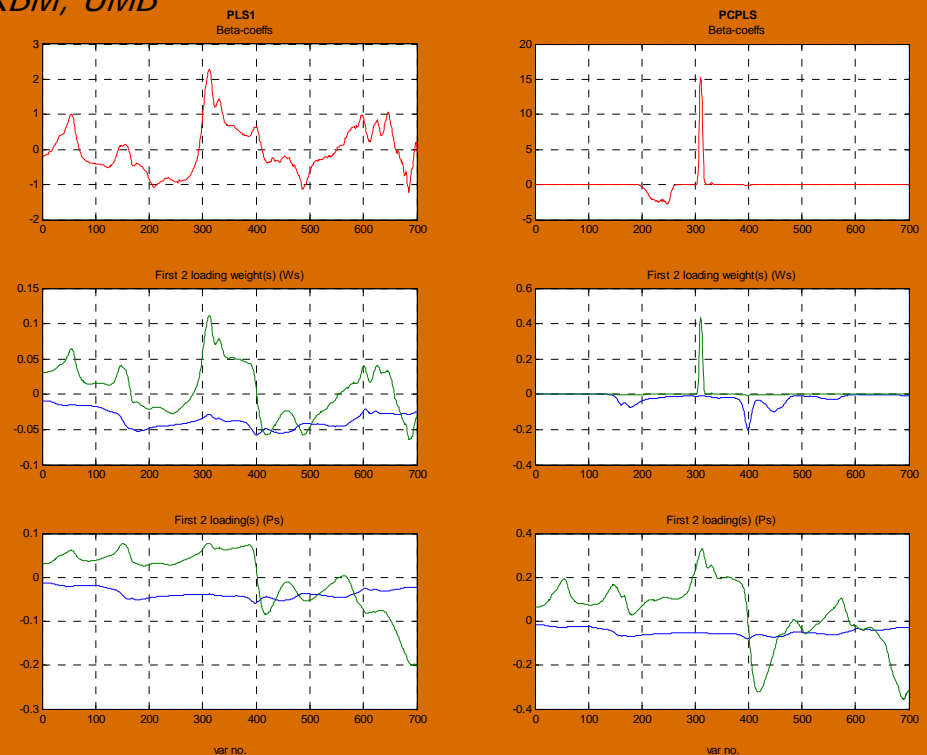
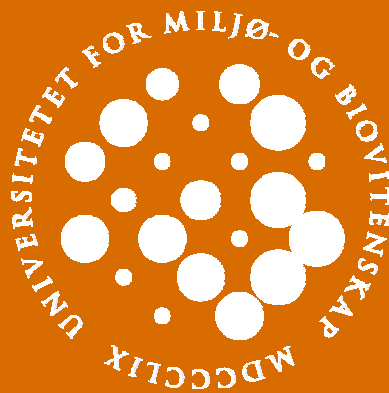


PPLS - en vri på klassisk PLS1 som gir muligheter for eksplorativ regresjonsanalyse og klassifikasjon med fokus på enklere modeller

Referanse: J. Chemometrics 2005; 19: 1–13

Ulf Indahl (ulf.indahl@umb.no)

IKBM, UMB



Figur: Regresjonskoeffisienter for Dough-data, prediksjon av fett med PLS og PPLS og PLS1

Klassisk PLS1 (Maksimering av kovarians)

1. Finn enhetsvektor \mathbf{w}_i som maksimerer $\mathbf{y}_{i-1}'\mathbf{X}_{i-1}\mathbf{w}_i$. Løsning:
 - a) $\mathbf{w}_i = \mathbf{Cov}(\mathbf{y}_{i-1}, \mathbf{X}_{i-1})' \equiv (\mathbf{X}_{i-1}' \mathbf{y}_{i-1})/n = [\text{Cov}(\mathbf{y}_{i-1}, \mathbf{x}_1), \dots, \text{Cov}(\mathbf{y}_{i-1}, \mathbf{x}_p)]$,
 (\mathbf{x}_k er k-te kolonne of \mathbf{X}_{i-1} for $k=1, \dots, p$).
 - b) $\mathbf{w}_i = \mathbf{w}_i / \|\mathbf{w}_i\|$ (skaleres til enhetsvektor).
2. Beregner scorerer (\mathbf{t}_i) og ladninger (\mathbf{p}_i & \mathbf{q}_i)
3. Deflaterer \mathbf{X}_{i-1} til \mathbf{X}_i og \mathbf{y}_{i-1} til \mathbf{y}_i .
4. Gjenta 1-3) til stoppkriterium.

Merk: I 1a. Er hver av koeffisientene i \mathbf{w}_i proporsjonal med kovariansen mellom \mathbf{y} og den korresponderende variable \mathbf{x}_i .

PPLS - Maksimering av korrelasjoner over en mengde med kandidat loadings

- **Idé – variere innflytelsen av varians og korrelasjon:**

$$w(\alpha, \beta) = K_{\alpha, \beta} [s_1 \cdot |\text{corr}(\mathbf{y}, \mathbf{x}_1)|^\alpha \cdot \text{std}(\mathbf{x}_1)^\beta, \dots, s_p \cdot |\text{corr}(\mathbf{y}, \mathbf{x}_p)|^\alpha \cdot \text{std}(\mathbf{x}_p)^\beta], \text{ der } \alpha, \beta \geq 0.$$

Med $\beta = \alpha = 1$ oppnås klassiske PLS-loading weights

- **Forenkling ved å kreve $\beta = 1/\alpha$:**

$$w(\alpha) = K_\alpha [s_1 \cdot |\text{corr}(\mathbf{y}, \mathbf{x}_1)|^\alpha \cdot \text{std}(\mathbf{x}_1)^{1/\alpha}, \dots, s_p \cdot |\text{corr}(\mathbf{y}, \mathbf{x}_p)|^\alpha \cdot \text{std}(\mathbf{x}_p)^{1/\alpha}]$$

- **Kan reparametrisere til $\gamma \in [0, 1]$ ved relasjonen $\alpha = \gamma/(1-\gamma)$:**

$$w(\gamma) = K_\gamma [s_1 \cdot |\text{corr}(\mathbf{y}, \mathbf{x}_1)|^{\gamma/(1-\gamma)} \cdot \text{std}(\mathbf{x}_1)^{(1-\gamma)/\gamma}, \dots, s_p \cdot |\text{corr}(\mathbf{y}, \mathbf{x}_p)|^{\gamma/(1-\gamma)} \cdot \text{std}(\mathbf{x}_p)^{(1-\gamma)/\gamma}]$$

Korrelasjon benyttes som relasjonsmål mellom x-er og y i vektingen av variable.

OBS: Klassisk PLS1 løsning når vi "fryser" $\gamma \equiv 1/2$

For $\gamma \rightarrow 1$ dominerer variablene med høyest korrelasjoner mot y.

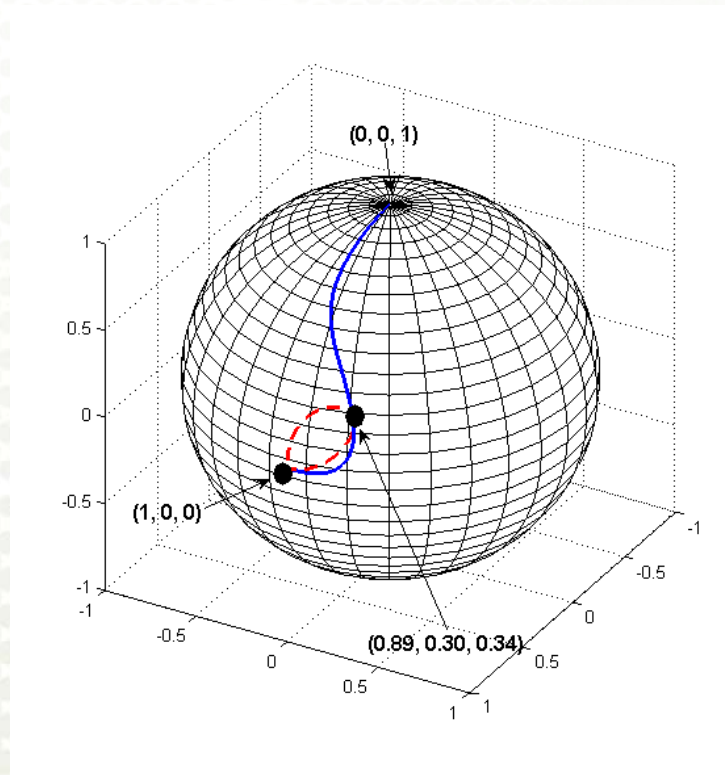
For $\gamma \rightarrow 0$ dominerer variablene med høyest varians.

Stegvis optimering av γ -verdi som maksimerer korrelasjon

Modifiserer 1a) i oppskriften på forrige side ved å søke γ -verdien som maksimerer korrelasjonen mellom \mathbf{y}_{i-1} og $\mathbf{X}_{i-1} \mathbf{w}_i(\gamma)$ (numerisk optimering)

Optimering av γ -parametre

- Optimale γ -verdier med tilhørende loadings $w(\gamma)$ på enhetskula i \mathbf{R}^p kan beregnes ved valg av en passende optimeringsalgoritme.
- Har selv benyttet en variant med såkalt **golden section search and parabolic interpolation**, designet til optimering av en funksjon over et angitt åpent intervall .



- Referanse:

Brent R P. Algorithms for Minimization without Derivatives, Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

La oss se på et eksempel

Myk tilnærming til variabelseleksjon

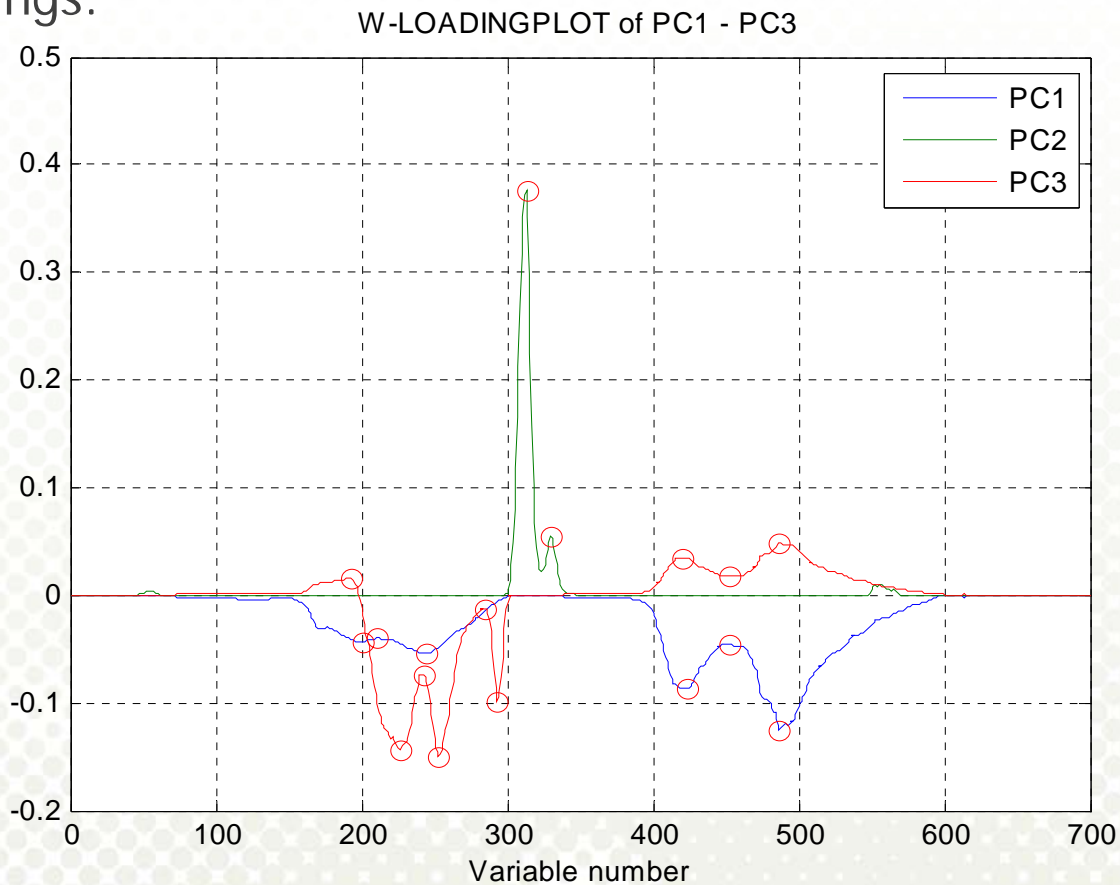
- NIR-datasett, (Discuit dough, Osborne et al. [1] and Brown et al. [2].)
 - Treningsdata med $N = 40$ prøver og $p = 700$ bølgelengder ($1100 - 2498 \text{ nm}$ in steps of 2 nm). Fire responsvariabler som for hver prøve angir prosentandeler i:
 - *Fett*
 - *Sukrose*
 - *Mel*
 - *Vann*
 - Testdata med $M = 32$ prøver.
 - Prøver og målinger ble laget ved ulike tidspunkter.

Vi skal fokusere på prediksjon av **fett**, og eksplorativt jobbe oss fram til "noen få" variable som gir god prediksjon både ved kryssvalidering og prediksjon på testdata.

1. Osborne B G, Fearn T, Miller A R and Douglas S. Application of Near Infrared Spectroscopy to the Compositional Analysis of Biscuits and Biscuit Doughs. *Journal of Scientific Food Agriculture* 1984; **35**: 99-105.
2. Brown P J, Fearn T and Vanucci M. Bayesian Wavelet Regression on Curves with Application to a Spectroscopic Calibration Problem. *Journal of the American Statistical Association* 2001; **96**: 398-408.

Parameterintervall: [0.90, 0.95] - optimal modell med 3 komponenter

W-loadings:



Eksplorativ PPLS der $\gamma \in [0.90, 0.95]$

RMSCV: 0.384

RMSEP: 0.291

Utvalgte variable:

- Eksplorativ PPLS med variabelseleksjon ($\gamma \equiv 1$) leder til: [226 313 487]
 - RMSCV: 0.377
 - RMSEP: 0.220
- Forslag fra Tom Fearn (dataleverandør): [226 318]
 - RMSCV: 0.456
 - RMSEP: 0.291

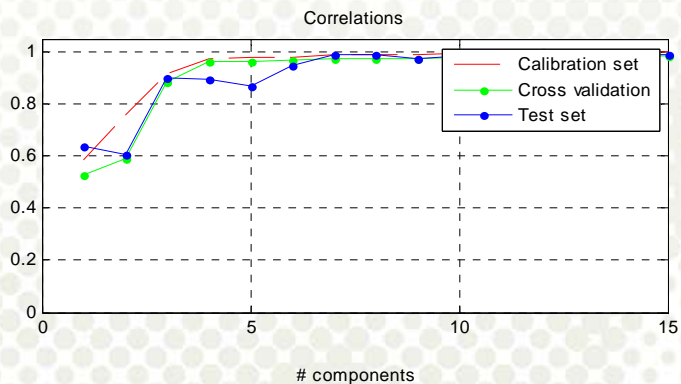
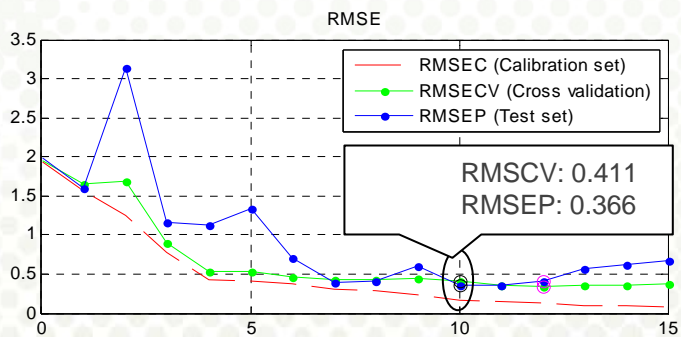
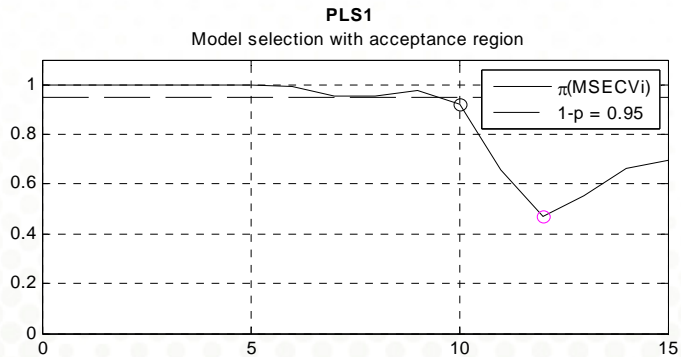
"Fulle" modeller:

- Ordinær PLS (10 komp):
 - RMSCV: 0.412
 - RMSEP: 0.366
- Eksplorativ PPLS der $\gamma \in [0.90, 0.95]$ (3 komp):
 - RMSCV: 0.384
 - RMSEP: 0.291

Ordinær PLS

RMSCV: 0.411

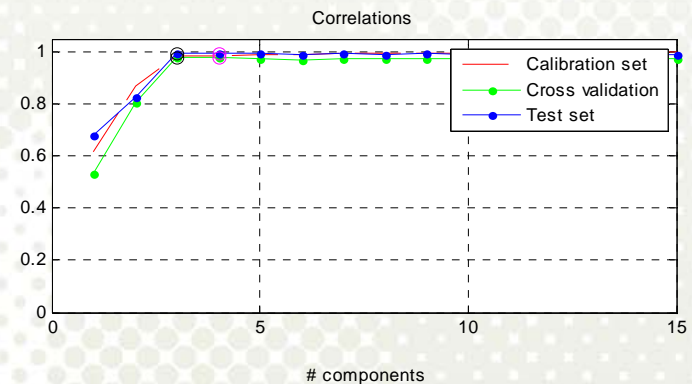
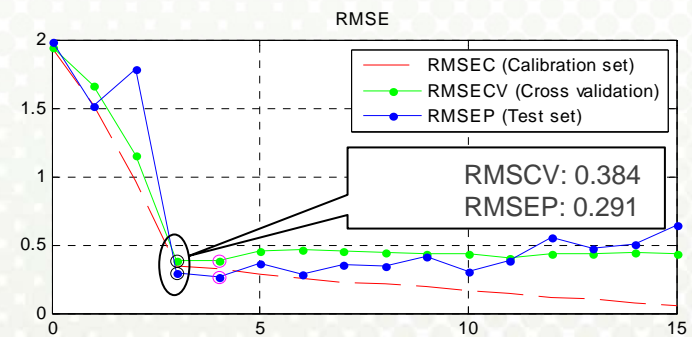
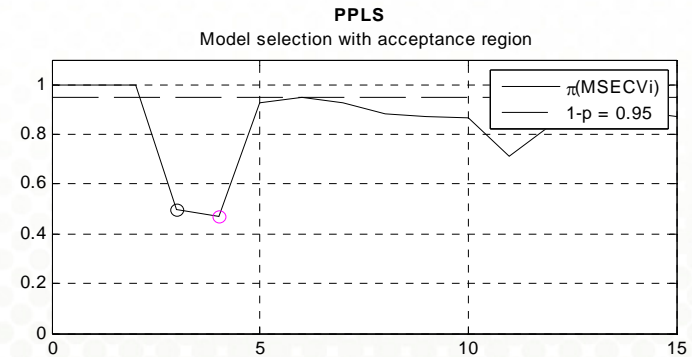
RMSEP: 0.366



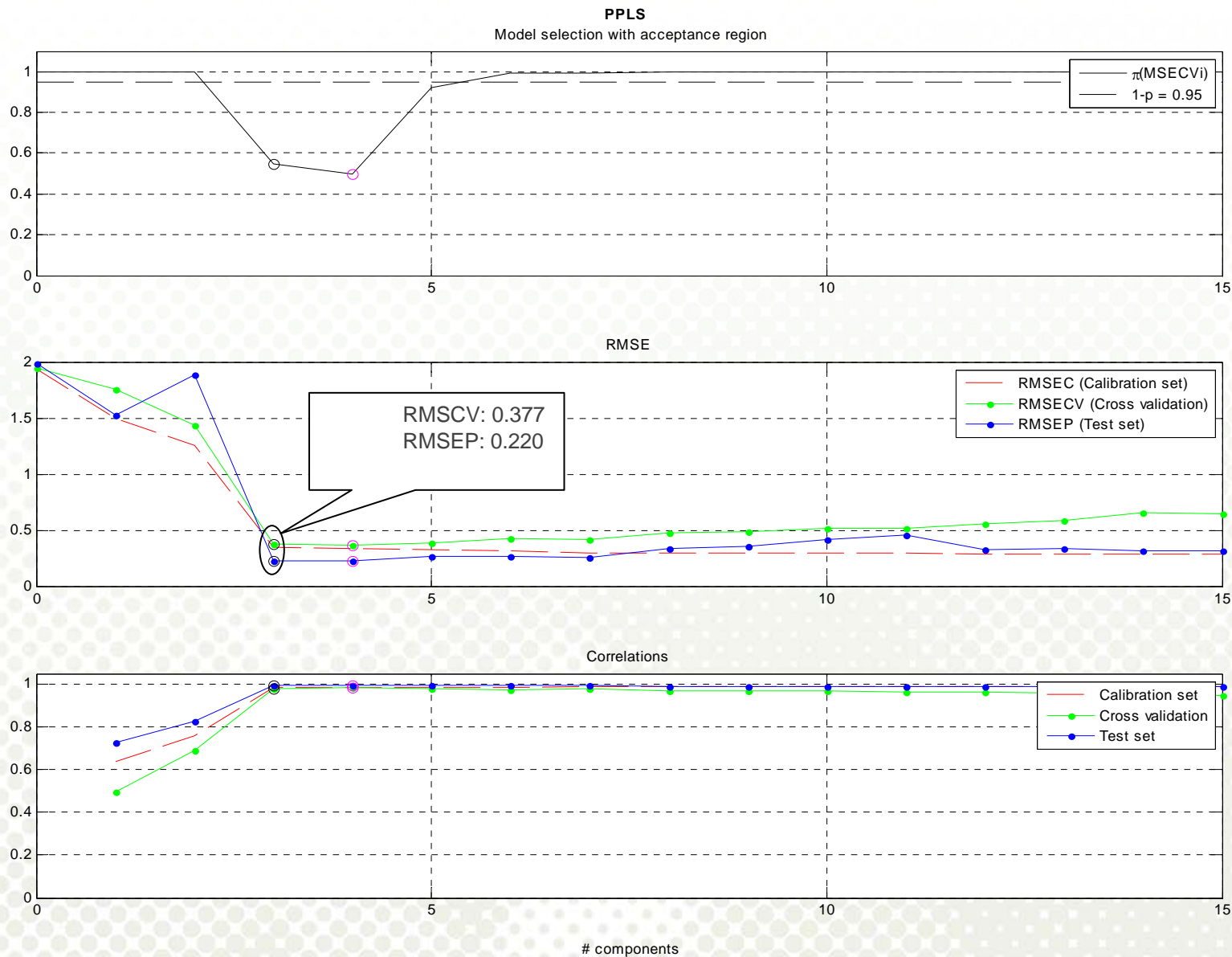
Eksplorativ PLS der $\gamma \in [0.90, 0.95]$

RMSCV: 0.384

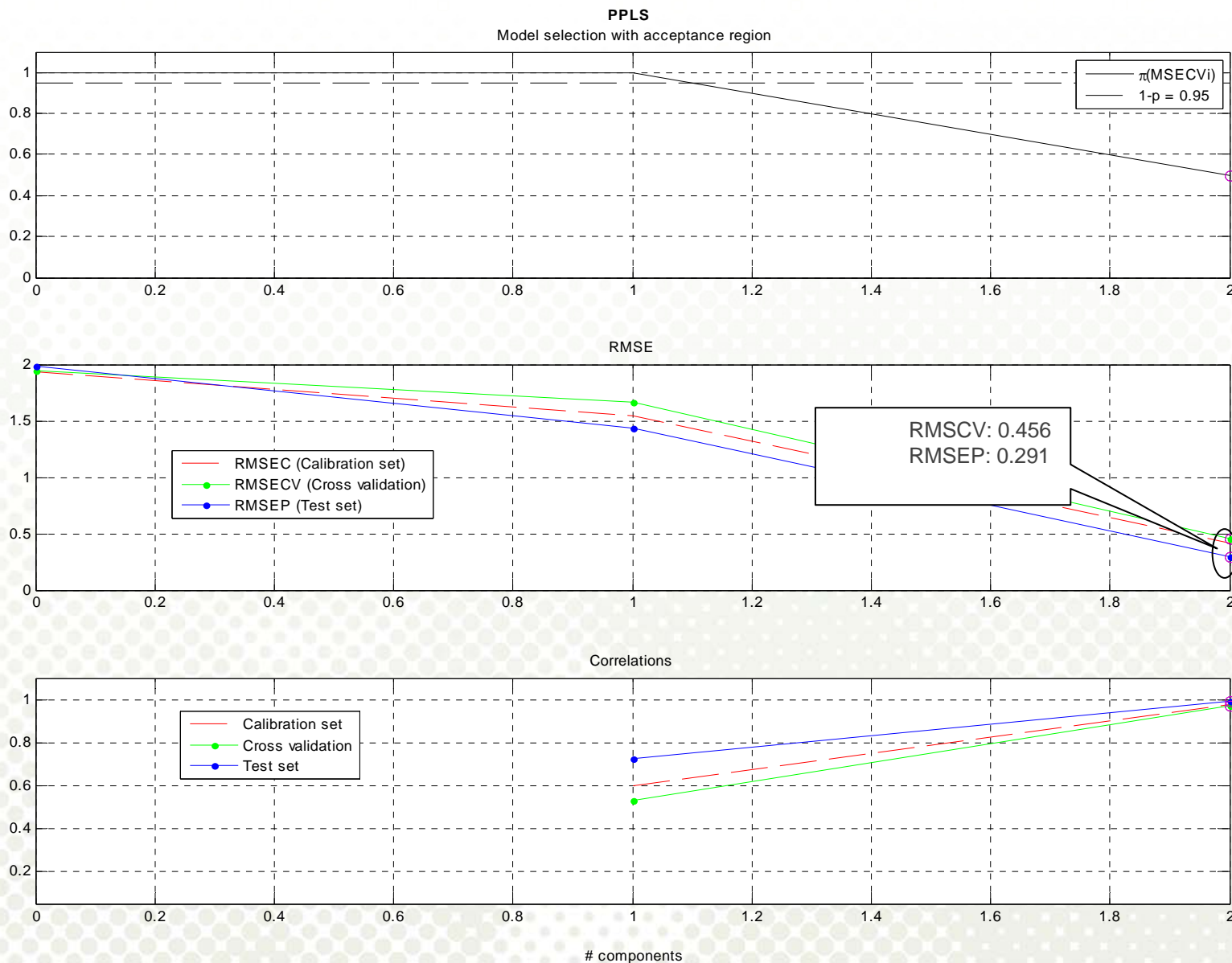
RMSEP: 0.291



Eksplorativ PPLS ledet til utvelgelse av variablene [226 313 487]



Tom Fearn foreslo: [226 318]



Vanlig PLS med 6 komponenter

