A comparetive study on PCR, PLS, Envelope and BayesPLS models

Raju Rimal

Supervisors
Solve Sæbø, Trygve Almøy
&
Joint work with
Inge Halland, UiO

4 September 2016

Norges miljø- og
biovitenskapelige
universitet

Overview

- Background
- Estimation methods under comparison
- Data Simulation
- Analysis, Results and Discussions

Background

- PLS **Population Model** [Helland, 1990] which further discussed by [Naes and Helland, 1993, Helland, 2001]

Background

- PLS **Population Model** [Helland, 1990] which further discussed by [Naes and Helland, 1993, Helland, 2001]
- PLS, *heavely developed* [Wold, 1985, Naes and Helland, 1993, De Jong, 1993], without addressing the population model [Cook et al., 2013]

Background

- PLS **Population Model** [Helland, 1990] which further discussed by [Naes and Helland, 1993, Helland, 2001]
- PLS, *heavely developed* [Wold, 1985, Naes and Helland, 1993, De Jong, 1993], without addressing the population model [Cook et al., 2013]
- Mostly popular among chemometrician

Background

- PLS **Population Model** [Helland, 1990] which further discussed by [Naes and Helland, 1993, Helland, 2001]
- PLS, *heavely developed* [Wold, 1985, Naes and Helland, 1993, De Jong, 1993], without addressing the population model [Cook et al., 2013]
- Mostly popular among chemometrician
- Was not very popular among statistician which has changed and is nowadays considered as an essential tool for multivariate analysis

Background

- PLS **Population Model** [Helland, 1990] which further discussed by [Naes and Helland, 1993, Helland, 2001]
- PLS, *heavely developed* [Wold, 1985, Naes and Helland, 1993, De Jong, 1993], without addressing the population model [Cook et al., 2013]
- Mostly popular among chemometrician
- Was not very popular among statistician which has changed and is nowadays considered as an essential tool for multivariate analysis
- Accounting the population model, new estimation methods have been purposed such as **Envelope** [Cook et al., 2010, Cook and Zhang, 2016] and **BayesPLS** [Helland et al., 2012] which are *closely related* to PLS

Background

- PLS **Population Model** [Helland, 1990] which further discussed by [Naes and Helland, 1993, Helland, 2001]
- PLS, *heavely developed* [Wold, 1985, Naes and Helland, 1993, De Jong, 1993], without addressing the population model [Cook et al., 2013]
- Mostly popular among chemometrician
- Was not very popular among statistician which has changed and is nowadays considered as an essential tool for multivariate analysis
- Accounting the population model, new estimation methods have been purposed such as **Envelope** [Cook et al., 2010, Cook and Zhang, 2016] and **BayesPLS** [Helland et al., 2012] which are *closely related* to PLS
- Cook et al. [2013] said that PLS is fundamentally an envelope in the population model

Background

- This study attempts to make an *emperial comparison* among PCR, PLS, Envelope and BayesPLS model on the basis of their **prediction ability**

Background

- This study attempts to make an *emperial comparison* among PCR, PLS, Envelope and BayesPLS model on the basis of their **prediction ability**
- Using `simrel` [Sæbø et al., 2015] R-package, data with diverse nature are simulated.

Background

- This study attempts to make an *emperial comparison* among PCR, PLS, Envelope and BayesPLS model on the basis of their **prediction ability**
- Using `simrel` [Sæbø et al., 2015] R-package, data with diverse nature are simulated.
- `simrel` allows to have control over latent structure (relevant component) of the data, fine analysis of strength and weakness of a models is possible

Statistical Model

The common ground of all the methods is to best describe (fit) the multivariate linear model below,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{1}$$

where,

$\boldsymbol{y}$	:	Response
$\boldsymbol{X}$	:	Matrix of p predictor variable
$\boldsymbol{\beta}$	:	Regression Coefficients
$\boldsymbol{\epsilon}$	:	Error $\epsilon \sim \text{NID}(0, \sigma^2)$

Here, both $\boldsymbol{y}$ and $\boldsymbol{X}$ are considered to be centered.

Statistical Model

All the models under this study consider a **subspace of predictor variables that is relevant for response**. They differ in the ways of finding the subspace and corresponding model estimates. The true estimates can also be written as,

$$\boldsymbol{\beta} = \Sigma_{XX}^{-1}\sigma_{Xy} = \sum_{j=1}^{p} \frac{1}{\alpha_j} \boldsymbol{e}_j \boldsymbol{e}_j^t \sigma_{Xy} = \sum_{j=1}^{p} \gamma_j \boldsymbol{e}_j$$

where,

γ_j	: $\frac{\boldsymbol{e}_j^t \sigma_{Xy}}{\lambda_j}$
$\boldsymbol{e}_j$	: Eigenvector of Σ_{xx}
λ_j	: Eigenvalue of Σ_{xx}
σ_{Xy}	: Covariance between y and X

So, True regression estimates are the space spanned by the eigenvectors of population covariance matrix Σ_{xx}.

Comparison of Methods

PCR	PLS
* Regression of response on latent space of predictor * No strict assumption	* Estimation through Iterative algorithm * No strict assumption

Comparison of Methods

PCR	PLS
* Regression of response on latent space of predictor * No strict assumption	* Estimation through Iterative algorithm * No strict assumption

Envelope (MLE)	Bayes
* Estimation using Maximum Likelihood	* Estimation through MCMC approach with rotation of relevant space
* Can not be used when predictor is larger than observations	* Heavy Computation when p is large

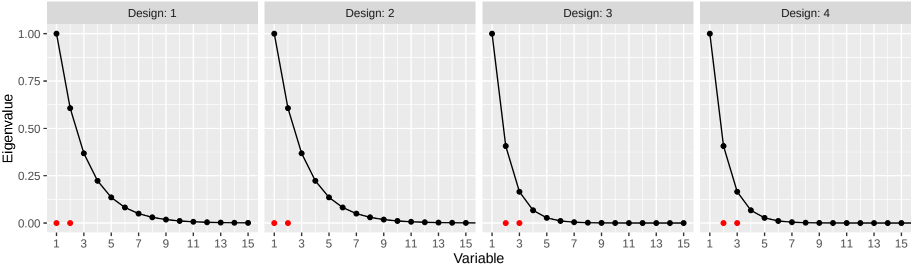
Data Simulation

Models are analysed under diverse nature of data. Data are simulated using `simrel` package (R). In this study, I have included following four design;

n	p	R2	relpos	gamma
50	15	0.5	1, 2	0.5
50	40	0.5	1, 2	0.5
50	15	0.9	2, 3	0.9
50	40	0.9	2, 3	0.9

`n`	:	Number of observations
`p`	:	Number of variables
`R2`	:	Variation explained by the model
`relpos`	:	Position of relevant components
`gamma`	:	Reduction factor of eigenvalue of X

Relevant Position and Eigenvalues



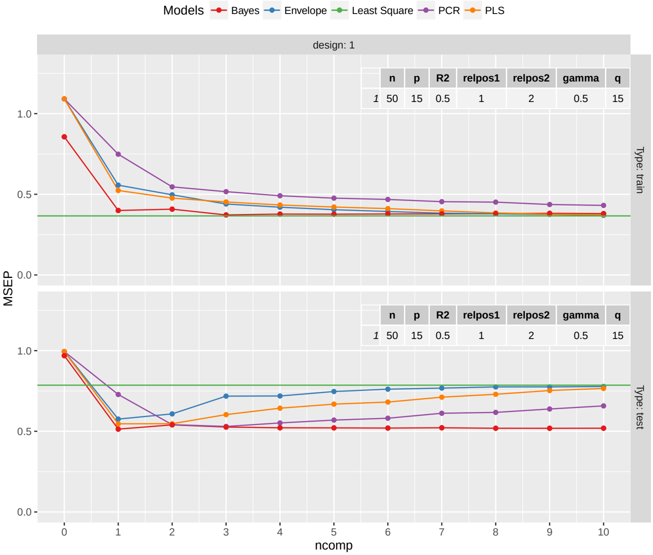
- When Relevant components are at the position of high eigenvalues, the situation is easier to model
- When Relevant components are at the position of low eigenvalues, for example 5, 10, then the most variation present in X are not relevant for Y and this will become a very difficult situation.

Model assessment

Models are compared on the basis of their prediction ability by measuring *test* and *training* **Mean Square Error of Prediction (*MSEP*)**. Mean prediction error is calculated as,

$$
\begin{aligned}
(\text{Prediction Error})_{\text{training}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i - \left(\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}\boldsymbol{X}_i\right)\right)^2 \\
(\text{Prediction Error})_{\text{test}} &= \frac{1}{n}\sum_{i=1}^{\text{ntest}}\left(\boldsymbol{y}_{i(\text{test})} - \hat{\boldsymbol{y}}_{i(\text{test})}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{\text{ntest}}\left(\boldsymbol{y}_{i(\text{test})} - \left(\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}\boldsymbol{X}_{i(\text{test})}\right)\right)^2
\end{aligned}
$$

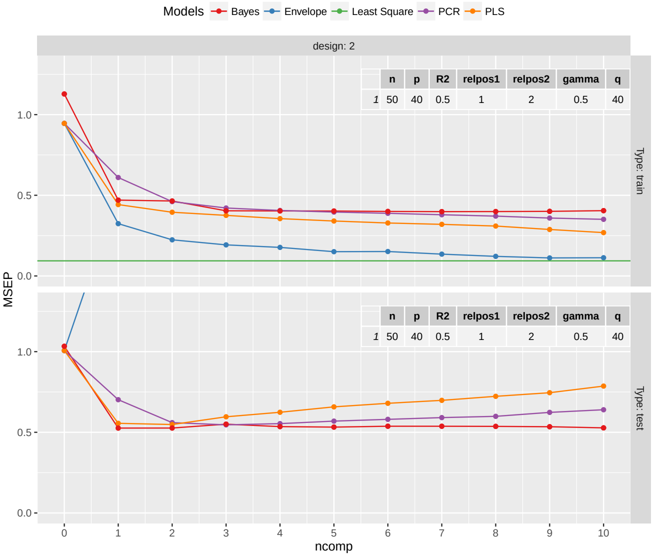
Analysis Results



	n	p	R2	relpos1	relpos2	gamma	q
1	50	15	0.5	1	2	0.5	15

	n	p	R2	relpos1	relpos2	gamma	q
1	50	15	0.5	1	2	0.5	15

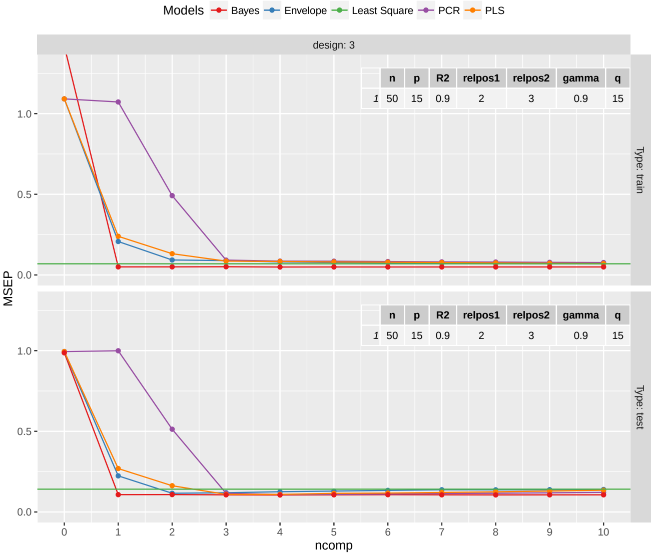
Analysis Results



	n	p	R2	relpos1	relpos2	gamma	q
1	50	40	0.5	1	2	0.5	40

	n	p	R2	relpos1	relpos2	gamma	q
1	50	40	0.5	1	2	0.5	40

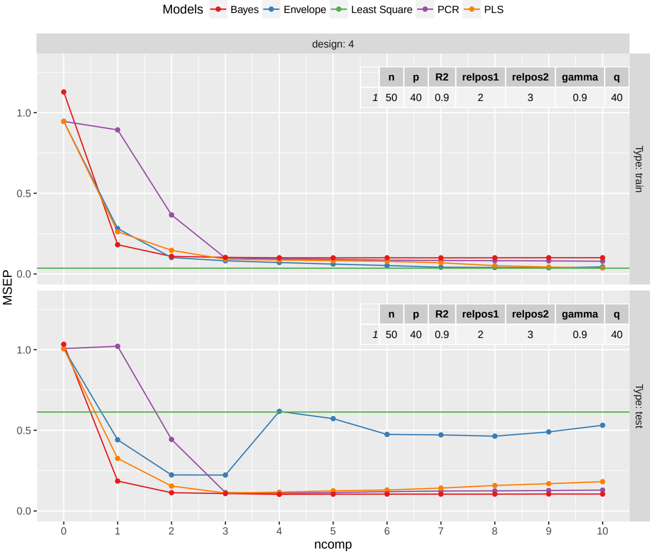
Analysis Results



	n	p	R2	relpos1	relpos2	gamma	q
1	50	15	0.9	2	3	0.9	15

	n	p	R2	relpos1	relpos2	gamma	q
1	50	15	0.9	2	3	0.9	15

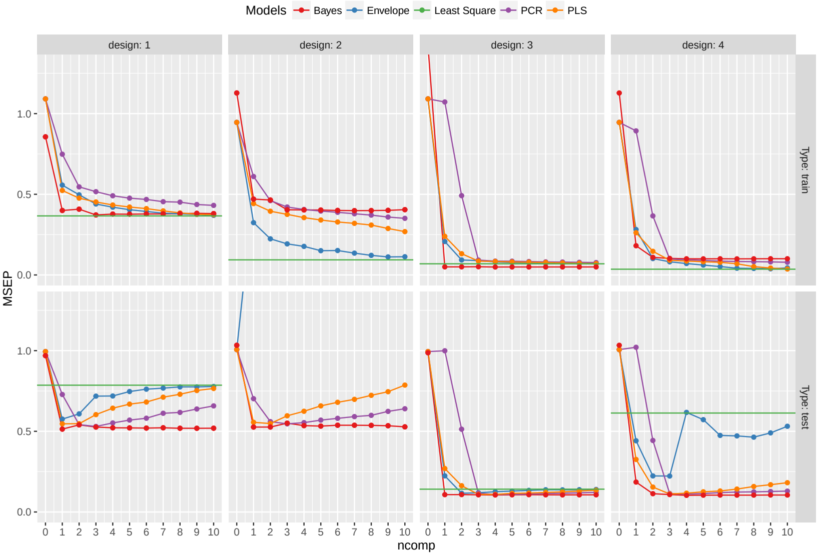
Analysis Results



	n	p	R2	relpos1	relpos2	gamma	q
1	50	40	0.9	2	3	0.9	40

	n	p	R2	relpos1	relpos2	gamma	q
1	50	40	0.9	2	3	0.9	40

Analysis Results



Conclusion

- New methods – Envelope and Bayes, as they claim, are performing better than algorithmic approach of PLS

Conclusion

- New methods – Envelope and Bayes, as they claim, are performing better than algorithmic approach of PLS
- However, the performance of MLE approach of Envelope is not satisfactory when number of variable is large

Conclusion

- New methods – Envelope and Bayes, as they claim, are performing better than algorithmic approach of PLS
- However, the performance of MLE approach of Envelope is not satisfactory when number of variable is large
- In the case of Bayes PLS, the prediction error does not raises noticably (test prediction) after capturing enough information with few components

Conclusion

- New methods – Envelope and Bayes, as they claim, are performing better than algorithmic approach of PLS
- However, the performance of MLE approach of Envelope is not satisfactory when number of variable is large
- In the case of Bayes PLS, the prediction error does not raises noticably (test prediction) after capturing enough information with few components
- This suggests that it is able to find the direction of maximum variation after successive rotations of predictor subspace

Conclusion

- New methods – Envelope and Bayes, as they claim, are performing better than algorithmic approach of PLS
- However, the performance of MLE approach of Envelope is not satisfactory when number of variable is large
- In the case of Bayes PLS, the prediction error does not raises noticably (test prediction) after capturing enough information with few components
- This suggests that it is able to find the direction of maximum variation after successive rotations of predictor subspace
- The computation regarding BayesPLS is intensive which will not be fisible in case of wide dataset (very common in genomic data)

Conclusion

- New methods – Envelope and Bayes, as they claim, are performing better than algorithmic approach of PLS
- However, the performance of MLE approach of Envelope is not satisfactory when number of variable is large
- In the case of Bayes PLS, the prediction error does not raises noticably (test prediction) after capturing enough information with few components
- This suggests that it is able to find the direction of maximum variation after successive rotations of predictor subspace
- The computation regarding BayesPLS is intensive which will not be fisible in case of wide dataset (very common in genomic data)
- All the models are performing better than the least square solution

salamat
Dakujem
teşekkür ederim
धन्यवाद
GRACiAS
TAK
SUWUN
TAKK
ASANTE
hvala
salamat
HVALA
mersi
شكرا
جزيل
Ευχαριστώ
감사합니다
GRAZZii
DANKE
GRAZAS
kiitos
merci
salamat
TAKK
MAHALO
Рахмет
Thank You
kiitos
arigato
ARiGATO
takk
suwun
धन्यवाद
HVALA
GRAZiE
DAKUJEM
teşekkür ederim
MERCi
kiitos
hvala
mahalo
GRACiAS
GRAZAS
TAKK
ありがとう
DANKE
Благодарам
ASANTE
salamat
SUWUN
grazie
TAKK
多謝
SALAMAT
спасибо
gracias

References

R Dennis Cook and Xin Zhang. Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics*, 25(1):284–300, 2016.

R Dennis Cook, Bing Li, and Francesca Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, pages 927–960, 2010.

RD Cook, IS Helland, and Z Su. Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):851–877, 2013.

Sijmen De Jong. Simpls: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems*, 18(3): 251–263, 1993.

Inge S Helland. Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, pages 97–114, 1990.

Inge S Helland. Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 58(2): 97–107, 2001.

Inge S Helland, Solve Saebø, Ha Tjelmeland, et al. Near optimal prediction from relevant components. *Scandinavian Journal of Statistics*, 39(4):695–713, 2012.

Tormod Naes and Inge S Helland. Relevant components in regression. *Scandinavian journal of statistics*, pages 239–250, 1993.

Solve Sæbø, Trygve Almøy, and Inge S Helland. simrel—a versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. *Chemometrics and Intelligent Laboratory Systems*, 146:128–135, 2015.

Herman Wold. Partial least squares. *Encyclopedia of statistical sciences*, 1985.