ETC3250

Business Analytics

Week 7

Other dimensionality reduction methods

17 April 2018

Outline

Week	Topic	Chapter	Lecturer
1	Introduction	1	Souhaib
2	Statistical learning	2	Souhaib
3	Regression	3	Souhaib
4	Classification	4	Souhaib
5	Clustering	10	Souhaib
	Semester break		
6	Model selection and resampling methods	5	Souhaib
7	**Dimension reduction**	6,10	Souhaib
	Principal Components Analysis		
	Advanced dimension reduction methods		
8	Advanced regression	6	Souhaib
9	Advanced regression	6	Souhaib
10	Advanced classification	9	Souhaib
11	Tree-based methods	8	Souhaib
12	Project presentation		Souhaib

Dimensionality reduction

- Wy dimensionality reduction?
  - Curse of dimensionality
  - Computational demand
  - Intrinsic dimensionality
  - Visualization
- Dimensionality reduction methods
  - Feature selection vs feature extraction (examples? advantages?)
  - Unsupervised vs supervised
  - Linear vs nonlinear
- Principal components analysis (PCA)
  - PCA finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
  - PCA: linear and unsupervised feature extraction

Dimensionality reduction

- Wy dimensionality reduction?
  - Curse of dimensionality
  - Computational demand
  - Intrinsic dimensionality
  - Visualization
- Dimensionality reduction methods
  - Feature selection vs feature extraction (examples? advantages?)
  - Unsupervised vs supervised
  - Linear vs nonlinear
- Principal components analysis (PCA)
  - PCA finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
  - PCA: linear and unsupervised feature extraction

Dimensionality reduction

- Wy dimensionality reduction?
  - Curse of dimensionality
  - Computational demand
  - Intrinsic dimensionality
  - Visualization
- Dimensionality reduction methods
  - Feature selection vs feature extraction (examples? advantages?)
  - Unsupervised vs supervised
  - Linear vs nonlinear
- Principal components analysis (PCA)
  - PCA finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
  - PCA: linear and unsupervised feature extraction

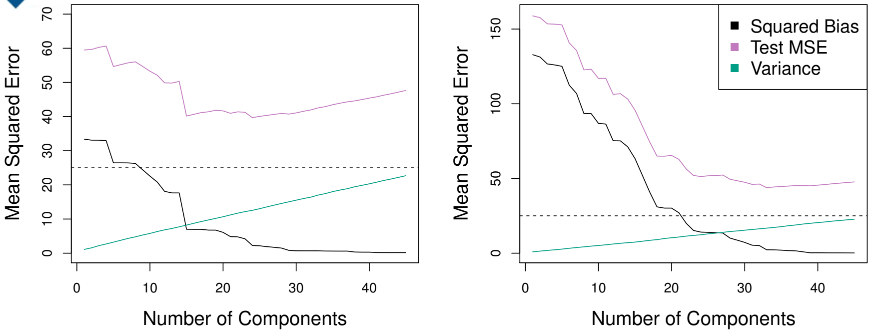
Dimensionality reduction in regression

- How would you reduce dimensionality in linear regression?
- *Principal components regression* (PCR): use PCA to construct the first $M \leq p$ principal components, $Z_1, \ldots, Z_M$, and then fit a linear regression model using these components.
- Assumption: *the directions in which $X_1, \ldots, X_p$ show the most variation are the directions that are associated with Y.*

Dimensionality reduction in regression

- How would you reduce dimensionality in linear regression?
- *Principal components regression* (PCR): use PCA to construct the first $M \leq p$ principal components, $Z_1, \ldots, Z_M$, and then fit a linear regression model using these components.
- Assumption: *the directions in which $X_1, \ldots, X_p$ show the <u>most variation</u> are the directions that are <u>associated</u> with Y.*

Principal components regression



- $n = 50$ observations and $p = 45$ predictors.
- left: Y is a function of **all** the predictors.
- right: Y is a function of **two predictors** only.

When is PCR better than traditional least squares?

Partial least squares

- With PCR, the components are identified in an **unsupervised way**. No guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response Y. How would you use the response Y to reduce dimensionality?
- Partial least squares (PLS) is a **supervised** alternative to PCR. Same procedure as PCR, but the new features are identified in a **supervised way**. Roughly speaking, PLS attempts to find directions that help explain both the response and the predictors.

Partial least squares

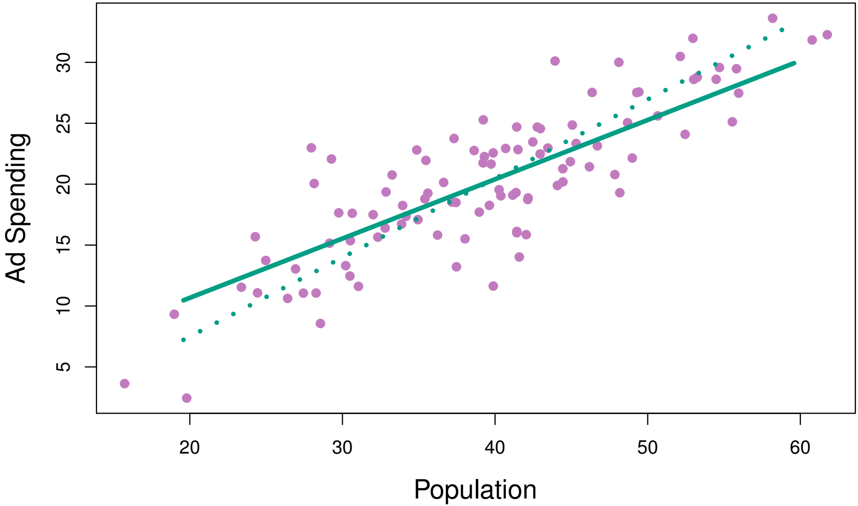
- With PCR, the components are identified in an **unsupervised way**. No guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response Y. How would you use the response Y to reduce dimensionality?

- Partial least squares (PLS) is a **supervised** alternative to PCR. Same procedure as PCR, but the new features are identified in a **supervised way**. Roughly speaking, PLS attempts to find directions that help explain both the response and the predictors.

Partial least squares

1. Standardize the p predictors
2. Compute the first direction Z_1 by setting each ϕ_{j1} in $Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$ equal to the coefficient from the simple linear regression of Y onto X_j
3. For the second direction Z_2, we first adjust each of the variables for Z_1, by regressing each variable on Z_1, and taking residuals, say X_j'. These residuals represent the remaining information that has not been explained by the the first PLS direction Z_1. We then compute Z_2 using X_j' as Z_1 was computed from X_j
4. We can compute $Z_1, \ldots, Z_M$ iteratively. Then we use least squares to fit a linear model using $Z_1, \ldots, Z_M$ as for PCR. How do we choose M?

Partial least squares



Solid line: first PLS direction. Dotted line: first PC.

Partial least squares

- There are two variants of PLS: PLS1 (one response variable) and PLS2 (at least two response variables).
- In practice, PLS1 is not better than ridge regression or PCR (PLS reduces bias but can potentially increase variance).
- PLS2 is a useful tool for multiresponse regression.
- Other supervised dimensionality reduction methods: CCA, LDA, etc.

Multidimensional scaling

Suppose that instead of measuring a $n \times p$ dataset $\boldsymbol{X} = [x_{ij}]$, we were only given a **pairwise distance matrix** Δ where

$$\Delta_{ij} = \|x_i - x_j\|, \quad i, j = 1, \ldots, N$$

i.e. we do not know the points themselves. Can we find a lower-dimension representation $\boldsymbol{Z}$ for $\boldsymbol{X}$ from Δ? (If we have only a distance matrix, we cannot perform PCA.)

Multidimensional scaling

Multidimensional scaling (MDS) attempts to find a lower dimensional space so that distances between points are **preserved** as well as possible.

MDS seeks values $z_1, z_2, \ldots, z_n \in \mathbb{R}^k$ that minimize the so-called stress function:

$$S_M(z_1, z_2, \ldots, z_n) = \sum_{i \neq j} (\|x_i - x_j\| - \|z_i - z_j\|)^2.$$

(**least squares** or *Kruskal-Shephard* scaling).
A variation on least squares scaling minimizes

$$S_{\mathrm{Sm}}(z_1, z_2, \ldots, z_n) = \sum_{i \neq j} \frac{(\|x_i - x_j\| - \|z_i - z_j\|)^2}{\|x_i - x_j\|}$$

(**Sammon** mapping) More emphasis is put on preserving smaller pairwise distances.

Link between MDS and PCA

Computation of PCs in PCA:

- Eigenvalue decomposition
  - $\boldsymbol{C} = \boldsymbol{X}'\boldsymbol{X}$ where the columns of $\boldsymbol{X}$ are scaled
  - $\boldsymbol{C} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}'$ with $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$
  - $\Phi = \boldsymbol{V} \rightarrow \boldsymbol{Z} = \boldsymbol{X}\Phi$
- Singular value decomposition
  - $\boldsymbol{X} = \boldsymbol{U}\Lambda\boldsymbol{V}'$ with $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$
  - $\Phi = \boldsymbol{V} \rightarrow \boldsymbol{Z} = \boldsymbol{X}\Phi$

Link between MDS and PCA

If Δ_{ij} are **Euclidean distances** between the rows of $\boldsymbol{X}$, then classical MDS is equivalent to PCA.
$\rightarrow$ Preserve Euclidean distances = retaining the maximum variance

$$\begin{aligned}\Delta_{ij}^2 &= \|x_i - x_j\|^2 \\ &= \|x_i - \bar{x}\|^2 + \|x_j - \bar{x}\|^2 - 2\langle x_i - \bar{x}, x_j - \bar{x}\rangle\end{aligned}$$

Classical multidimensional scaling minimizes

$$S_C(z_1, z_2, \ldots, z_n) = \sum_{i \neq j} \left(\langle x_i - \bar{x}, x_j - \bar{x}\rangle - \langle z_i - \bar{z}, z_j - \bar{z}\rangle\right)^2$$

Link between MDS and PCA

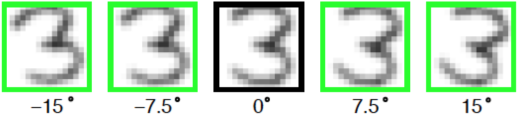
Given distance matrix $\Delta \in \mathbb{R}^{n \times n}$,

1. Recover the inner product $B = \boldsymbol{XX'}$ from Δ
   - Compute $A_{ij} = -\frac{1}{2}\Delta_{ij}^2$
   - Double center A and compute: $B = (I - M)A(I - M)$ where $M = \frac{1}{n}\mathbb{1}\mathbb{1}' \in \mathbb{R}^{n \times n}$ where $B_{ij} = \langle x_i - \bar{x}, x_j - \bar{x} \rangle$
2. Factorize B to gest the first k principal component scores

- PCA uses the $d \times d$ covariance matrix: $\boldsymbol{C} = \frac{1}{n-1}\boldsymbol{X}^{'}\boldsymbol{X}$
- MDS uses the $n \times n$ Gram (inner product) matrix: $\boldsymbol{K} = \boldsymbol{XX}^{'}$

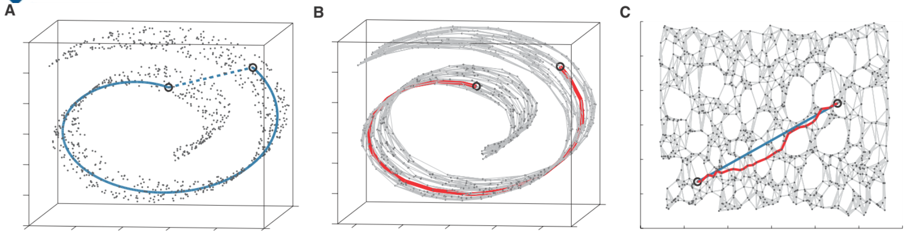
Multidimensional scaling

- Multidimensional scaling can be applied to any Δ_{ij}, not just Euclidean distances
- In this case, we don't compute principal component scores, and the lower-dimensional representation can be a nonlinear function of the data
- When do we need to use non-Euclidean distances?



We wish to remove the effect of rotation in measuring distances between two digits

Isometric feature mapping



(From Tenenbaum et al. (2000), “A global geometric framework for nonlinear dimensionality reduction”)

- Construct a graph $G = (V, E)$ based on the structure between $x_1, \ldots, x_n$.
- Then, define a graph distance $\Delta_{ij}^{\text{Isomap}}$ between i and j, and use MDS for the low-dimensional representation

Dimensionality reduction methods

- Feature selection vs feature extraction
- Linear and nonlinear
- Unsupervised and supervised
- Low-dimensional representation with maximum variance, that retains local properties of the data, etc.
- **Linear PCA**, Nonlinear PCA, Kernel PCA, Sparse PCA, etc.
- **MDS**, ICA, LDA, etc.
- **PLS**, CCA, FA, etc.
- **Isomap**, diffusion maps, MVU, LLE, t-SNE, autoencoders, etc.