemometric models – A tutorial. *Analytica Chimica Acta* 893, 14–24. https://doi.org/10.1016/j.aca.2015.06.056

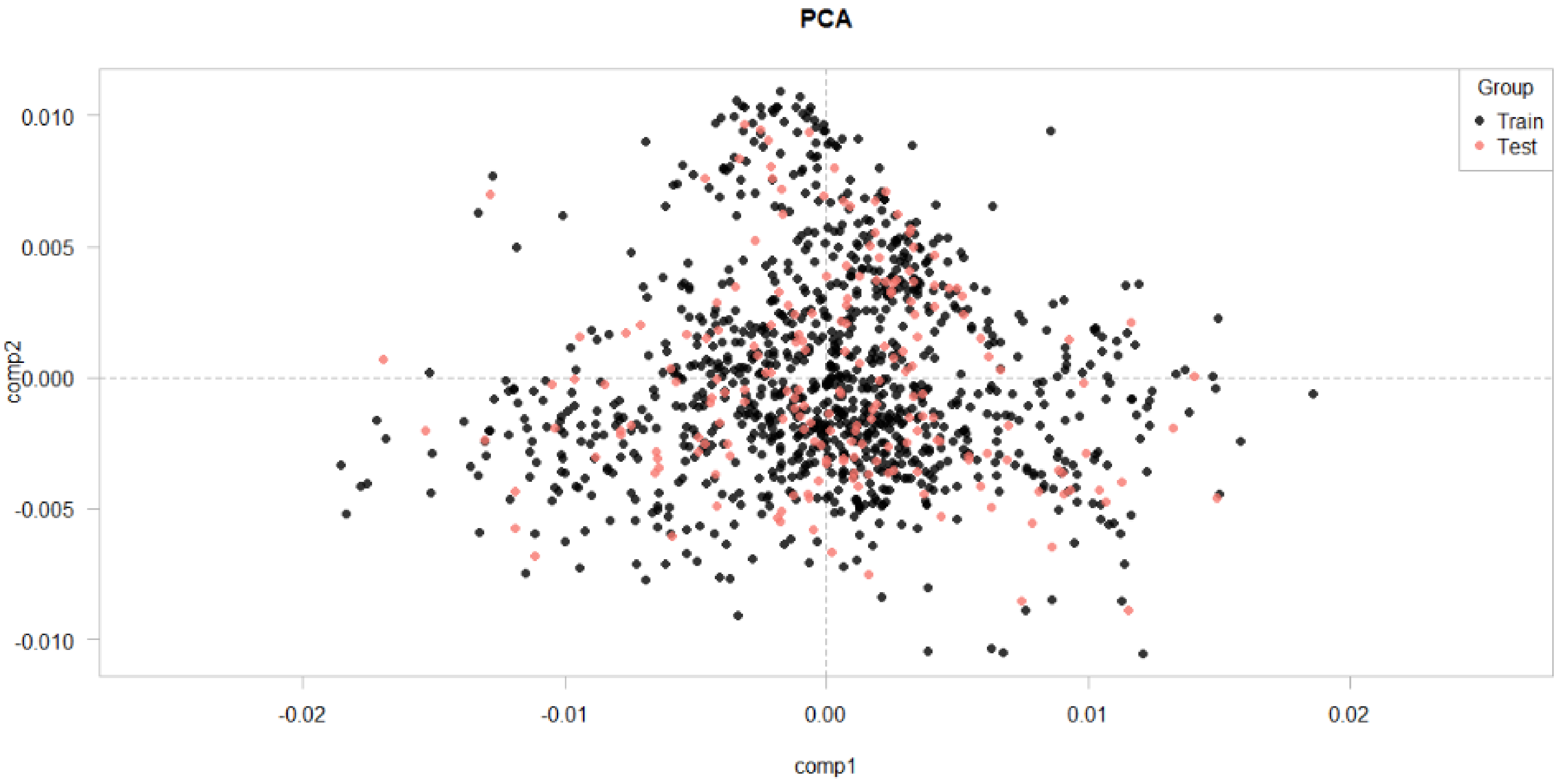
More representative

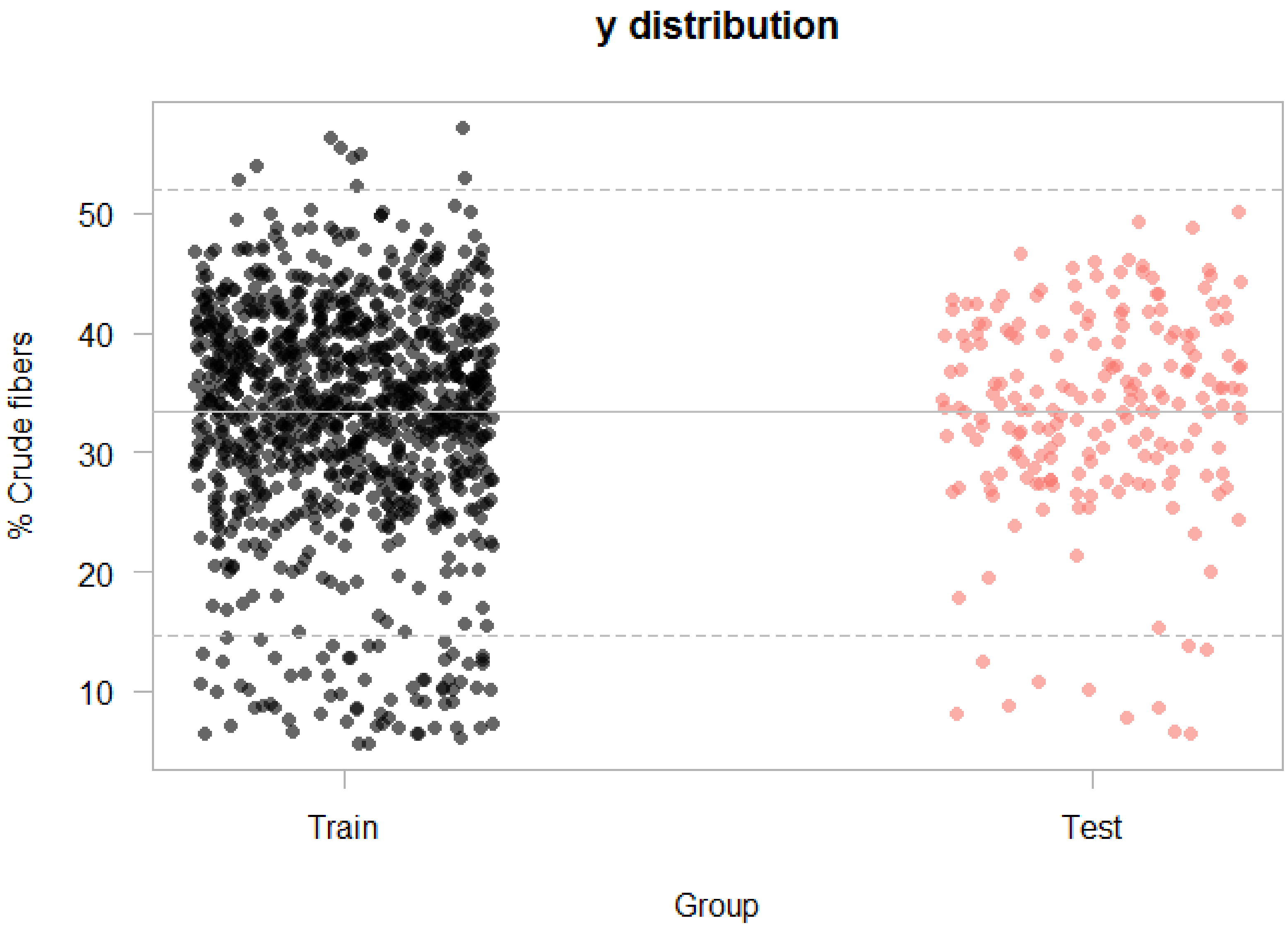
- **Uniform random sampling** Unbiased for F but variable
- **Representative stratified sampling** Less variable
  - Clustering and then proportional intra-cluster sampling
- **Duplex (Snee 1977)** Alernate Kennard-Stone
- **Latin Hypercube sampling (LHS)** Uniform over the margins

Data separation for this example: uniform sampling

n = **1206**

Training set	1006 samples
Test	200 samples (15%) **randomly selected**


PCA
Group
● Train
● Test
0.010
0.005
0.000
-0.005
-0.010
comp2
-0.02
-0.01
0.00
0.01
0.02
comp1


y distribution
% Crude fibers
50
40
30
20
10
Train
Test
Group

Cross-validation

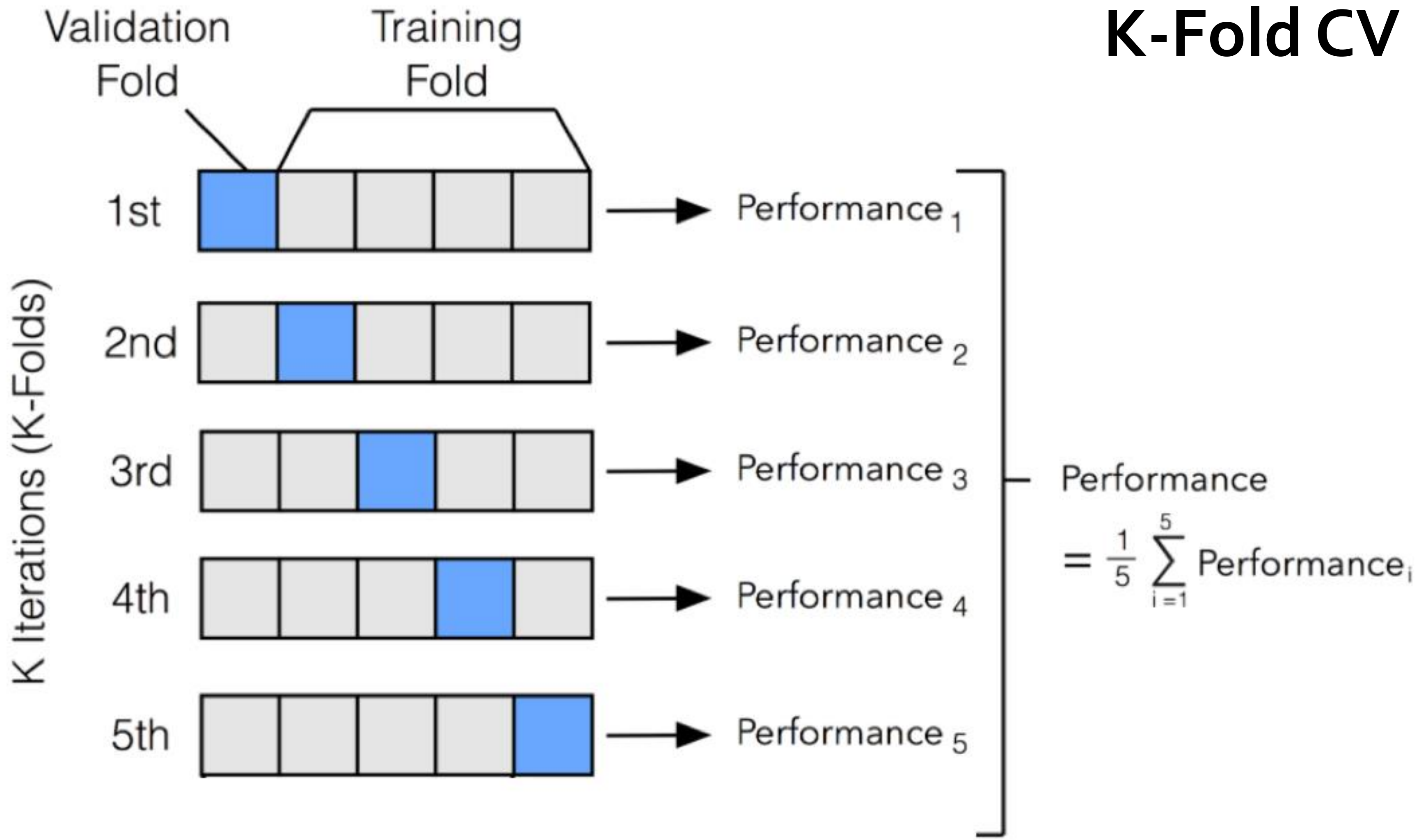
1) **K-Fold CV** ← In the present example

- Can be repeated

2) **Test-set (or Monte Carlo) CV**

- Alternative when K-Fold not possible (large datasets)
- In general, returns models slightly more parcimonious
- All samples are not seen in the VAL

K-Fold CV



http://ethen8181.github.io/machine-learning/model_selection/model_selection.html

Under hypothesis $F_{\text{Train}} = F_{\text{Future}}$

Err = **Expected prediction (or test) error** (See Appendix)

CV is a non parametric method for estimating Err

- $\widehat{\text{M}}\text{SEP}_{\text{CV}} = \widehat{\text{E}}\text{rr}$

Particular K-Fold CV

- Leave-One-Out LOO-CV

$K = n$

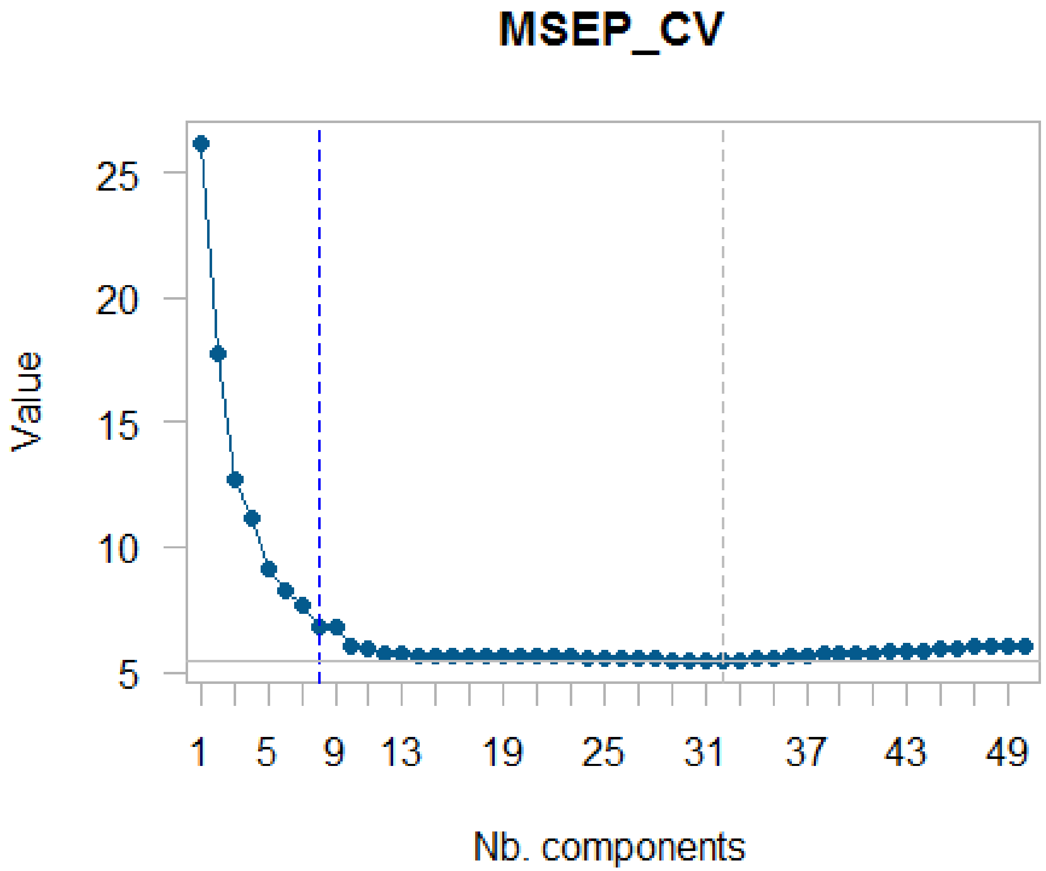
Within the K-Fold strategy

- LOO-CV ($K = n$) is the less biased for Err

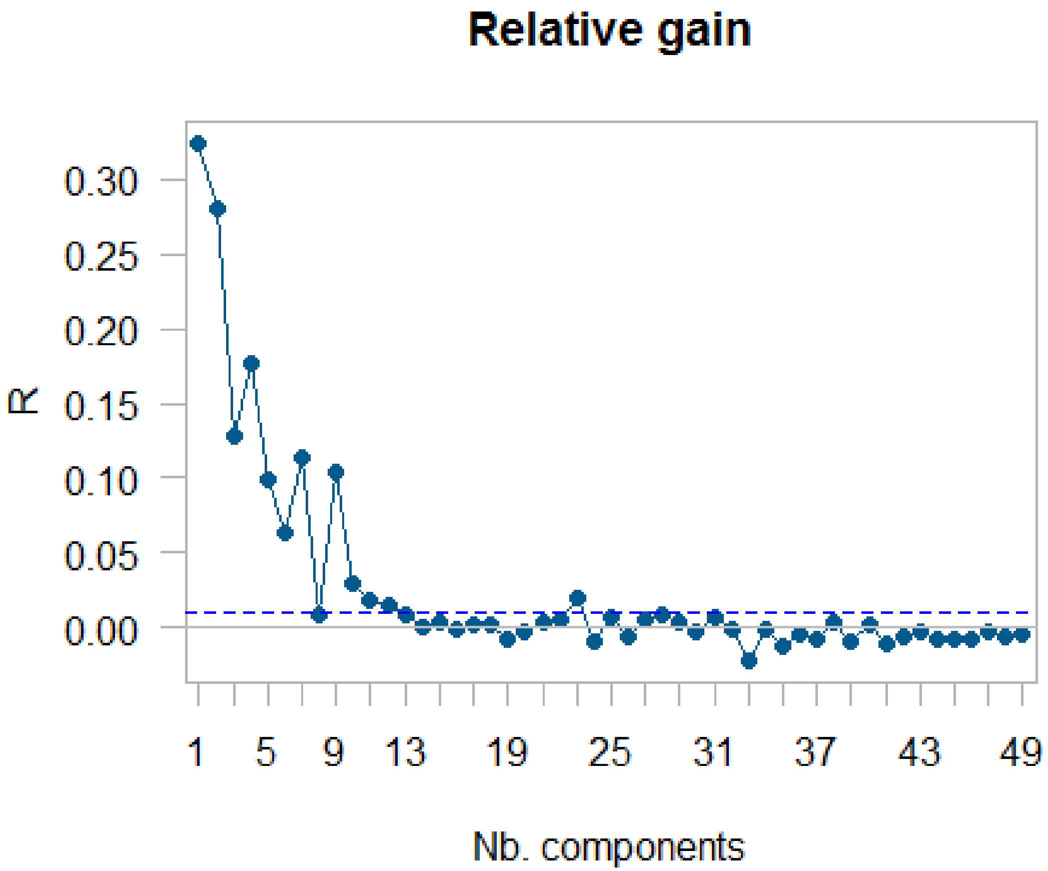
  but can have high variability with the training $\boldsymbol{\tau}$

(*) If other training sets $\boldsymbol{\tau}$ are generated from distribution F, LOO-CV can return quite different models

$K = n$



MSEP_CV



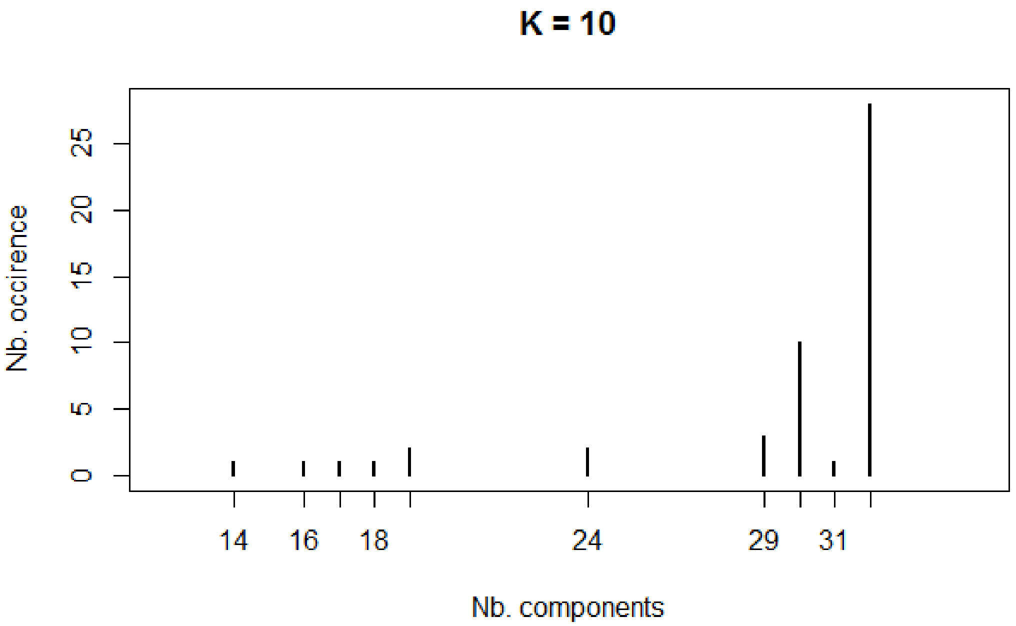
Relative gain

```
ncomp nbpred  msep   mad rmsep   sep      b    r2  cor2
   32   1006 5.441 1.555 2.333 2.333 -0.005 0.939 0.939
```

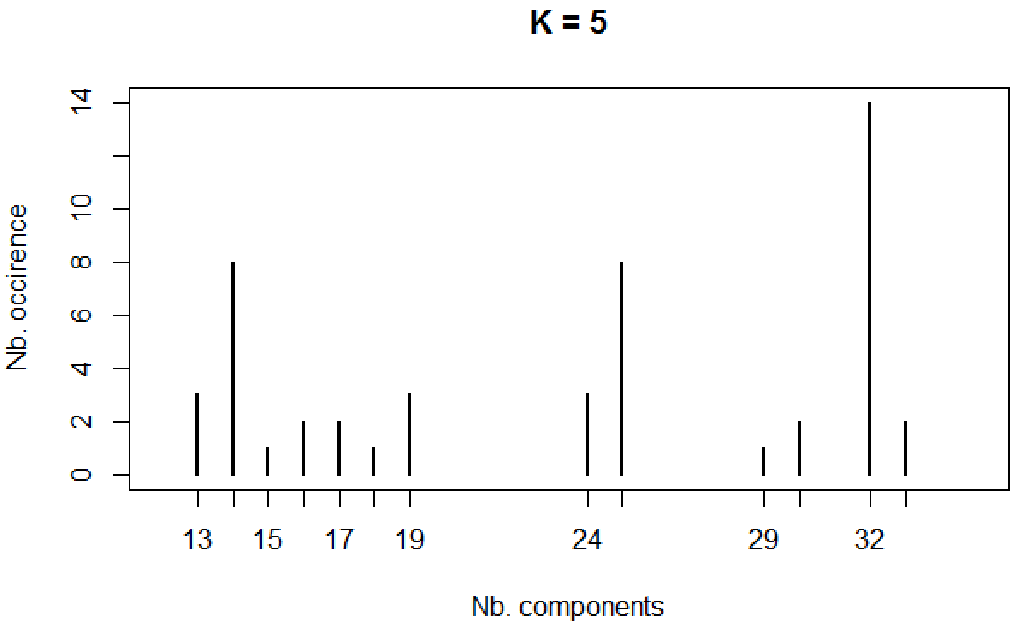
Consensus in the literature (for low to moderate biases)

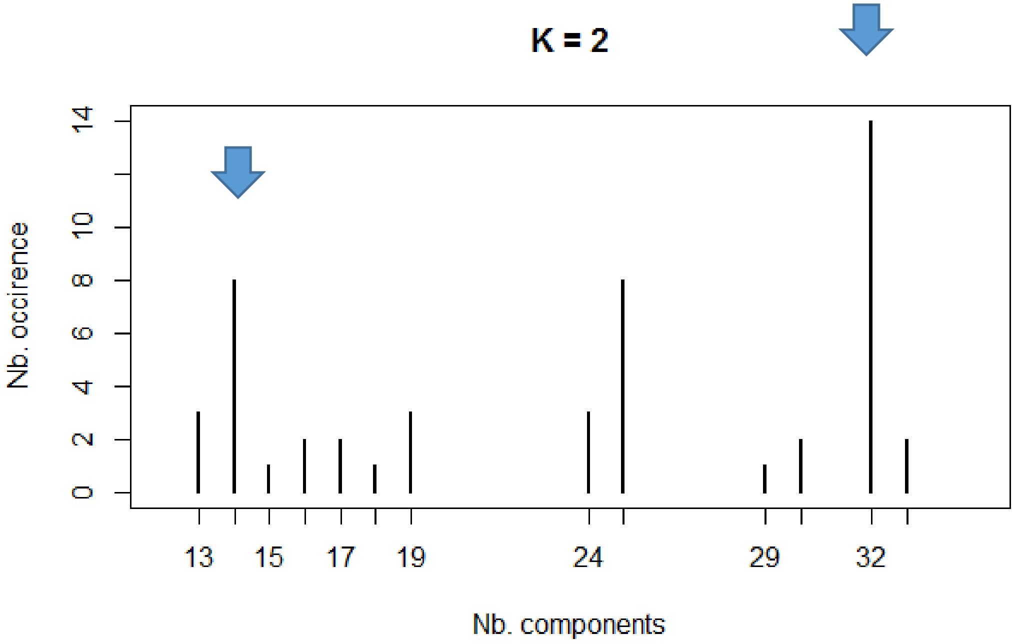
- K = 5-10

- $K = n$ opt = 32 Wold1% = 8
- $K < n$ With 50 repetitions
  - $K = 10$ opt = 32 Wold1% = 13
  - $K = 5$ opt = 32 Wold1% = 13
  - $K = 2$ opt = 13 Wold1% = 13



For the 50 repetitions, nb. occurrence for opt





Permutation tests on CV predictions

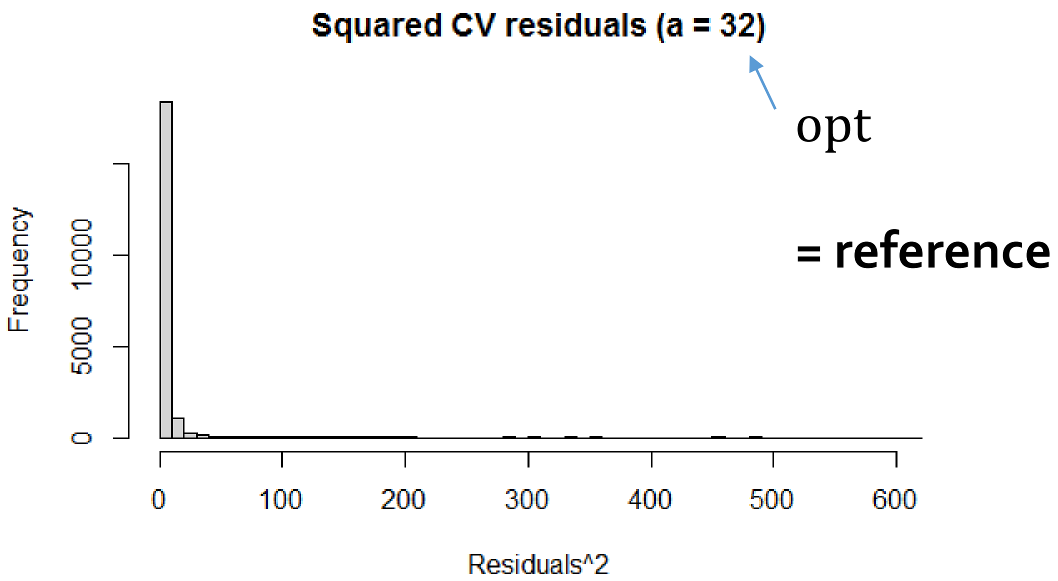
In the chemometrics literature, one implementation (within many other possible) is given in

van der Voet, H., 1994. Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and Intelligent Laboratory Systems* 25, 313–323. https://doi.org/10.1016/0169-7439(94)85050-X

Test if the squared CV residuals for a components and for opt components have the same distribution

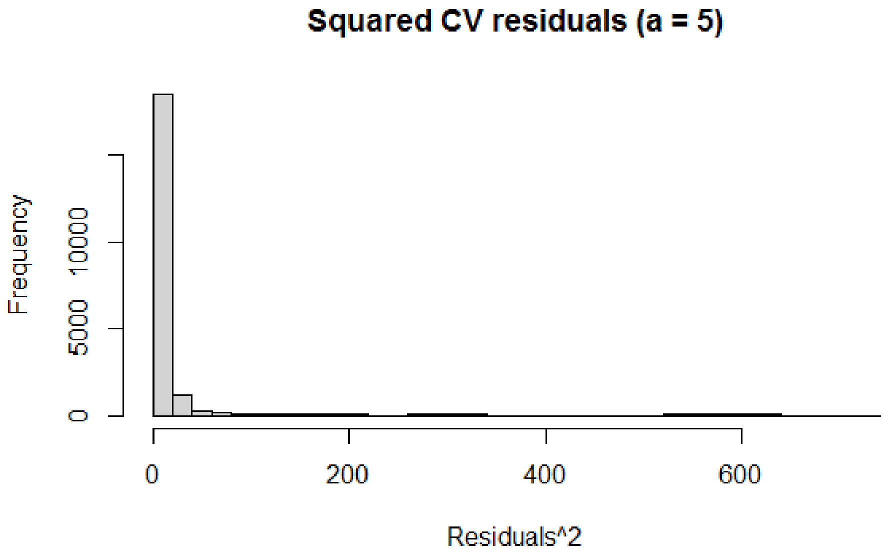
→ Non parametric permutation tests for matched pairs

(eventually with randomization)

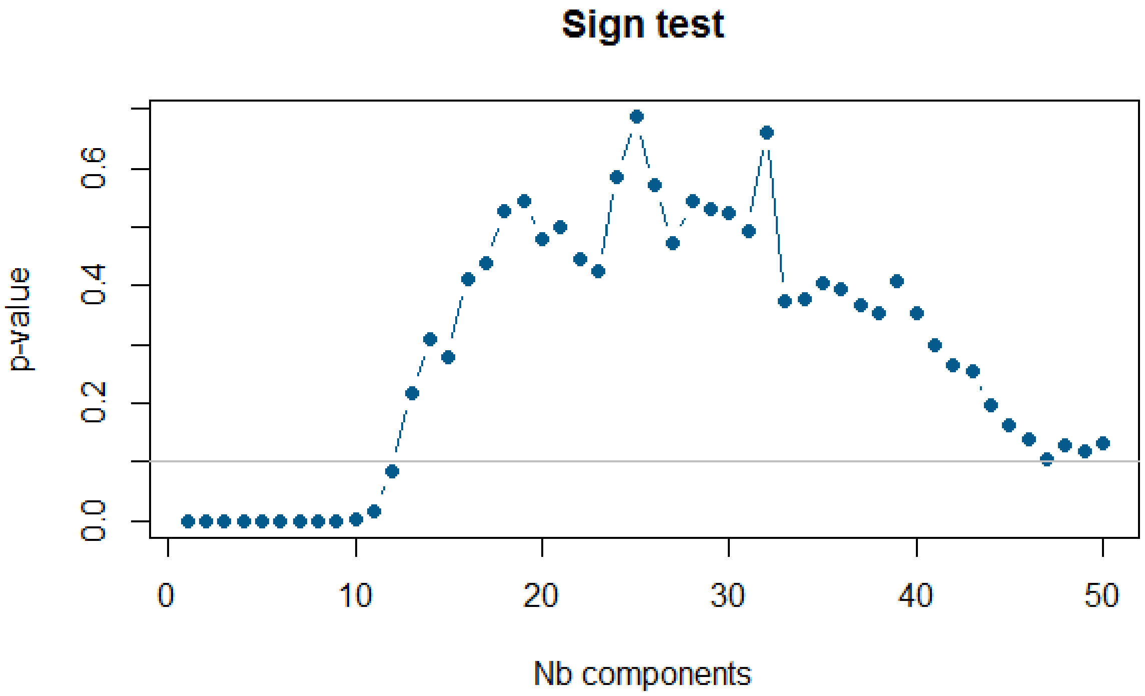

Squared CV residuals (a = 32)

opt

= reference


Squared CV residuals (a = 5)

Example with the **Sign test** for $K = 5$ (with 50 repetitions)



Other usual tests for matched pairs

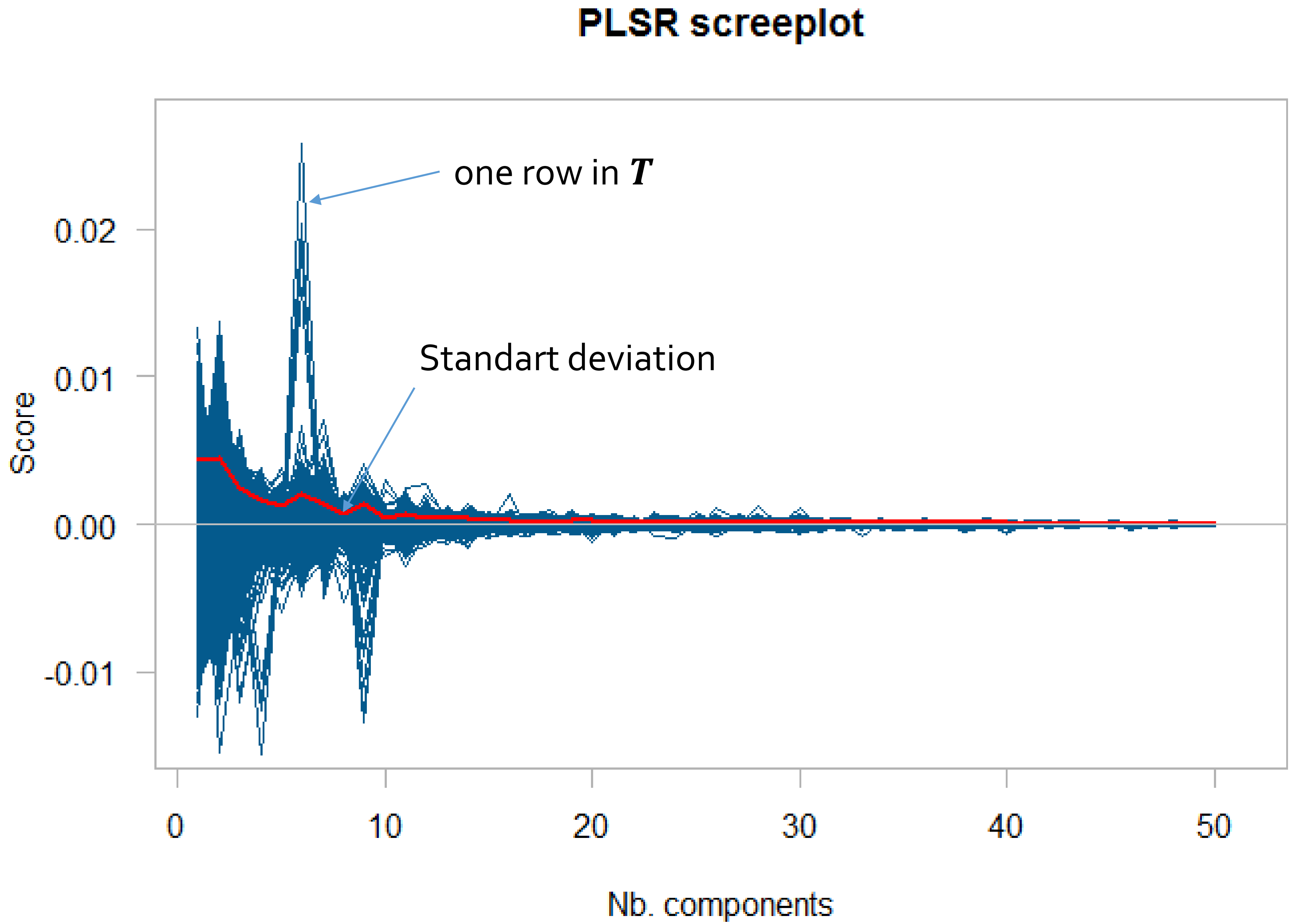
- Wilcoxon signed-rank test
- Randomized permutation test (this method is used in **van der Voet 1994**)

Alternative = F-tests

- Asumme Gaussian residuals
- Require calculating the good *dfs*!!

Examination of the PLS scores

based on training set



one row in $\boldsymbol{T}$

Standart deviation

Randomized permutation test of Wiklund et al. 2007

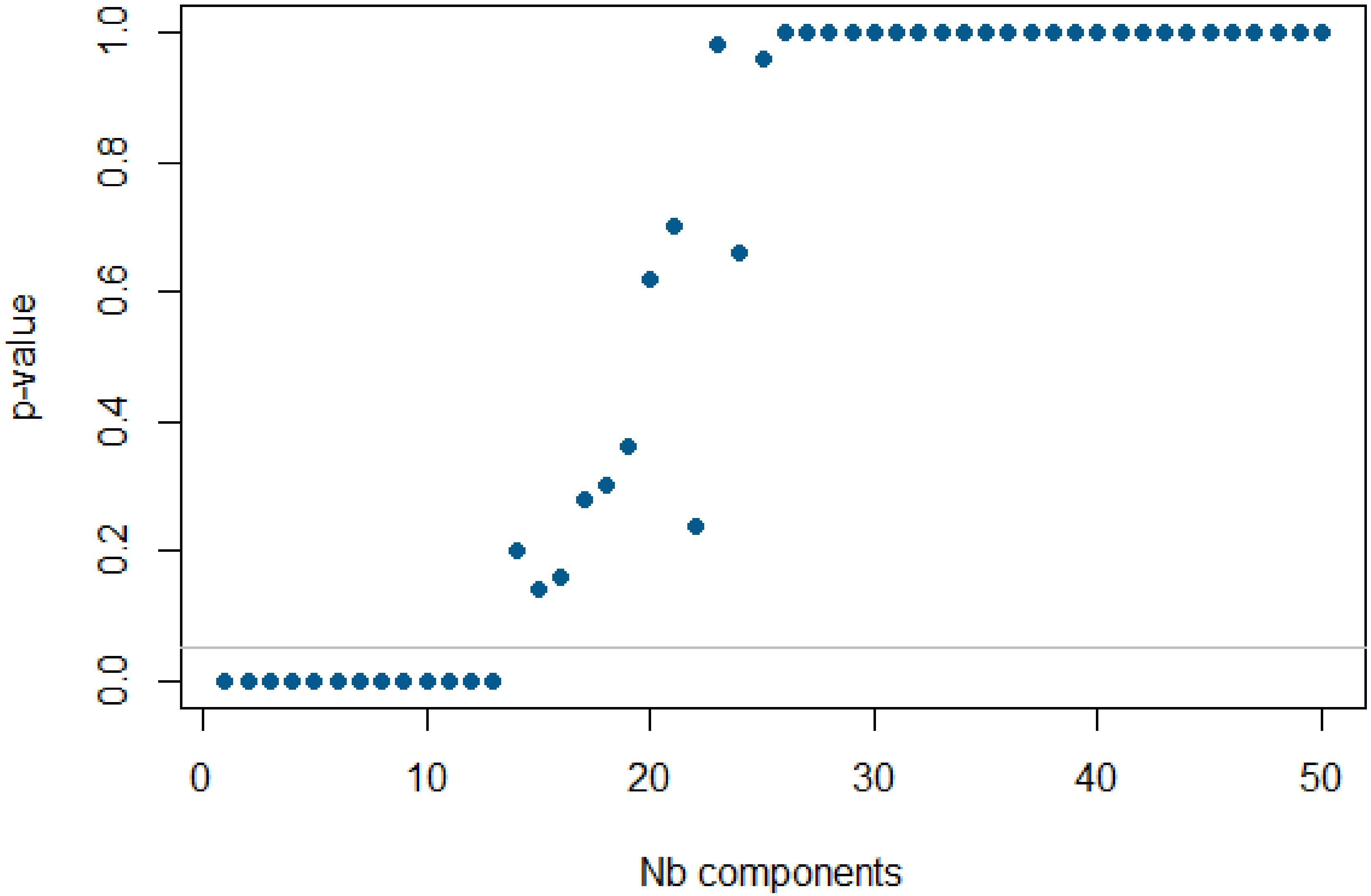
- $\mathrm{Corr}(\boldsymbol{t}_{\mathrm{a}}, \mathbf{y})$

Compare the observed distribution with a Null (H_0) distribution
H_0: $\boldsymbol{y}$ is randomly permuted (${\boldsymbol{y}}$ “scrambling”)

Long calculation time for large data
since conditional permutations for each component (successive PLS1)
$a = 1, \ldots, A$

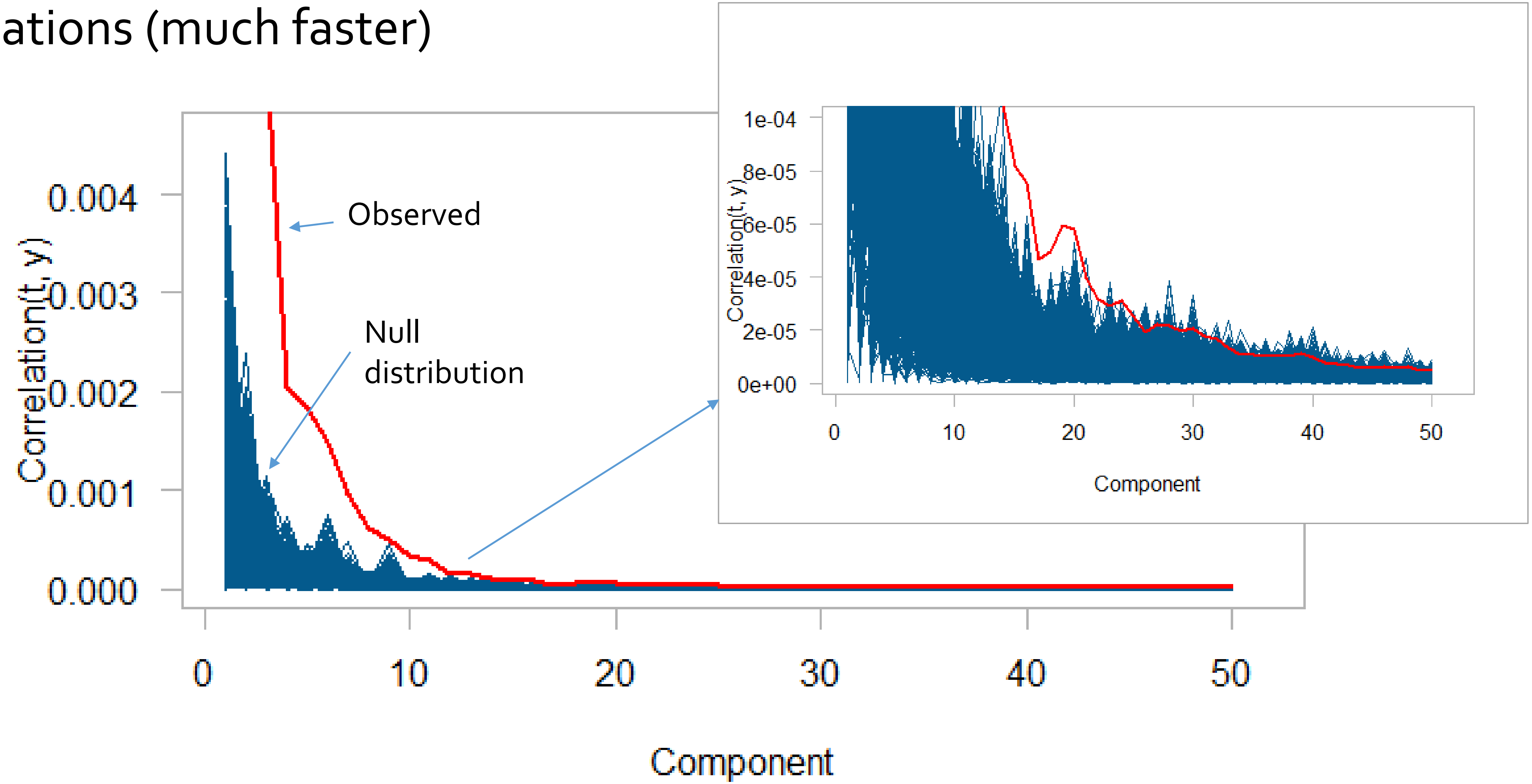
Wiklund, S., Nilsson, D., Eriksson, L., Sjöström, M., Wold, S., Faber, K., 2007. A randomization test for PLS component selection. *Journal of Chemometrics* 21, 427–439. https://doi.org/10.1002/cem.1086

Wiklund et al test



Another approach

Unconditionnal randomized permutations (much faster)



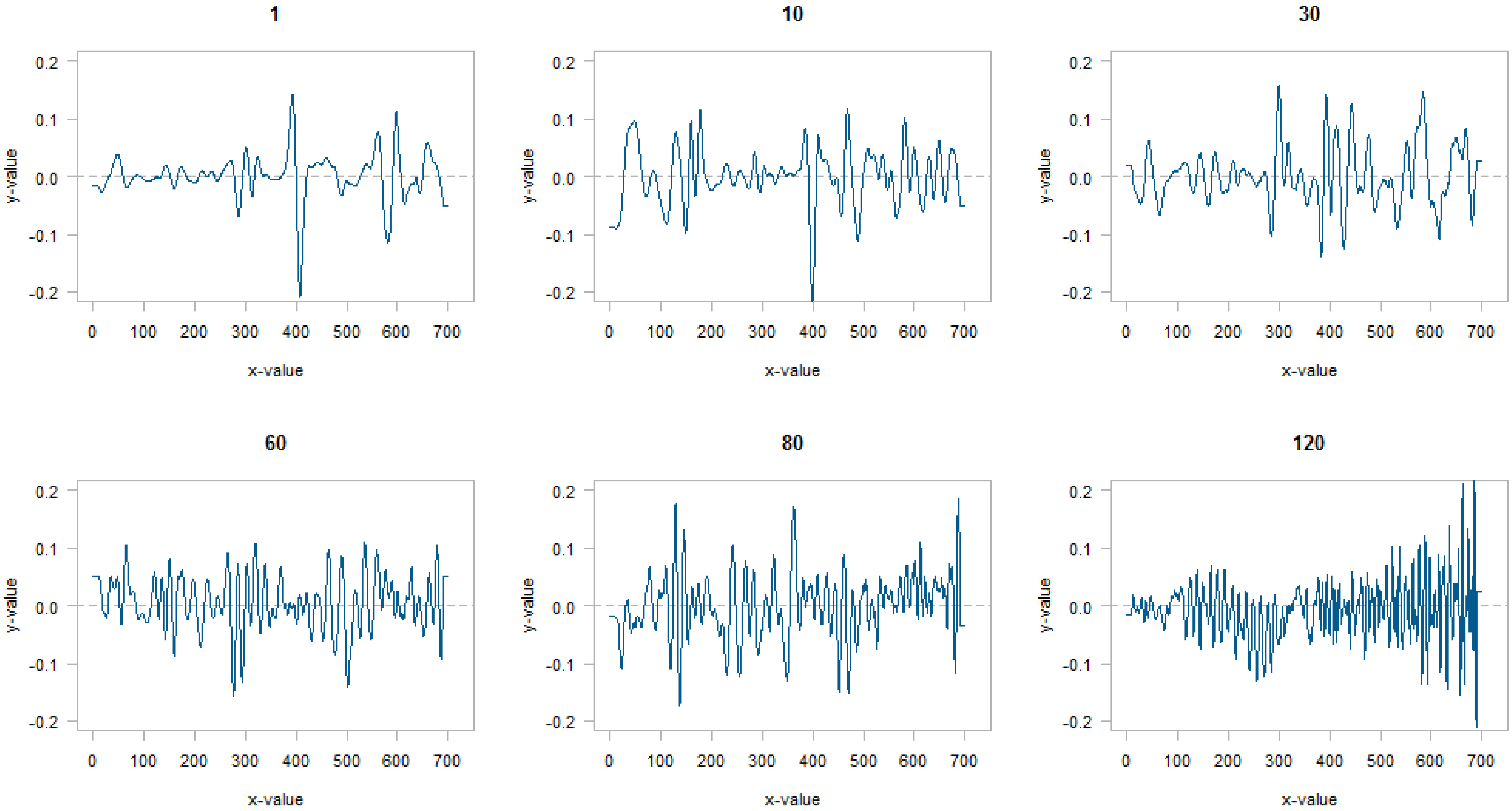
Observed

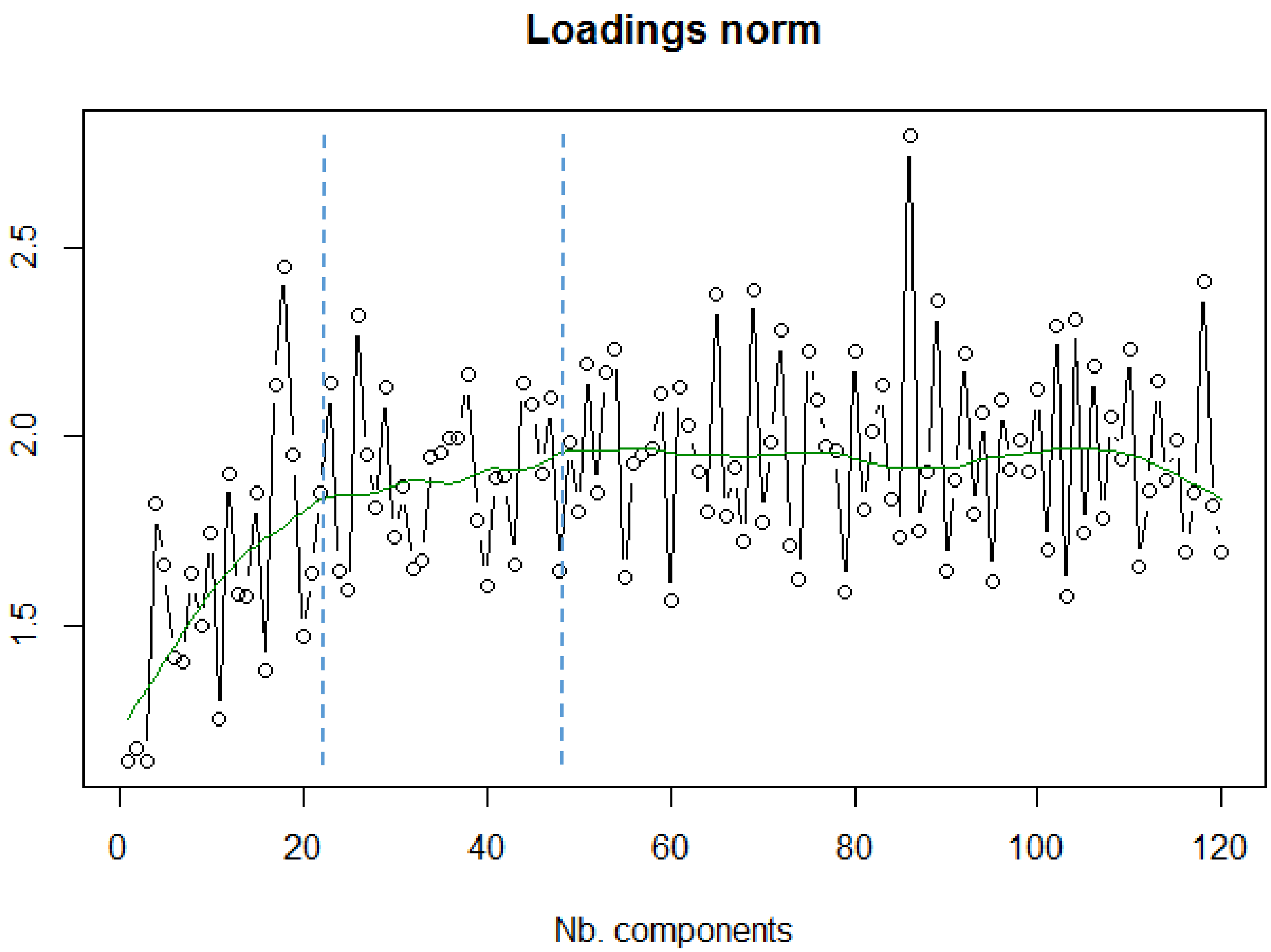
Null distribution

Examination of the PLS loadings and b-coefficients

based on training set

Loadings


1
10
30
60
80
120
0.2
0.1
0.0
-0.1
-0.2
y-value
0
100
200
300
400
500
600
700
x-value


Loadings norm
1.5
2.0
2.5
0
20
40
60
80
100
120
Nb. components

Stability of loadings when bootstrapping

Used in PCA contexts

- **Ye, Z., Weiss, R.E., 2003**. Using the Bootstrap to Select One of a New Class of Dimension Reduction Methods. *Jasa* 98, 968–979. https://doi.org/10.1198/016214503000000927

- **Luo, W., Li, B., 2016**. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika* 103, 875–887. https://doi.org/10.1093/biomet/asw051

- Training $\boldsymbol{\tau}$ → Loadings $\boldsymbol{P}_{\mathbf{obs}}$ $p \times a$ matrix
- Non parametric bootstrap $\boldsymbol{\tau}^{*(b)}$ → $\boldsymbol{P}^{*(\mathrm{b})}$ $p \times a$ matrix

  → Angle ($\boldsymbol{P}_{\mathbf{obs}}$, $\boldsymbol{P}^{*(b)}$)

When the dimension a of $\boldsymbol{P}$ increases,
the last columns of $\boldsymbol{P}^{*}$ become instable with variation of
the bootstraped training $\boldsymbol{\tau}*$ (related to increasing noise)

b = 1, ..., B → Mean of the B angles

The mean angle will tend to increase toward *pi* / 2

How measuring angles between matrices?

- Vector correlation coefficient q (Hotelling 1936) (Ye & Weiss 2003, Luo & Li 2016)
- Maxsub angle (Krzanowski 1979, Hubert et al 2005, Engelen et al. 2005)
- Multivariate coefficient correlation (El Ghaziri, E.M. Qannari 2015)
- Etc.

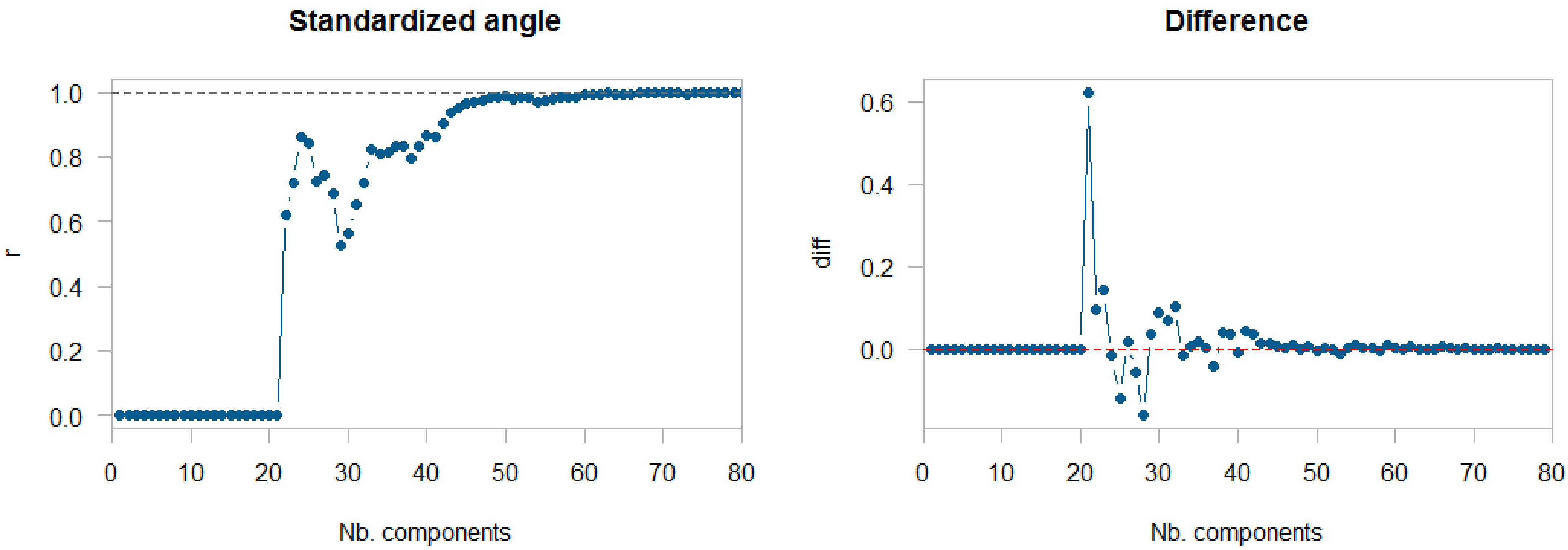
Krzanowski, W.J., 1979. Between-Groups Comparison of Principal Components. *Journal of the American Statistical Association* 74, 703–707. https://doi.org/10.1080/01621459.1979.10481674

Engelen, S., Hubert, M., Branden, K.V., 2005. A Comparison of Three Procedures for Robust PCA in High Dimensions. *Austrian Journal of Statistics* 34, 117–126–117–126. https://doi.org/10.17713/ajs.v34i2.405

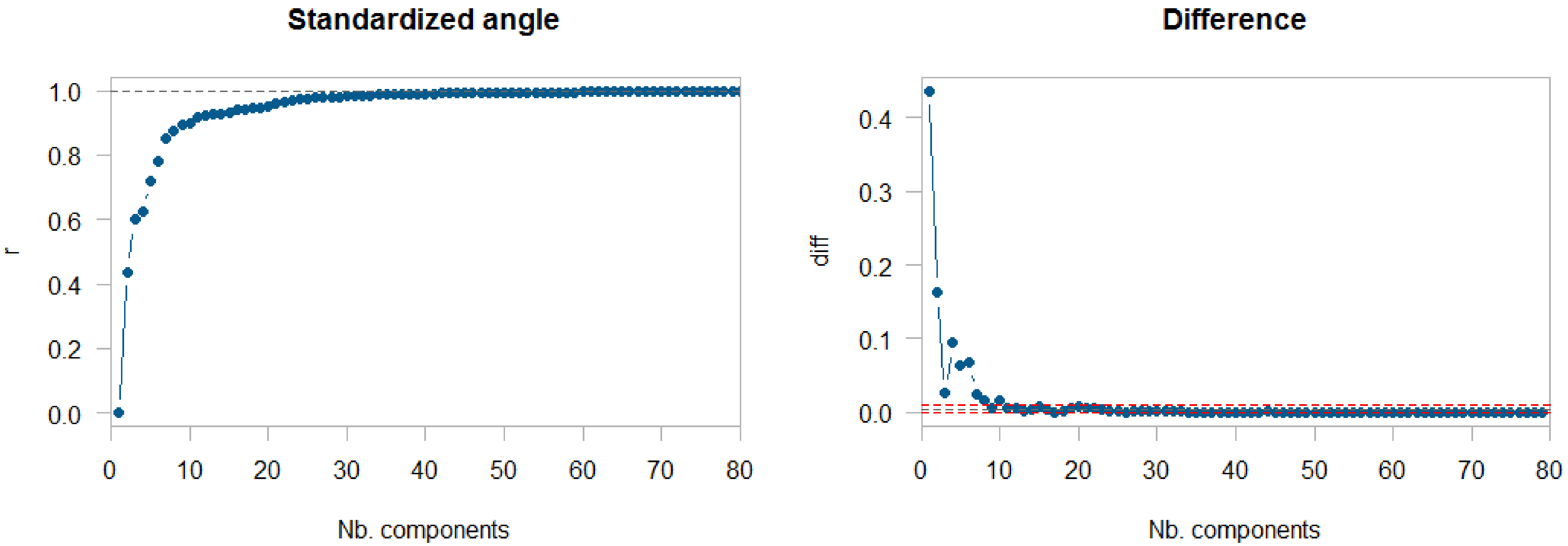
Hubert, M., Rousseeuw, P.J., Vanden Branden, K., 2005. ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics* 47, 64–79. https://doi.org/10.1198/004017004000000563

El Ghaziri, A., Qannari, E.M., 2015. Measures of association between two datasets; Application to sensory data. *Food Quality and Preference* 40, 116–124. https://doi.org/10.1016/j.foodqual.2014.09.010

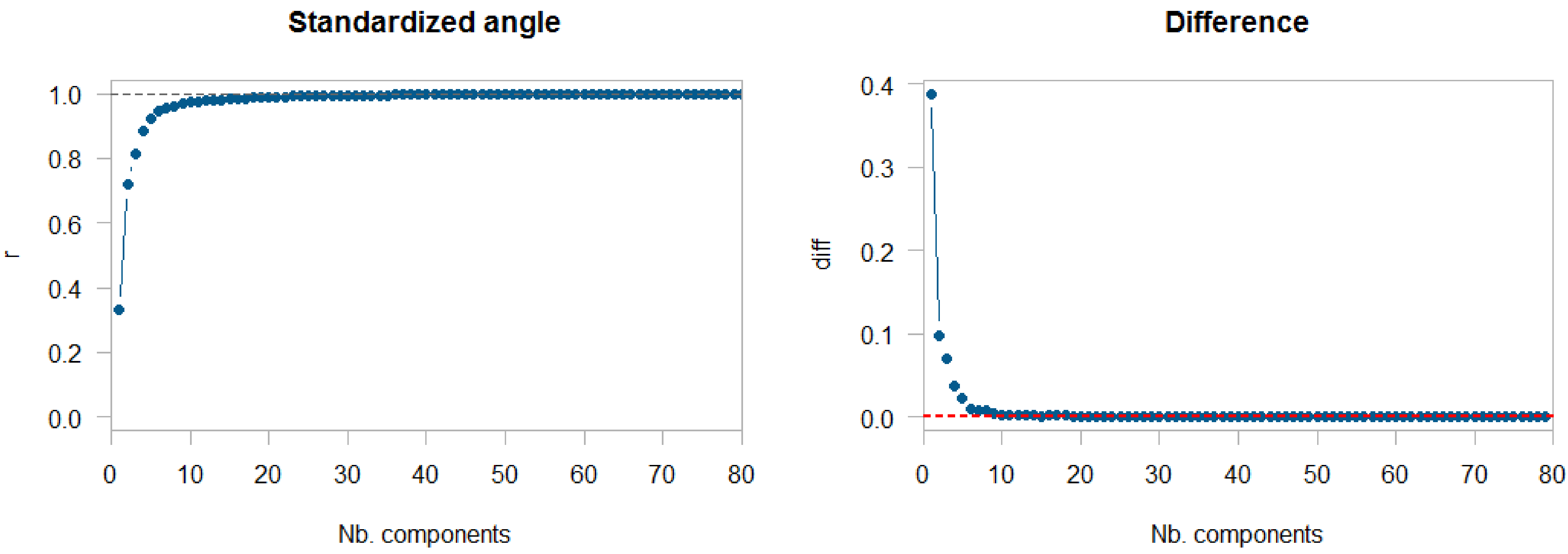
Hotelling q


Standardized angle
r
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
80
Nb. components
Difference
diff
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
80
Nb. components

Maxsub

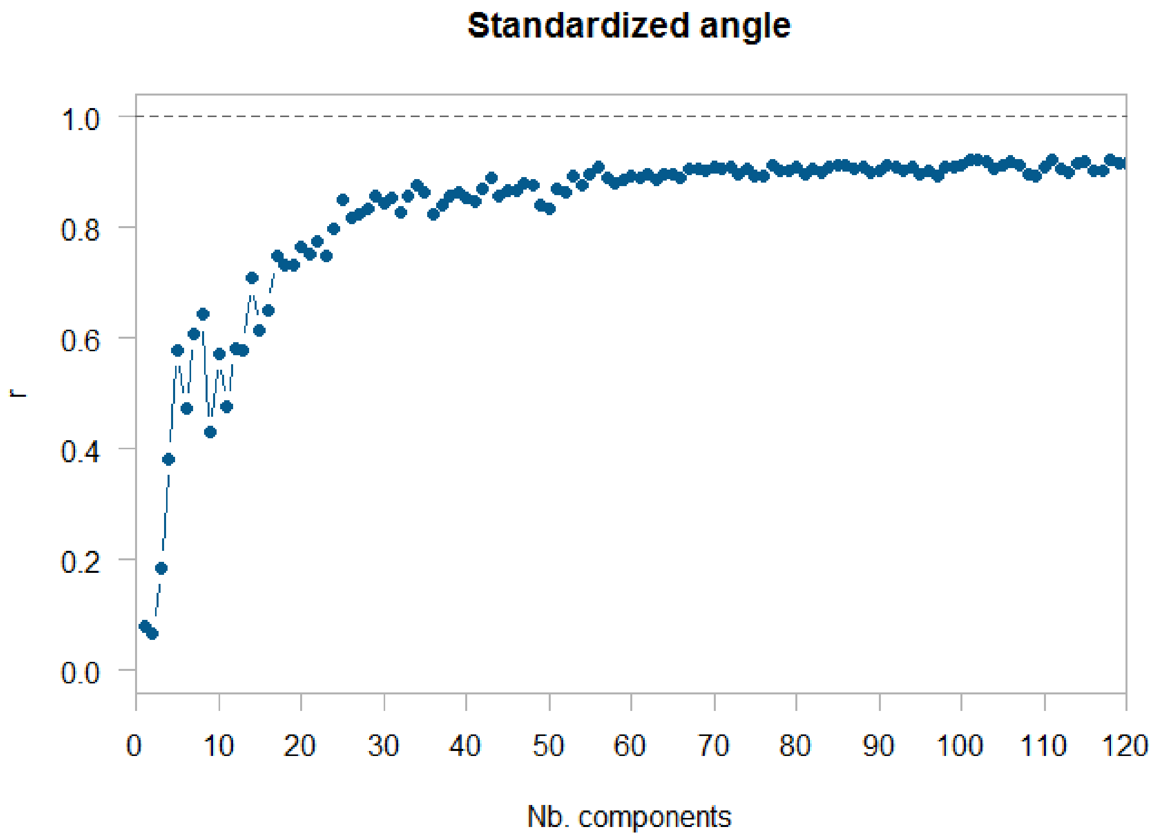

Standardized angle
r
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
80
Nb. components
Difference
diff
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
Nb. components

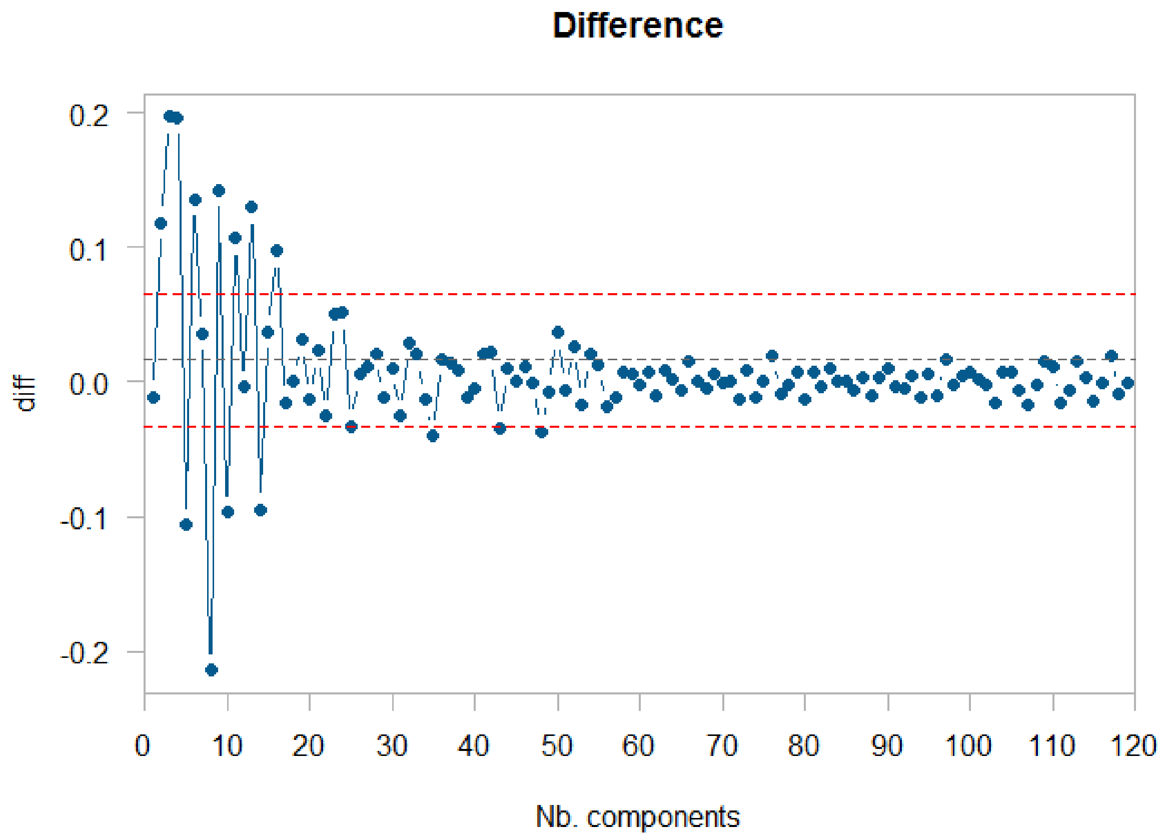
Mult corr


Standardized angle
r
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
80
Nb. components
Difference
diff
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
Nb. components

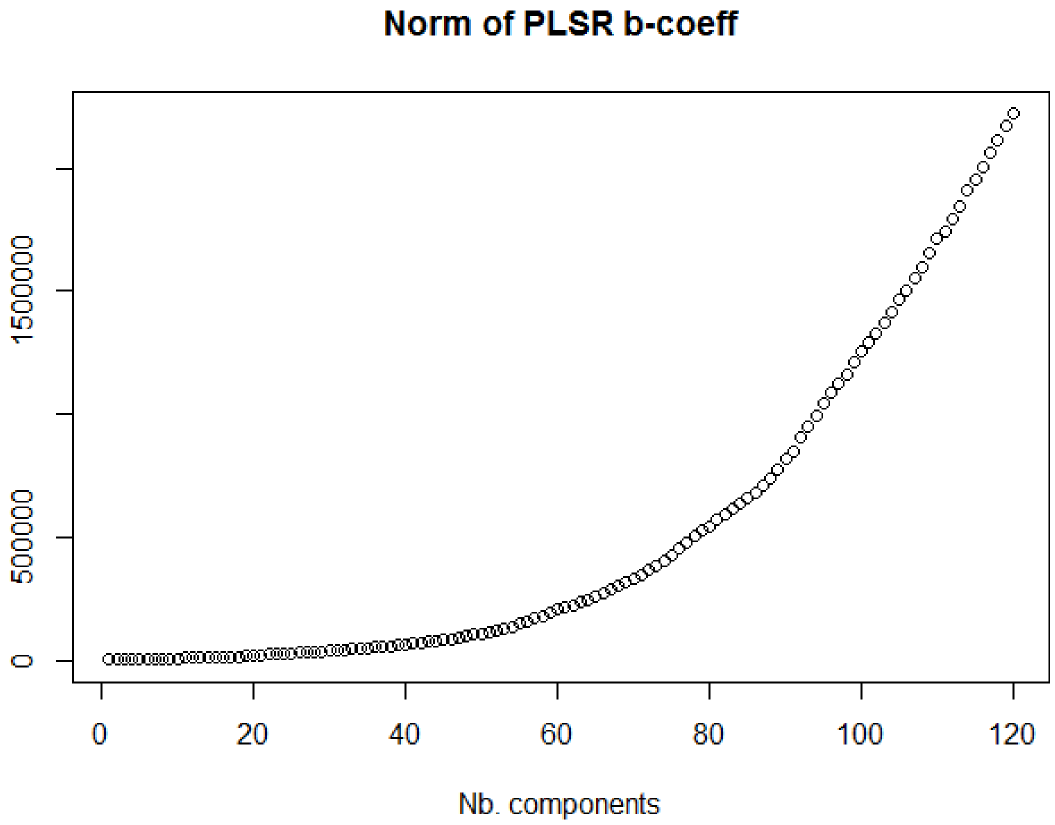
Another possible method!! $\mathrm{SVD}(\,[\boldsymbol{p}_a^{*(1)}, \ldots, \boldsymbol{p}_a^{*(B)}]\,)$ $\lambda_1 / \mathrm{sum}(\boldsymbol{\lambda})$

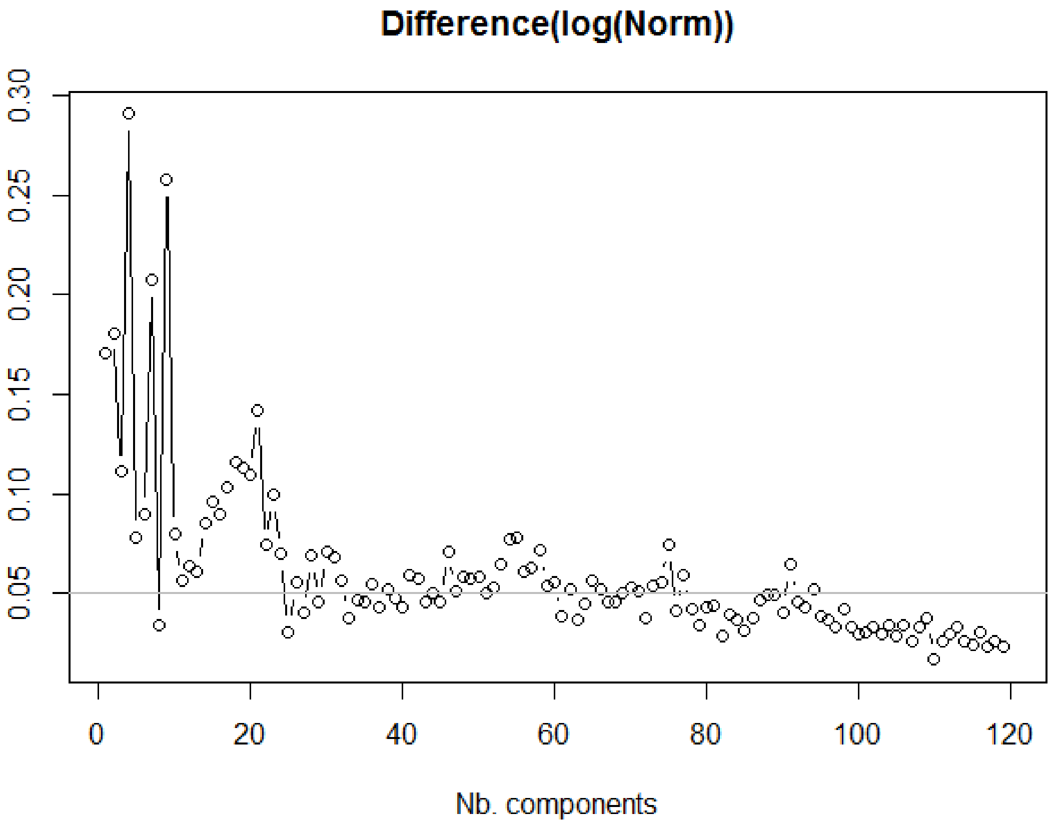
1 – Colinearity index



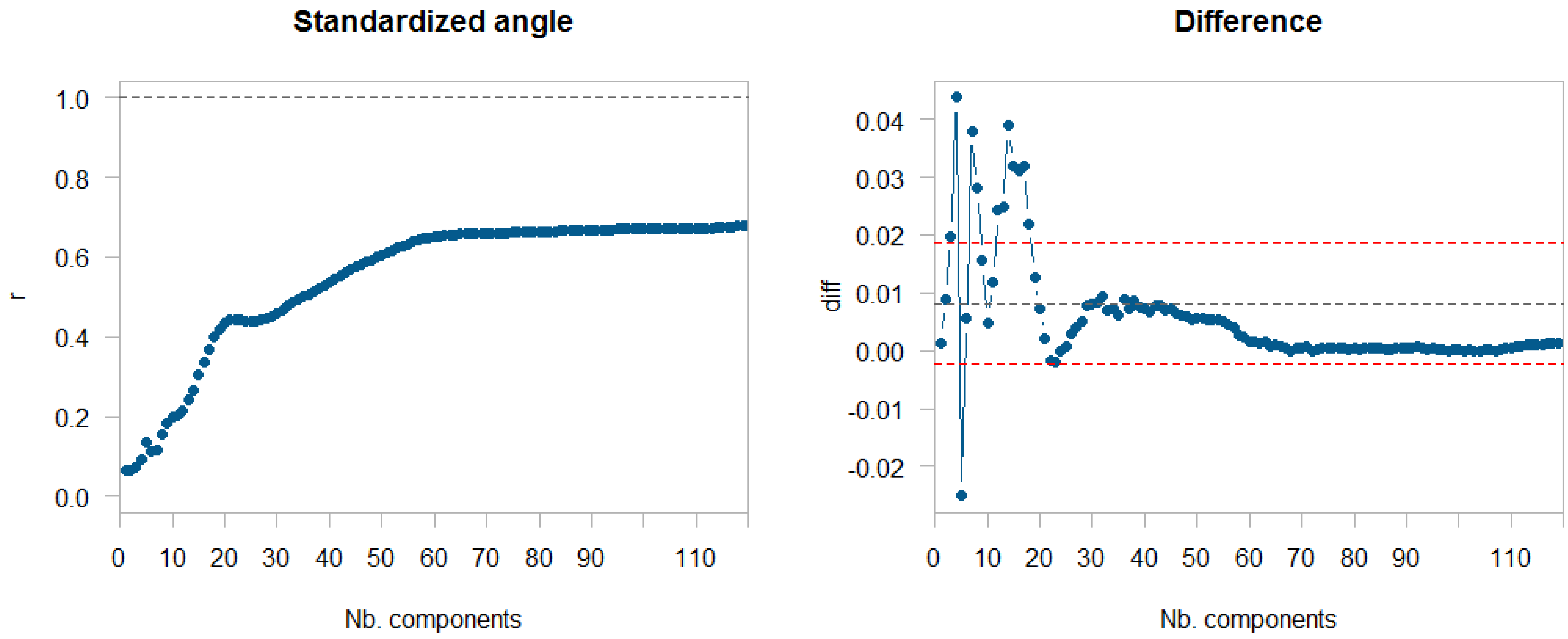


b-Coefficients



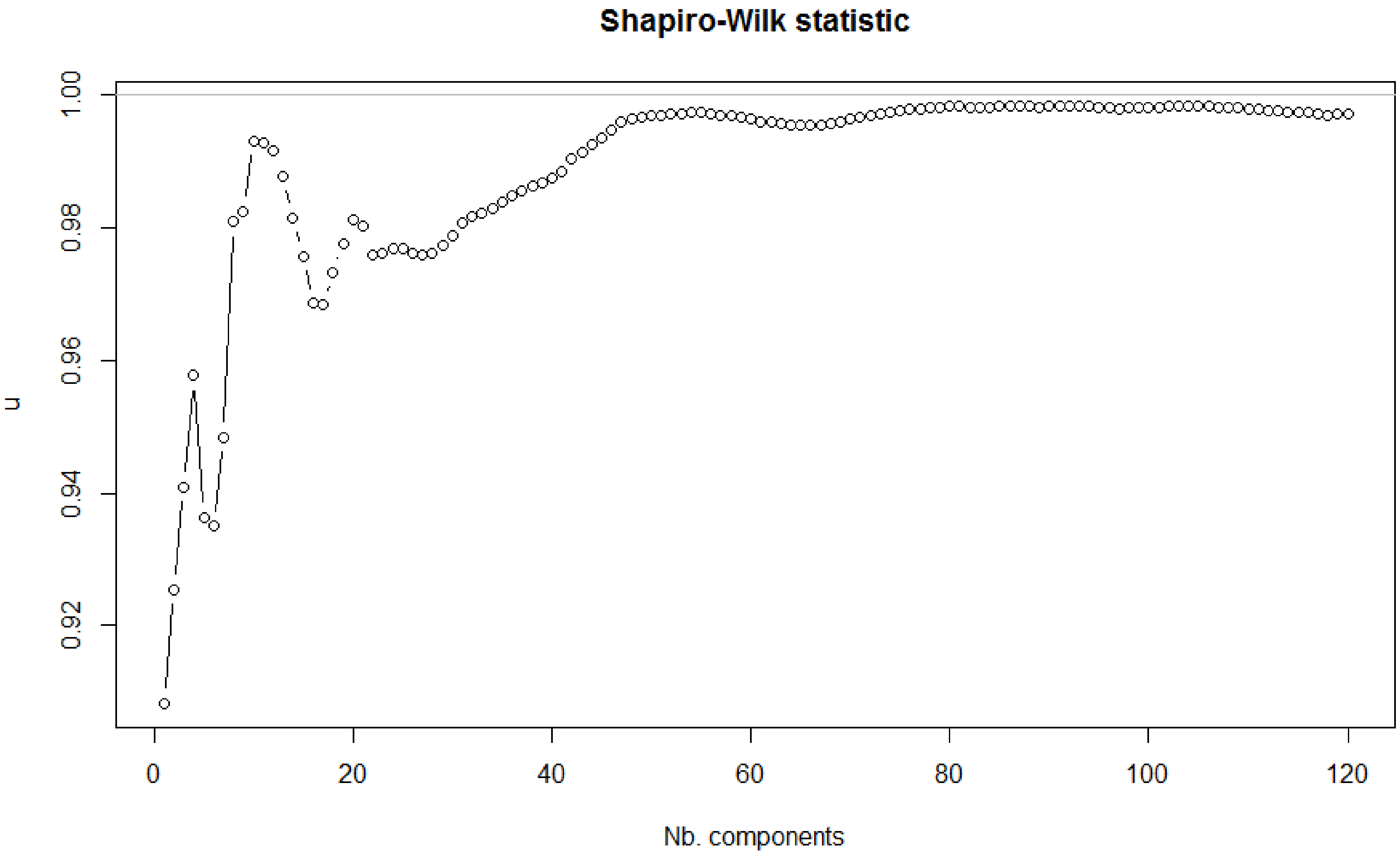


Colinearity method (λ_1 / $\mathrm{sum}(\lambda)$) as previous but now on the b-coefficient vector



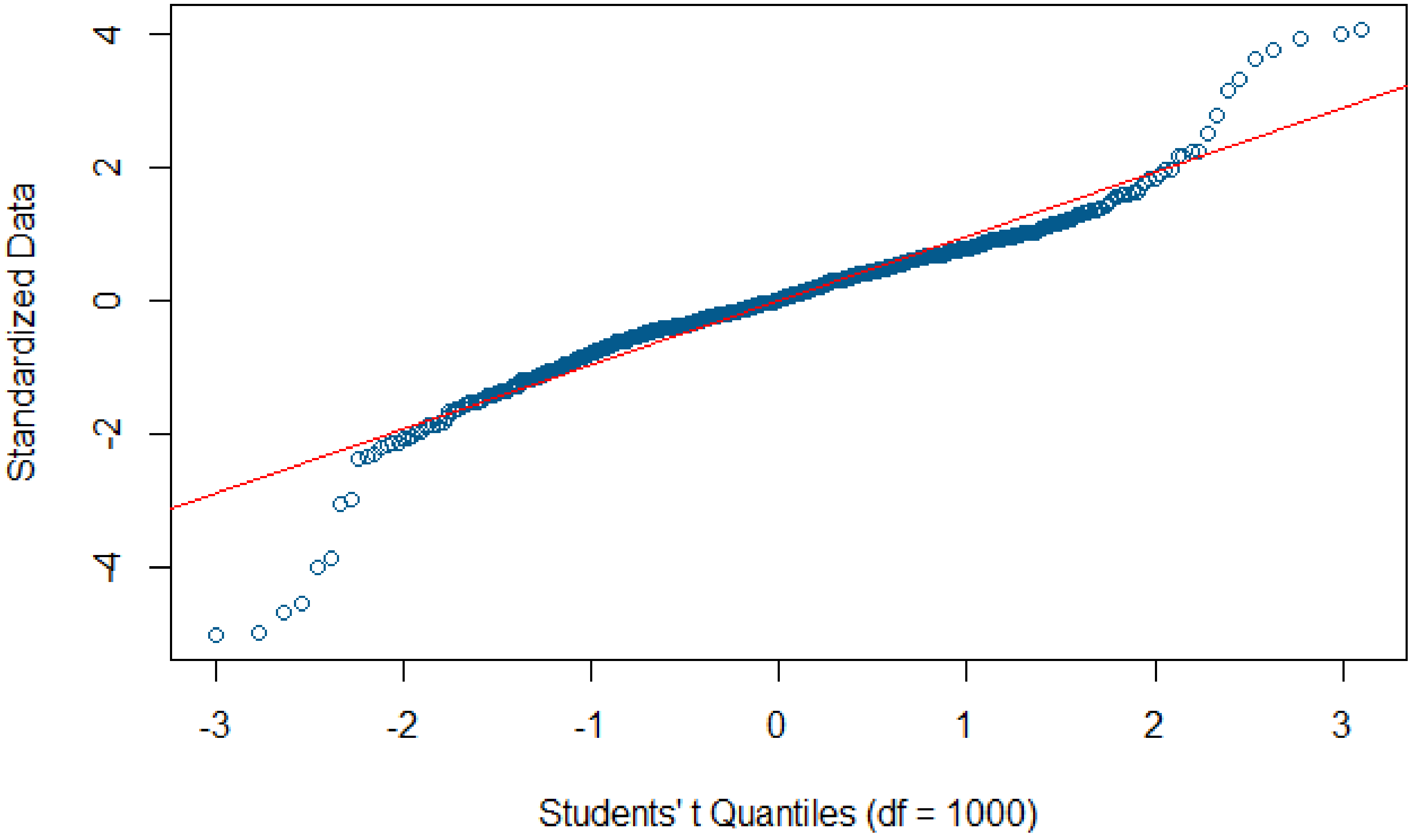
Gaussian b?

Shapiro-Wilk statistic

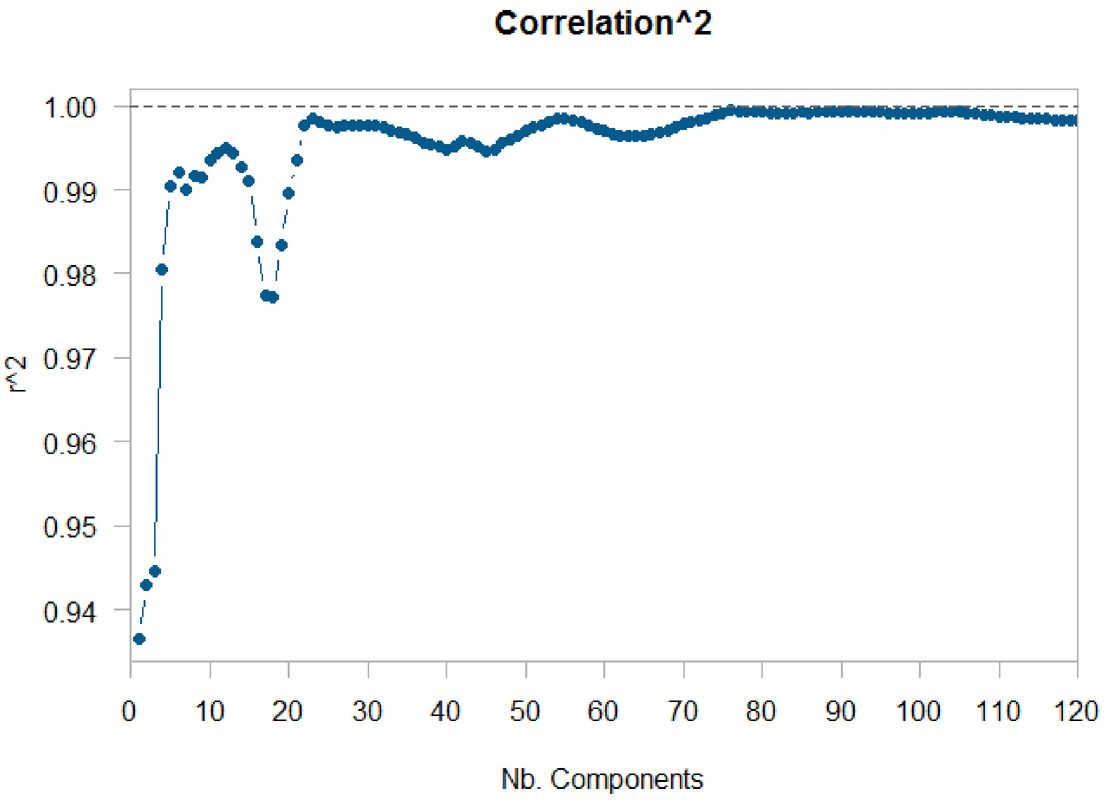


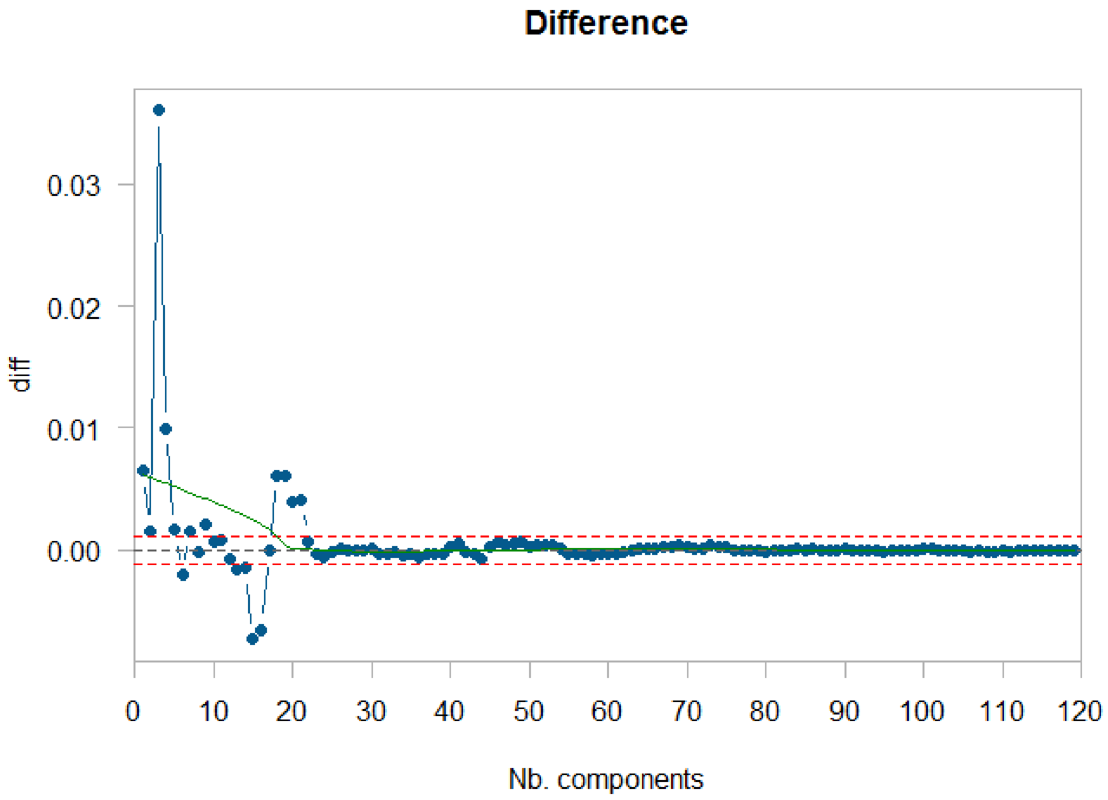
QQplot method

QQnorm plot for $\boldsymbol{b}_{\mathrm{a\,=\,5\;LV}}$

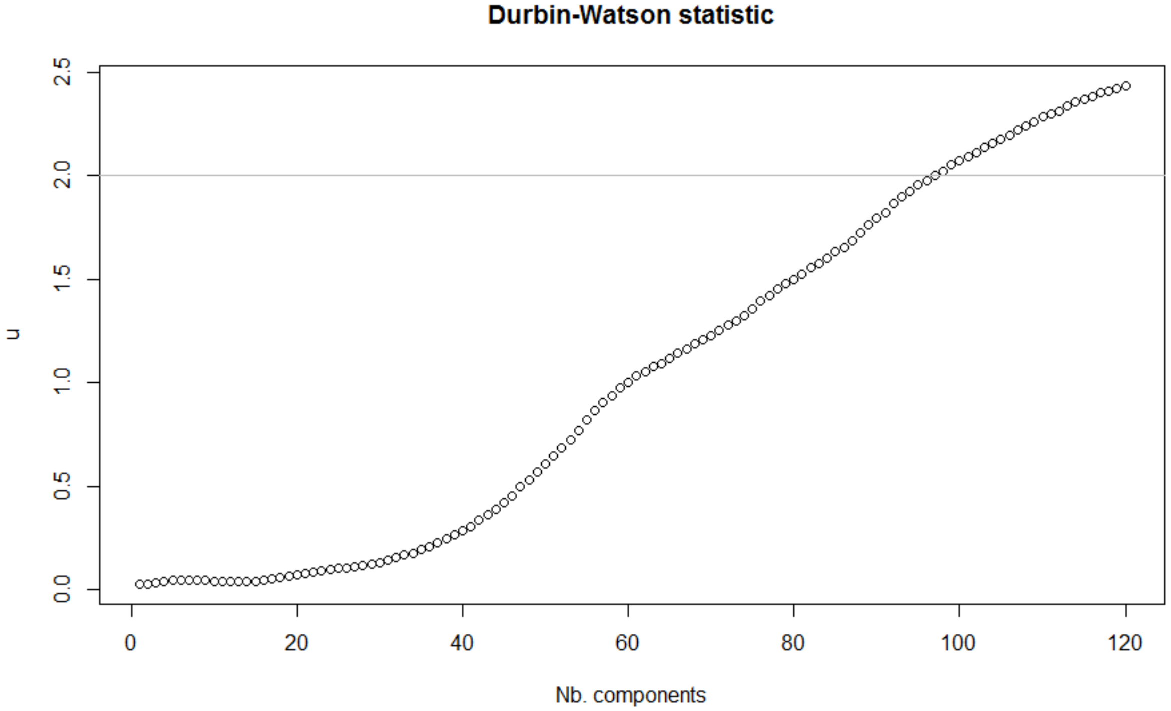


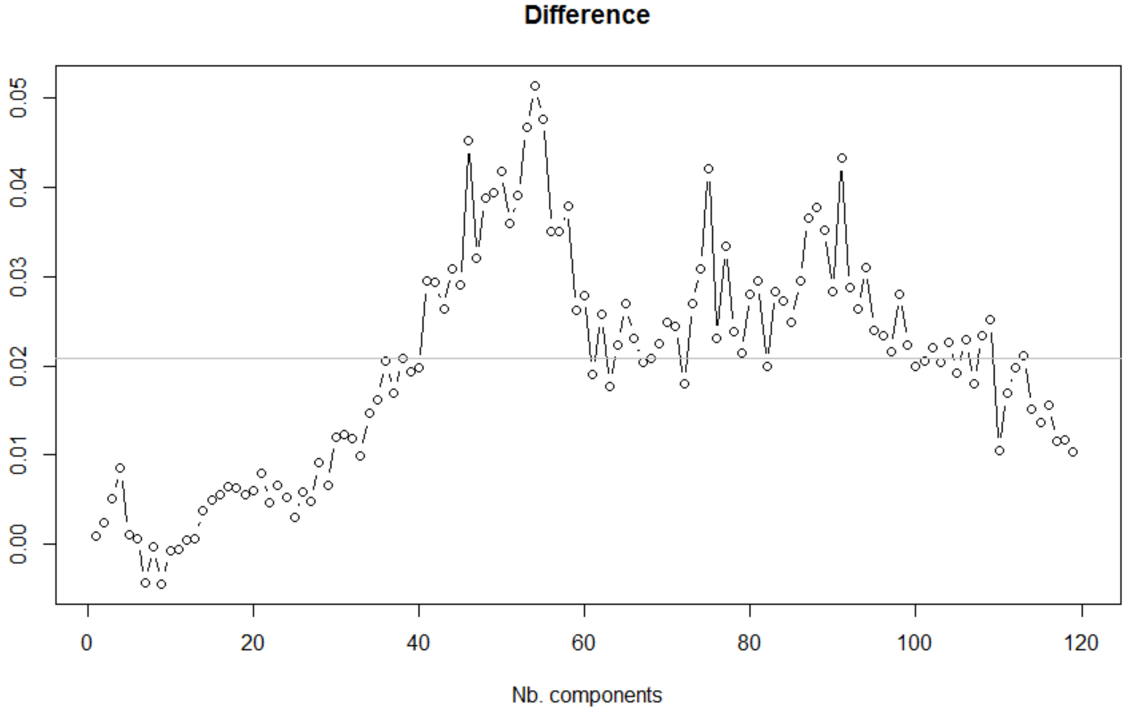
→ Linearity index in QQnorm plots



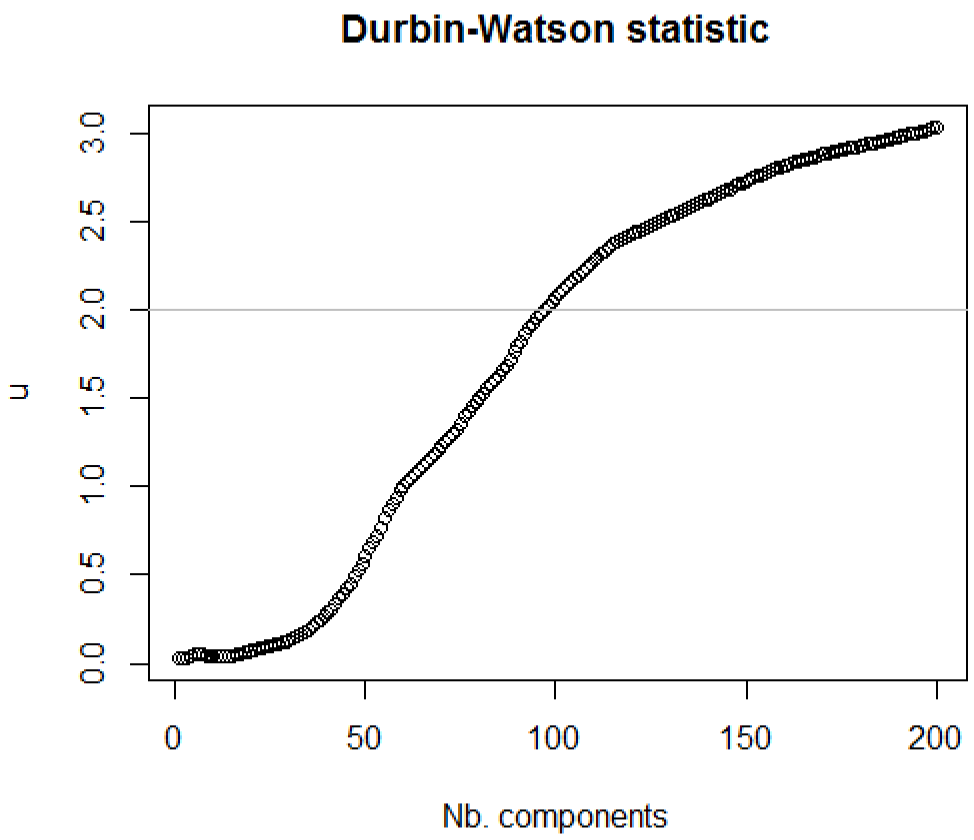


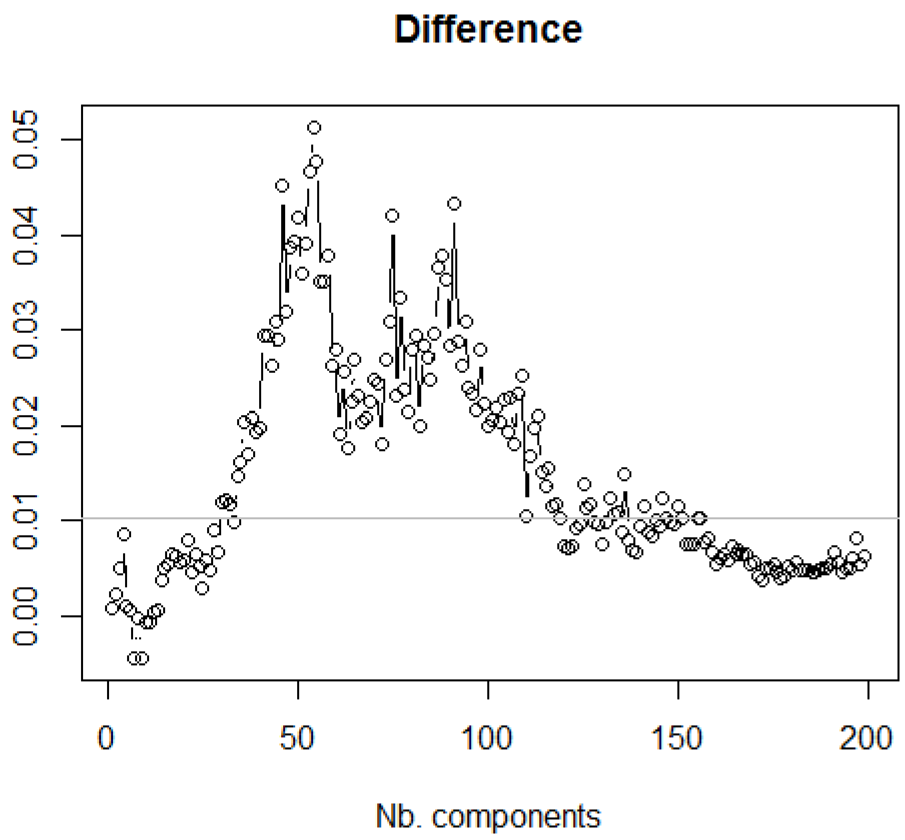
Autocorrelated b? (order 1)





On this example DW seems very not informative

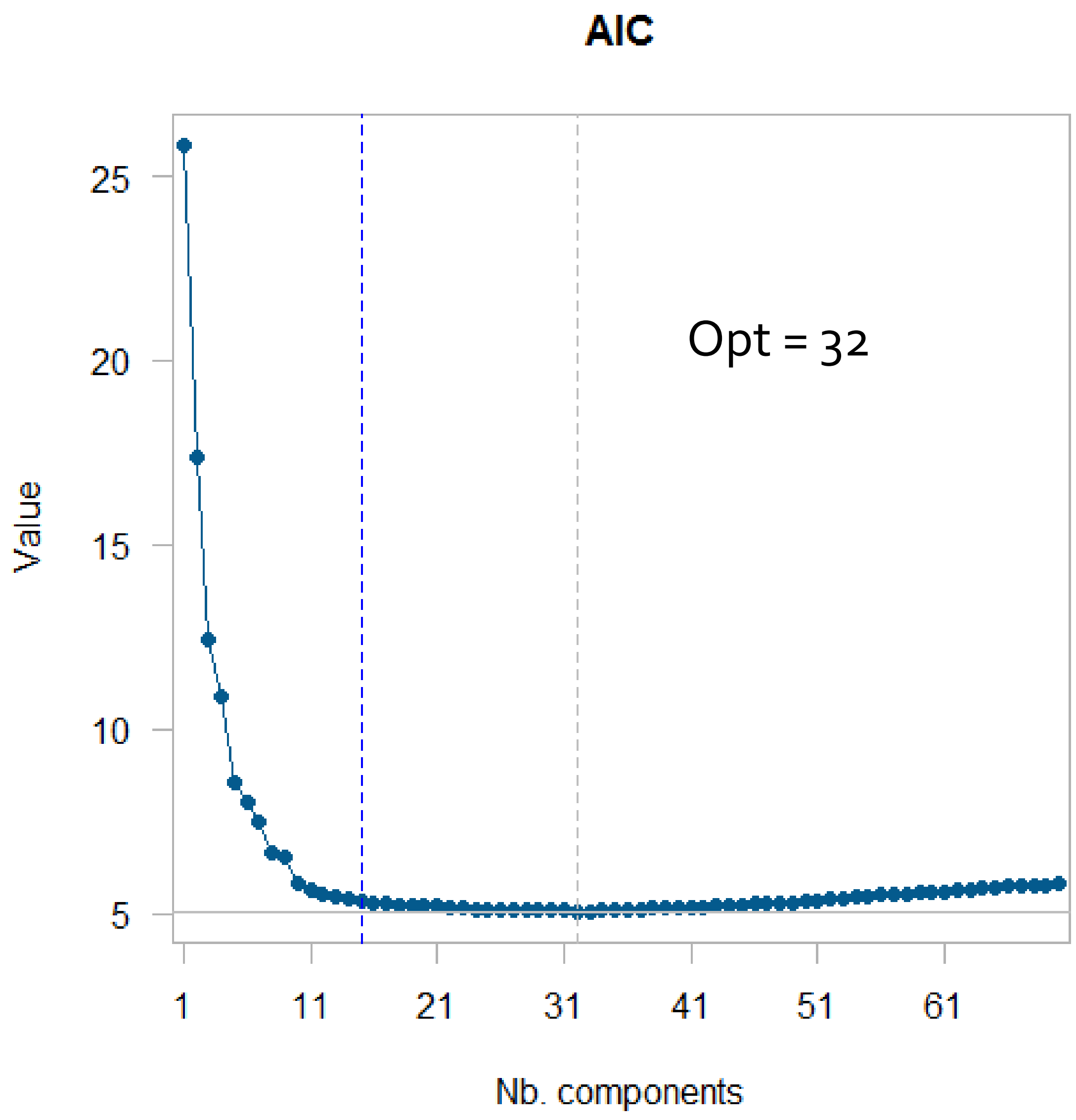




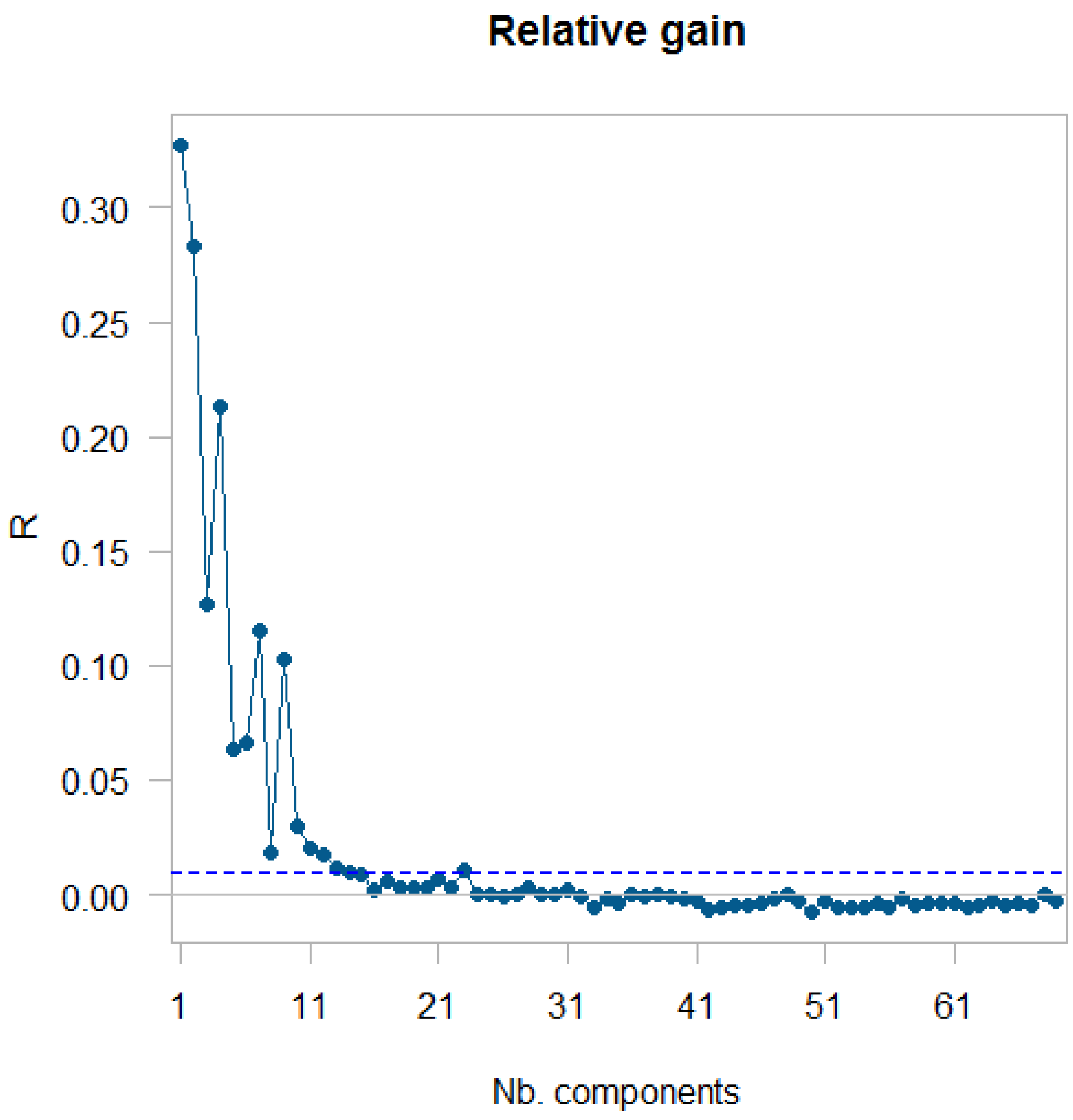
Cp

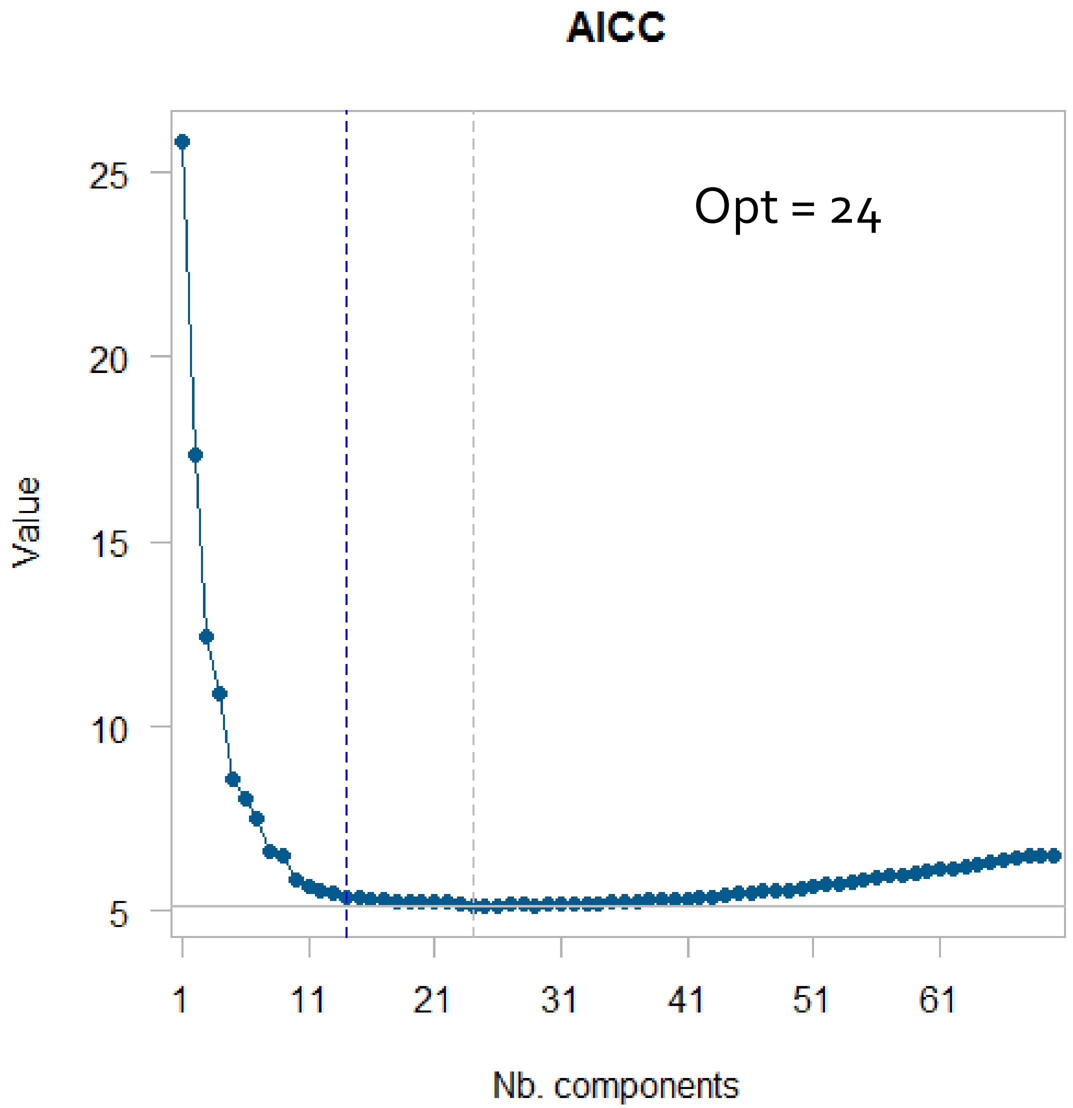
- $\mathrm{Cp} = \dfrac{\mathrm{SSR}}{n} + \frac{2}{n}\sum_{i=1}^{n} \hat{C}ov_{\varepsilon}(y_i, \hat{y}_i)$

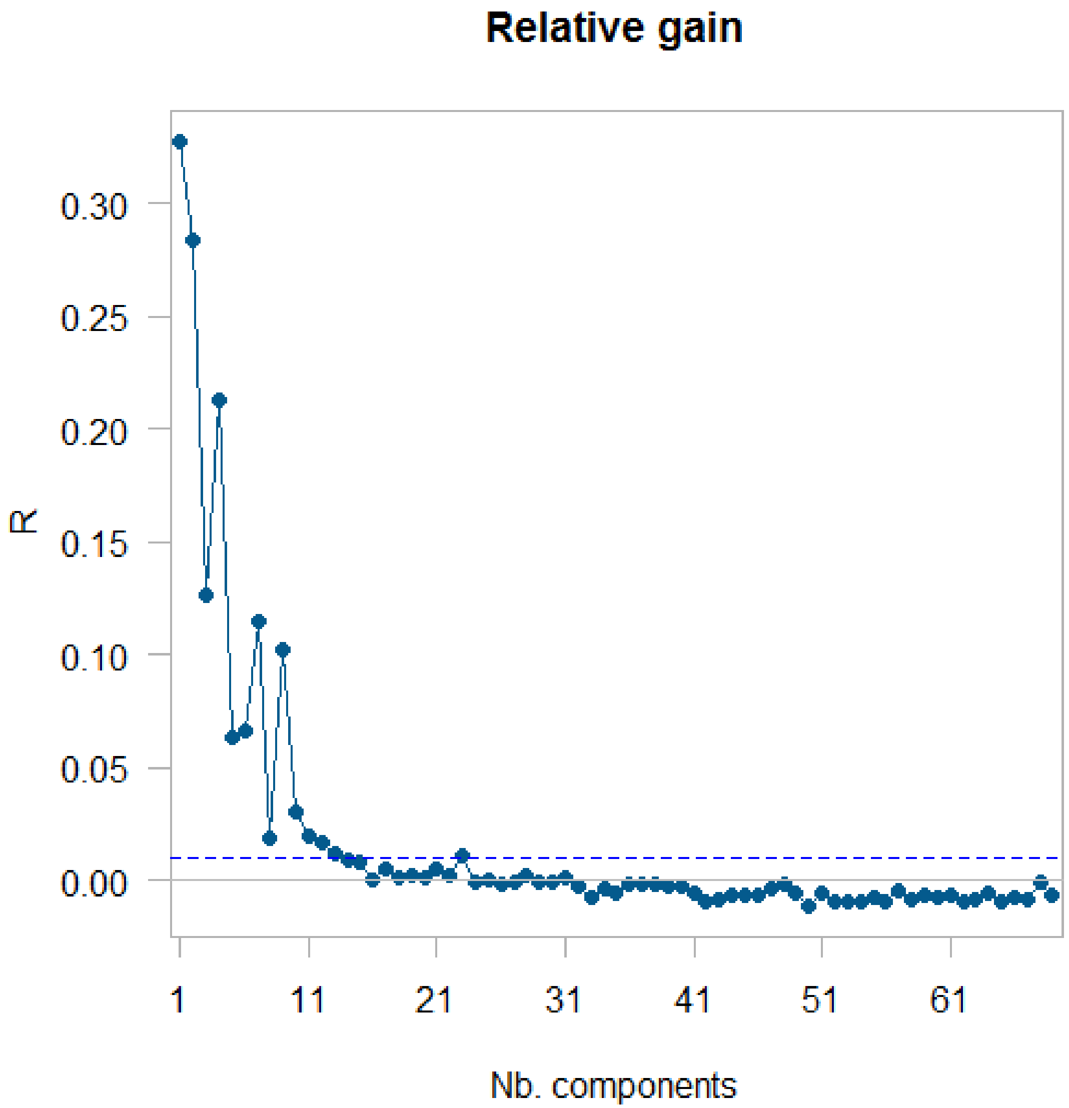
30 replications bootsrap

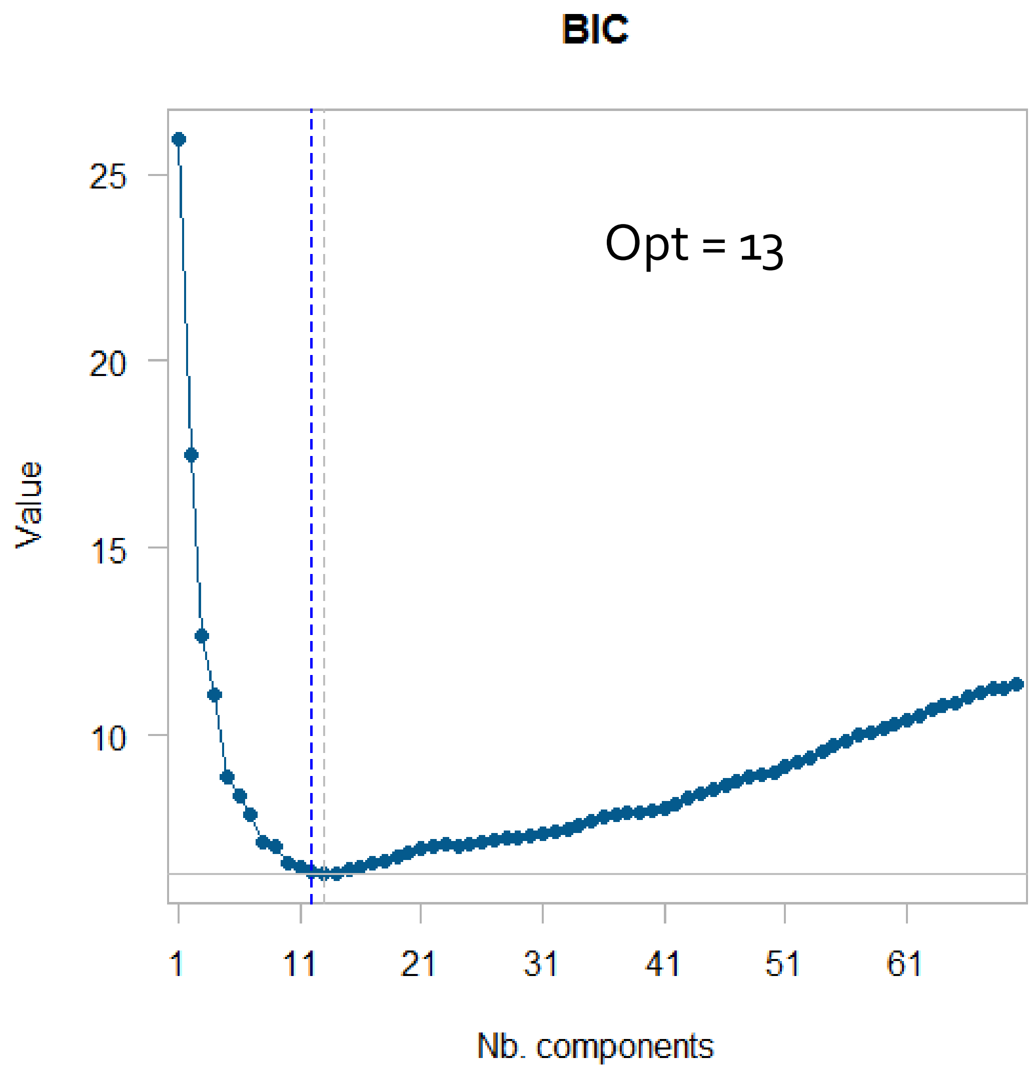

AIC
Value
25
20
15
10
5
1
11
21
31
41
51
61
Nb. components

Opt = 32


Relative gain
R
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1
11
21
31
41
51
61
Nb. components

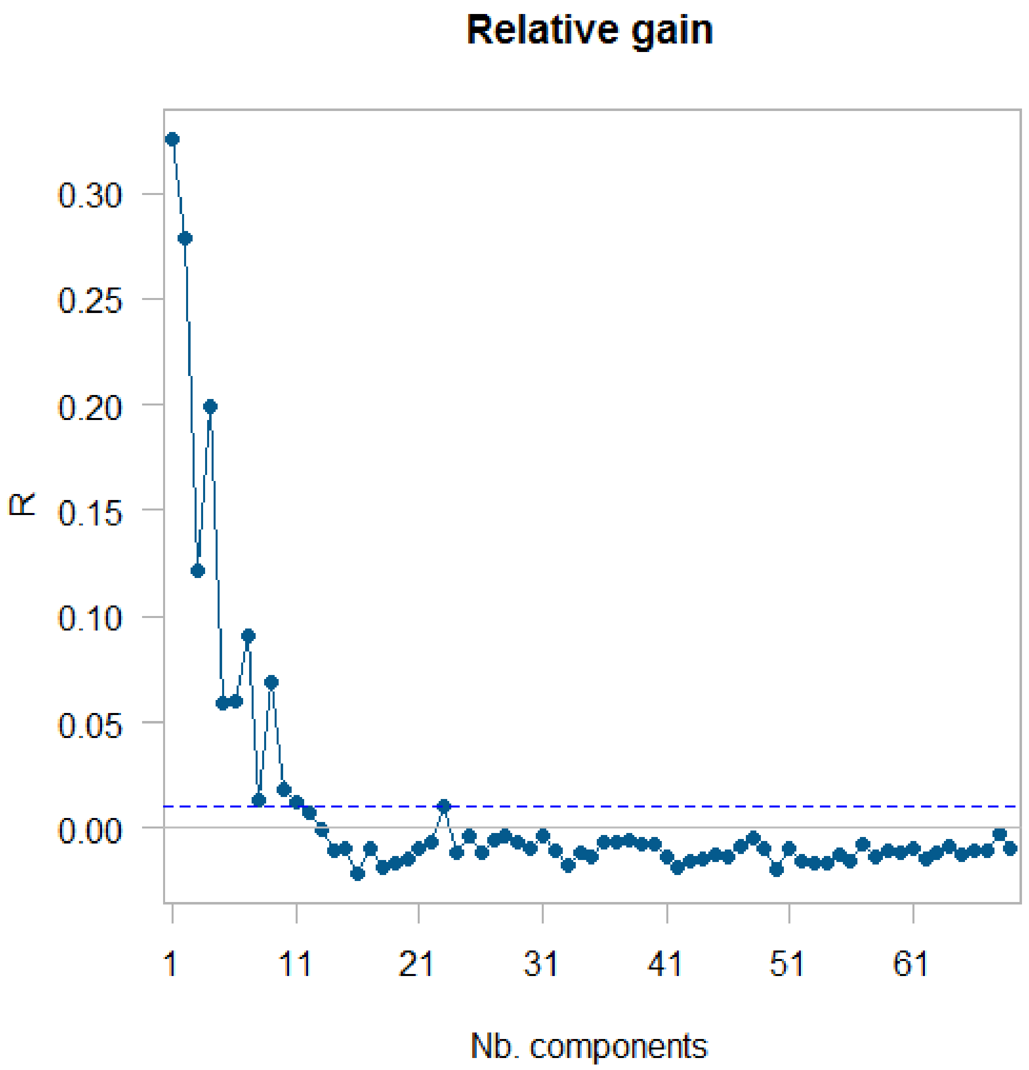

AICC
Opt = 24
Value
25
20
15
10
5
1
11
21
31
41
51
61
Nb. components


Relative gain
R
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1
11
21
31
41
51
61
Nb. components

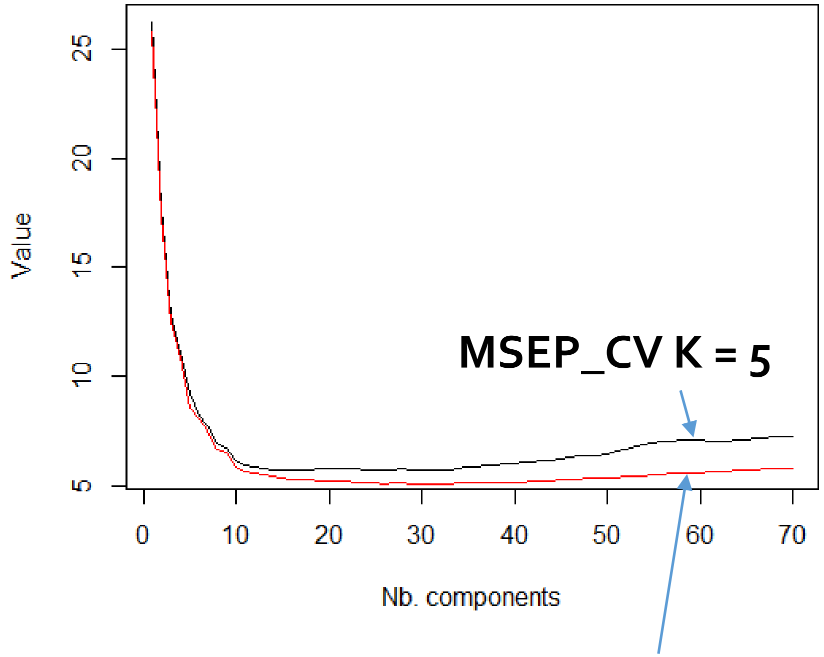


BIC

Opt = 13

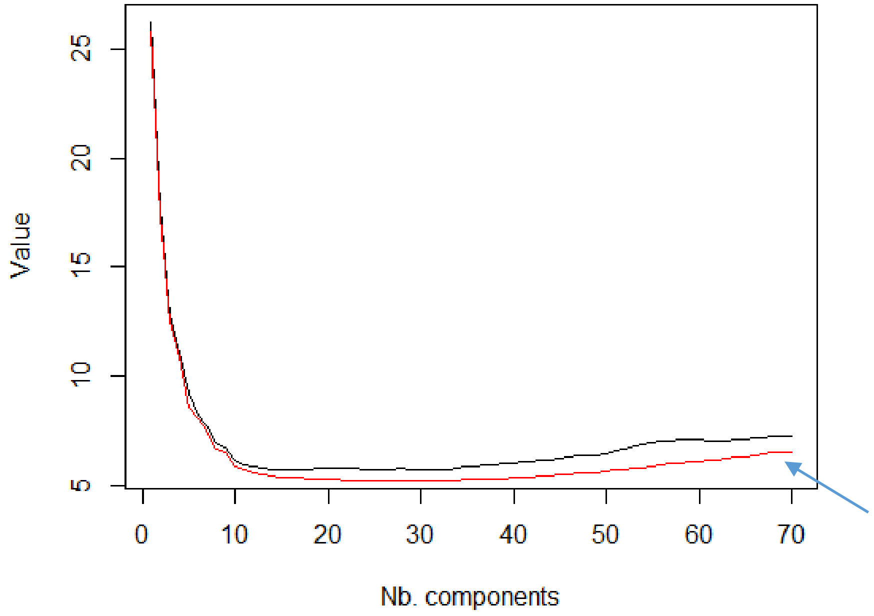


Relative gain

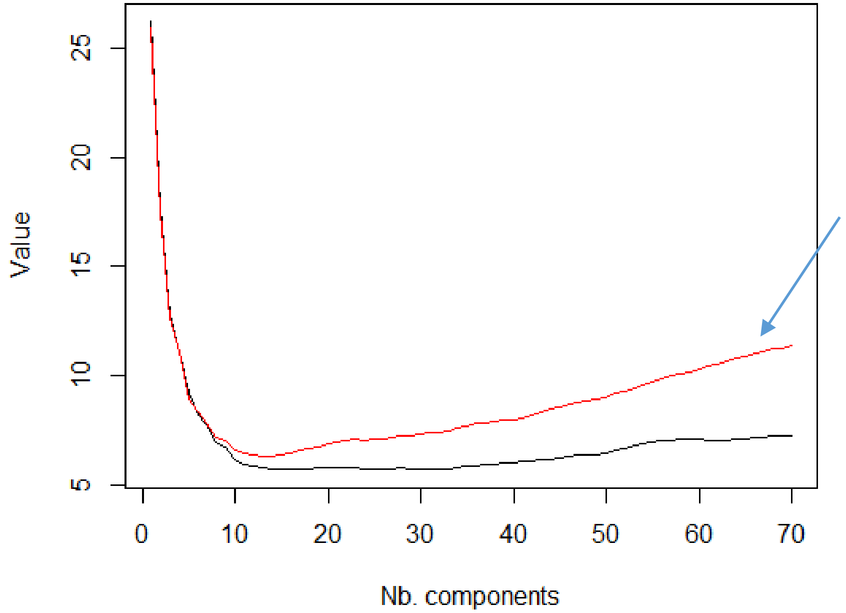


MSEP_CV K = 5

AIC



AICc



BIC

- CV
  - LOO opt = 32 Wold1% = 8
  - $K = 10$ opt = 32 Wold1% = 13
  - $K = 5$ opt = 32 Wold1% = 13
  - $K = 2$ opt = 13 Wold1% = 13

- Covariance penalty Cp
  - AIC opt = 32 Wold1% = 13
  - AICc opt = 24 Wold1% = 14
  - BIC opt = 32 Wold1% = 12

Remark Weights for model averaging (stacking)

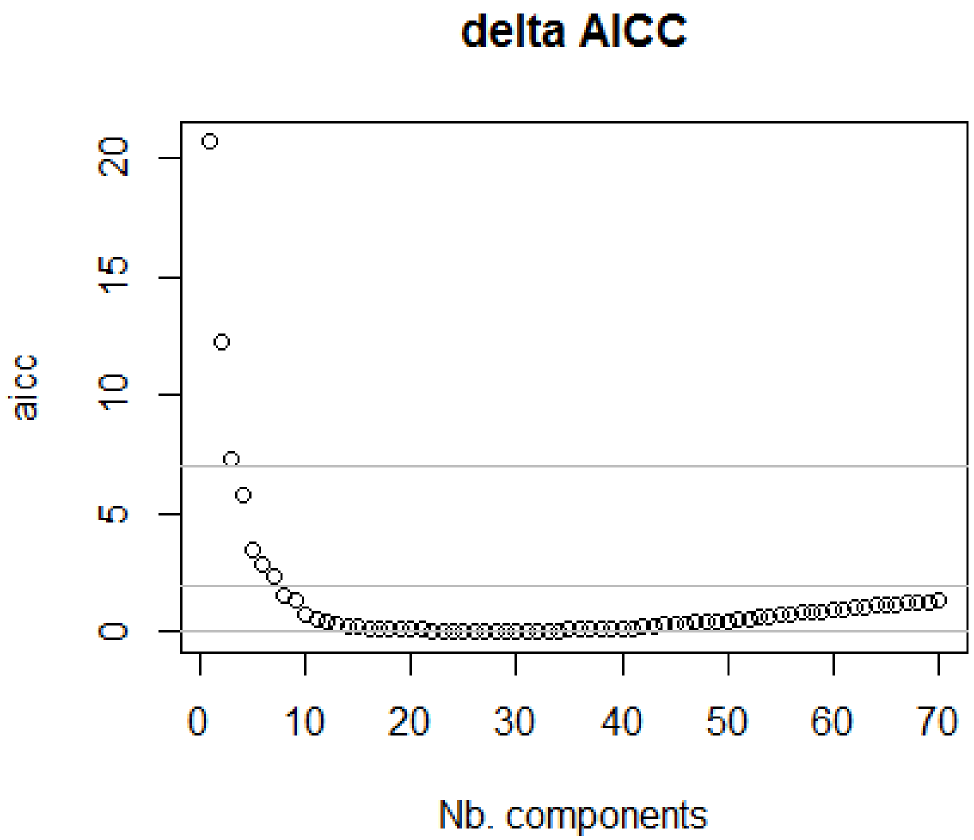
AIC weights (e.g. Burnham & Anderson 2002, 2004; Hastie et al 2009)

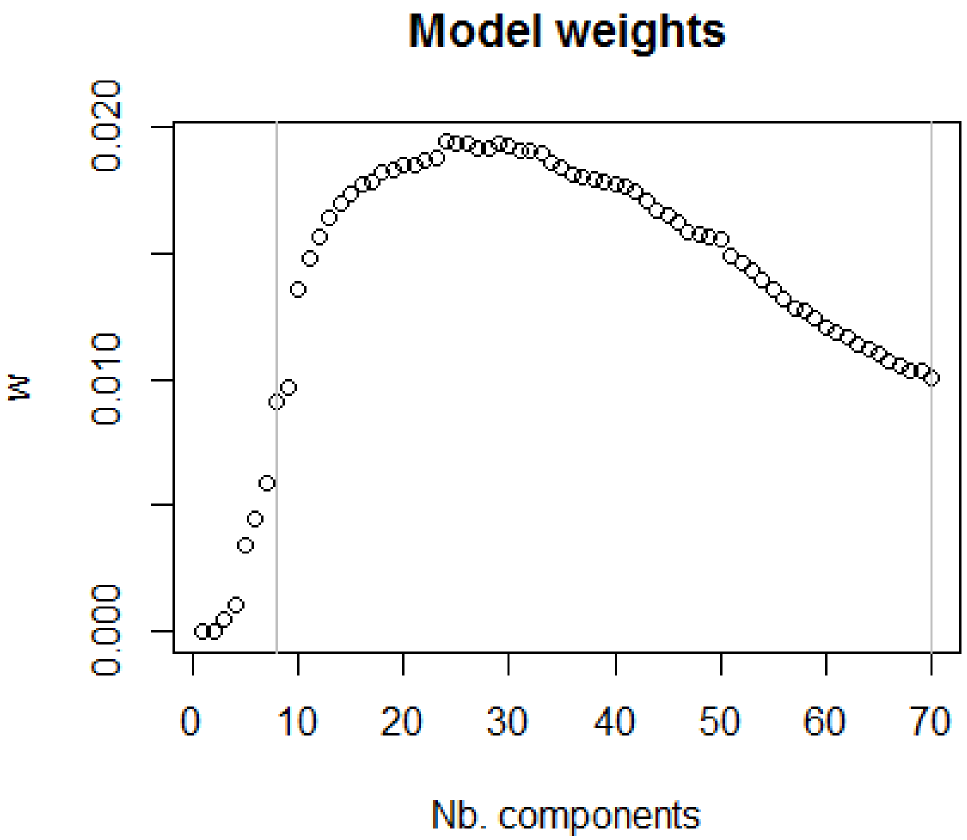
- $\mathit{\Delta}\mathrm{AIC}_a = \mathrm{AIC}_a - \min(\mathrm{AIC})$

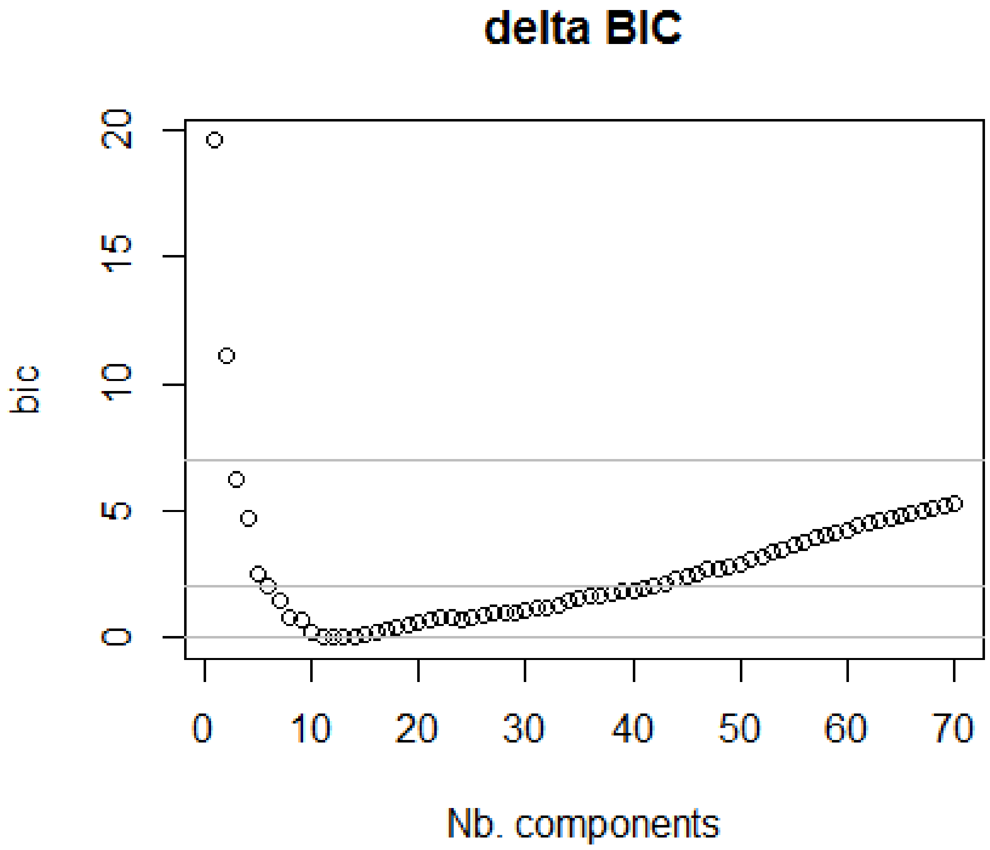
- $w_a = \exp(-.5\ \mathit{\Delta}\mathrm{AIC}_a)\ /\ \mathrm{sum}(\ \exp(-.5\ \mathit{\Delta}\mathrm{AIC}_a))$

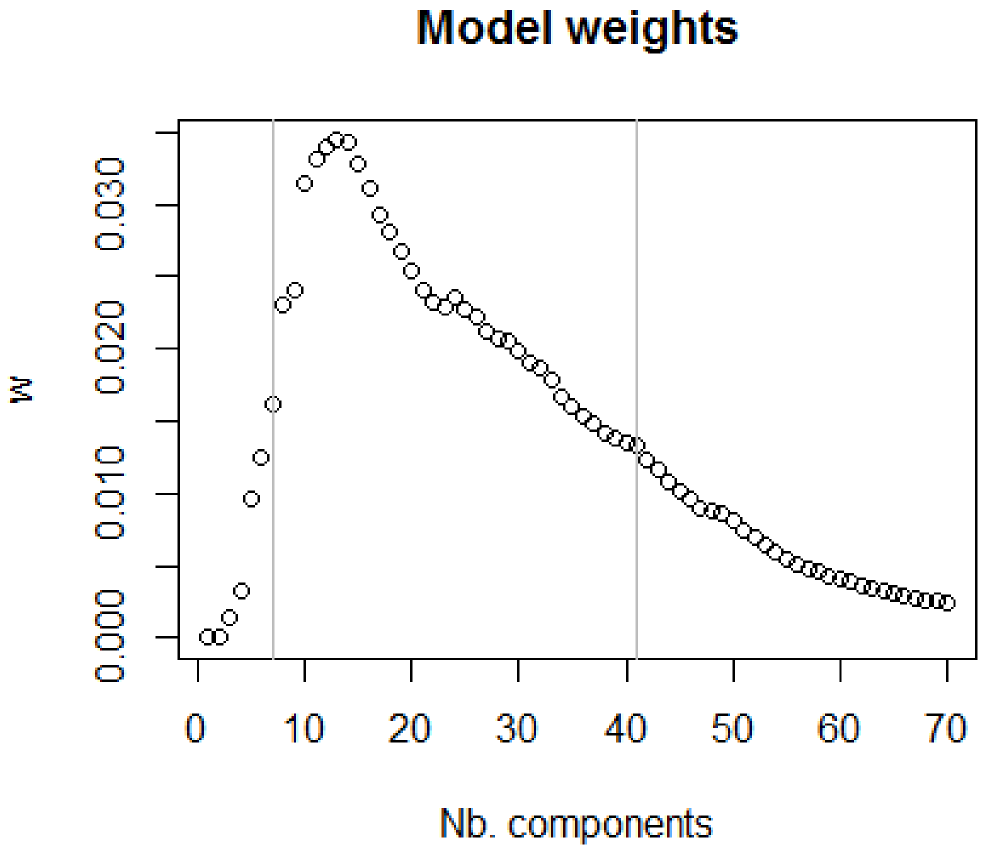
Same for AICc, BIC, $\mathrm{MSEP}_{\mathrm{CV}}$, etc.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed. *Springer*, New York, NY, USA.
Burnham, K.P., Anderson, D.R., 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* 33, 261–304. https://doi.org/10.1177/0049124104268644

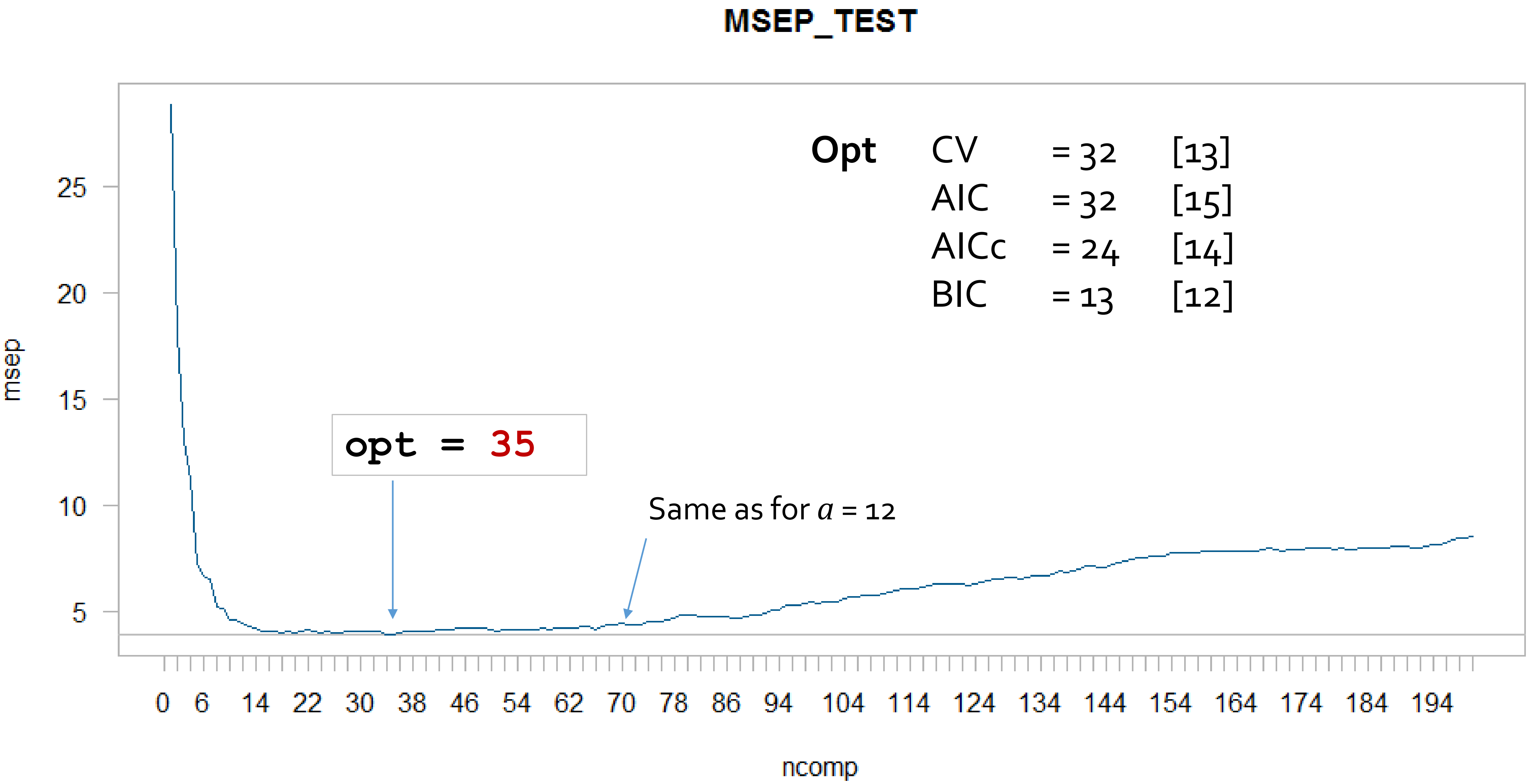

delta AICC
aicc
0
5
10
15
20
0
10
20
30
40
50
60
70
Nb. components


Model weights
w
0.000
0.010
0.020
0
10
20
30
40
50
60
70
Nb. components


delta BIC
bic
0
5
10
15
20
0
10
20
30
40
50
60
70
Nb. components


Model weights
w
0.000
0.010
0.020
0.030
0
10
20
30
40
50
60
70
Nb. components

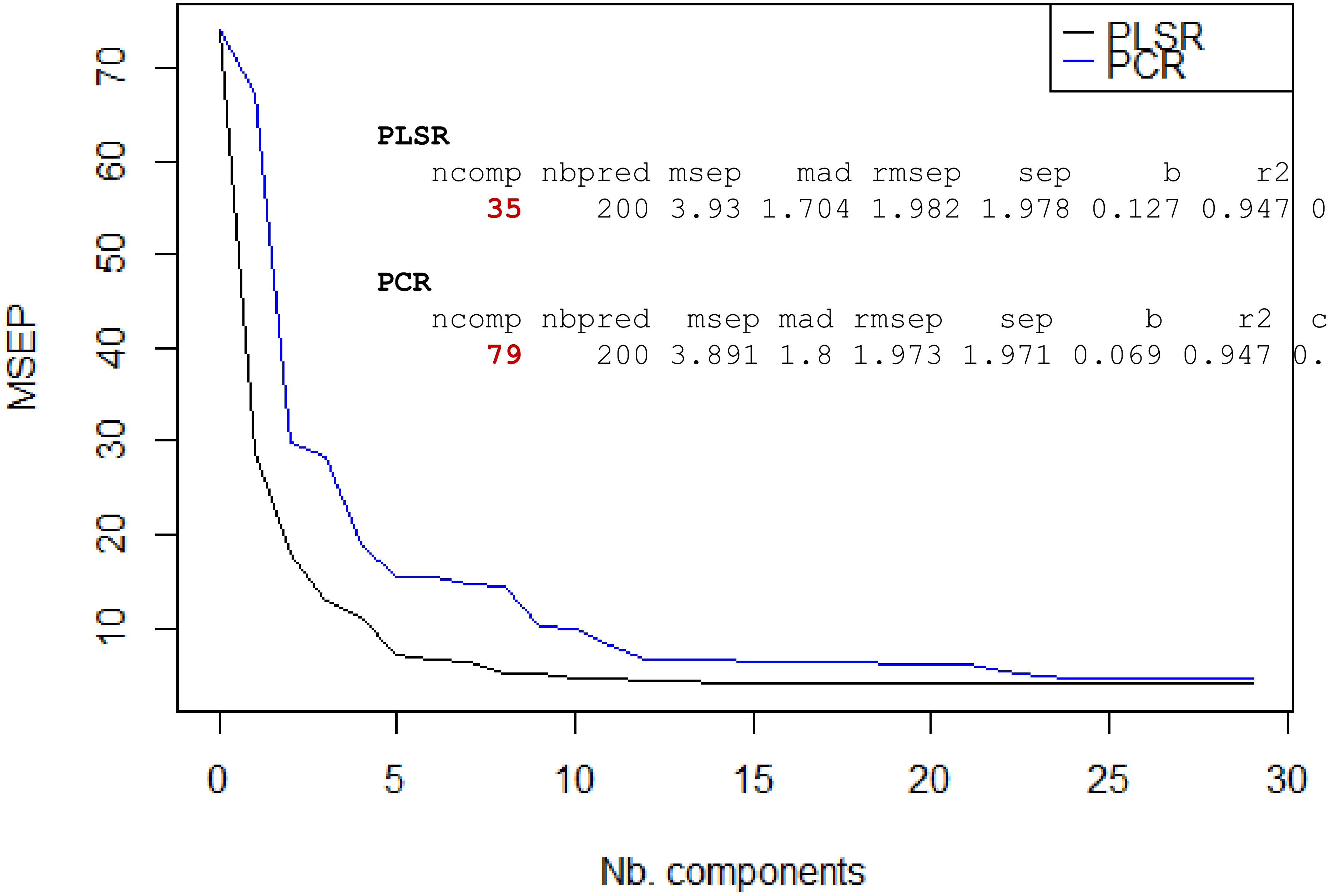
Results on the test set



Opt	CV	= 32	[13]
	AIC	= 32	[15]
	AICc	= 24	[14]
	BIC	= 13	[12]

`opt = 35`

Same as for a = 12



```
PLSR
   ncomp nbpred msep   mad rmsep   sep     b    r2  cor2
      35    200 3.93 1.704 1.982 1.978 0.127 0.947 0.947

PCR
   ncomp nbpred  msep mad rmsep   sep     b    r2  cor2
      79    200 3.891 1.8 1.973 1.971 0.069 0.947 0.947
```

Sensitivity to the test set

Hypothesis Future $\neq$ Mechanism (probability distribution) generating the training

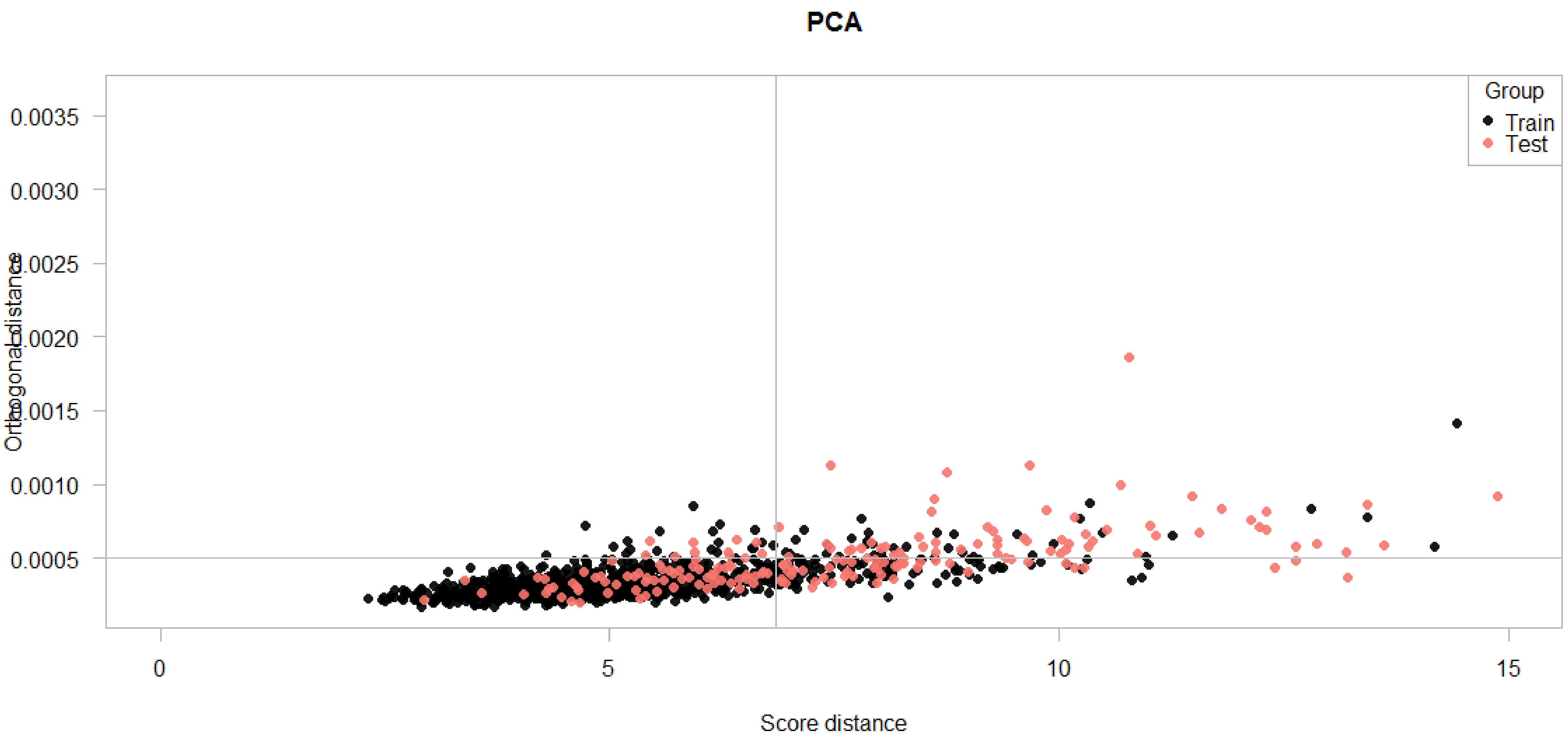
- **Training set** $F \rightarrow \boldsymbol{\tau} = \{ (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \}$

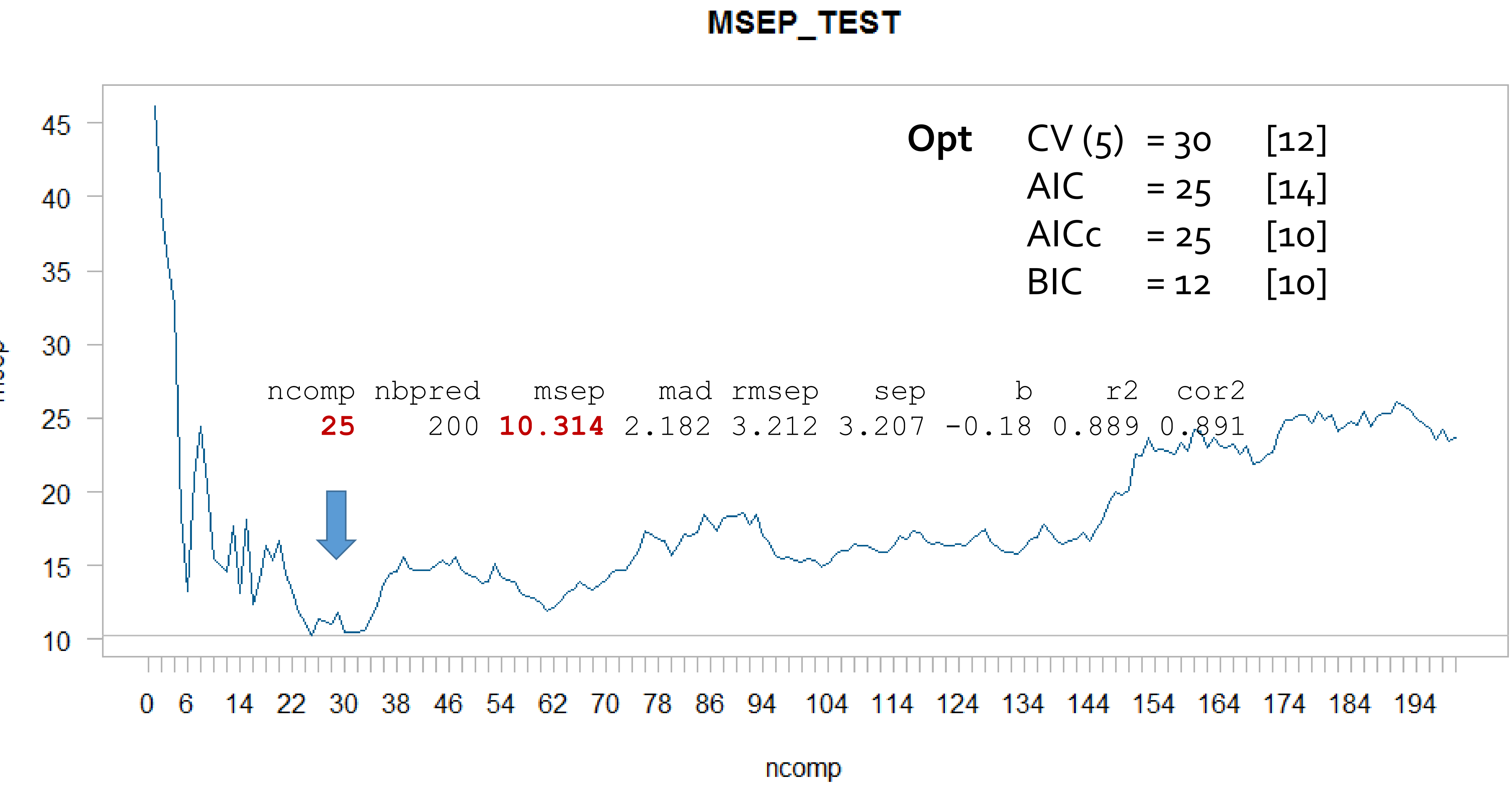
- **New observation** $F^* \, (\neq F) \rightarrow (\boldsymbol{x}^*, y^*)$

Example 1

Data split by the Kennard-Stone algorithm (inverted)

- **Training set** Lower dispersion n = 1006
- **New observation** Higher dispersion m = 200


PCA
Group
● Train
● Test
Orthogonal distance
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0
5
10
15
Score distance



Opt

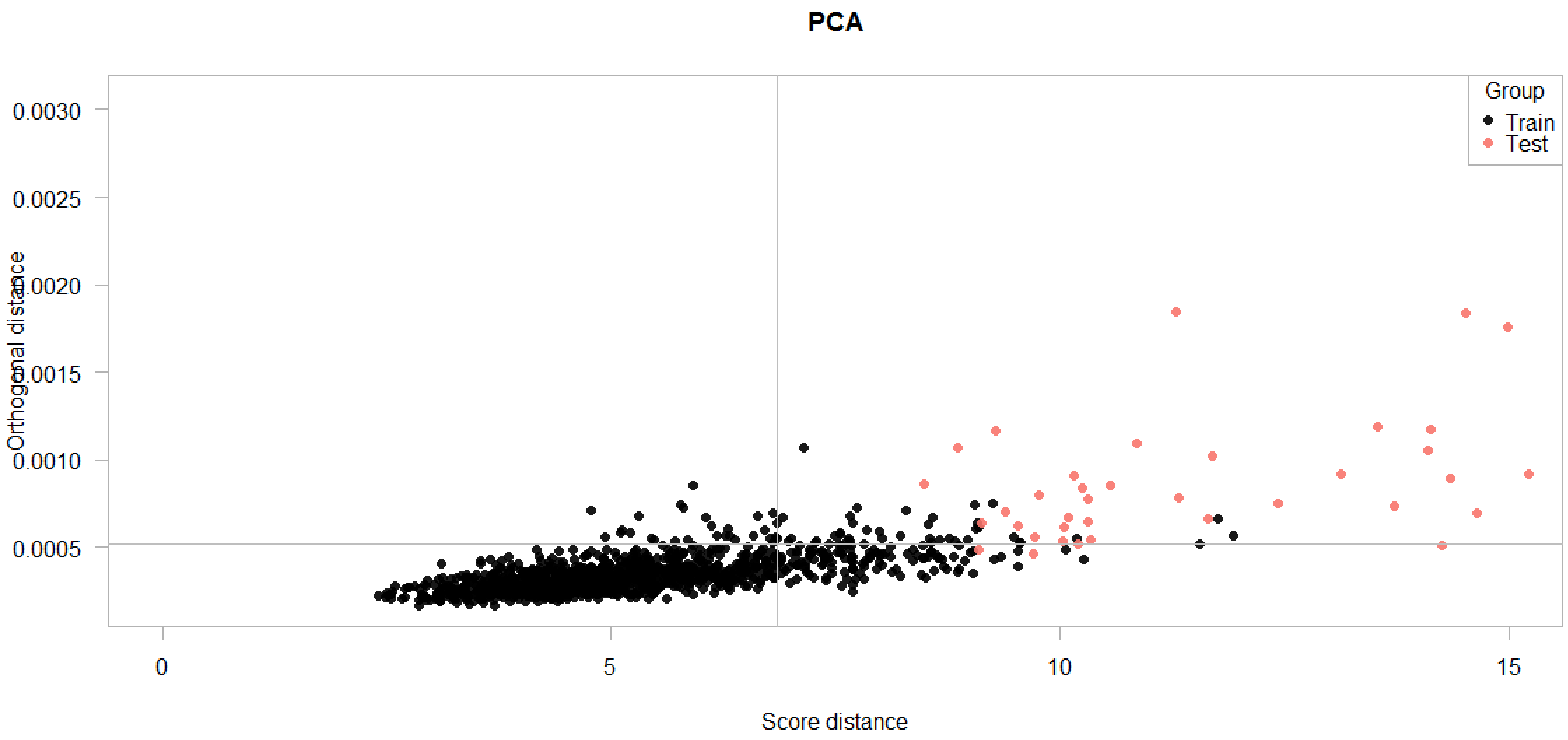
CV (5)	= 30	[12]
AIC	= 25	[14]
AICc	= 25	[10]
BIC	= 12	[10]

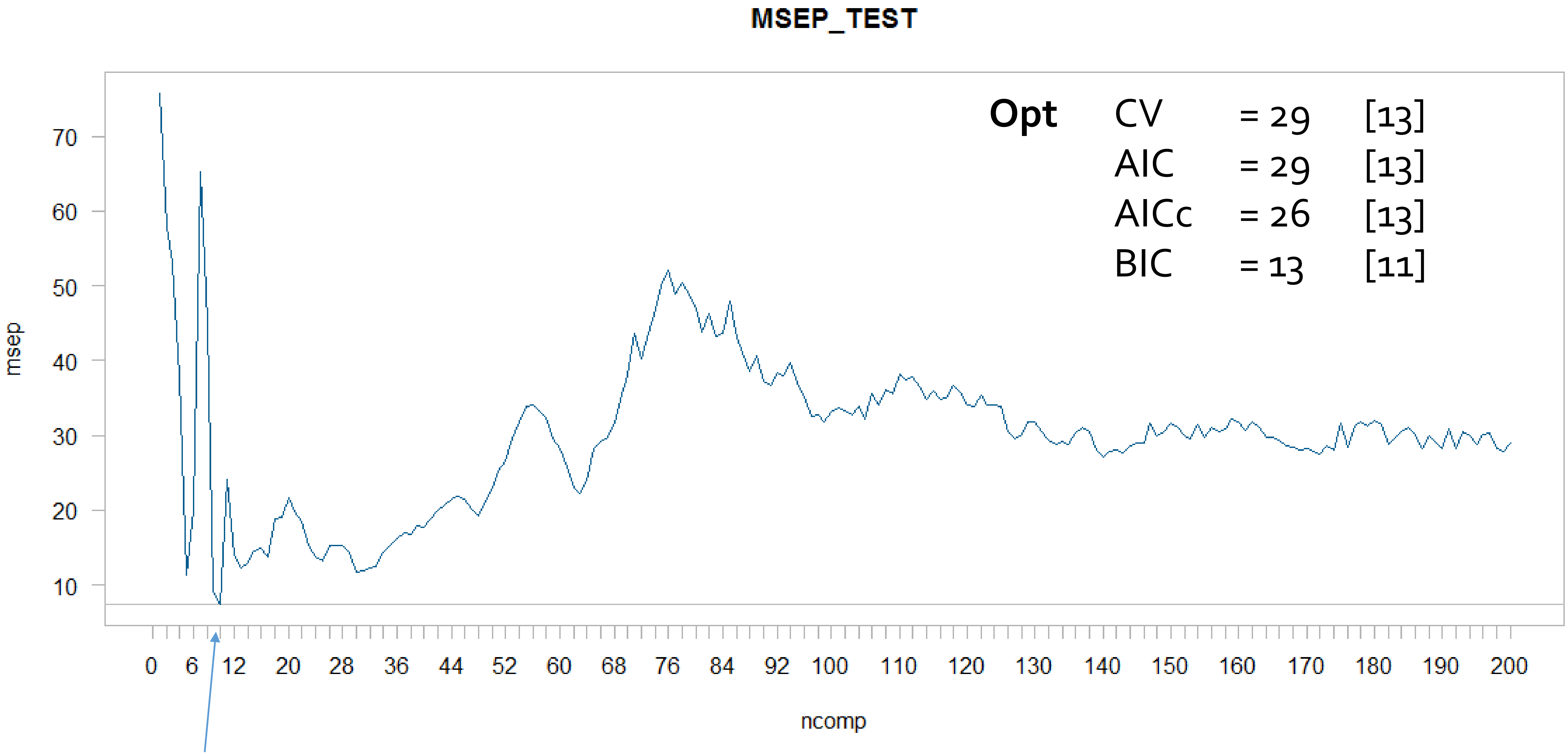
```
ncomp  nbpred    msep    mad  rmsep    sep      b     r2   cor2
   25     200  10.314  2.182  3.212  3.207  -0.18  0.889  0.891
```

Example 2

Data split based on the Score-Orthogonal distance

- **Training set** Lower distances $n = 1156$
- **New observation** Higher distances $m = 50$


PCA
Group
● Train
● Test
Orthogonal distance
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0
5
10
15
Score distance



Opt	CV	= 29	[13]
	AIC	= 29	[13]
	AICc	= 26	[13]
	BIC	= 13	[11]

```
ncomp nbpred  msep   mad rmsep   sep      b    r2  cor2
   10     50 7.408 2.102 2.722 2.711 -0.241 0.942 0.942
```

Some conclusion points

1) Cp (AIC**, etc.) was consistent** with CV and other preliminary analyses

2) Also with results on the **test sets**

3) For the sorghum data, **if n increases**

- the variance-bias compromise will probably allow to increase the nb. components

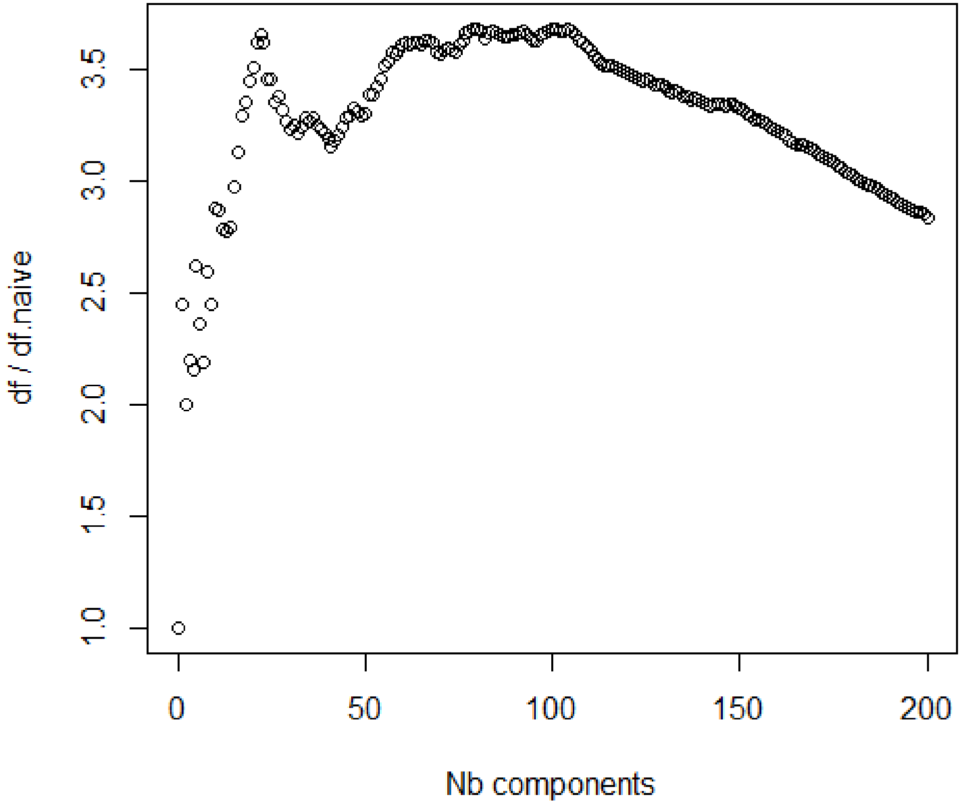
- Guess: opt would probably increase up to around 70-80 (if one believes the graphical methods on loadings and b-coefficents)

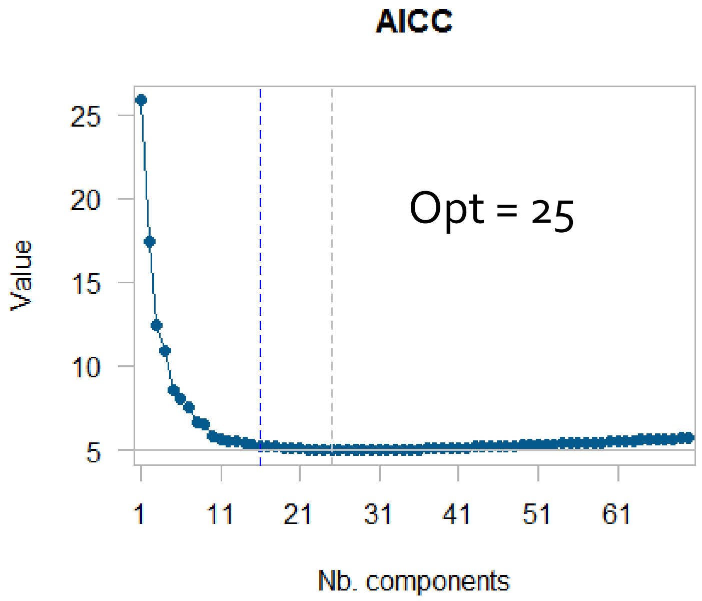
4) **Choice between** AIC vs BIC penalty

5) Cp **requires Monte Carlo for estimating** df_{PLSR}

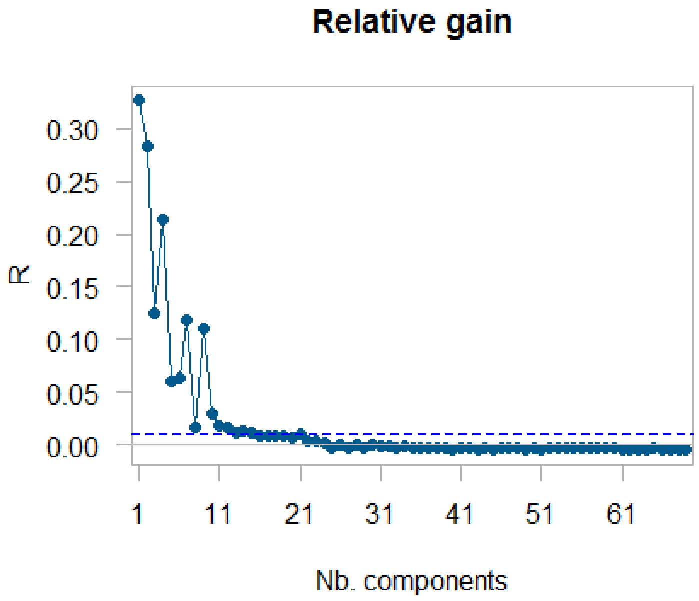
- Could we use a rule of thumb df_{approx} for escaping simulations?

- This would be a very fast procedure for approximating what returns $\mathrm{MSEP}_{\mathrm{CV}}$ selection

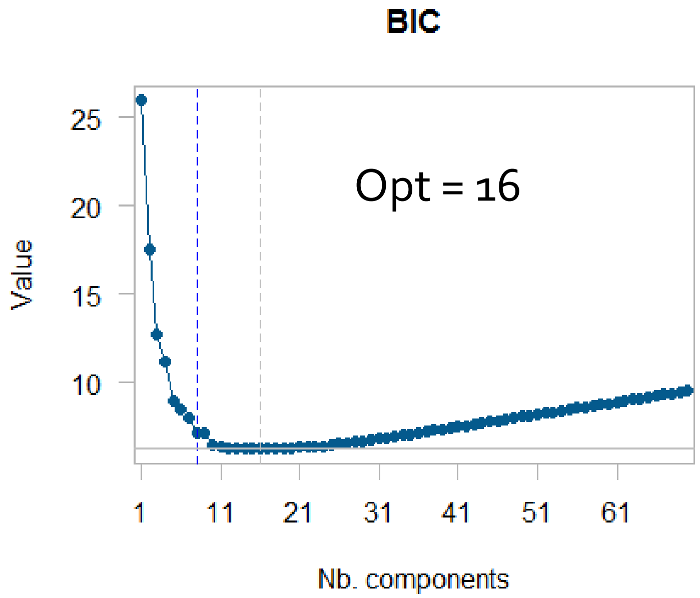




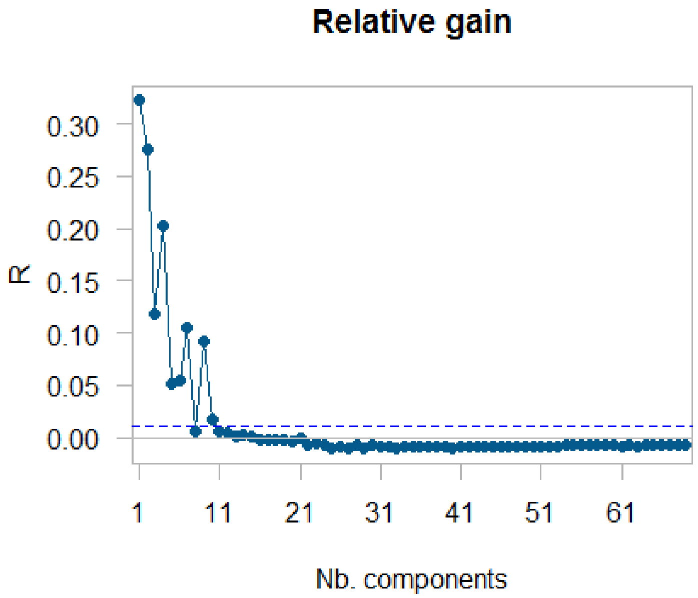
Opt = 25



with df_{approx} = 2.8 a + 1

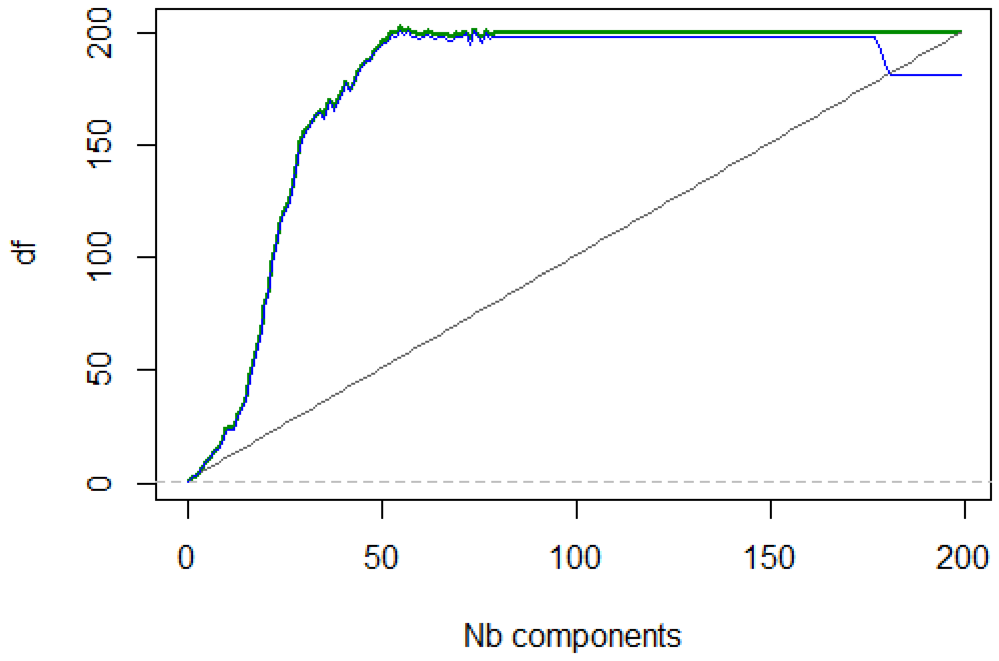


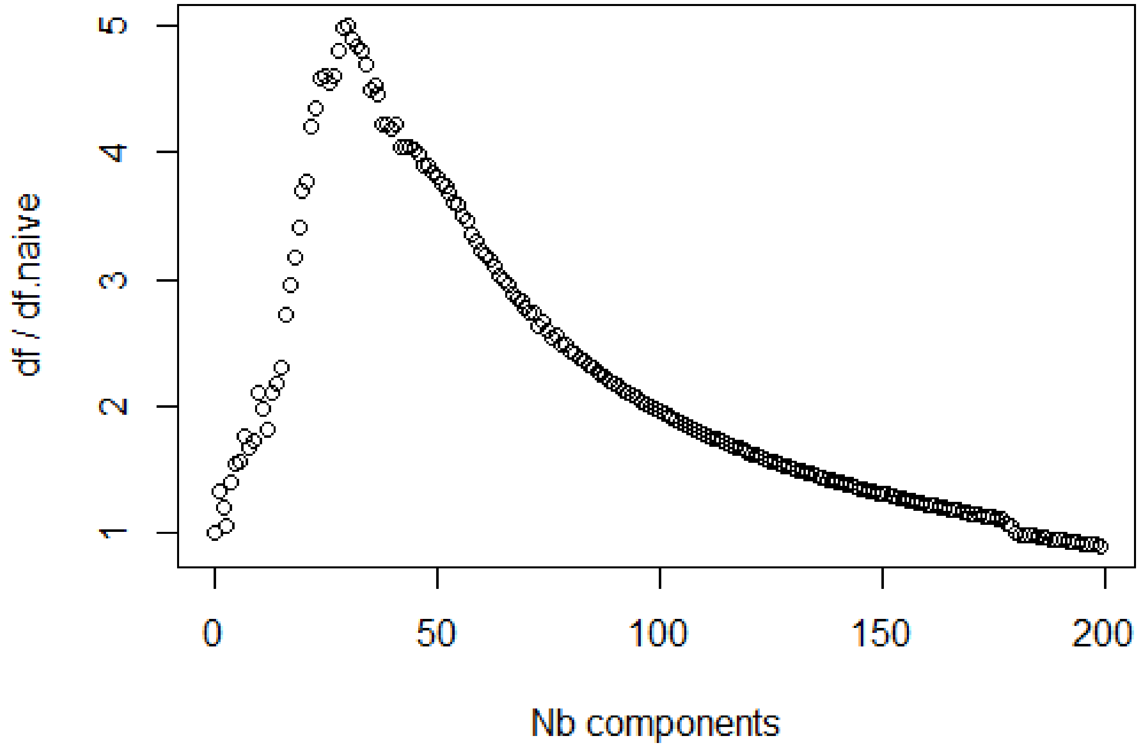
Opt = 16



But the ratio pattern can vary with the data …

Ex: Cassava data (n = 200)





6) **Generating distributions**

- If hypothesis $F = F^*$ is not accepted,

  one should specify F^*

  and optimize the model for this specific future

  (personal opinion: there is no generic rule for model selection if F^* is not specified)

APPENDIX – Some details

Statistical model

“True” generating distribution F

- **Training set** $F \rightarrow \quad \boldsymbol{\tau} \ = \{ (\boldsymbol{x}_1 , y_1) , \ldots , (\boldsymbol{x}_n , y_n) \}$
- **New observation** $F \rightarrow \quad (\boldsymbol{x}^*, y^*)$

Two sources of variations $\boldsymbol{\tau}$ and $(\boldsymbol{x}^*, y^*)$

Same distribution for training set and new observations

A convenient way is to consider a “true” generating model $\boldsymbol{g}$

- **Training set** $y \mid \boldsymbol{x} = g(\boldsymbol{x}, \gamma) + \varepsilon$

- **New observation** $y^* \mid \boldsymbol{x}^* = g(\boldsymbol{x}^*, \gamma) + \varepsilon^*$

Irreductible error

ε	iid	$E(\varepsilon)=0$	$Var(\varepsilon) = \sigma^2$
ε^*	iid	$E(\varepsilon^*)=0$	$Var(\varepsilon^*) = \sigma^2$

g unknown (and will stay unkown)

Prediction model $f(\boldsymbol{x}, \boldsymbol{\theta})$

Ex: f = PLSR model
$f \neq g$ (unknown)

- Training set $\boldsymbol{\tau} \rightarrow \widehat{\boldsymbol{\theta}}$
- Predictions $\hat{y} \mid \boldsymbol{x} = f(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})$

$$\hat{y}^* \mid \boldsymbol{x}^* = f(\boldsymbol{x}^*, \widehat{\boldsymbol{\theta}})$$

- Residual $e\,|\boldsymbol{x}_i = y|\boldsymbol{x}_i - f(\boldsymbol{x}_i, \hat{\theta})$ calibration error

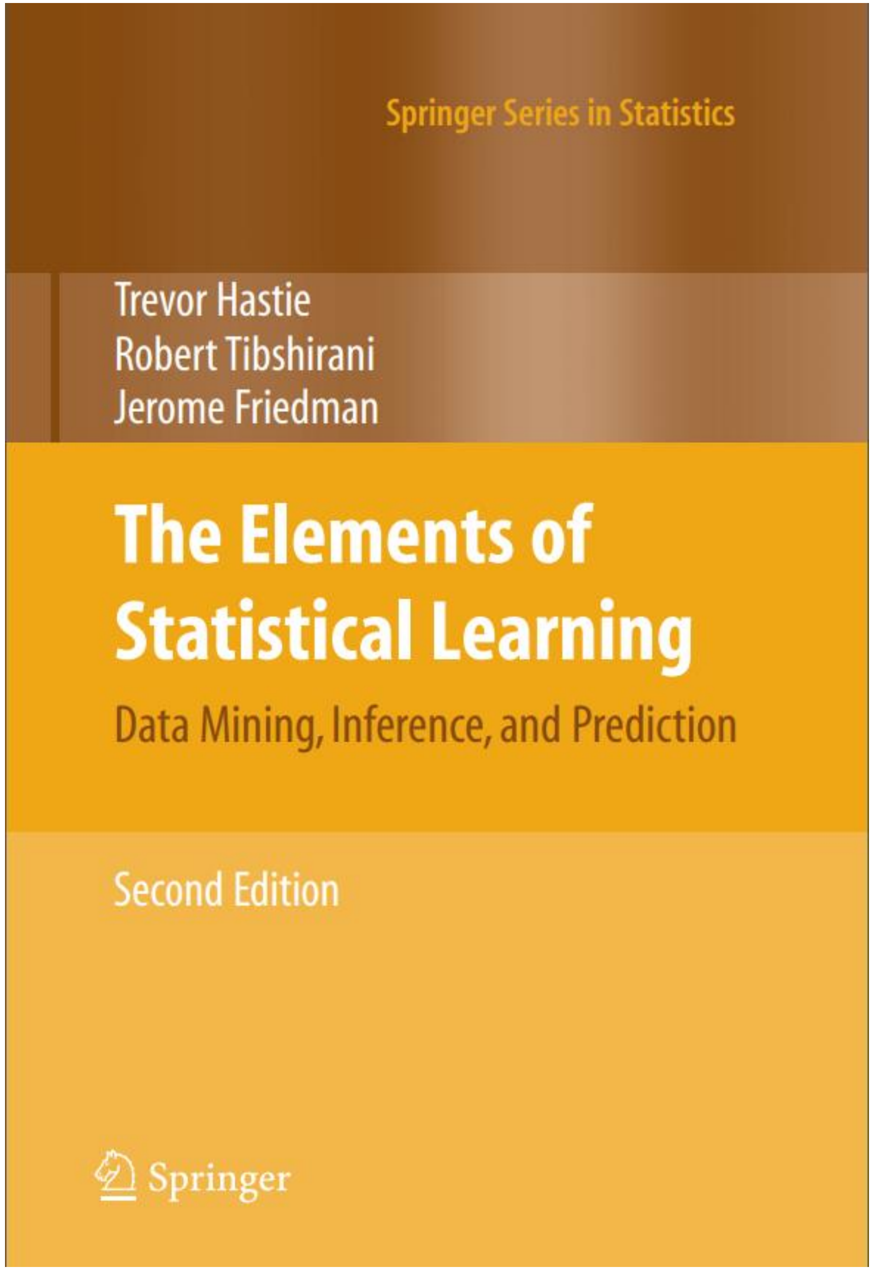
- Prediction error $e^*|\,\boldsymbol{x}^* = y^*|\boldsymbol{x}^* - f(\boldsymbol{x}^*, \widehat{\boldsymbol{\theta}})$ non-observable

  $= (g(\boldsymbol{x}^*, \gamma) + \varepsilon^*) - f(\boldsymbol{x}^*, \widehat{\boldsymbol{\theta}})$ two sources of variation

Prediction errors for model selection

Notations used in

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction, 2nd ed. *Springer*, New York.



Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

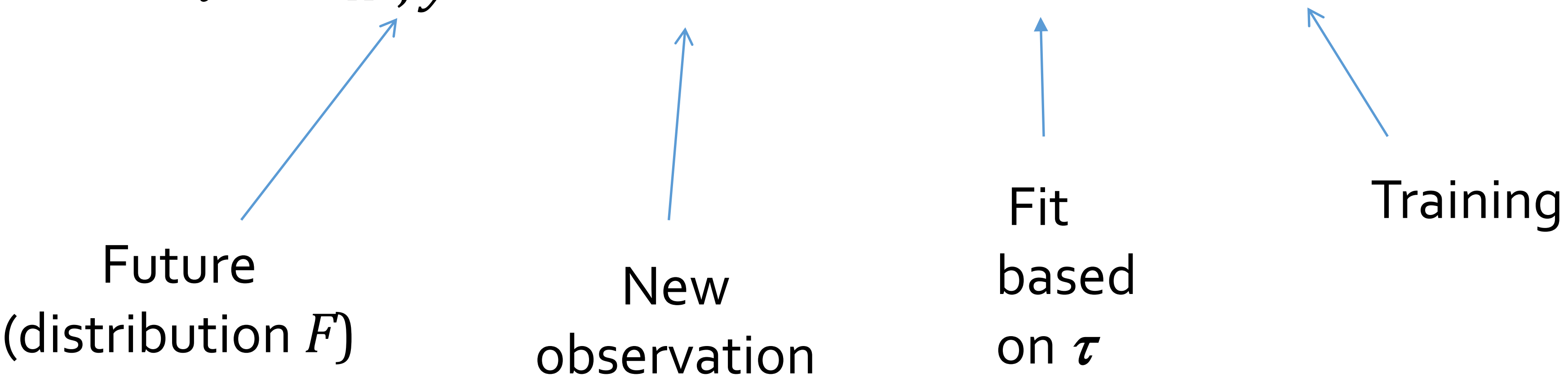
Springer

Conditional extra-sample error $\mathrm{Err}_{\boldsymbol{\tau}}$

Other names: ***Conditional generalization (or test) error***

For a quadratic loss:

$$\mathrm{Err}_{\boldsymbol{\tau}} = E_{\boldsymbol{x}^*,\, y^*}\left(\{ y^* | \boldsymbol{x}^* - f(\boldsymbol{x}^*, \widehat{\boldsymbol{\theta}}) \}^2 \;|\; \boldsymbol{\tau} \right)$$



Future (distribution F)

New observation

Fit based on $\boldsymbol{\tau}$

Training

Ideally for model selection, one would expect comparing the models based on their Err_{τ}

Problem: Err_{τ} is very difficult to estimate from the training data $\boldsymbol{\tau}$

In general, Err_τ is estimated ***a posteriori*** **from a test set** (data not used in the training!)

- $\widehat{\mathrm{MSEP}}_{\mathrm{Test}} = \widehat{\mathrm{Err}}_\tau$

- Can not be used for model selection

Note: Double repeated CV expects to estimate both the errors Err_τ and Err (see thereafter) in the same time

Instead of Err_{τ} , statistical methods for model selection target two other error measures, e.g.

1) CV, Bootstrap Err

2) Covariance penalty criteria (Cp, AIC, BIC, etc.) $\mathrm{Err}_{\mathrm{in}}$

Both measures are derived from Err_{τ}

1) **Expected prediction (or test) Error** $\mathrm{Err} = E_{\tau}(\mathrm{Err}_{\tau})$

- CV, Bootstrap $\widehat{\mathrm{MSEP}}_{\mathrm{CV}}$, $\widehat{\mathrm{MSEP}}_{\mathrm{Boot}} = \widehat{\mathrm{Err}}$

2) **Conditional in-sample Error** $\mathrm{Err}_{\mathrm{in}} = \sum_{i=1}^{n} \mathrm{Err}_{\tau}(\boldsymbol{x}_i)/n$

- Simplif. 1: Plug-in of Err_{τ} on the training set $\boldsymbol{\tau}$
- Simplif. 2: $\boldsymbol{\tau}$ variations comes only from $\boldsymbol{\varepsilon}$ (design $\boldsymbol{X}$ assumed fixed)

- Covariance penalty criteria Cp, etc. = $\widehat{\mathrm{Err}}_{\mathrm{in}}$

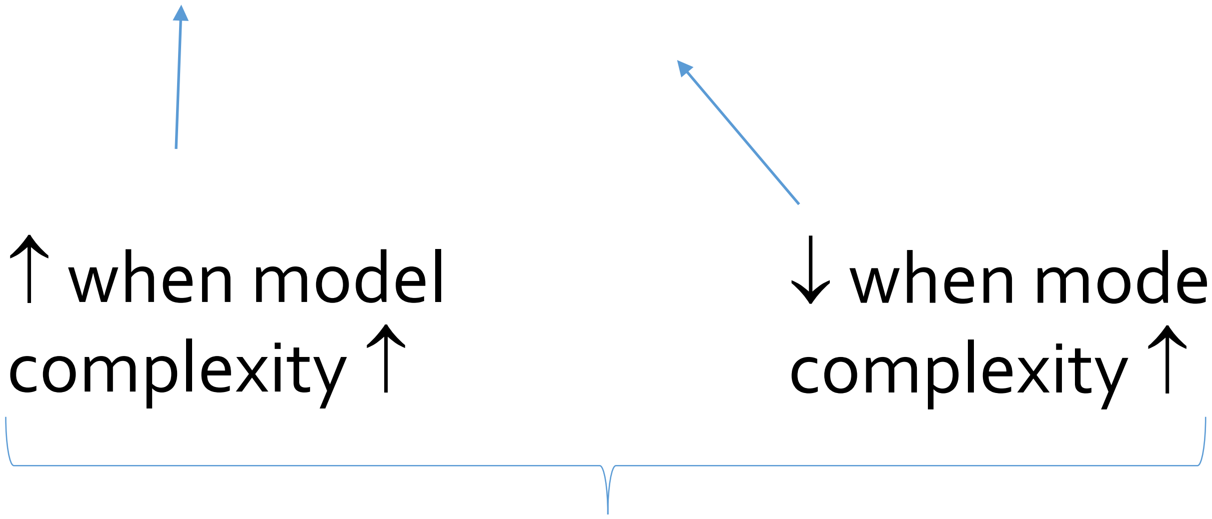
Note: $\mathrm{Err}_{\mathrm{in}}$ is noted "Err" in Efron Jasa 2004 (!!)

Both Err and $\mathrm{Err}_{\mathrm{in}}$ have been shown

- effective for model selection
- estimable from the training set $\boldsymbol{\tau}$
- compromises between variance and bias

Variance-bias compromise for Err

$$\mathrm{Err} = \sigma^2 + \mathrm{Var}_\tau(\widehat{\boldsymbol{y}}) + \mathrm{Bias}_\tau(\widehat{\boldsymbol{y}})^2$$



↑ when model complexity ↑

↓ when model complexity ↑

Model selection expects to find a compromise well afforded by the training of size n

Variance-bias compromise for $\mathrm{Err}_{\mathrm{in}}$

$E_{\boldsymbol{\varepsilon}}(\mathrm{Err}_{\mathrm{in}})$ = MSEP $= \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSEP}(\boldsymbol{x}_i)$ Only $\boldsymbol{\varepsilon}$ varies (X assumed fixed)

$= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}(\boldsymbol{x}_i) = \sigma^2 + \mathrm{MSE}$

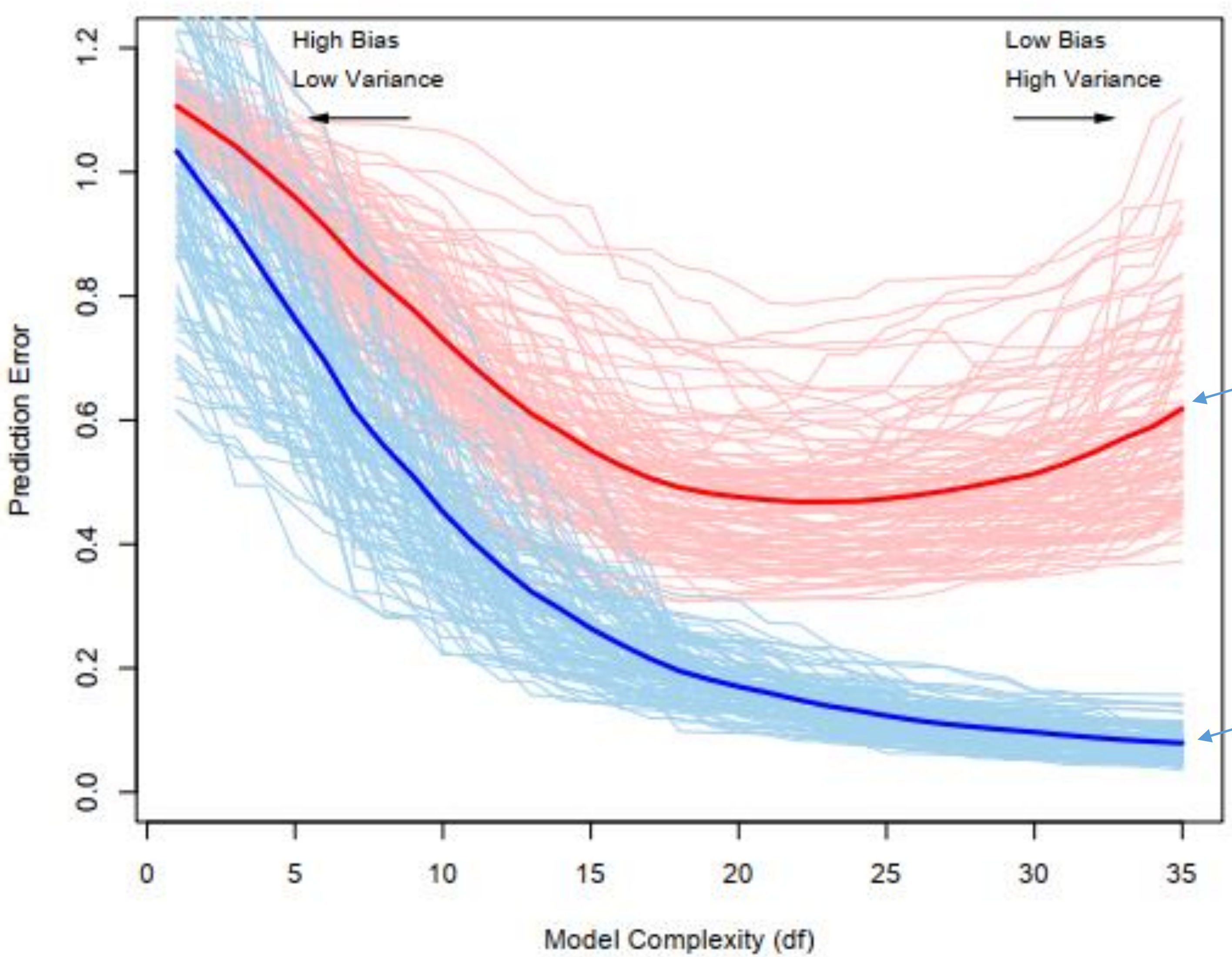
$= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} Var_{\boldsymbol{\varepsilon}}(f(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})) + \frac{1}{n}\sum_{i=1}^{n} \alpha(\boldsymbol{x}_i)^2$

with $\alpha(x_i)= g(\boldsymbol{x}_i, \gamma) - E_{\boldsymbol{\varepsilon}}(f(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})) = -\mathrm{Bias}_{\boldsymbol{\varepsilon}}(f(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}))$

$= \overline{\sigma}_*{}^2 + \frac{1}{n}\boldsymbol{\alpha}'\boldsymbol{\alpha}$

with $\overline{\sigma}_*{}^2 = \frac{1}{n}\sum_{i=1}^{n} \sigma_*{}^2(\boldsymbol{x}_i) = \sum_{i=1}^{n}\left(\sigma^2 + Var_{\boldsymbol{\varepsilon}}(f(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})\right)$

Hastie *et al* 2009 Fig 7.1



Err

$\overline{err}$ $= \mathrm{SSR}/n$
$= \mathrm{MSEC}$

Important relation for $\mathrm{Err}_{\mathrm{in}}$

Taking the expectation over $\boldsymbol{\tau}$ for $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_n$ fixed

Expected **model optimism** ω

$$E_{\boldsymbol{\varepsilon}}(\mathrm{Err}_{\mathrm{in}}) = E_{\boldsymbol{\varepsilon}}\left(\frac{\mathrm{SSR}}{n}\right) + \boxed{\frac{2}{n}\sum_{i=1}^{n} Cov_{\boldsymbol{\varepsilon}}(y_{i\,,}\, \hat{y}_i)}$$

$\boldsymbol{\tau}$

$= \mathrm{MSEP}$

$= \mathrm{MSEC}$

$f(\boldsymbol{x}_{\mathrm{i}}, \widehat{\boldsymbol{\theta}})$

Efron, B., 2004. The Estimation of Prediction Error. *Journal of the American Statistical Association* 99, 619–632. https://doi.org/10.1198/016214504000000692

$$\omega = \frac{2}{n}\sum_{i=1}^{n} Cov_{\boldsymbol{\varepsilon}}(y_i, \widehat{y}_i)$$

Higher is the effect of a variation of y_i on its prediction $\hat{y}_i$

higher is ω

For Gaussian errors $\boldsymbol{\varepsilon}$ (Stein 1981, Efron 2004)

$$Cov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i) \quad = \quad \sigma^2 E_{\boldsymbol{\varepsilon}}\left(\frac{\partial \hat{y}_i}{\partial y_i}\right)$$

$$\Rightarrow \quad \omega = \frac{2}{n}\sum_{i=1}^{n} Cov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i) = \frac{2}{n}\sigma^2 \sum_{i=1}^{n} E_{\boldsymbol{\varepsilon}}\left(\frac{\partial \hat{y}_i}{\partial y_i}\right)$$

Stein, C.M., 1981. Estimation of the Mean of a Multivariate Normal Distribution.
The Annals of Statistics 9, 1135–1151.

Covariance penalty criteria

$$E_{\boldsymbol{\varepsilon}}(\mathrm{Err}_{\mathrm{in}}) = E_{\boldsymbol{\varepsilon}}\left(\frac{\mathrm{SSR}}{n} + \frac{2}{n}\sum_{i=1}^{n} Cov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i)\right)$$

$\Rightarrow$ **Natural unbiased estimator** for $\mathrm{Err}_{\mathrm{in}}$

(in the sense $\mathrm{E}_{\boldsymbol{\varepsilon}}(\widehat{\mathrm{Err}}_{\mathrm{in}})$ = $\mathrm{E}_{\boldsymbol{\varepsilon}}(\mathrm{Errin})$)

$$\widehat{\mathrm{Err}}_{\mathrm{in}} = \underset{\substack{\uparrow \\ \mathrm{MSEC}}}{\frac{\mathrm{SSR}}{n}} \quad + \quad \underbrace{\frac{2}{n}\sum_{i=1}^{n} \hat{C}ov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i)}_{\text{Covariance penalty}}$$

→ = Covariance penalty criteria

→ = Mallows' Cp (or "AIC", Hastie et al 2009 p.231) family criteria

Using Stein’s equality when assuming Gaussian error

$$\widehat{\mathrm{Err}}_{\mathrm{in}} = \frac{\mathrm{SSR}}{n} \quad + \quad \frac{2}{n}\hat{\sigma}^2 \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i}$$

Divergence *i*

→ = Stein Unbiased Risk Estimate (SURE) for $\mathrm{Err}_{\mathrm{in}}$

Can be estimated using Monte Carlo simulations

- $\widehat{\mathrm{Err}}_{\mathrm{in}} = \frac{\mathrm{SSR}}{n} \quad + \quad \frac{2}{n}\sum_{i=1}^{n} \hat{C}ov_{\varepsilon}(y_i, \hat{y}_i)$ ← Parametric bootstrap (See e.g. Efron 2004)

- $\widehat{\mathrm{Err}}_{\mathrm{in}} = \frac{\mathrm{SSR}}{n} \quad + \quad \frac{2}{n}\hat{\sigma}^2\sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i}$ (SURE) ← Sensitivity analysis (perturbations)

Model’s degrees of freedom

Simple case: Linear smoothers

- $\widehat{\boldsymbol{y}} = \boldsymbol{S}\,\boldsymbol{y}$ $\boldsymbol{y}$ not involved in $\boldsymbol{S}$

- Ex: OLS regression $\widehat{\boldsymbol{y}} = \boldsymbol{S}\,\boldsymbol{y} = \boldsymbol{H}\,\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\boldsymbol{y}$

Then

$$\sum_{i=1}^{n} Cov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i) = tr(2\boldsymbol{S} \; - \; \boldsymbol{S}\boldsymbol{S}')\sigma^2$$

If $\boldsymbol{S}$ is idempotent $p \times p$

$$\sum_{i=1}^{n} Cov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i) = tr(\boldsymbol{S})\sigma^2 = p\sigma^2$$

- $\widehat{\mathrm{Err}}_{\mathrm{in}} = \frac{\mathrm{SSR}}{n} \quad + \quad \frac{2}{n} p \hat{\sigma}^2$ Mallows' Cp

- $\widehat{\mathrm{Err}}_{\mathrm{in}} = \frac{\mathrm{SSR}}{n} \quad + \quad \frac{\log(n)}{n} p \hat{\sigma}^2$ BIC ("Cp" form)

For orthogonal linear smoother such as OLS regression, it is accepted that $df = p$

If df is defined by

$$df \; = \sum\nolimits_{i=1}^{n} Cov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i)/\sigma^2$$

then

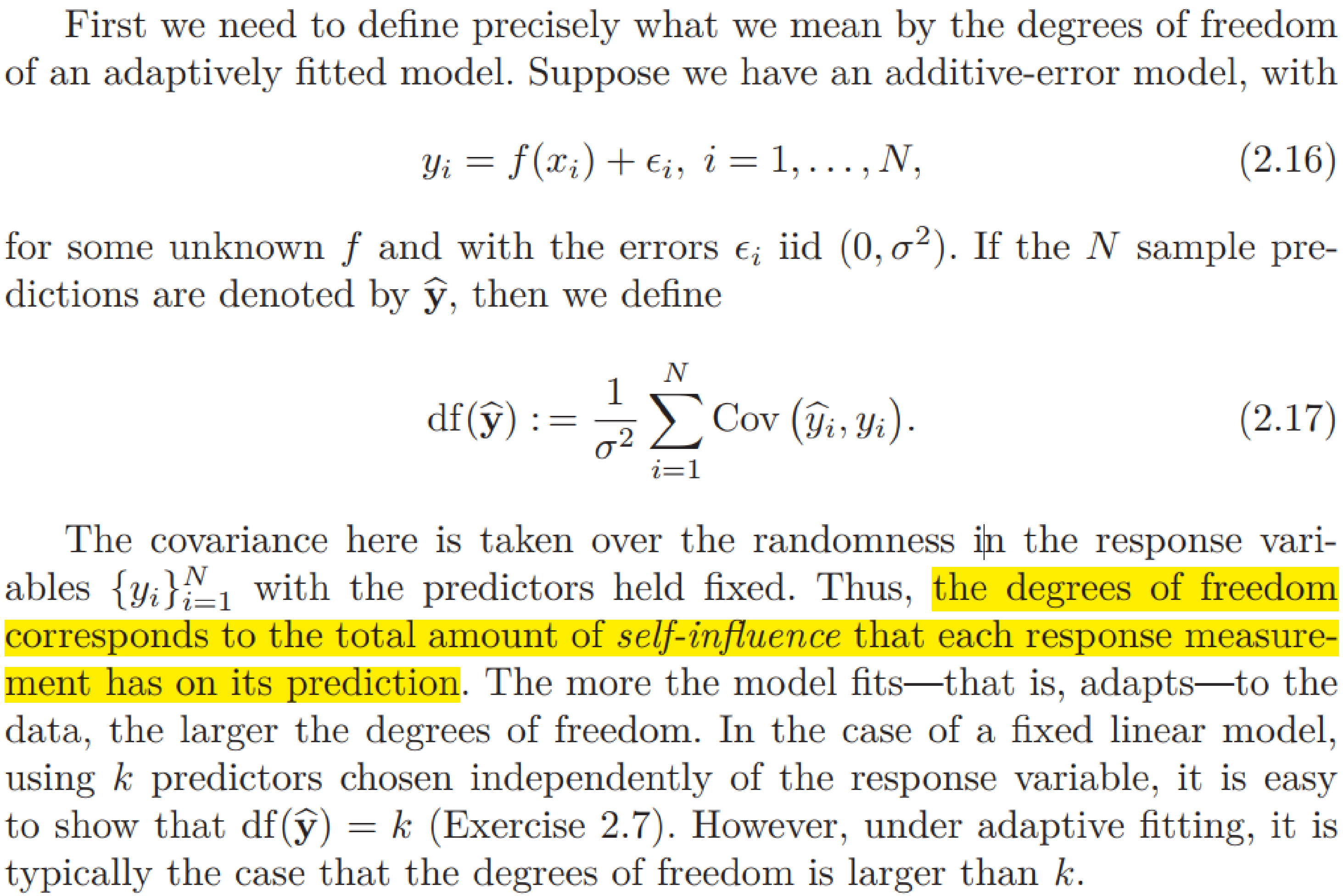
$$df \, = (p\sigma^2)/\sigma^2 = p$$

which is consistent with the usual accepted df

This has led to the general definition

For any model (linear or not), the consensus is to consider that

$$df \ = \sum_{i=1}^{n} Cov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i) / \sigma^2$$



First we need to define precisely what we mean by the degrees of freedom of an adaptively fitted model. Suppose we have an additive-error model, with

$$y_i = f(x_i) + \epsilon_i,\ i = 1, \ldots, N, \tag{2.16}$$

for some unknown f and with the errors ϵ_i iid $(0, \sigma^2)$. If the N sample predictions are denoted by $\widehat{\mathbf{y}}$, then we define

$$\mathrm{df}(\widehat{\mathbf{y}}) := \frac{1}{\sigma^2} \sum_{i=1}^{N} \mathrm{Cov}\left(\widehat{y}_i, y_i\right). \tag{2.17}$$

The covariance here is taken over the randomness in the response variables $\{y_i\}_{i=1}^N$ with the predictors held fixed. Thus, the degrees of freedom corresponds to the total amount of *self-influence* that each response measurement has on its prediction. The more the model fits—that is, adapts—to the data, the larger the degrees of freedom. In the case of a fixed linear model, using k predictors chosen independently of the response variable, it is easy to show that $\mathrm{df}(\widehat{\mathbf{y}}) = k$ (Exercise 2.7). However, under adaptive fitting, it is typically the case that the degrees of freedom is larger than k.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. *CRC Press*.

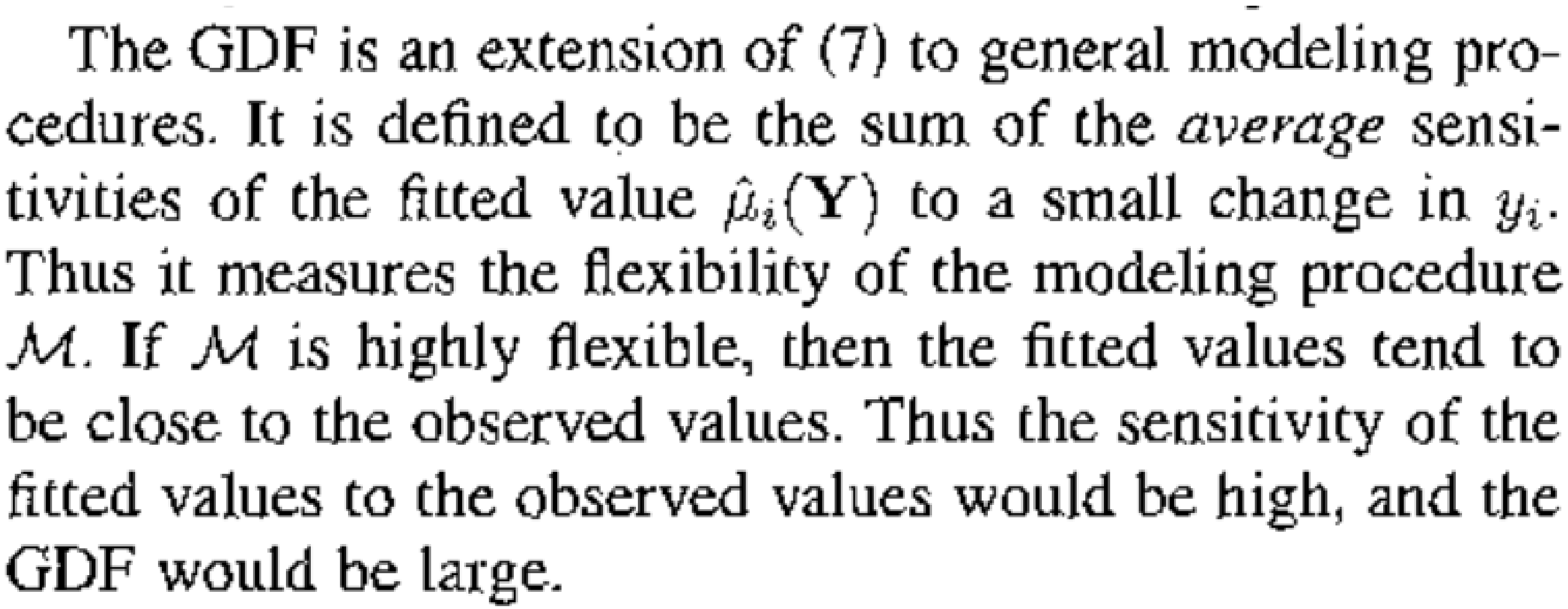
As for covariance penalty, *df* can be estimated from parametric bootstrap or sensitivity analysis

- $\widehat{df} = \sum_{i=1}^{n} \hat{C}ov_{\boldsymbol{\varepsilon}}(y_i, \hat{y}_i)/\widehat{\sigma}^2$ **Efron Jasa 2004**

- $\widehat{df} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i}$ (SURE) **Ye Jasa 1998** *Generalized df*

Ye, J., 1998. On Measuring and Correcting the Effects of Data Mining and Model Selection. *Journal of the American Statistical Association* 93, 120–131. https://doi.org/10.1080/01621459.1998.10474094

Ye 1998 p.122



The GDF is an extension of (7) to general modeling procedures. It is defined to be the sum of the *average* sensitivities of the fitted value $\hat{\mu}_i(\mathbf{Y})$ to a small change in y_i. Thus it measures the flexibility of the modeling procedure $\mathcal{M}$. If $\mathcal{M}$ is highly flexible, then the fitted values tend to be close to the observed values. Thus the sensitivity of the fitted values to the observed values would be high, and the GDF would be large.