A primer on Partial Least Squares Regression

Dennis Cook

School of Statistics
University of Minnesota

Collaborating at times with
Liliana Forzani, Inge Helland, and Zhihua Su

PLS Regression $\in$ PLS Family of Methods

Partial Least Squares (PLS), aka Projection onto Latent Structures, and PLS regression are not synonyms.

Herman Wold (1908–1992)

- PhD in 1938, Harald Cramér, advisor
- Contributions: Cramér-Wold theorem, time series, casual modeling, econometrics, . . . originator of PLS family of methods

Wold's algorithms leading to PLS

- fixed-point algorithm, early 1960's: iterative OLS regressions to estimate the parameters of multi-equation systems
- NILES, Nonlinear Iterative Least Squares: modification for principal components and canonical correlations
- NIPALS, late 1960's: Nonlinear Iterative Partial Least Squares: refinement of NILES.
- PLS-PM, 1970's: Path Modeling, aka Structural Equation Modeling, via PLS

PLS Regression Algorithm

Developed around 1980 by the Scandinavian chemometrics community – S. Wold, H. Martens & H. Wold – mainly for prediction in high-dimensional multi-response regressions, without requiring that the sample size n be larger than the number of predictors, p.

PLS regression often does better than OLS when $n > p$, as judged empirically by cross validation.

1. Cursory introduction to PLS regression
2. History of PLS regression
3. PLS regression formulation via envelopes, recent model-based dimension-reduction methodology that can greatly reduce estimative and predictive variation relative to standard methods.
4. PLS n, p-asymptotics

1. Introduction to PLS regression

Context

Consider the usual linear regression model

$$Y_i = \boldsymbol{\alpha} + \boldsymbol{\beta}^T \mathbf{X}_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

- Y univariate (Why?)
- $\mathbf{X}$: $p \times 1$, $\text{var}(\mathbf{X}) = \boldsymbol{\Sigma}_{\mathbf{X}}$, sample version $\mathbf{S}_{\mathbf{X}}$.
- Errors $\boldsymbol{\varepsilon}$ have mean 0 and variance τ^2, $\varepsilon \perp\!\!\!\perp \mathbf{X}$
- $\boldsymbol{\beta}$: $p \times 1$, unknown coefficients
- Goal: estimate $\boldsymbol{\beta}$, prediction.
- $\mathbf{B}$ = OLS estimator of $\boldsymbol{\beta}$ when $n > p$.

PLS regressions – algorithms for reducing the predictor dimension

1. Select a matrix $\widehat{\mathbf{\Phi}} \in \mathbb{R}^{p\times q}$, with $q \ll p$, $q =$ number of components, and then reduce $\mathbf{X} \mapsto \widehat{\mathbf{\Phi}}^T\mathbf{X}$.
2. Fit $Y = \alpha + \boldsymbol{\eta}^T\{\widehat{\mathbf{\Phi}}^T\mathbf{X}\} + \varepsilon$ using OLS when $n > q$.
3. Estimator is of the form

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \widehat{\mathbf{\Phi}}\hat{\boldsymbol{\eta}} = \widehat{\mathbf{\Phi}}(\widehat{\mathbf{\Phi}}^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Phi}})^{-1}\widehat{\mathbf{\Phi}}^T\mathbf{S}_{\mathbf{X}Y} \\ &= \mathbf{P}_{\widehat{\mathbf{\Phi}}(\mathbf{S}_{\mathbf{X}})}\mathbf{B} \text{ when } n > p\end{aligned}$$

$\mathbf{S}_{\mathbf{X}}$: Sample variance of $\mathbf{X}$; $\mathbf{S}_{\mathbf{X}Y}$: sample covariance between $\mathbf{X}$ and Y.

How do we get an estimator span$(\widehat{\mathbf{\Phi}})$? PLS regression algorithm.

SIMPLS algorithm (de Jong, 1993), $n > q$

Set $\widehat{\varphi}_0 = 0$ and $\widehat{\Phi}_0 = (\widehat{\varphi}_0)$. For $k = 0, \ldots, q-1$, set

Version I

$$\begin{aligned} \widehat{\varphi}_{k+1} &= \arg\max_{\mathbf{w}} \mathbf{w}^T \mathbf{S}_{\mathbf{X}Y} \mathbf{S}_{\mathbf{X}Y}^T \mathbf{w}, \text{ subject to} \\ & \quad \mathbf{w}^T \mathbf{S}_{\mathbf{X}} \widehat{\Phi}_k = 0 \text{ and } \mathbf{w}^T \mathbf{w} = 1 \\ \widehat{\Phi}_{k+1} &= (\widehat{\varphi}_0, \ldots, \widehat{\varphi}_k, \widehat{\varphi}_{k+1}) \end{aligned}$$

Version II With $\mathbf{Q}_{\mathcal{S}_k} = \mathbf{I} - \mathbf{P}_{\mathcal{S}_k}$,

$$\begin{aligned} \mathcal{S}_k &= \text{span}(\mathbf{S}_{\mathbf{X}} \widehat{\Phi}_k) \\ \widehat{\varphi}_{k+1} &= \mathbf{Q}_{\mathcal{S}_k} \mathbf{S}_{\mathbf{X}Y} / \|\mathbf{Q}_{\mathcal{S}_k} \mathbf{S}_{\mathbf{X}Y}\| \\ \widehat{\Phi}_{k+1} &= (\widehat{\varphi}_0, \ldots, \widehat{\varphi}_k, \widehat{\varphi}_{k+1}). \end{aligned}$$

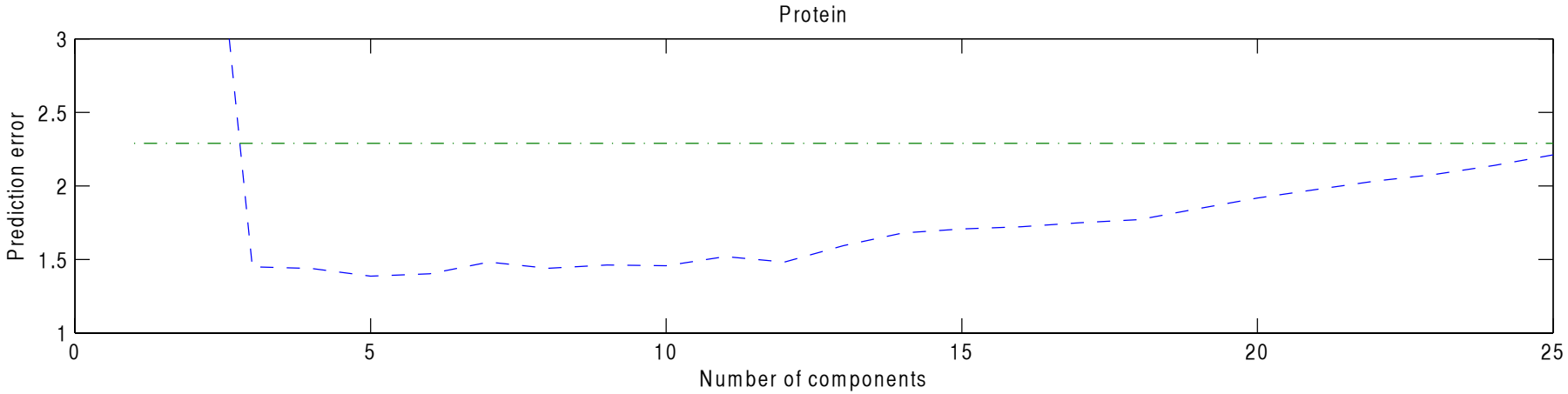
At termination $\widehat{\Phi}_{\text{pls}} = \widehat{\Phi}_q$.

Version III

$$\text{span}(\widehat{\Phi}_{\text{pls}}) = \text{span}(\mathbf{S}_{\mathbf{X}Y}, \mathbf{S}_{\mathbf{X}} \mathbf{S}_{\mathbf{X}Y}, \ldots, \mathbf{S}_{\mathbf{X}}^{q-1} \mathbf{S}_{\mathbf{X}Y})$$

Beef protein

Predict protein content (Y) of beef based on spectral measurements at $p = 50$ wave lengths, $n = 103$. 5-fold cross validation used for root mean squared prediction error. Blue, PLS; Green, OLS.



2. Brief History of PLS regression, Mixed with Personal Encounters

- S. Wold, H. Martens and H. Wold (1983).
- Los Alamos National Labs, ~1980.
- Gordon Research Conferences on Statistics in Chemistry and Chemical Engineering, 1980–mid 1990's
- Statistics: Helland (1990, 1992, 2001), Næs and Helland (1993), Frank and Friedman (1993), Garthwaite (1994)
- Chemometrics: de Jong's (1993) SIMPLS algorithim
- Chemometrics conference early 2000's
- Outside of Chemometrics: Micro-array data, FMRI data, biomedical analyses, tumor classification, bioprocesses, forecasting, characteristics of craft beer, ...
  Selling point – serviceable when $n < p$.

- Chun and Keleş (2010): PLS regression is consistent for $\boldsymbol{\beta}$ only if $p/n \to 0$ as n & $p \to \infty$. That is, $\widehat{\boldsymbol{\beta}}_{\text{pls}} \xrightarrow{pr} \boldsymbol{\beta}$ only if only if $p/n \to 0$. (same context)
  1.
  2.
  3.
  4.

- Chun and Keleş (2010): PLS regression is consistent for $\boldsymbol{\beta}$ only if $p/n \to 0$ as n & $p \to \infty$. That is, $\widehat{\boldsymbol{\beta}}_{\text{pls}} \xrightarrow{p} \boldsymbol{\beta}$ only if only if $p/n \to 0$. (same context)

  1. Consistency doesn't always signal the usefulness of a method in practice.
  2.
  3.
  4.

- Chun and Keleş (2010): PLS regression is consistent for $\boldsymbol{\beta}$ only if $p/n \to 0$ as n & $p \to \infty$. That is, $\widehat{\boldsymbol{\beta}}_{\text{pls}} \xrightarrow{p} \boldsymbol{\beta}$ only if only if $p/n \to 0$. (same context)

  1. Consistency doesn't always signal the usefulness of a method in practice.
  2. Operationally easy way to fit $n < p$ regressions that performed well against the competing methods of the day.
  3.
  4.

- Chun and Keleş (2010): PLS regression is consistent for $\boldsymbol{\beta}$ only if $p/n \to 0$ as n & $p \to \infty$. That is, $\widehat{\boldsymbol{\beta}}_{\text{pls}} \xrightarrow{p} \boldsymbol{\beta}$ only if only if $p/n \to 0$. (same context)

  1. Consistency doesn't always signal the usefulness of a method in practice.
  2. Operationally easy way to fit $n < p$ regressions that performed well against the competing methods of the day.
  3. The technical restrictions (eg. $\Sigma_{\mathbf{X}}$ bdd as $p \to \infty$) of Chun and Keleş ruled out important practical contexts.
  4.

- Chun and Keleş (2010): PLS regression is consistent for $\boldsymbol{\beta}$ only if $p/n \to 0$ as n & $p \to \infty$. That is, $\widehat{\boldsymbol{\beta}}_{\text{pls}} \xrightarrow{p} \boldsymbol{\beta}$ only if only if $p/n \to 0$. (same context)

  1. Consistency doesn't always signal the usefulness of a method in practice.
  2. Operationally easy way to fit $n < p$ regressions that performed well against the competing methods of the day.
  3. The technical restrictions (eg. $\Sigma_{\mathbf{X}}$ bdd as $p \to \infty$) of Chun and Keleş ruled out important practical contexts.
  4. Chun and Keleş used their result to motivate a sparse version of PLS, but H. Wold, et al. (circa 1988) & S. Wold, et al. (1996) argued against sparse versions of PLS.

- Inge Helland's 2011 visit: Envelopes & PLS regression algorithms
- Cook, Helland and Su (2013):
  - gave first firm statistical model for PLS regression
  - proved that PLS regression provides a root-n consistent moment-based envelope method (p fixed)
  - showed that likelihood-based envelope methods dominate PLS when $n \gg p$ (p fixed)
- Cook and Forzani (2017, 2018) studied PLS in high-dimensional ($n < p$) regressions.

3. PLS regression via Envelopes

(Cook, Helland & Su, 2013)

Motivation

Driving question: Are there linear combinations $\mathbf{\Phi}^T\mathbf{X}$ of $\mathbf{X}$ that carry all the information that $\mathbf{X}$ has about Y? So, changing $\mathbf{X}$ while holding $\mathbf{\Phi}^T\mathbf{X}$ fixed has no impact on the distribution of Y.

Motivation

Driving question: Are there linear combinations $\mathbf{\Phi}^T\mathbf{X}$ of $\mathbf{X}$ that carry all the information that $\mathbf{X}$ has about Y? So, changing $\mathbf{X}$ while holding $\mathbf{\Phi}^T\mathbf{X}$ fixed has no impact on the distribution of Y.

Let $\mathcal{E} = \text{span}(\mathbf{\Phi})$. Formally, we seek the smallest subspace $\mathcal{E} \subseteq \mathbb{R}^p$ so that

$$(Y, \mathbf{P}_{\mathcal{E}}\mathbf{X}) \perp\!\!\!\perp \mathbf{Q}_{\mathcal{E}}\mathbf{X}.$$

Consequently, $\mathbf{X}$ affects Y only via $\mathbf{P}_{\mathcal{E}}\mathbf{X}$, the material part of $\mathbf{X}$.

The condition $(Y, \mathbf{P}_{\mathcal{E}}\mathbf{X}) \perp\!\!\!\perp \mathbf{Q}_{\mathcal{E}}\mathbf{X}$ holds iff

1. $\mathcal{B} := \text{span}(\boldsymbol{\beta}) \subseteq \mathcal{E}$, so $\mathcal{E}$ envelops $\mathcal{B}$
2. $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{P}_{\mathcal{E}} + \mathbf{Q}_{\mathcal{E}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{Q}_{\mathcal{E}}$ so $\mathcal{E}$ is a reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{X}}$

- The intersection of all such subspaces $\mathcal{E}$ is called the $\boldsymbol{\Sigma}_{\mathbf{X}}$-envelope of $\mathcal{B}$, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ with semi-orthogonal basis $\boldsymbol{\Phi}$
- Model:

$$\begin{aligned} Y &= \alpha + \boldsymbol{\eta}^T(\boldsymbol{\Phi}^T\mathbf{X}) + \varepsilon \\ \boldsymbol{\Sigma}_{\mathbf{X}} &= \boldsymbol{\Phi}\boldsymbol{\Delta}\boldsymbol{\Phi}^T + \boldsymbol{\Phi}_0\boldsymbol{\Delta}_0\boldsymbol{\Phi}_0^T \end{aligned}$$

  $\boldsymbol{\beta} = \boldsymbol{\Phi}\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$ are still of primary interest. $\mathbf{X}$ not ancillary.
- SIMPLS return a basis for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ in the population and $q = \dim(\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B}))$, the *number of components*

How do we get an estimator of $\mathcal{E}_{\mathbf{\Sigma}_X}(\mathcal{B})$? Two options for now:

- Use MLE when $n \gg p$, giving $\widehat{\mathbf{\Phi}}_{\text{mle}}$
- Use SIMPLS since $\widehat{\mathbf{\Phi}}_{\text{pls}}$ is a root-n consistent estimator of a basis for $\mathcal{E}_{\mathbf{\Sigma}_X}(\mathcal{B})$ when q is known.

In either case, as we saw, we estimate

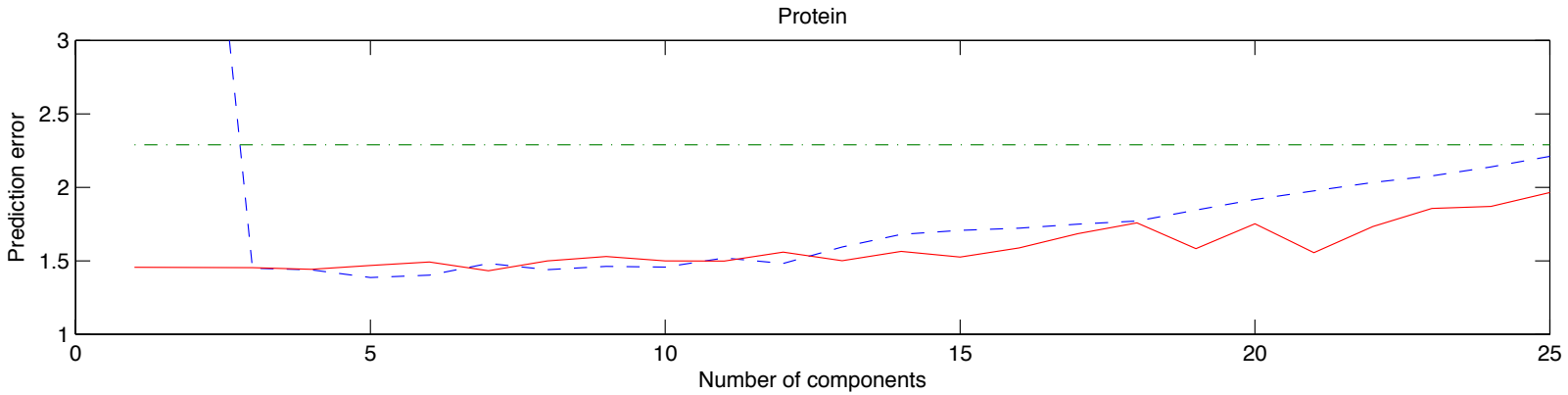
$$\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{\Phi}}\hat{\boldsymbol{\eta}} = \widehat{\mathbf{\Phi}}(\widehat{\mathbf{\Phi}}^T\mathbf{S}_X\widehat{\mathbf{\Phi}})^{-1}\widehat{\mathbf{\Phi}}^T\mathbf{S}_{XY}$$

with $\widehat{\mathbf{\Phi}} = \widehat{\mathbf{\Phi}}_{\text{pls}}$ or $\widehat{\mathbf{\Phi}} = \widehat{\mathbf{\Phi}}_{\text{mle}}$, both of which provide root-n consistent estimators of $\boldsymbol{\beta}$ when q is assumed known.

Asymptotically, $\widehat{\boldsymbol{\beta}}_{\text{mle}}$ is better than $\widehat{\boldsymbol{\beta}}_{\text{pls}}$, but perhaps not when $n < p$ or when $n \not\gg p$.

Beef protein

Predict protein content (Y) of beef based on spectral measurements at $p = 50$ wave lengths, $n = 103$. 5-fold cross validation used for prediction errors. Blue, PLS; Green, OLS, Red, Env.



Comparison of PLS and Envelopes

Rimal, Almøy and Sæbø(2019). Comparison of Multi-response Prediction Methods. *Chemometrics and Intelligent Laboratory Systems*, to appear.

> *Analysis using both simulated data and real data has shown that the envelope methods are more stable, less influenced by . . . [predictor collinearity] and in general, performed better than PCR and PLS methods. These methods are also found to be less dependent on the number of components.*

When $n < p$ they used PCA to reduce the predictors prior to applying envelope methodology.

> *The results from this study will . . . encourage [researchers] to use newly developed methods such as the envelopes.*

4. PLS n, p Asymptotics

(Cook & Forzani, 2017, 2018)

Same context with

- $q = \dim(\mathcal{E}_{\mathbf{\Sigma}_X}(\mathcal{B}))$ fixed. $\mathbf{\Phi}$ semi-orthogonal basis for $\mathcal{E}_{\mathbf{\Sigma}_X}(\mathcal{B})$. $\mathbf{\Phi}^T\mathbf{X}$ true active predictors.
- $\tau^2 = \text{var}(Y \mid \mathbf{X})$ bounded away from 0
- $\boldsymbol{\beta} \neq 0$, so $q \geqslant 1$

Gauges: Fitted value $\widehat{Y}_N$ at a new independent $\mathbf{X}_N$: $\widehat{Y}_N - E(Y|\mathbf{X}_N)$ & Estimation: $\|\widehat{\boldsymbol{\beta}}_{\text{pls}} - \boldsymbol{\beta}\|_{\mathbf{\Sigma}_X}$

Governing Quantities

Let $\mathbf{\Phi}^T\mathbf{X}$ and $\mathbf{\Phi}_0^T\mathbf{X}$ be true active and inactive predictors, $(\mathbf{\Phi}, \mathbf{\Phi}_0)$ orthogonal.

- Colinearily: $\rho(p) =$ sum of population VIFs for the reg. of Y on $\mathbf{\Phi}^T\mathbf{X}$.
- Signal: $\eta(p) \asymp \text{trace}(\mathbf{\Phi}^T\mathbf{\Sigma_X}\mathbf{\Phi}) = \text{trace}\{\text{var}(\mathbf{\Phi}^T\mathbf{X})\}$
- Noise: $\kappa(p) \asymp \text{trace}(\mathbf{\Phi}_0^T\mathbf{\Sigma_X}\mathbf{\Phi}_0) = \text{trace}\{\text{var}(\mathbf{\Phi}_0^T\mathbf{X})\}$

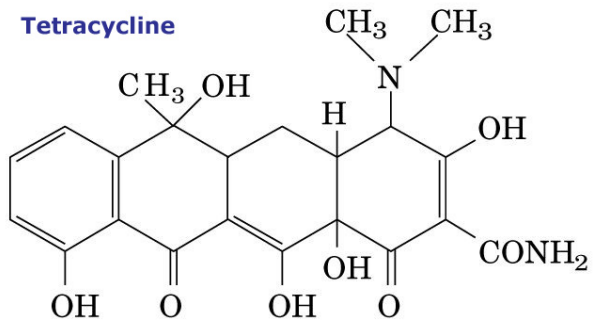
Special cases for reference:

- Abundance: If, as $p \to \infty$, $\|\mathbf{\Sigma_{XY}}\|^2 \to \infty$ then $\eta \to \infty$
- Sparsity: If the regression is sparse (only q predictors are material) then η is bounded.

Conclusions, all assuming bounded $\rho(p)$ and noise $\kappa(p) = \text{trace}\{\text{var}(\mathbf{\Phi}_0^T\mathbf{X})\} \asymp p$

- If the signal $\eta(p) = \text{trace}\{\text{var}(\mathbf{\Phi}^T\mathbf{X})\}$ is bounded then PLS is not consistent and may not work well in high dimensional regressions unless $n \gg p$. (sparsity, as argued against by H. Wold, S. Wold et al.)
- If the signal $\eta(p)$ is unbounded then PLS is consistent and should work well in high dimensional regressions, even when $n < p$. (abundance, as argued for by H. Wold, S. Wold et al.)
- Chun-Keleş' assumption that the eigenvalues of $\text{var}(\mathbf{X})$ are bounded away from 0 and ∞ implies sparsity, which motivated their sparse methodology.

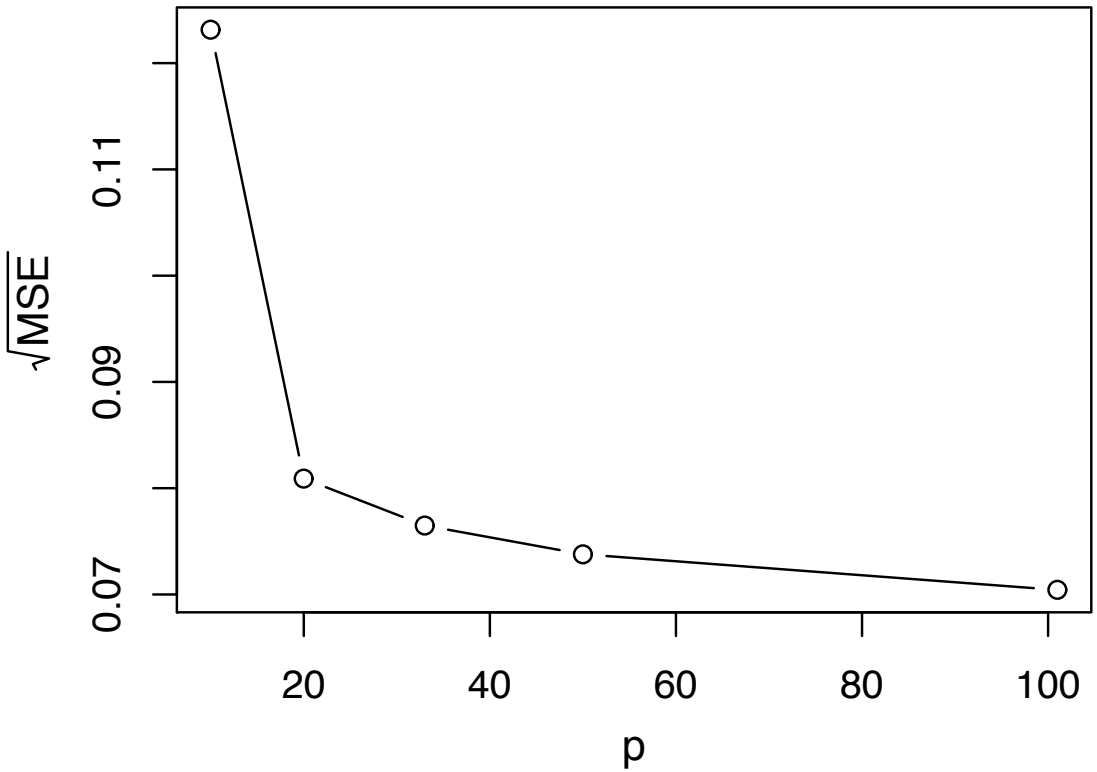
Tetracycline



Goal: Predict Tetracycline concentration in blood. $n = 50$ training samples constructed by spiking blank sera. $m = 57$ validation samples constructed in the same way.

Predictors: fluorescence intensity at $p = 101$ equally spaced points in $450 - 550$ nm. Original PLS analysis indicated that $q = 4$ (Goicoechea and Olivieri, 2009) .

We repeated the analysis by selecting $p = 10, 20, 33, 50, 101$ equally spaced points, computing the validation root-MSE prediction in each case.


$\sqrt{\text{MSE}}$
0.07
0.09
0.11
20
40
60
80
100
p

Overall conclusions

- Envelope methodology and SIMPLS estimate the same population quantity, $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B})$.
- Envelope methodology generally gives better results than SIMPLS, even when $n < p$ (Rimal, et al.).
- SIMPLS estimates are generally easier to compute.
- SIMPLS remains a serviceable method in abundant regressions.

Computing: z.umn.edu/envelopes

Papers: www.stat.umn.edu/~dennis

Book: Cook, R. D. (2018). *An introduction to Envelopes*. Wiley

Thank you!