Bi-cross-validation of the SVD and the Nonnegative Matrix Factorization

Art Owen and Patrick Perry
presented by Linbo Wang

April 18th, 2013
Biostat 572: Advanced Regression Methods: Project

Introduction: Statistical modeling [1]

- Statistical model: $\mathbf{Y} = \boldsymbol{\Theta} + \mathbf{E}$
  - $\mathbf{Y}$: Observed data, potentially a matrix (e.g. subject $\times$ academic fields)
  - $\boldsymbol{\Theta}$: Mean model: a fixed pattern we want to recover
  - $\mathbf{E}$: Covariance Model: $E[\mathbf{E}] = 0$
- Mean model
  1. Regression model: $\boldsymbol{\Theta} = \boldsymbol{\Theta}(\mathbf{B}, \mathbf{X})$, $\mathbf{X}$ observed (given)
     - Supervised learning problem
  2. Rank/factor model: $\boldsymbol{\Theta} = \boldsymbol{\Theta}(\mathbf{B}, \mathbf{F})$, $\mathbf{F}$ latent
     - Unsupervised learning problem

[1] From Peter Hoff's notes on CSSS 594

Introduction: Statistical modeling [1]

- Statistical model: $\mathbf{Y} = \mathbf{\Theta} + \mathbf{E}$
  - $\mathbf{Y}$: Observed data, potentially a matrix (e.g. subject $\times$ academic fields)
  - $\mathbf{\Theta}$: Mean model: a fixed pattern we want to recover
  - $\mathbf{E}$: Covariance Model: $E[\mathbf{E}] = 0$
- Mean model
  1. Regression model: $\mathbf{\Theta} = \mathbf{\Theta}(\mathbf{B}, \mathbf{X})$, $\mathbf{X}$ observed (given)
     - Supervised learning problem
  2. Rank/factor model: $\mathbf{\Theta} = \mathbf{\Theta}(\mathbf{B}, \mathbf{F})$, $\mathbf{F}$ latent
     - Unsupervised learning problem

[1] From Peter Hoff's notes on CSSS 594

Example [2]

- Outcome: examination scores from each of 10 different academic fields of 1000 students
- Latent covariates: "verbal intelligence", "mathematical intelligence", "EQ", etc.
- Build a model with the latent covariates

$$\mathbf{Y} = \mathbf{BF} + \mathbf{E}$$

  - $\mathbf{Y} \in \mathbf{R}^{n \times s}, \mathbf{F} \in \mathbf{R}^{n \times r}, \mathbf{B} \in \mathbf{R}^{r \times s}$
  - we would want $rank(\mathbf{BF}) < rank(\mathbf{Y})$
  - Mathematically: lower-rank approximation to $\mathbf{Y}$

[2] From wikipedia item on factor analysis

Example [2]

- Outcome: examination scores from each of 10 different academic fields of 1000 students
- Latent covariates: "verbal intelligence", "mathematical intelligence", "EQ", etc.
- Build a model with the latent covariates

$$\mathbf{Y} = \mathbf{BF} + \mathbf{E}$$

  - $\mathbf{Y} \in \mathbf{R}^{n \times s}, \mathbf{F} \in \mathbf{R}^{n \times r}, \mathbf{B} \in \mathbf{R}^{r \times s}$
  - we would want $rank(\mathbf{BF}) < rank(\mathbf{Y})$
  - Mathematically: lower-rank approximation to $\mathbf{Y}$

[2] From wikipedia item on factor analysis

Recall: linear regression

- Projection of Y to $R(X) \triangleq \text{span}\{\text{column space of } \mathbf{X}\}$

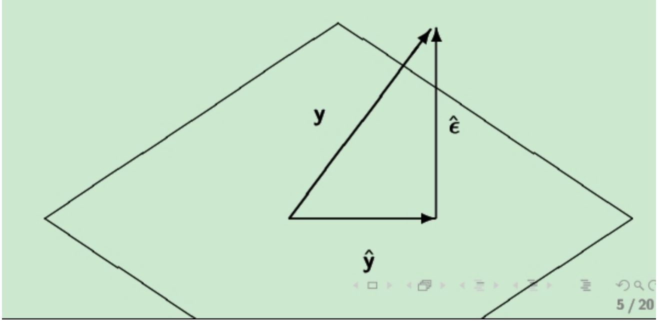

Figure: OLS estimation in linear regression is a projection.[3]

[3] From Daniela Witten's 533 lecture notes.

Estimation method: Truncated SVD

- Singular Value Decomposition (SVD):

$$\mathbf{Y} = \mathbf{UDV}^T = \sum_{k=1}^{n} d_k u_k v_k^T$$

  - $d_1 \geq d_2 \geq \cdots \geq d_n \geq 0$

- Eckart and Young(1936): Given $rank(\mathbf{BF}) = r$, the following truncated SVD minimized square error $\|\mathbf{Y} - \mathbf{BF}\|_2^2$

$$\hat{\mathbf{B}}\mathbf{F} = \sum_{k=1}^{r} d_k u_k v_k^T = U_r D_r V_r^T$$

Estimation method: Truncated SVD

- $\mathbf{Y}_{n\times s} = \mathbf{U}_{n\times s}\mathbf{D}_{s\times s}\mathbf{V}^T_{s\times s} \approx \mathbf{U}1_{n\times r}\mathbf{D}1_{r\times r}\mathbf{V}1^T_{r\times r}$

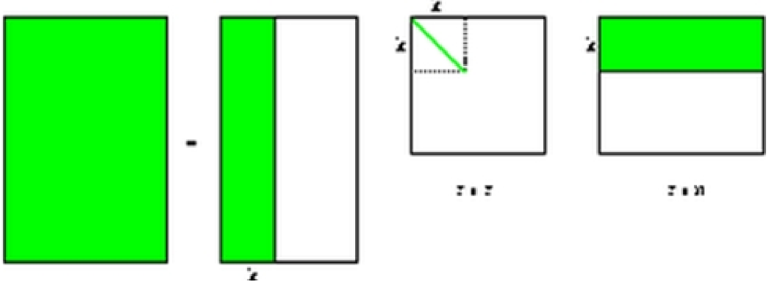

Figure: Comparison of SVD and truncated SVD [4]

- Question: How to choose r?

[4] Figure adapted from http://web.eecs.utk.edu/ berry/lsi++/node8.html

Estimation method: Truncated SVD

- $\mathbf{Y}_{n\times s} = \mathbf{U}_{n\times s}\mathbf{D}_{s\times s}\mathbf{V}^T_{s\times s} \approx \mathbf{U}1_{n\times r}\mathbf{D}1_{r\times r}\mathbf{V}1^T_{r\times r}$

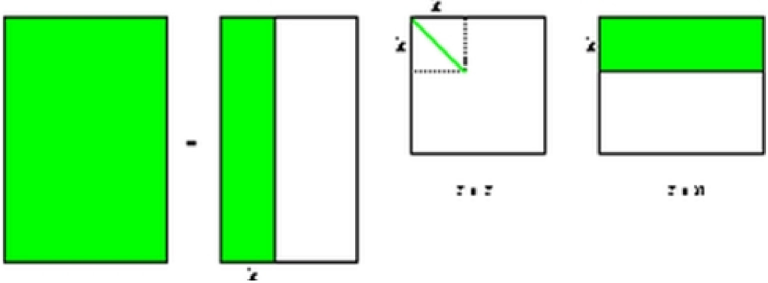

Figure: Comparison of SVD and truncated SVD [4]

- Question: How to choose r?

[4] Figure adapted from http://web.eecs.utk.edu/ berry/lsi++/node8.html

How to choose r?

- F-test(Dias and Krzanowski(2003)): standard approach for model selection, but F-tests are not reliable here.
- Square error $\|\mathbf{Y} - \mathbf{BF}\|_2^2$
  - Prone to overfitting
- Usual practice (Hoff(2007)): look for where the last large gap or elbow appears in a plot of singular values
  - Lack of numerical standards
- Wold(1978) : add terms until the residual standard error matches the noise level
  - requires knowledge of noise level, which we usually don't have
- Cross validation: usual practice for supervised learning, as well as providing numerical standards
  - non-trivial under unsupervised learning settings

Cross validation under unsupervised learning

- "The" way to choose r if we know how to do it, especially for prediction purpose
- We don't know covariates as in regression/supervised learning setting, then what can we do?
  - $$\mathbf{Y}_{n\times s} = \mathbf{X}_{n\times p}\mathbf{B}_{p\times s} + \mathbf{E}_{n\times s}$$
  - $$(\mathbf{Y}, \mathbf{X}) = \begin{pmatrix} \mathbf{Y}_{1:r,1:s} & \mathbf{X}_{1:r,1:p} \\ \mathbf{Y}_{(r+1):n,1:s} & \mathbf{X}_{(r+1):n,1:p} \end{pmatrix}$$
  - What can we do without knowing X?
- How to choose the left-out portion? (i.e. r in the above example)

Summary

- Factor model is a popular alternative to regression model
- Lower rank approximation to the observed data can be obtained via truncated SVD
- Current practice of choice of k is arbitrary
- Problem: cross-validation for unsupervised learning

Questions?

Bi-cross-validation

- Usual cross-validation leaves rows out
  - $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_{1:r,1:s} \\ \mathbf{Y}_{(r+1):n,1:s} \end{pmatrix}$
  - doesn't work here!
- Bi-cross-validation (BCV): leaves out rows and columns simultaneously
  - $\mathbf{Y}_{n \times s} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$
    - example: test scores from 100 students on 10 academic fields
  - Most authors consider leaving out a 1×1 matrix

Bi-cross-validation

- Usual cross-validation leaves rows out
  - $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_{1:r,1:s} \\ \mathbf{Y}_{(r+1):n,1:s} \end{pmatrix}$
  - doesn't work here!
- Bi-cross-validation (BCV): leaves out rows and columns simultaneously
  - $\mathbf{Y}_{n \times s} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$
    - example: test scores from 100 students on 10 academic fields
  - Most authors consider leaving out a 1×1 matrix