Outline: Linear Models in Large Feature Spaces (supervised) Dimension Reduction

(F. Chiaromonte)

Back to more traditional statistical approaches:

- **Feature (Subset) Selection**
  - Best Subset Selection
  - Step-wise Selection

  Followed by LS fit of the (smaller) model comprising the selected features.

- **Dimension Reduction**
  - Principal Components (unsupervised reduction)
  - Sufficient Dimension Reduction (supervised reduction)

  Followed by LS fit of the (smaller) model comprising the selected linear combinations

  also, in the case of a binary or categorical response (Generalized Linear Models)

  - Linear Discriminant Analysis (seen as supervised reduction)

  Followed by ML fit of the (smaller) model comprising the selected linear combinations

Also these can be though of in the framework of constrained LS: *Linear constraints* to force β in a coordinate space, or a generic linear subspace, of R^p. ***BUT WE NEED AN ADDITIONAL "INGREDIENT"!***

==

Dimension Reduction: ISLR also describes Partial Least Squares. But does not describe Sufficient Dimension Reduction techniques.

Dimension Reduction: How do we produce a pxp matrix V expressing an O.N. basis in the feature space?

(we then focus on its last (p-d) rows to create the linear constraints V_{DR})

Dimension Reduction:

Take the O.N. basis provided by the Eigen decomposition of an appropriate positive definite matrix. The eigenvectors express the directions, and the corresponding eigenvalues their ordering.

This corresponds to a rotation of the elements of the Canonical O.N. basis $\{e_1,\dots\ e_p\}$, and provides

V – a rotation matrix

Other approaches to dimension reduction exist, but some of the most popular (e.g. **Principal Components Analysis**, unsupervised; **Sliced Inverse Regression** supervised) use Eigen decompositions.

<u>Principal Components</u> (see also ISLR Chapter 10, Sections 1,2)

Units\Features	X_1	X_2	…	X_p
Unit 1	x_{11}	x_{12}		x_{1p}
Unit 2	x_{21}	x_{22}		x_{2p}
. . .				
Unit n	x_{n1}	x_{n2}		x_{np}

p features measured on n units:

- An n x p data matrix X
- A data cloud of n points in $\mathbf{R}^p$.

Location is irrelevant; assume each feature is centered, mean 0.

$\bar{X} = 0_p$ mean vector in R^p; cloud is centered/located at origin

$S \propto X'X$ (sample) p x p variance/covariance matrix

Create orthogonal directions in $\mathbf{R}^p$ (and corresponding linear combinations) ranked by variability of the data cloud.

In many applications (but not all) these are the most informative, capturing structure in the data. ***Here, we think of them as generating composite features to be used in a linear model – but we do not consider Y in creating them!***

Take the Eigen decomposition of S:

$$S = \sum_{m=1}^{p} \lambda_m \phi_m \phi_m^T$$

$$\lambda_1 \geq \ldots \geq \lambda_p \geq 0 \quad \text{(eigenvalues)}$$

$$\| \phi_m \| = \phi_m^T \phi_m = 1 \, , \, \phi_m^T \phi_k = 0 \; m,k = 1,2\ldots p \quad \text{(eigenvectors)}$$

eigenvalues are variances along the directions identified by eigenvectors; in non-increasing order.

For any given dimension *m*, identify the linear subspaces closest to the data:

$$\| P_{Span(\phi_1 \ldots \phi_p)} X \|^2 = \max$$

m-dimensional representation most likely to be useful in capturing structure, most informative (not always!)

Here, the reduced feature space we use to formulate a linear model once we chose m*=d.

$V = (\phi_1,\ \phi_2 \dots \phi_p) \quad (p \times p)$ **ROTATION MATRIX** provided by the eigenvectors of S

LOADINGS: coefficients expressing the m-th component in terms of the original features

$$\phi_m^T = (\phi_{m1},\ \ \phi_{m2} \dots \phi_{mp})$$

coordinates of each element (m-th) of the new rotated basis in terms of the original Canonical basis.

SCORES: values of the m-th component on the n units

$$z_{im} = \phi_{m1} x_{i1} + \ \phi_{m2} x_{i2} \dots + \phi_{mp} x_{ip} \quad , \quad i = 1, 2 \dots n$$

coordinates of the n data points in terms of each element (m-th) of the new rotated basis. These express the values of the new composite features on each unit.

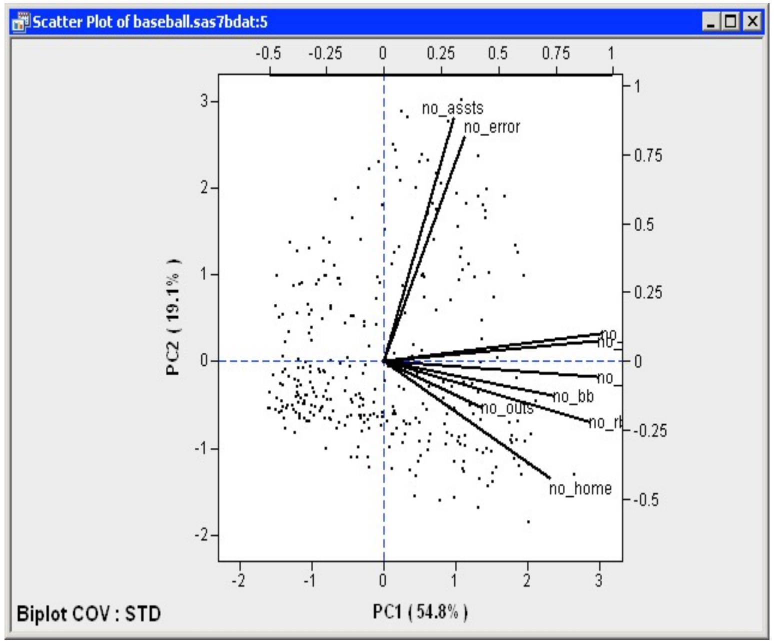
PERCENTAGE OF VARIANCE EXPLAINED (PVE) by the m-th component

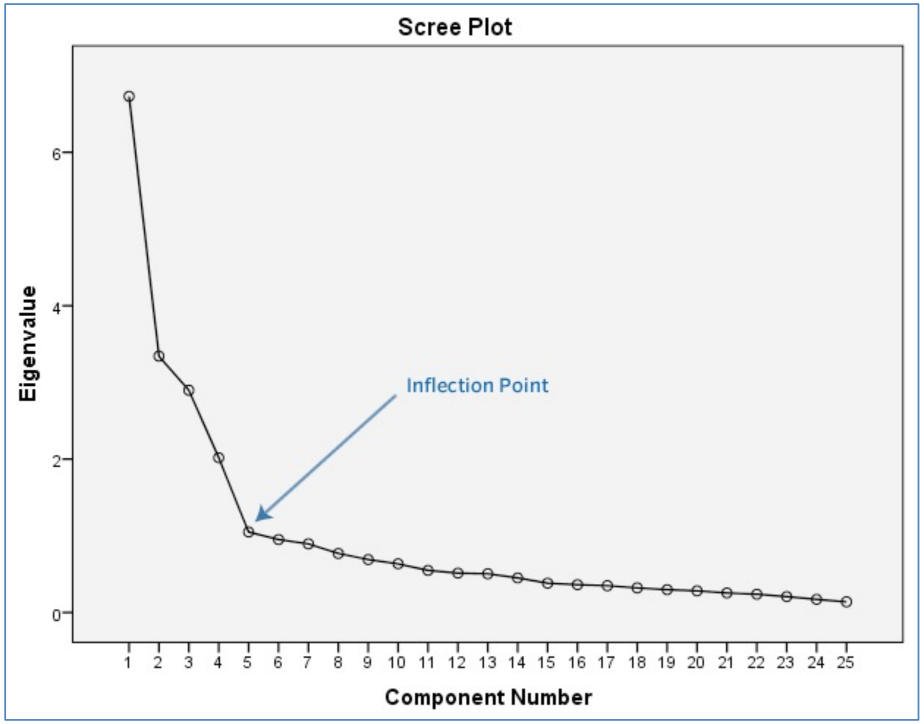
$$PVM_m = \frac{\lambda_m}{\sum_{k=1}^{p} \lambda_k} = \frac{\operatorname{var}(\phi_m^T X)}{\sum_{k=1}^{p} \operatorname{var}(\phi_k^T X)}$$

Also, **CUMULATIVE PVE** up to the m-th component

$$CPVM_m = \sum_{j=1}^{m} PVM_j = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{k=1}^{p} \lambda_k}$$

Biplot: shows the scores for two specified components (e.g. 1st and 2nd; projection of the points on the first PCA plane) along with the loadings represented by arrows.





Scree plot: Show the variances (eigenvalues) or PVEs in non-increasing order (graphical diagnostic to aid in the selection of m*=d)

Sliced Inverse Regression (SIR):

Method for supervised (sufficient) dimension reduction.

Main limitation of PC regression is that the composite features are identified as to capture variability in the feature space; the response Y plays no role.

Directions of maximal variability are not necessarily directions of maximal explanatory power with respect to Y!

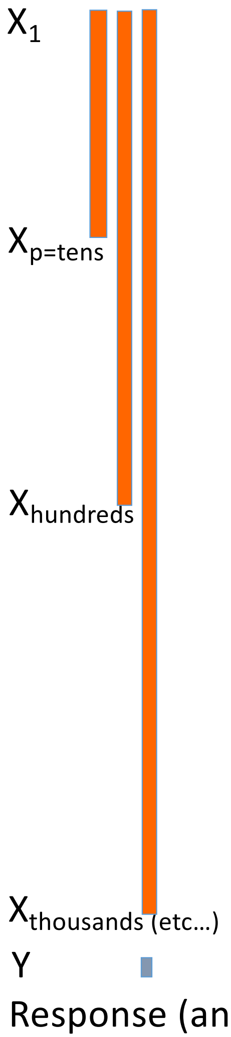
Can we perform ***supervised*** dimension reduction?
Yes, ~30 years old field of statistical research named ***Sufficient Dimension Reduction*** (SDR)

Some references:

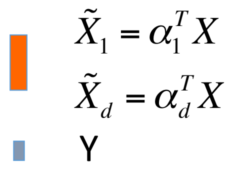
- Li, K. C. (1991) Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association, 86, 316–327.
- Cook, R. D. (1998) Regression Graphics: Ideas for Studying Regressions through Graphics. New York: Wiley.
- Ma, Y. and Zhu, L. (2013) A review on dimension reduction. International Statistical Review, 81, 134–150.

R package:
https://cran.r-project.org/web/packages/dr/index.html
https://cran.r-project.org/web/packages/dr/vignettes/overview.pdf

high dimensional feature vector (p)



Small composite feature vector
(d=1,2,3)



(1) Sufficient Dimension Reduction (SDR): inference on the **Central Subspace** (CS), which retains information on the dependence between Y and X

$$S_{Y|X} \text{ smallest s.t. } Y \perp X \mid P_{S_{Y|X}} X$$

$$S_{Y|X} = Span(A_{p\times d}),\ \ Y \perp X \mid A^T X \ \text{ basis version}$$

$$d = \dim(S_{Y|X}) \ \text{ structural dimension}$$

(2) Modeling: Y as a function of the composite feature vector and random error

$$Y = m(\tilde{X};\varepsilon) \quad e.g. \quad Y = \beta_o + \beta^T \tilde{X} + \varepsilon$$

need methods for inference on the CS (on the spanning matrix A) and on d!

SIR: simplest SDR method, based on an Eigen decomposition

- Consider the regression of X on Y (this is the *inverse regression*)
- *Slice* the range of Y in h=1,2… H slices (if continuous, otherwise use classes) and form the sample covariance matrix of E(X|Y)

$$\bar{X}_h = \frac{1}{n_h} \sum_{y_i \in h} X_i \quad h = 1,2 \ldots H$$

(we always assume overall mean vector = 0)

$$M = \frac{n_h}{n} \sum_{h=1 \ldots H} \bar{X}_h \bar{X}_h^T$$

- Take the Eigen decomposition of

$$S^{-1} M = \sum_{m=1}^{p} \lambda_m \phi_m \phi_m^T$$

Rescaling by S^{-1} very important! Weighing directions in the feature space

(also, cannot apply as is if n < p, rank(S) cannot exceed n-1)

$$\lambda_1 \geq \ldots \geq \lambda_p \geq 0 \quad \text{(eigenvalues)}$$

Note: only H-1 of the eigenvalues can be > 0, rank(M) cannot exceed H-1

$$\| \phi_m \| = \phi_m^T \phi_m = 1 \,,\ \phi_m^T \phi_k = 0 \ m,k = 1,2 \ldots p \quad \text{(eigenvectors)}$$

- Under conditions on the joint distribution of X (linearity), for any given dimension *m*, we identify the linear subspaces with highest explanatory power (directions within the CS).
- If we know d, we estimate the whole CS.

Remarks:

- Can make bi-plots and scree plots exactly as for PCA.

Also, If d=1,2

- Can plot Y vs the composite features to visualize the data and create a satisfactory model.
- Can use non-parametric regression fits.

- Like in feature selection (LASSO and related techniques), large literature. Some relevant developments:
  - Chiaromonte et al. (2002). *Sufficient dimension reduction in regressions with categorical predictors*. Annals of Statistics.
  - Li et al. (2010). *Groupwise dimension reduction*. JASA.
  - Guo et al. (2014). *Groupwise dimension reduction via envelope methods*. JASA.
  - Liu et al. (2017). *Structured Ordinary Least Squares: A Sufficient Dimension Reduction approach for regressions with partitioned predictors and heterogeneous units*. Biometrics.

Dimension Reduction: the relevant (structural) dimension d is selected

- Based on diagnostic statistics and their plots; e.g., for methods based on an Eigen decomposition, statistics and plots can be derived from eigenvalues.

- Using tests. For methods based on an Eigen decomposition, one can test how many tail eigenvalues are significantly > 0. More generally, one can use a sequence of tests, e.g. for any q=0,1... (p-1) test whether Ho: d=q vs Ha: d>q.

- Minimizing BIC or BIC-like criteria.

- Assessing stability through the Bootstrap (rather different approach; see Ye and Weiss (2003))

Linear Discriminant Analysis (LDA):
as a method for supervised dimension reduction when Y is categorical, again based on an Eigen decomposition.

Works just like SIR, but uses a *different rescaling*.

- Consider predicting Y based on X looking at how X varies in each of the Y classes (inverse approach)
- Say Y has H levels h=1,2… H. Without having to slice, form the sample covariance matrix of E(X|Y)

$$\bar{X}_h = \frac{1}{n_h}\sum_{y_i \in h} X_i \quad h = 1,2\ldots H$$

(we always assume overall mean vector = 0)

$$M = \frac{n_h}{n}\sum_{h=1\ldots H} \bar{X}_h \bar{X}_h^T$$

This is the between var/cov matrix S_B

- Take the Eigen decomposition of

$$\boxed{S_W^{-1}}\,\cancel{S}^{-1}M = \sum_{m=1}^{p} \lambda_m \phi_m \phi_m^T$$

In rescaling, instead of S (<u>overall</u> sample var/cov matrix) use S_W (<u>within</u> sample var/cov matrix). A different way of weighing directions in the feature space! (cannot apply as is if rank(S_W) < p)

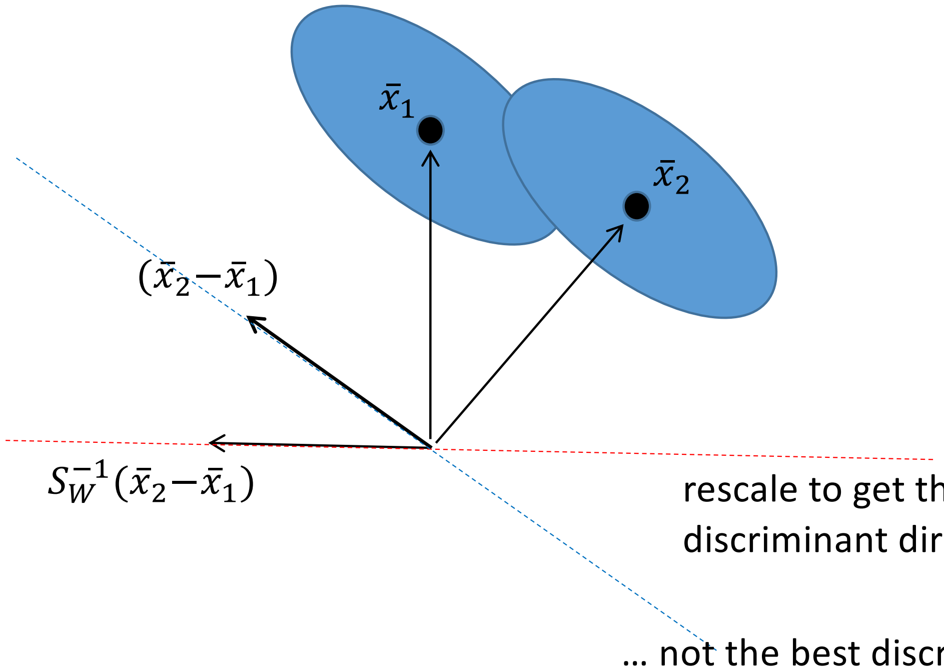
$$\lambda_1 \geq \ldots \geq \lambda_p \geq 0 \quad \text{(eigenvalues)}$$

<u>Note:</u> only H-1 of the eigenvalues can be > 0, rank(M) cannot exceed H-1

$$\|\phi_m\| = \phi_m^T \phi_m = 1\,,\ \phi_m^T \phi_k = 0 \ \ m,k = 1,2\ldots p \quad \text{(eigenvectors)}$$

Decomposition of the var/cov matrix; within and between variation

$$S = S_B + S_W \propto \sum_{k=1\ldots H} \sum_{i:y_i \in k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)' + \sum_{k=1\ldots H} n_k\,(\bar{x}_k - \bar{x}\ \ )(\bar{x}_k - \bar{x}\ \ )'$$



rescale to get the best discriminant direction

… not the best discriminant direction

LDA discriminant function (classify to highest)

$$\hat{\delta}_k(x) = (S_W^{-1}\bar{x}_k)'x - \frac{1}{2}\bar{x}_k{}'S_W^{-1}\bar{x}_k + log\frac{n_k}{n}$$

$$\hat{\delta}_2(x) - \hat{\delta}_1(x) = (S_W^{-1}\,(\bar{x}_2 - \bar{x}_1))'x + const$$

for simplicity H=2