

Partial Least Squares Regression

Introduction/Motivation

Examples

- **Generic Spectroscopy Example**

1. Predict chemical composition, y , of a compound based on signals for a particular wavelength (x 's).
2. 317 wavelengths $\longrightarrow x_1, x_2, \dots, x_{317}$.

- **Spectrometric Example—Observed Data**

1. Predict the amounts of three compounds present in samples from Baltic Sea:

$y_1 = LS =$ lignin sulfonate: pulp industry pollution

$y_2 = HA =$ humic acids: natural forest products

$y_3 = DT =$ optical whitener from detergent

2. 16 samples (n) of known concentration of LS, HA, DT.
3. Spectra/signals from 27 frequencies (wavelengths) $\longrightarrow x$'s.

- **DNA Microarray Expression Data**

The Problem

Consider the standard regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (1)$$

where

$\mathbf{X} : n \times p$ matrix of p predictors

$\mathbf{y} : n \times 1$ vector of responses

$\boldsymbol{\beta} : p \times 1$ vector of regression coefficients

$\boldsymbol{\epsilon} : n \times 1$ vector of iid $(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ errors.

- Define a K -dimensional subspace of p -dim Euclidean space.
- Perform the regression under the restriction that the coefficient vector $\boldsymbol{\beta}$ lies in that subspace; i.e.,

$$\boldsymbol{\beta} = \sum_{k=1}^K a_k \mathbf{c}_k \quad (2)$$

where $\{\mathbf{c}_k\}_1^K$ span the subspace with $\mathbf{c}'_k \mathbf{c}_k = 1$.

- How should the subspace be define/constructed?
- The $\{\mathbf{c}_k\}_1^K$ are constructed to satisfy some optimality condition.
- RR, PCR, PLS, CR (and other methods) fit into this context.
- Purpose here is to introduce PLS and its relationship to RR PCR and CR.
- Throughout assume $\text{rank}(\mathbf{X}'\mathbf{X}) = p$.

Ordinary Least Squares (OLS)

- **The Subspace** : is defined by the (single) unit vector that maximizes the squared sample correlation between the response \mathbf{y} and the linear combination (L.C.) of \mathbf{X} , \mathbf{Xc} .
- **The Problem**: Find \mathbf{c} s.t.

$$T_{ols}(\mathbf{c}) \equiv \text{corr}^2(\mathbf{Xc}, \mathbf{y}) = \frac{\mathbf{c}'\mathbf{M}\mathbf{c}}{z\mathbf{c}'\mathbf{S}\mathbf{c}} \quad (3)$$

is maximized (where $\mathbf{M} = \mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X} \equiv \mathbf{s}\mathbf{s}'$, $\mathbf{S} = \mathbf{X}'\mathbf{X}$, $z = \mathbf{y}'\mathbf{y}$) subject to $\mathbf{c}'\mathbf{c} = 1$.

- **The Solution** : is obtained by solving the eigen-equation

$$\mathbf{S}^{-1}\mathbf{M}\mathbf{c} = \lambda\mathbf{c} \quad (4)$$

which yields (the solution)

$$\mathbf{c}_{ols} = \operatorname{argmax}_{\mathbf{c}'\mathbf{c}=1} \operatorname{corr}^2(\mathbf{X}\mathbf{c}, \mathbf{y}) = \frac{\hat{\boldsymbol{\beta}}_{ols}}{\|\hat{\boldsymbol{\beta}}_{ols}\|} \quad (5)$$

where $\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

- Thus, the L.C. of \mathbf{X} , $\mathbf{X}\mathbf{c}$ ($\mathbf{c}'\mathbf{c} = 1$) s.t. $\operatorname{corr}^2(\mathbf{X}\mathbf{c}, \mathbf{y})$ is maximized is $\mathbf{X}\mathbf{c}_{ols}$.
- **Prediction** : $\hat{\mathbf{y}}_{ols} = \mathbf{X}\hat{\boldsymbol{\beta}}_{ols}$ (projection of \mathbf{y} onto $\mathbf{X}\mathbf{c}_{ols}$)

Principal Component Regression (PCR)

- Spectral decomposition:

$$\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' \quad \mathbf{\Lambda} = \text{diag}\{e_1^2, \dots, e_p^2\} \quad (6)$$

where $\{e_i^2, \mathbf{v}_i\}_1^p$ are the (ordered) eigenvalues and corresponding (normalized) eigenvectors of \mathbf{S} .

- The PCs of \mathbf{X} : $\boldsymbol{\xi}_i = \mathbf{X}\mathbf{v}_i$, or $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_p) = \mathbf{X}\mathbf{V}$ where

$$\mathbf{S}\mathbf{v}_i = e_i^2 \mathbf{v}_i \quad i = 1, \dots, p.$$

- Retain only the first K ($< p$) PCs leads to the regression model $\mathbf{y} = \boldsymbol{\Xi}_K \boldsymbol{\beta}_{pcr}^K + \boldsymbol{\epsilon}$ where $\boldsymbol{\Xi}_K = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_K)$.
- OLS of $\boldsymbol{\beta}_{pcr}^K$: $\hat{\boldsymbol{\beta}}_{pcr}^K = (\boldsymbol{\Xi}_K' \boldsymbol{\Xi}_K)^{-1} \boldsymbol{\Xi}_K' \mathbf{y} = \boldsymbol{\Lambda}_K^{-1} \boldsymbol{\Xi}_K' \mathbf{y}$.
- **The (Sequence of K -dim) Subspaces** : The K th PCR model is based on the subspace spanned by the first K eigenvectors of \mathbf{S} ($K = 1, \dots, p$).

- **The Problem** : For each k ($k = 1, \dots, p$) find \mathbf{c}_k , s.t.

$$T_{pcr}(\mathbf{c}) \equiv \text{var}(\mathbf{X}\mathbf{c}) = \mathbf{c}'\mathbf{S}\mathbf{c} \quad (7)$$

is maximized subject to $\mathbf{c}'\mathbf{c} = 1$ and $\text{cov}(\mathbf{X}\mathbf{c}, \mathbf{X}\mathbf{c}_j) = \mathbf{c}'\mathbf{S}\mathbf{c}_j = 0$, $j = 1, \dots, k - 1$ (“**S**-orthogonality” to $\{\mathbf{c}_j\}_1^{k-1}$).

- **The Solution**: results from solving the eigen-equation

$$\mathbf{S}\mathbf{c} = \lambda\mathbf{c}.$$

Thus, each \mathbf{c}_k ($k = 1, \dots, p$) have the solution

$$\mathbf{c}_{k,pcr} = \underset{\{\mathbf{c}'\mathbf{S}\mathbf{c}_j\}_1^{k-1}, \mathbf{c}'\mathbf{c}=1}{\text{argmax}} \text{var}(\mathbf{X}\mathbf{c}) = \mathbf{v}_k. \quad (8)$$

- **Prediction** : $\hat{\mathbf{y}}_{pcr}^K = \mathbf{\Xi}_K \hat{\boldsymbol{\beta}}_{pcr}^K$.

(projection of \mathbf{y} onto the space spanned by the first K PCs)

Ridge Regression (RR)

- **RR Estimate** : $\hat{\beta}_{rr} = (\mathbf{S} + \theta\mathbf{I})^{-1}\mathbf{s}$ where θ is the ridge parameter. Or in terms of OLS estimate,

$$\hat{\beta}_{rr} = [\mathbf{I} + \theta\mathbf{S}^{-1}]^{-1}\hat{\beta}_{ols}.$$

- **The Subspace** : is defined by a (single) unit vector (as in OLS) obtained by using the criterion

$$\mathbf{c}_{rr} = \underset{\mathbf{c}'\mathbf{c}=1}{\operatorname{argmax}} \operatorname{corr}^2(\mathbf{X}\mathbf{c}, \mathbf{y}) \frac{\operatorname{var}(\mathbf{X}\mathbf{c})}{\operatorname{var}(\mathbf{X}\mathbf{c}) + \theta}. \quad (9)$$

Partial Least Squares (PLS)

- **Construction** of $\{\mathbf{c}_k\}_1^K$: is as in PCR except the maximization criterion is different.
- **First PLS dimension** : Find \mathbf{c} , say \mathbf{c}_1 , s.t.

$$T_{pls}(\mathbf{c}) = cov^2(\mathbf{X}\mathbf{c}, \mathbf{y}) = (\mathbf{c}'\mathbf{s})^2 (= \mathbf{c}'\mathbf{M}\mathbf{c}) \quad (10)$$

is maximized subject to $\mathbf{c}'\mathbf{c} = 1$.

The Solution : is obtained by solving the eigen-equation

$$\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}\mathbf{c} = \lambda\mathbf{c} \quad (11)$$

which yields (the solution)

$$\mathbf{c}_{1,pls} = \frac{\mathbf{s}}{\|\mathbf{s}\|} \equiv \mathbf{w}_1. \quad (12)$$

Thus, the L.C. of \mathbf{X} s.t. $cov^2(\mathbf{X}\mathbf{c}, \mathbf{y})$ is maximized is $\mathbf{X}\mathbf{w}_1 \equiv \mathbf{t}_1$.

- **Second PLS dimension** : say \mathbf{t}_2 , is the vector in the space

orthogonal to \mathbf{t}_1 whose squared sample covariance with \mathbf{y} is maximum.

That is, Find \mathbf{c} , say \mathbf{c}_2 , s.t. $cov(\mathbf{X}\mathbf{c}, \mathbf{y})$ is maximized subject to $\mathbf{c}'\mathbf{c} = 1$, and $cov(\mathbf{X}\mathbf{c}, \mathbf{X}\mathbf{c}_1) = \mathbf{c}'\mathbf{S}\mathbf{c}_1 = 0$.

The Solution : is obtained by solving

$$\mathbf{E}'_1 \mathbf{y} \mathbf{y}' \mathbf{E}_1 \mathbf{c} = \lambda \mathbf{c} \quad (13)$$

(where $\mathbf{E}_1 = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}$ with $\mathbf{P}_1 = (\mathbf{t}'_1 \mathbf{t}_1)^{-1} \mathbf{t}_1 \mathbf{t}'_1$) which yields (the solution)

$$\mathbf{w}_2 = \frac{\mathbf{E}'_1 \mathbf{y}}{\|\mathbf{E}'_1 \mathbf{y}\|} \equiv \mathbf{c}_{2,pls}. \quad (14)$$

- **The k th PLS dimension** : Find \mathbf{c} , say \mathbf{c}_k , s.t. $cov^2(\mathbf{X}\mathbf{c}, \mathbf{y})$ is maximized subject to $\mathbf{c}'\mathbf{c} = 1$, and $\{cov^2(\mathbf{X}\mathbf{c}, \mathbf{X}\mathbf{c}_j) = 0\}_{j=1}^{k-1} = \{\mathbf{c}'\mathbf{S}\mathbf{c}_j = 0\}_{j=1}^{k-1}$.

That is, find $\mathbf{t}_k = \mathbf{X}\mathbf{c}_k$ that maximizes $cov(\mathbf{t}_k, \mathbf{y}) = \mathbf{t}'_k \mathbf{y}$ subject

to $\mathbf{t}'_1 \mathbf{t}_k = \dots = \mathbf{t}'_{k-1} \mathbf{t}_k = 0$ and $\mathbf{w}'_k \mathbf{w}_k = 1$.

The Solution : is obtained by solving

$$\mathbf{E}'_{k-1} \mathbf{y} \mathbf{y}' \mathbf{E}_{k-1} \mathbf{c} = \lambda \mathbf{c} \quad (15)$$

(where $\mathbf{E}_{k-1} = (\mathbf{I}_n - \sum_{i=1}^{k-1} \mathbf{P}_i) \mathbf{X}$ with $\mathbf{P}_i = (\mathbf{t}'_i \mathbf{t}_i)^{-1} \mathbf{t}_i \mathbf{t}'_i$) which yields (the solution)

$$\mathbf{w}_k = \frac{\mathbf{E}'_{k-1} \mathbf{y}}{\|\mathbf{E}'_{k-1} \mathbf{y}\|} \equiv \mathbf{c}_{k,pls}. \quad (16)$$

- Since $cov^2(\mathbf{X}\mathbf{c}, \mathbf{y}) = corr^2(\mathbf{X}\mathbf{c}, \mathbf{y}) var(\mathbf{X}\mathbf{c}) z$ where $z = var(\mathbf{y}) = \mathbf{y}'\mathbf{y}$ does not depend on \mathbf{c} we can summarize for a K component PLS model as follows:

For $k = 1, \dots, p$,

$$\mathbf{w}_k \equiv \mathbf{c}_{k,pls} = \underset{\{\mathbf{c}'\mathbf{S}\mathbf{c}_j\}_1^{k-1}, \mathbf{c}'\mathbf{c}=1}{\operatorname{argmax}} \quad corr^2(\mathbf{X}\mathbf{c}, \mathbf{y}) var(\mathbf{X}\mathbf{c}). \quad (17)$$

- **The (Sequence of K -dim) Subspaces** : The K th PLS model is based on the subspace spanned by the first K PLS components (L.C. of \mathbf{X}), $\{\mathbf{t}_1, \dots, \mathbf{t}_K\}$ ($K = 1, \dots, p$).
Or equivalently, by the span of $\{\mathbf{w}_1, \dots, \mathbf{w}_K\}$
- Retain only the first K ($< p$) PLS components :
 $\mathbf{T}_K = (\mathbf{t}_1, \dots, \mathbf{t}_K)$ leads to the regression model $\mathbf{y} = \mathbf{T}_K \boldsymbol{\beta}_{pls}^K + \boldsymbol{\epsilon}$.
- OLS of $\boldsymbol{\beta}_{pls}^K$: $\hat{\boldsymbol{\beta}}_{pls}^K = (\mathbf{T}'_K \mathbf{T}_K)^{-1} \mathbf{T}'_K \mathbf{y} = \mathbf{D}_K^{-1} \mathbf{T}'_K \mathbf{y}$
(where $\mathbf{D}_K = \text{diag}\{\mathbf{t}'_1 \mathbf{t}_1, \dots, \mathbf{t}'_K \mathbf{t}_K\}$).
- **Prediction** : $\hat{\mathbf{y}}_{pls}^K = \mathbf{T}_K \hat{\boldsymbol{\beta}}_{pls}^K$.
(projection of \mathbf{y} onto the space spanned by $\{\mathbf{t}_1, \dots, \mathbf{t}_K\}$)

• Notes :

1. The projection matrix is

$$\mathbf{P}_{pls}^{(K)} \equiv \mathbf{T}_K (\mathbf{T}'_K \mathbf{T}_K)^{-1} \mathbf{T}'_K = \sum_{i=1}^K \mathbf{P}_i$$

2. Thus, $\hat{\mathbf{y}}_{pls}^K = \sum_{i=1}^K \mathbf{P}_i \mathbf{y}$ (is sum of the sequential orthogonal projection of \mathbf{y} onto \mathbf{t}_i).

3. The “residual” matrix at step i , \mathbf{E}_i :

$$\mathbf{E}_i = \mathbf{X} - \sum_{j=1}^i (\mathbf{t}'_j \mathbf{t}_j)^{-1} \mathbf{t}_j \mathbf{t}'_j \mathbf{X} = \left(\mathbf{I}_n - \sum_{j=1}^i \mathbf{P}_j \right) \mathbf{X}$$

represents a “deflation” of the original \mathbf{X} -space by “removing” from \mathbf{X} the space spanned by $\{\mathbf{t}_1, \dots, \mathbf{t}_i\}$.

4. \mathbf{E}_i is obtained by projecting \mathbf{X} onto the orthocomplement of the space spanned by $\{\mathbf{t}_1, \dots, \mathbf{t}_i\}$.

Summary

Method : Criterion

- OLS : $T_{ols}(\mathbf{c}) \equiv \text{corr}^2(\mathbf{X}\mathbf{c}, \mathbf{y})$
- PCR : $T_{pcr}(\mathbf{c}_k) \equiv \text{var}(\mathbf{X}\mathbf{c}_k)$
- RR : $T_{rr}(\mathbf{c}) \equiv \text{corr}^2(\mathbf{X}\mathbf{c}, \mathbf{y}) \frac{\text{var}(\mathbf{X}\mathbf{c})}{\text{var}(\mathbf{X}\mathbf{c}) + \theta}$
- PLS : $T_{pls}(\mathbf{c}_k) \equiv \text{corr}^2(\mathbf{X}\mathbf{c}_k, \mathbf{y}) \text{var}(\mathbf{X}\mathbf{c}_k)$

Continuum Regression (CR)

(of Stone and Brooks (“SB”) (1990))

- SB proposed a generalized criterion to be maximized which encompasses OLS, PLS, and PCR.
- That is, a criterion, $T(\mathbf{c}, \alpha)$ is maximized w.r.t. \mathbf{c} for values of $\alpha \in [0, 1)$.
- The resulting L.C, $\mathbf{X}\mathbf{c}_\alpha$, span the continuum linking OLS, PLS, and PCR.
- The previous summary suggests using the following criterion

$$T(\mathbf{c}, \alpha) \equiv cov^2(\mathbf{X}\mathbf{c}, \mathbf{y})(var(\mathbf{X}\mathbf{c}))^{\frac{\alpha}{1-\alpha}-1} = (\mathbf{c}'\mathbf{s})^2(\mathbf{c}'\mathbf{S}\mathbf{c})^{\frac{\alpha}{1-\alpha}-1}. \quad (18)$$

Maximization of $T(\mathbf{c}, \alpha)$

- **The Problem** : Determine \mathbf{c}_{k+1} maximizing $T(\mathbf{c}, \alpha)$ subject to $\mathbf{c}'_{k+1} \mathbf{c}_{k+1} = 1$ and $\{\mathbf{c}'_j \mathbf{S} \mathbf{c}_{k+1} = 0\}_{j=1}^k$.
- $\mathbf{c}_{k+1} \in C(\mathbf{S})$, i.e. $\mathbf{c}_{k+1} = \sum_{i=1}^p z_i \mathbf{v}_i$, so just need to determine coordinates $\{z_i\}_1^p$.
- Substituting into T gives the criterion in z -coordinates:

$$T(\mathbf{c}_{k+1}, \alpha) = \left(\sum_{i=1}^p z_i d_i \right)^2 \left(\sum_{i=1}^p e_i z_i^2 \right)^{\gamma-1} \equiv f(z_1, \dots, z_p) \quad (19)$$

where $d_i = \mathbf{s}' \mathbf{v}_i$, and $\gamma = \alpha / (1 - \alpha)$.

- The constraints in z -coordinates are

$$\mathbf{c}'_{k+1} \mathbf{c}_{k+1} = 1 \iff g_0(z_1, \dots, z_p) \equiv \sum_{i=1}^p z_i^2 - 1 = 0 \quad (20)$$

$$\{\mathbf{c}'_j \mathbf{S} \mathbf{c}_{k+1} = 0\}_{j=1}^k \iff \{g_j(z_1, \dots, z_p) = 0\}_{j=1}^k \equiv \{\sum_{i=1}^p a_{ij} z_i = 0\}_{j=1}^k \quad (21)$$

where $a_{ij} = e_i^2 \mathbf{c}'_j \mathbf{v}_i$.

- The maximizing $\{z_i\}_1^p$ is the solution of the Lagrange multiplier equations :

$$\frac{\partial F}{\partial z_i} = \frac{\partial f}{\partial z_i} - \sum_{j=0}^k \lambda_j \frac{\partial g_j}{\partial z_i} = 0, \quad i = 1, \dots, p \quad (22)$$

where $F = f - \sum_{j=0}^k \lambda_j g_j$.

- Combining (19), (20), and (21) into one matrix equation and using the standard formula for inverting a partitioned matrix gives the solution for $\{z_i\}_1^p$, and obtaining $\mathbf{c}_{k+1} = \sum_{i=1}^p z_i \mathbf{v}_i$.

How to Choose K

- K is usually chosen by cross-validation (CV):

$$\hat{K} = \operatorname{argmin}_{0 \leq K \leq p} \sum_{i=1}^n L(y_i, \hat{y}_{K \setminus i}). \quad (23)$$

where $\hat{y}_{K \setminus i}$ is the K th model computed from the training sample with the i th observation removed and $L(,)$ is a loss function, usually taken to be quadratic loss.

- For CR (of SB) both α and K are chosen using CV:
Find a value of (α, K) s.t. $C_{\alpha, K}$ is minimized, where

$$C_{\alpha, K} = (1/n) \sum_{i=1}^n L(y_i, \hat{y}_{K \setminus i}). \quad (24)$$

Continuum Regression (CR)

(of Lorber, Wangen, and Kowalski (“LWK”) (1987))

Geometry of PLS

(Adopted from Phatak, Reilly, and Penlidis (1992))

- Let $\tilde{\mathbf{a}} : p \times 1$, and $\mathbf{a} : n \times 1$ be vectors and $\mathbf{Z} : n \times p$ matrix.
- The column space of \mathbf{X} , say $C(\mathbf{X})$ is a p -dim linear subspace in n -dim. Transforming this space to a p -dim ellipsoid space corresponds to the following transformation :
- **Transformation** : $\tilde{\mathbf{a}} = \mathbf{Z}'\mathbf{a}$
where $\mathbf{a} = \mathbf{X}\mathbf{c}$ is a L.C. of \mathbf{X} and $\mathbf{Z} = \mathbf{X}\mathbf{V}\mathbf{\Lambda}^{-\frac{1}{2}}$.
Thus $\mathbf{c}'\mathbf{c} = 1 \iff \tilde{\mathbf{a}}'\mathbf{\Lambda}^{-1}\tilde{\mathbf{a}} = 1$ which is an equation (in $\tilde{\mathbf{a}}$) of a p -dim ellipsoid. (Fig.1)
- The \sim denotes the image of an n -dim vector in p -dim.
- The inverse mapping from $\tilde{\mathbf{a}} : p \times 1$ to $\mathbf{a} : n \times 1$ is $\mathbf{a} = \mathbf{Z}\tilde{\mathbf{a}}$.
- The axes of the ellipsoid correspond to the PCs of \mathbf{X} (Fig.2a).
- **Examples** :

1. $\hat{\tilde{y}}_{ols} = \mathbf{Z}'\hat{\mathbf{y}}_{ols} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}'\mathbf{X}'\mathbf{y}(= \mathbf{Z}'\mathbf{y})$ so \mathbf{y} and $\hat{\mathbf{y}}_{ols}$ yields the same mapping.
 2. For the K th PCR model the projection of \mathbf{y} onto $C(\mathbf{\Xi}_K)$ is the same as the projection of $\hat{\mathbf{y}}_{ols}$ onto $C(\mathbf{\Xi}_K)$. (Fig.2b, two PCs retained)
- **PLS** : Geometrical construction of the PLS components $\{\mathbf{t}_i\}_1^K$ is based on the following facts:

$$cov^2(\mathbf{X}\mathbf{w}_1, \mathbf{y}) = cov^2(\mathbf{X}\mathbf{w}_1, \hat{\mathbf{y}}_{ols}) \quad (25)$$

and is true for succeeding dimensions ($i = 0, 1, \dots, p - 1$):

$$cov^2(\mathbf{E}_i\mathbf{w}_{i+1}, \mathbf{y}) = cov^2(\mathbf{E}_i\mathbf{w}_{i+1}, \hat{\mathbf{y}}_{ols}) = cov^2(\mathbf{E}_i\mathbf{w}_{i+1}, \hat{\mathbf{y}}_i) \quad (26)$$

where $\hat{\mathbf{y}}_i$ is the orthogonal projection of \mathbf{y} or $\hat{\mathbf{y}}_{ols}$ onto $C(\mathbf{E}_i)$, and ($i = 0, \mathbf{E}_0 = \mathbf{X}, \hat{\mathbf{y}}_0 = \hat{\mathbf{y}}_{ols}$)

Geometrical Construction of PLS Components, $\{\tilde{\mathbf{t}}_i\}_1^K$

- **The first PLS component, $\tilde{\mathbf{t}}_1$:**
 1. given a p -dim ellipsoid, $\tilde{\mathbf{a}}' \boldsymbol{\Lambda}^{-1} \tilde{\mathbf{a}} = 1$ and the vector of OLS fitted values, $\hat{\tilde{\mathbf{y}}}_{ols}$ (transformed), find the $(p - 1)$ -dim hyperplane orthogonal to $\hat{\tilde{\mathbf{y}}}_{ols}$ and tangent to the ellipsoid.
 2. The vector from the origin of the ellipsoid to the point of tangency is the first PLS component, $\tilde{\mathbf{t}}_1$. (Fig.3a)
- **The second PLS component, $\tilde{\mathbf{t}}_2$:**
 1. Recall that $\tilde{\mathbf{t}}_2 \in C(\mathbf{E}_1)$ orthogonal to $\tilde{\mathbf{t}}_1$ and was obtained as a L.C., $\mathbf{E}_1 \mathbf{c}$ ($\mathbf{c}' \mathbf{c} = 1$) that maximize one of (25).
 2. The subspace is a $(p - 1)$ -dim ellipsoid, $\tilde{\mathbf{a}}' \boldsymbol{\Lambda}_1^{-1} \tilde{\mathbf{a}} = 1$, whose axes are the $(p - 1)$ PCs of \mathbf{E}_1 , and $\boldsymbol{\Lambda}_1$ is diagonal whose $(p - 1)$ elements are the non-zero eigenvalues of $\mathbf{E}_1' \mathbf{E}_1$.
 3. The point of tangency of the $(p - 2)$ -dim hyperplane, orthogonal to $\hat{\tilde{\mathbf{y}}}_1$ (= orthogonal projection of \mathbf{y} or $\hat{\tilde{\mathbf{y}}}_{ols}$ onto

$C(\mathbf{E}_1)$ gives $\tilde{\mathbf{t}}_2$. (Fig.3b)

- **Successive dimensions**, say at step $k + 1$, $\tilde{\mathbf{t}}_{k+1}$:
 1. The subspace is a $(p - k)$ -dim ellipsoid, $\tilde{\mathbf{a}}' \mathbf{\Lambda}_k^{-1} \tilde{\mathbf{a}} = 1$, whose axes are the $(p - (k - 1))$ PCs of \mathbf{E}_k , and $\mathbf{\Lambda}_k$ is diagonal whose $(p - (k - 1))$ elements are the non-zero eigenvalues of $\mathbf{E}'_k \mathbf{E}_k$.
 2. The point of tangency of the $(p - (k + 1))$ -dim hyperplane, orthogonal to $\hat{\mathbf{y}}_k$ (= orthogonal projection of \mathbf{y} or $\hat{\mathbf{y}}_{ols}$ onto $C(\mathbf{E}_k)$) gives $\tilde{\mathbf{t}}_{k+1}$.
- **Fitted values from PLS** : obtained by projecting $\hat{\mathbf{y}}_{ols}$ onto the space spanned by $\{\tilde{\mathbf{t}}_i\}$ (Fig.4).

Comparison of PCR, PLS, and RR

(Frank and Friedman (1993))

- **Notation** : $\text{ave}(\eta) = \sum_{i=1}^n \eta_i$, \mathbf{x} : $p \times 1$ vector.
- **Goal** : RR, PCR, and PLS shrink the solution coefficient $\hat{\beta}_*$ (* = rr, pcr, or pls) vector away from from OLS solution, $\hat{\beta}_{ols}$ toward directions in the predictor-variable space of larger sample variance.

The is, the goal is to bias the solution coefficient vector $\hat{\beta}_*$ away from directions for which the projected sample predictor variables have small sample variance:

$$\text{var} \left(\frac{\mathbf{X}\hat{\beta}_*}{\|\hat{\beta}_*\|} \right) = \text{ave}(\hat{\beta}'_* \mathbf{x})^2 = \textit{small}.$$

- The objective criterion for RR, PCR, and PLS all involve a function of the scale factor $var(\mathbf{X}\mathbf{c})$ ($\mathbf{c} = \hat{\boldsymbol{\beta}}_*/\|\hat{\boldsymbol{\beta}}_*\|$) therefore producing biased estimates.
- Bias is regulated by λ for RR and K for PCR and PLS.
- Effect of decreasing K is to attract the coefficient vector towards larger values of $var(\mathbf{X}\mathbf{c})$.
- For a given K , the degree of this attraction depends, for:
 1. **PLS** : jointly on covariance structure of the predictor variables as well as the OLS solution ($\hat{\boldsymbol{\beta}}_{ols}$ which depends on responses $\{y_i\}$.)
 2. **PCR** : only on covariance structure of the predictor variables.

A Bayesian Motivation to Explain Shrinkage Structure

- Since estimators here are equivariant w.r.t. rotations in predictor variable space (standardization), for convenience let $\mathbf{S} = \text{diag}\{e_1^2, \dots, e_p^2\}$.
- Consider an unrestrictive prior probability distribution that that considers all coefficient vector directions, $\boldsymbol{\beta}/\|\boldsymbol{\beta}\|$ equally likely (i.e., depending only on the norm $\boldsymbol{\beta}'\boldsymbol{\beta}$),

$$\pi(\boldsymbol{\beta}) = \pi(\boldsymbol{\beta}'\boldsymbol{\beta}). \quad (27)$$

- Assume simple linear shrinkage estimates of the form

$$\hat{\beta}_j = f_j \hat{\beta}_{j,ols} \quad j = 1, \dots, p \quad (28)$$

where $\hat{\beta}_{j,ols}$ is the OLS estimate of the j th regression coefficient and $\{f_j\}_1^p$ are shrinkage factors (independent of $\{y_i\}_1^n$).

- For a given prior $\pi(\boldsymbol{\beta})$ the mean squared prediction error is

$$\begin{aligned}
 MSE[\hat{y}(\mathbf{x})] &= E_{\boldsymbol{\beta}} E_{\epsilon} [\boldsymbol{\beta}' \mathbf{x} - \hat{\boldsymbol{\beta}}' \mathbf{x}]^2 \\
 &= E_{\boldsymbol{\beta}} E_{\epsilon} \left[\sum_{j=1}^p (\hat{\beta}_j - f_j \hat{\beta}_{j,ols}) x_j \right]^2. \\
 &= \sum_{j=1}^p \left[(1 - f_j)^2 \frac{1}{p} E_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|^2 + f_j^2 \frac{\sigma^2}{ne_j^2} \right] x_j^2.
 \end{aligned} \tag{29}$$

1. First term is the bias (squared) which depends only on the prior distribution; does not depend on the error variance and the predictor-variable distribution.
2. Second term is the variance of the estimate which depends on the experimental error variance and the predictor-design sample; does not depend on the nature of the true coefficient vector $\boldsymbol{\beta}$.

- Setting
 1. $\{f_j = 1\}$ \longrightarrow OLS with variance given by 2nd term.
 2. $\{f_j < 1\}$ \longrightarrow increases bias, decreases variance.
 3. $\{f_j > 1\}$ \longrightarrow increases both bias and variance.
- From 2nd term, the variance of the model estimate for a given (eigen) direction (x_j) is inversely proportional to the sample predictor variance e_j^2 associated with that direction; therefore, directions with small spread in predictor variables give rise to high variance in the model estimate.
- $\{f_j\}_1^p$ minimizing $MSE[\hat{y}(\mathbf{x})]$ is

$$f_j^* = \frac{e_j^2}{e_j^2 + \theta}, \quad j = 1, \dots, p \quad (30)$$

where

$$\theta = \left(\frac{p}{n}\right) \frac{\sigma^2}{E_{\beta} \|\beta\|^2} = \left(\frac{p}{n}\right) \frac{\text{noise}}{\text{signal}}. \quad (31)$$

- So the optimal (minimal MSE) linear shrinkage estimates are

$$\hat{\beta}_j = \hat{\beta}_{j,ols} \frac{e_j^2}{e_j^2 + \theta} \quad (32)$$

which also turns out to be the RR estimates expressed in the coordinate system defined by $\mathbf{S} = \text{diag}\{e_1^2, \dots, e_p^2\}$, $\hat{\beta}_j = \hat{\beta}_{j,rr}$.

- PCR is also a linear shrinkage estimator,

$$\hat{\beta}_{j,pcr} = \hat{\beta}_{j,ols} I(e_j^2 \geq e_K^2) \quad (33)$$

where $I(\cdot)$ is the indicator function for (\cdot) .

- PLS is not linear shrinkage estimator.

Comparing shrinkage Structures of RR, PCR, and PLS

- **Goal** : Comparing shrinkage structures of RR, PCR, and PLS in various situations.
- To do this, expand the solution vector $\hat{\boldsymbol{\beta}}_*$ in terms of $\{\mathbf{v}_j\}_1^p$, and $\hat{\boldsymbol{\beta}}_{ols}$,

$$\hat{\boldsymbol{\beta}}_* = \sum_{j=1}^p f_{j,*}^K \hat{\alpha}_j \mathbf{v}_j \quad (34)$$

where $\hat{\alpha}_j$ is the projection of the OLS solution on \mathbf{v}_j and $\{f_{j,*}^K\}_{j=1}^p$ ($*$ = rr, pcr, or pls) are factors along each of the eigendirections that scale the OLS solution. (Note that $f_{j,rr}$ does not depend on K .)

- The scale factors $\{f_{j,*}^K\}$ are
 1. **RR** : $f_{j,rr} = e_j^2 / (e_j^2 + \theta)$
 2. **PCR** : $f_{j,pcr}^K = I(e_j^2 \geq e_K^2)$
 3. **PLS** : $f_{j,pls}^K = \sum_{k=1}^K \tau_k e_j^{2k}$ where

$$\boldsymbol{\tau} = \{\tau_k\}_1^K = \mathbf{W}^{-1} \mathbf{w}$$

$$\mathbf{w} = \{w_k\}_{k=1}^K = \left\{ \sum_{j=1}^p \hat{\alpha}_j^2 e_j^{2(k+1)} \right\}_{k=1}^K$$

$$\mathbf{W} = \{W_{kl}\}_{k=1}^K \quad K \quad K = \left\{ \sum_{j=1}^p \hat{\alpha}_j^2 e_j^{2(k+l+1)} \right\}_{k=1}^K \quad K \quad K$$

- Scale factors for PLS depends on OLS solution $\{\hat{\alpha}_j\}$, hence on the responses $\{y_i\}$.
- For a given K compute (3) for PLS and compare to RR and PCR for corresponding situations.

Simulation Results of FF : Fig.1-Fig.4

- For each figure : $p = 10, k = 1, \dots, K = 6$.
- Overall shrinkage : $sh = \|\hat{\beta}_{pls}\| / \|\hat{\alpha}\|$.
- RR and PCR solutions normalized so that they give the same overall shrinkage, i.e.,
 1. RR : choose λ so that $\|\hat{\beta}_{rr}\| = \|\hat{\beta}_{pls}\|$.
 2. PCR : choose K so that $\|\hat{\beta}_{pcr}\| \simeq \|\hat{\beta}_{pls}\|$.
- Plotted on each figure are :
 1. **PLS** : solid line(k) : plotted $\{f_{j,pls}\}_{j=1}^{10}$ vs. j .
 2. **RR** : dashed line(k) : plotted $\{f_{j,rr}\}_{j=1}^{10}$ vs. j .
 3. **PCR** : dotted line(k): plotted $\{f_{j,pcr}\}_{j=1}^{10}$ vs. j .

- Four situations simulated:

	OLS Solution : $\{\hat{\alpha}_j\}_1^p$	Eigen-structure : $\{e_j^2\}_1^p$
Fig.1	neutral $\hat{\alpha}_j$'s: $\{\hat{\alpha}_j = 1\}_1^p$	hi collin: $\{e_j^2 \sim 1/j^2\}_1^p$
Fig.2	neutral $\hat{\alpha}_j$'s: $\{\hat{\alpha}_j = 1\}_1^p$	mod collin: $\{e_j^2 \sim 1/j\}_1^p$
Fig.3	fav $\hat{\alpha}_j$'s: $\{\hat{\alpha}_j = 1/j\}_1^p$	hi collin: $\{e_j^2 \sim 1/j^2\}_1^p$
Fig.4	unfav $\hat{\alpha}_j$'s: $\{\hat{\alpha}_j = j\}_1^p$	hi collin: $\{e_j^2 \sim 1/j^2\}_1^p$

(collin=collinearity, hi=high, mod=moderate, fav=favorable, unfav=unfavorable)

- Neutral : OLS solution is taken to project equally in all eigendirections.
- Favorable : OLS solution is taken to align with the major axes of predictor design.
- Unfavorable : OLS solution is taken to align in orthogonal direction to the major axes of the predictor design.

Simulation Results

- For the same overall shrinkage, the relative shrinkage of RR tracks that of PLS but is somewhat more moderate.
- PLS used fewer components to achieve the same overall shrinkage—roughly half as many components.
- PLS reached OLS solution with about 5-6 components, whereas PCR requires all 10 components.
- PLS shrinks ($f_{j,pls} \leq 1$) OLS solution in some eigendirections as well as expands ($f_{j,pls} > 1$) in other directions—for a K -component solution, the OLS solution is expanded in the subspace defined by the eigendirections associated with the eigenvalues closest to the K th eigenvalue.
- Situation 2 : same patterns but less shrinkage due to moderate collinearity.

- Situation 3 (and 4) : same patterns but less (more) overall shrinkage because of favorable (unfavorable) alignment of OLS solution.

- Clearly shows that they all penalize the solution coefficient $\hat{\beta}_*$ for projecting onto the low-variance subspace of the predictor design.
 1. PLS and PCR : penalty decreases as K increases.
 2. RR : penalty increases as θ increases.
 3. RR : strength of penalty is monotonically increasing for directions of decreasing variance.
 4. PCR : strength of penalty is a sharp threshold function.
 5. PLS : strength of penalty is relatively smoothed but not monotonic.

Appendix

Back to Bayesian

Power Ridge Regression

Examples with Real Data

Multivariate Extensions