Last lecture is today

- Exam date? Tuesday 15 May? Or 22 May?
- 15 min: Oral presentation of one project of your own choice (from one of the lectures). Outline required background + present results and interpret.
- 25 min: General questions from the curriculum (main ideas from papers, lecture notes and project work). 30 min to work on questions (before presentation).

Today's topic : Useful "dimension reduction techniques"

- Partial least squares (PLS) regression and related methods.
- Multidimensional scaling (MDS) and related methods.

Regression

- Response or dependent variable Y. Dimension K for I realizations. ($K = 1$ most common.)
- Explanatory or predictor variable X. Dimension J for I realizations.
- $\hat{Y}_i = X_i \hat{\beta} = x_{i1}\hat{\beta}_1 + \ldots + x_{iJ}\hat{\beta}_J$

Subset regression

- Often J very large.
- Avoid overfitting by wisely reducing regressors $p << J$:

$$\hat{y}_p = x_{i(1)}\hat{\beta}_{(1)} + \ldots + x_{i(p)}\hat{\beta}_{(p)}$$

Linear combinations regression

- Work on derived (linear) features of the data $\tilde{X} = Xw$
- Wisely select the (linear) predictors and associated regression parameters:

$$\hat{y}_p = \tilde{x}_{i(1)}\tilde{\beta}_{(1)} + \ldots + \tilde{x}_{i(p)}\tilde{\beta}_{(p)}$$

PLS - background

Define

$$t = Xw, \quad u = Yc,$$

such that

$$tt' = 1, \quad ww' = 1, \quad \max tu'$$

Developed by Herman and Svante Wold (Sweden, 1960-70). Many statistical properties derived by Inge Helland, Tormod Naess, Harald Martens, and others (Ås, Norway).

PLS - idea

- t is score matrix (composed of latent vectors)
- w is loading matrix
- $t = Xw$ are 'optimal' linear combination of predictors.
- PLS is done iteratively, projecting residual variation, and searching for optimal order of latent vectors.

Maximize covariance of the linear combinations of response and predictors!

PLS regression - algorithm

- Initialize $u = Y_k$, for some k
    1. $w = X'u/|X'u|$
    2. $t = Xw$
    3. $c = Y't/|Y't|$
    4. $u = Yc$
    5. $X = X - tw'$, $Y = Y - uc'$.
    6. Go to 1 if t has not converged.

(Similar to Conjugate Gradients - find optimal projections iteratively.)

Software

R: pls
https://cran.r-project.org/web/packages/pls/vignettes/pls-manual.pdf
MATLAB: plsregress

PLS regression - example

Prediction of values from massive image data.

$$\{(X_{1,i},\ldots,X_{J,i})Y_i\},\quad i=1,\ldots,1000$$

$J \sim 10\ 000$.

PLS regression - predictors

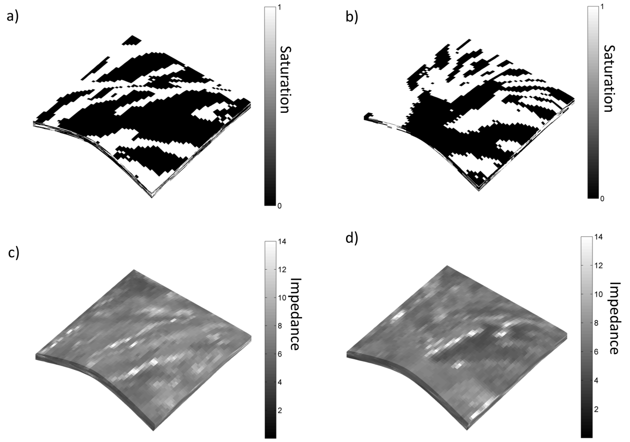
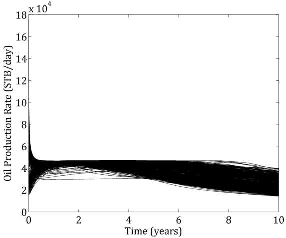


Figure: Two of 1000 realizations of seismic variables (explanatory variables).

PLS regression - dependent variables

a)



b)

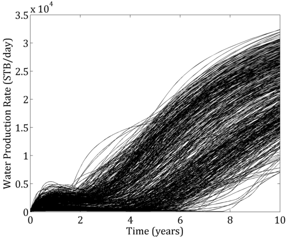


Figure: 1000 realizations of value variables (response).

PLS regression - PRESS

Aim to find structural relationship but not overfitting the noise.

- Use k-fold partitioning of ensemble into test and training sets.
- Evaluation statistic for order predicted residual sum of squares (PRESS):

$$\text{PRESS}(p) = \sum_{i=1}^{I_{\text{test}}} \left\| Y_{i,\text{test}} - \hat{Y}_{i,\text{test}} \right\|^2$$

- Randomisation over repeated random splits

$$p^* = \text{argmin}_p \text{PRESS}(p)$$

PLS regression - PRESS

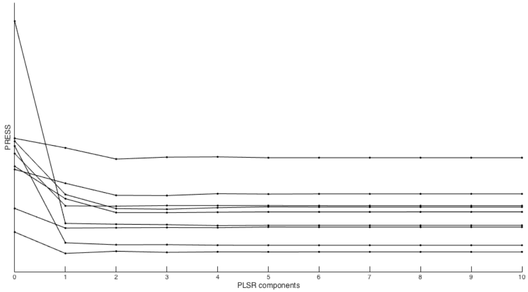


Figure: PRESS for different datasets and a range of PLS orders.

PLS vs PCR

- Principal component (PC) regression selects components of X to maximize $\mathrm{Var}(X)$ (singular values).
- PLS regression selects components of X to maximize $\mathrm{Cov}(Y, X)$.
- Results are often similar: PLSR might have more predictive power, while PCR could be easier to interpret.

MDS idea

Datasets $Y_1, \ldots, Y_N$, $Y_i = (Y_{i1}, \ldots, Y_{iK})$ for large K.
Embed the dissimilarity of data $D_{ij} = \text{distance}(Y_i, Y_j)$ in a smaller dimension (typically 2) such that close points in the 2 dimensional plane are also close in the K dimensional space.
Idea goes back to Torgerson (1950s) and Kruskal (1960-70s).

MDS mathematics

$$\text{Stress}(x_1, \ldots, x_N) = \sqrt{\sum_{i \neq j} (D_{ij} - |x_i - x_j|)^2}$$

Find $x_1, \ldots, x_N$, locations in 2 dimensional space that best **visualize differences and clusters in the data**. (x attributes are centered at the origin.)

MDS plot in 2 dimensions

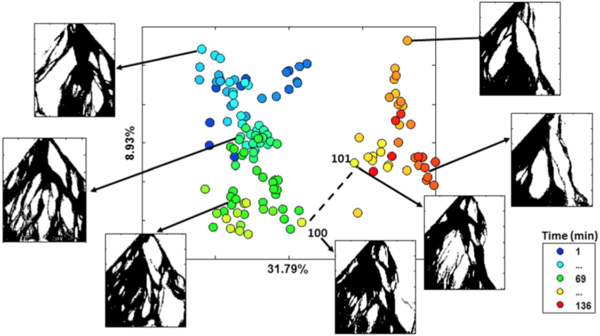


Figure: MDS of physical modeling data of drainage data (Scheidt et al. (2017).

MDS algorithm

Gradient descent is one method to solve for x_i, $i = 1, \ldots, N$.

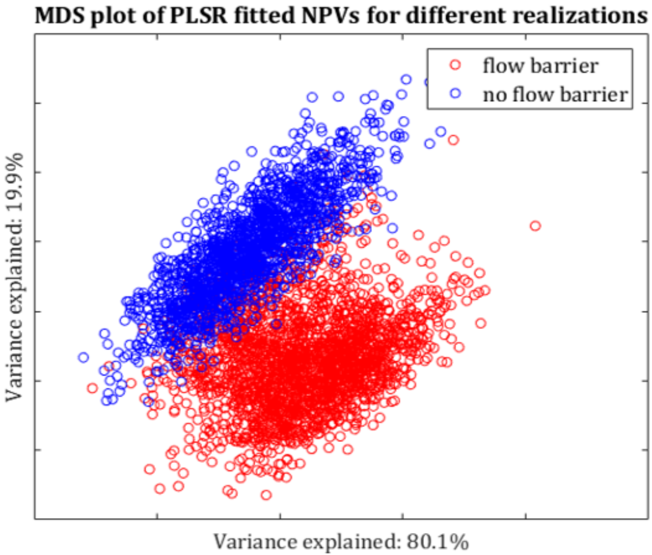
$$\text{argmin}_{x_1,\ldots,x_N} \sqrt{\sum_{i \neq j} (D_{ij} - |x_i - x_j|)^2}$$

Find $x_1, \ldots, x_N$, locations in 2 dimensional space that best visualize differences in the data.
Actual implementation depends on distance measures (Sect 6.2 in Buja et al.)

MDS extensions

- Dissimilarities can be metric or nonmetric distance between data inner-products (Sect 4.2 in Buja et al.)
- Dissimilarities use PCA (or PLS) with smoothing kernels.
- Dissimilarities based on neighborhood embeddings (tSNE), using conditional probabilities, kernels with heavy tails and divergence measures.

MDS for datasets



Tasks:

PLS in R: Run and interpret the gasoline example of the tutorial *pls*
`https://cran.r-project.org/web/packages/pls/vignettes/pls-manual.pdf`
PLS in MATLAB: Run and interpret the gasoline example used to explain *plsregress*
MDS
Generate 50 datasets of zero-mean Gaussian vectors $Y_i = (Y_{i1}, \ldots, Y_{i,10})$, $i = 1, \ldots, 50$ with independent variance 1 components.
Generate 50 datasets of zero-mean Gaussian vectors $Y_i = (Y_{i1}, \ldots, Y_{i,10})$, $i = 51, \ldots, 100$ with unit diagonal and off-diagonal $0 < r < 1$.
Use MDS with Euclidean distance and another constructive measure to cluster the two datasets in a two-dimensional display $x_i = (x_{i1}, x_{i2})$, $i = 1, \ldots, 100$. Compare for different r.
In R or MATLAB: *cmdscale*.