

PPLS2

Multirespons variant av PPLS

Alf Synstad og Ulf Indahl IKBM, UMB



Foredragsholderen

- Sisteårs masterstudent i statistikk ved IKBM, UMB på Ås
- Veileder er Ulf Indahl
- Presentasjonen er et lite sammendrag av masteroppgaven.

Les mer om PPLS på denne referansen:
<http://www3.interscience.wiley.com/cgi-bin/abstract/110557324/ABSTRACT>

Motivasjon

- Man måler gjerne flere størrelser når man gjør et laboratorieforsøk.
- Dette resulterer i en rekke responser Y_1, Y_2, \dots, Y_m
- Bør/Kan man utnytte denne informasjonen til å lage et felles antall komponenter?

Utvidelse av PPLS1

- Ingen intuitive relasjonsmål til utvidelse til multirespons problemstilling?
- Henter ut w med SVD på X matrisa og deler vekk standardavviket.
- Bytter ut korrelasjonskriteriet
- $w(\gamma) = K_\gamma [s_1 \cdot |\text{corr}(y, X_1)|^{\gamma/(1-\gamma)} \cdot \text{std}(x_1)^{(1-\gamma)/\gamma}, \dots, s_p \cdot |\text{corr}(y, X_p)|^{\gamma/(1-\gamma)} \cdot \text{std}(x_p)^{(1-\gamma)/\gamma}]$
- med resten "rest" etter å ha utført SVD.
- $w(\gamma) = K_\gamma [s_1 \cdot |\text{rest}_1|^{\gamma/(1-\gamma)} \cdot \text{std}(x_1)^{(1-\gamma)/\gamma}, \dots, s_p \cdot |\text{rest}_p|^{\gamma/(1-\gamma)} \cdot \text{std}(x_p)^{(1-\gamma)/\gamma}]$
- Har valgt å bruke produkt av korrelasjoner til optimering av potensen.
- Vi skal da søke den γ verdien som maksimerer produktet:

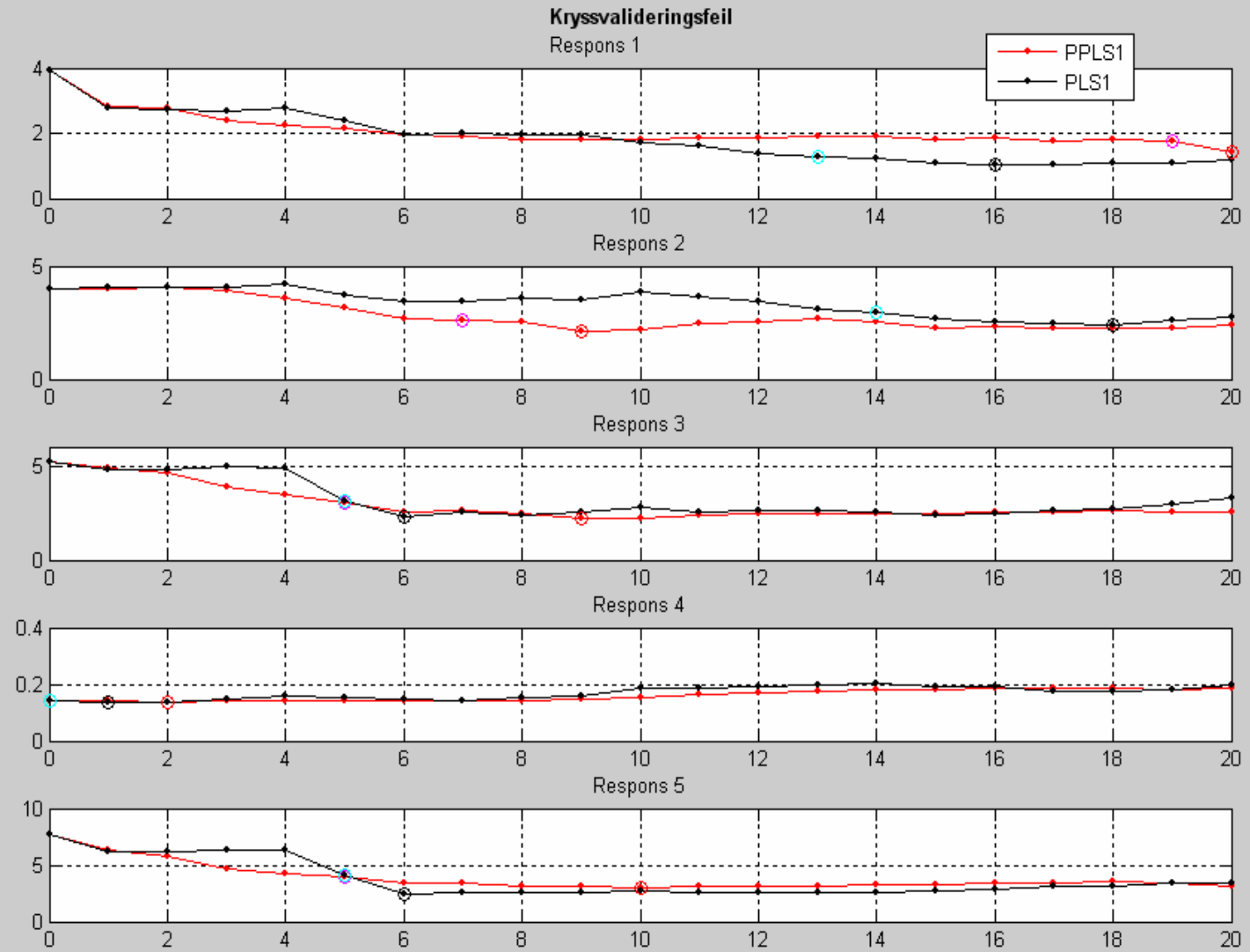
$$\text{korr}(y_1, \underline{X}_{i-1} \underline{w}_i(\gamma))^* \dots \cdot \text{korr}(y_m, \underline{X}_{i-1} \underline{w}_i(\gamma))$$



Hvordan bestemme metodenes godhet?

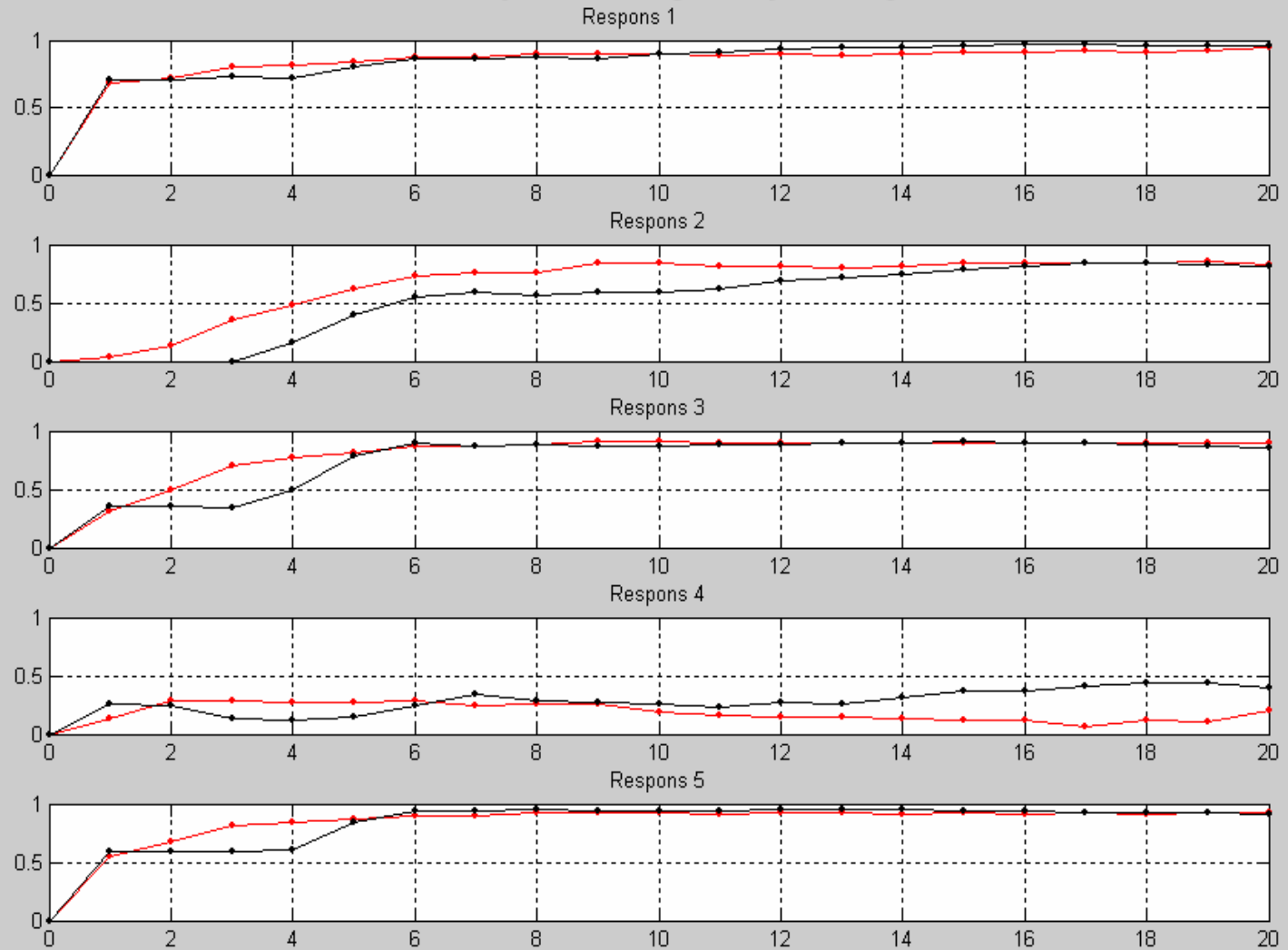
- Tradisjonell metode for utvelgelse av antall komponenter i unirespons problemstilling er minimum RMSE (Root mean square error)

RMSE verdier



Korrelasjoner

Korrelasjoner mellom Y og \hat{Y} i kryssvalidering



Hvordan bestemme modellens godhet gitt at vi er i en multirespons kontekst?

- Noen forslag:
 - Størst korrelasjonssum
 - Størst korrelasjonsprodukt
 - Størst felles R-sq
 - Størst gjennomsnittlig R-sq

Mitt kandidater

- Felles forklart variasjon:

$$R^2_{\text{Felles}} = 1 - \frac{\sum_{i=1}^m (\hat{\mathbf{e}}_{ii}^t \hat{\mathbf{e}}_{ii})}{\sum_{i=1}^m (\mathbf{Y}_{ii}^t \mathbf{Y}_{ii})}$$

- Gjennomsnittlig forklaringskraft

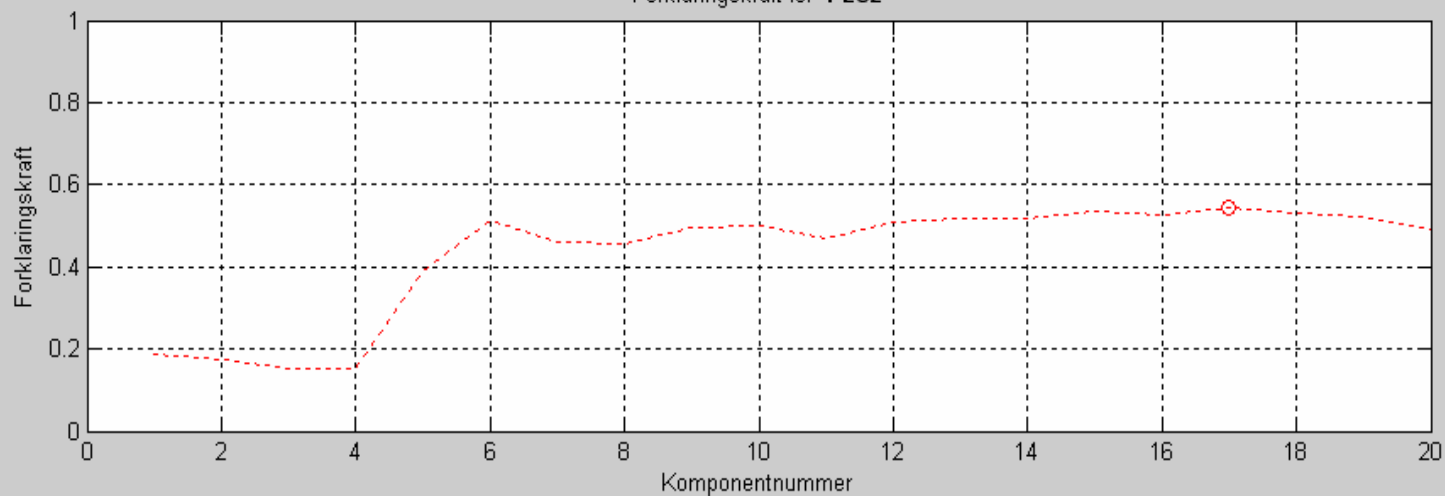
$$R^2_{\text{Snitt}} = \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\hat{\mathbf{e}}_i^t \hat{\mathbf{e}}_i}{\mathbf{Y}_i^t \mathbf{Y}_i} \right)$$

Hvor m er antall responser, \mathbf{Y} er responsmatrisa og $\hat{\mathbf{e}}$ er residualmatrisa.

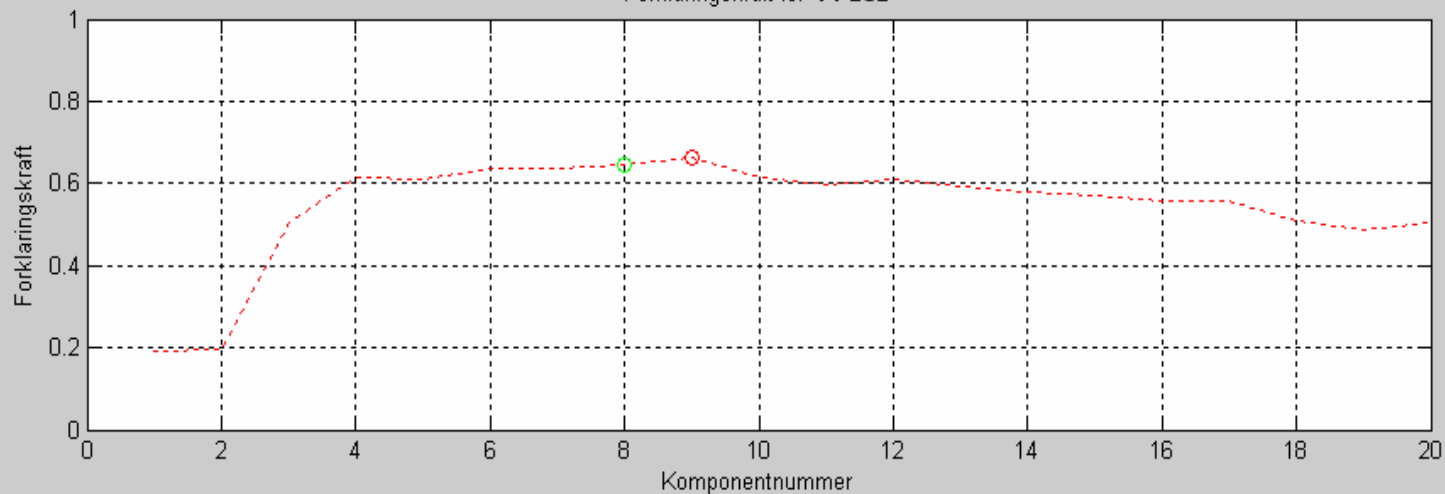
- Egenlig det samme som å se på sum av forklaringskraften for hver av responsene.

Illustrasjon

Forklaringskraft for **PLS2**



Forklaringskraft for **PPLS2**



Justert utvelgelse av antall komponenter

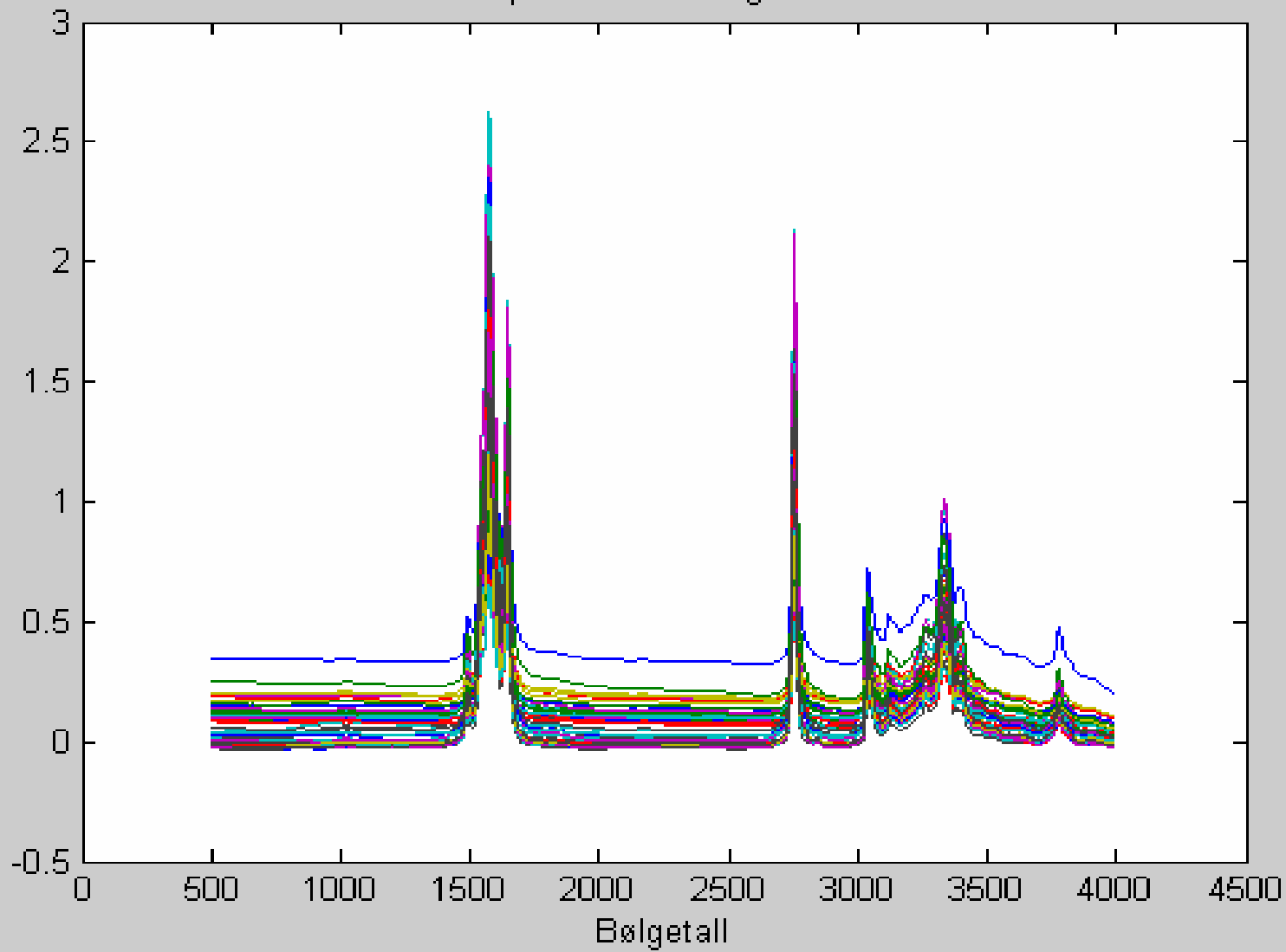
- Utfører et kji-kvadrat test for å sjekke om jeg kan godta en modell med færre antall komponenter som ikke er signifikant dårligere.
- Antar at $n * \frac{MSECV_{min}}{\sigma^2}$ følger en χ^2 fordeling med n frihetsgrader.
- For multirespons vil man ha denne situasjonen: $\sum_{i=1}^p n * \frac{MSECV_{i min}}{\sigma_i^2}$ som følger en χ^2 fordeling med n*p frihetsgrader.
- Må ha uavhengighet for at dette skal gjelde, er ikke helt oppfylt.
- Definerer et akseptabelt område: $AO_{\alpha} = \{ \sigma^2 : \pi(\sigma^2) > \alpha \}$
 hvor α er valgt signifikansnivå, $\pi(\sigma^2) = F_{np}(\sum_{i=1}^p n * \frac{MSECV_{i min}}{\sigma_i^2})$
 og F_{np} er den kumulative fordelingsfunksjonen for kji-kvadrat fordelingen med n*p frihetsgrader.
- Sørger for en mer robust modell og minsker faren for overfitting.

Presentasjon av datasett

- Marine fettsyrer i svinefett
- FT-IR biomodulen utført på utsmeltet fett
- 5 responser:
 - SFA = Saturated fatty acids. Mettete fettsyrer
 - MUFA = Monounsaturated fatty acids. Enumettete fettsyrer
 - PUFA = Polyunsaturated fatty acids. Flerumettete fettsyrer
 - C22:5/C22:6. Dette er marine fettsyrer som man vet kommer fra foret.
 - Iodtall. Grad av umettethet
- 3631 bølgenummer (variable)
- 73 prøver hvorav 49 i treningsdata og 24 i testdata.
- Det er brukt full kryssvalidering

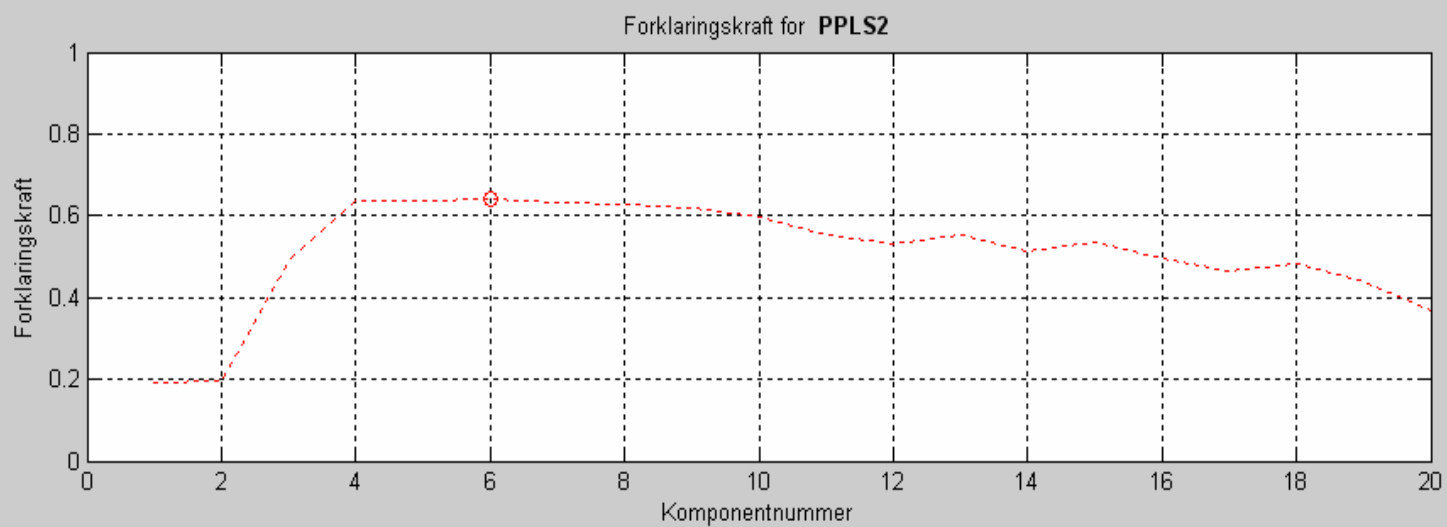
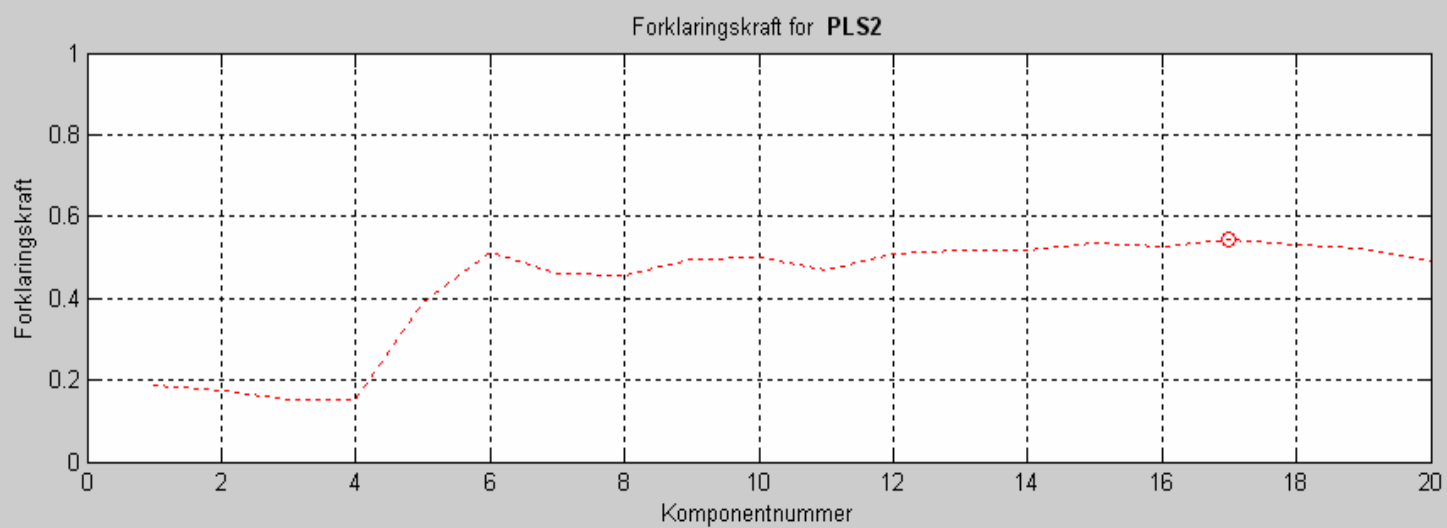
Spekteret

Spekter av treningsdata

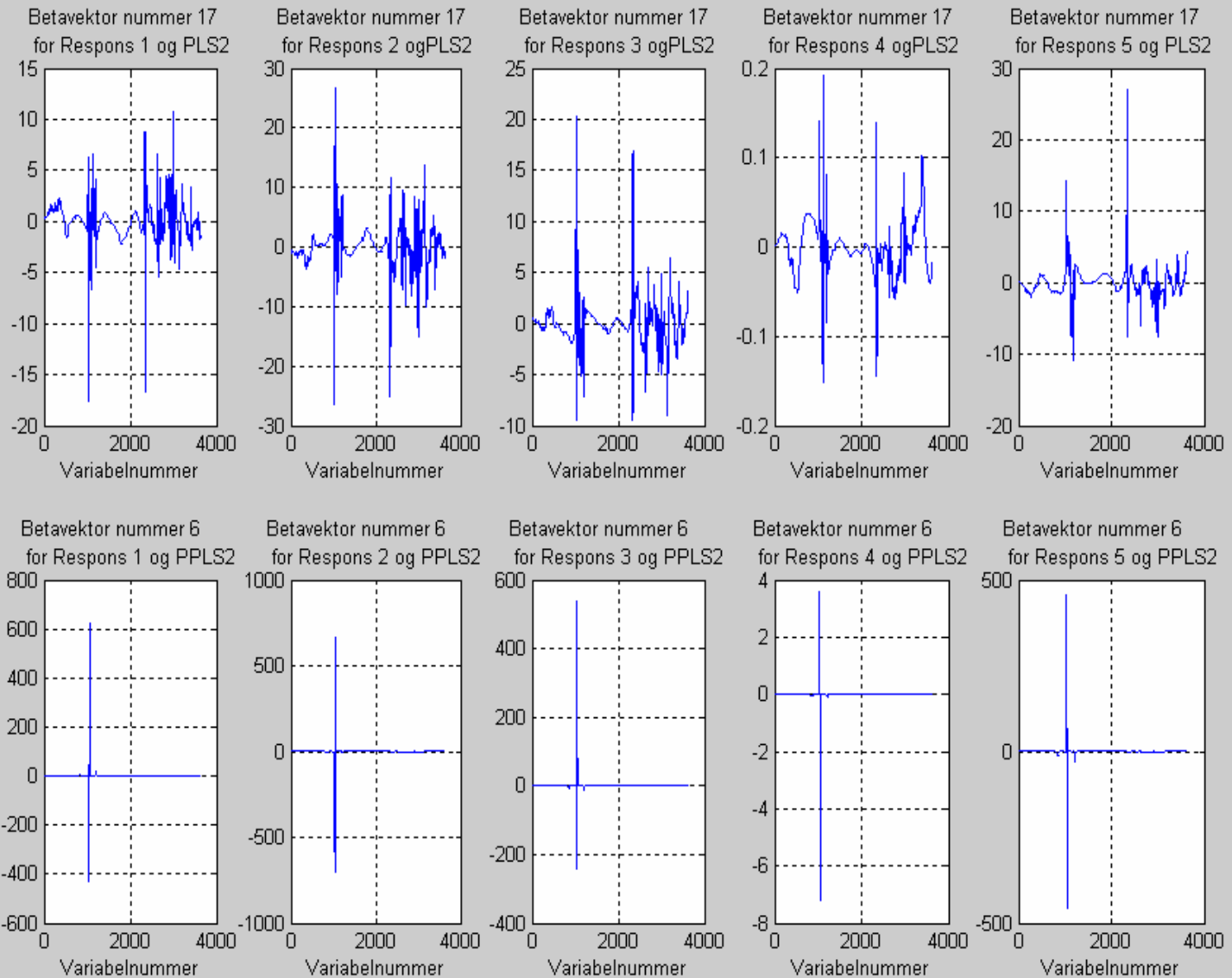




Endelig valg av komponenter



Regresjonskoeffisienter



Konklusjon/Oppsummering

- Produkt av korrelasjoner mellom en variabel og alle responsene brukes som alternativt relasjonsmål.
- Til å velge ut antall komponenter brukes gjennomsnittlig forklaringskraft.
- Bruker kji-kvadrat justering til å søke en enklere modell (hvis mulig) .