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

Art Owen and Patrick Perry
presented by Linbo Wang

May 9th, 2013
Biostat 572: Advanced Regression Methods: Project

Problem

- Problem: lower-rank approximation to matrix $\mathbf{Y}$

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

  - $\mathbf{Y} \in \mathbf{R}^{m \times n}, \mathbf{M} \in \mathbf{R}^{m \times n}$
  - $k = rank(\mathbf{M}) < rank(\mathbf{Y}) = min(m, n)$

- Motivation: factor analysis

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

- What is special: unsupervised learning

Given k (rank of $\mathbf{M}$): Truncated SVD

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

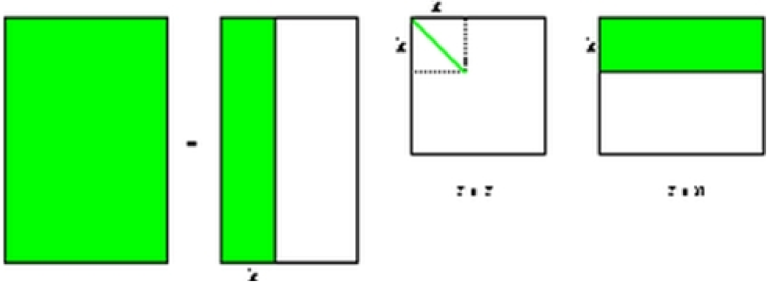

Figure: Comparison of SVD and truncated SVD [1]

- Question: How to choose k?

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

How to choose k?

- Usual practice (Hoff(2007)): look for where the last large gap or elbow appears in a plot of singular values
  - Lack of numerical standards
- Cross validation: usual practice for supervised learning, as well as providing numerical standards
  - non-trivial under unsupervised learning settings

Bi-cross-validation

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

Eastment and Krzanowski (1982)

- Recall from SVD:
  - $\mathbf{Y}_{m\times n} = \mathbf{U}_{m\times n}\mathbf{D}_{n\times n}\mathbf{V}^T_{n\times n} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$
  - U represents row information, and V represents column information.
- Eastment and Krzanowski (1982):
  - $(C, D)_{(m-1)\times n} = \mathbf{U}1_{(m-1)\times n}\mathbf{D}1_{n\times n}\mathbf{V}1^T_{n\times n} \approx \bar{\mathbf{U}}1_{(m-1)\times k}\bar{\mathbf{D}}1_{k\times k}\bar{\mathbf{V}}1^T_{k\times k}$
  - $\begin{pmatrix} B \\ D \end{pmatrix}_{m\times(n-1)} = \mathbf{U}2_{m\times(n-1)}\mathbf{D}2_{(n-1)\times(n-1)}\mathbf{V}2^T_{(n-1)\times(n-1)} \approx \bar{\mathbf{U}}2_{m\times k}\bar{\mathbf{D}}2_{k\times k}\bar{\mathbf{V}}2^T_{k\times k}$
- $U_{m\times n} \approx U2_{m\times(n-1)} \approx \bar{\mathbf{U}}2_{m\times k}$, $V^T_{n\times n} \approx V1^T_{n\times n} \approx \bar{\mathbf{V}}1^T_{k\times k}$
- $D_{k\times k} \approx \sqrt{\bar{D}1_{k\times k}\bar{D}2_{k\times k}}$
- $A_{1\times 1} = \hat{Y}[1, 1]$

Eastment and Krzanowski (1982)

- Recall from SVD:
  - $\mathbf{Y}_{m\times n} = \mathbf{U}_{m\times n}\mathbf{D}_{n\times n}\mathbf{V}^T_{n\times n} = \begin{pmatrix} & B \\ C & D \end{pmatrix}$
  - U represents row information, and V represents column information.
- Eastment and Krzanowski (1982):
  - $(C,D)_{(m-1)\times n} = \mathbf{U}1_{(m-1)\times n}\mathbf{D}1_{n\times n}\mathbf{V}1^T_{n\times n} \approx \bar{\mathbf{U}}1_{(m-1)\times k}\bar{\mathbf{D}}1_{k\times k}\bar{\mathbf{V}}1^T_{k\times k}$
  - $\begin{pmatrix} B \\ D \end{pmatrix}_{m\times(n-1)} = \mathbf{U}2_{m\times(n-1)}\mathbf{D}2_{(n-1)\times(n-1)}\mathbf{V}2^T_{(n-1)\times(n-1)} \approx \bar{\mathbf{U}}2_{m\times k}\bar{\mathbf{D}}2_{k\times k}\bar{\mathbf{V}}2^T_{k\times k}$
- $U_{m\times n} \approx U2_{m\times(n-1)} \approx \bar{\mathbf{U}}2_{m\times k},\ V^T_{n\times n} \approx V1^T_{n\times n} \approx \bar{\mathbf{V}}1^T_{k\times k}$
- $D_{k\times k} \approx \sqrt{\bar{D}1_{k\times k}\bar{D}2_{k\times k}}$
- $A_{1\times 1} = \hat{Y}[1,1]$

Eastment and Krzanowski (1982)

- Best known method up to date (cited by 237 up to date)
- Critiques
  - cross-validation errors decrease monotonically with k
    - some awkward adjustments based on estimated degree of freedom are used in practice.
  - sign for singular vectors $\mathbf{u}_i, \mathbf{v}_i^T$ is not determined
  - Linbo: lack of theoretical justification

Bi-cross-validation (BCV)

- Motivation: cross-validation of principal component regression (PCR) [2]
- First studied by Gabriel (2002) in 1×1 case.
- avoids drawbacks of Eastment and Krzanowski (1982)

[2] This is a made-up motivation by the presenter.

Bi-cross-validation (BCV)

- $$\mathbf{Y} = \begin{pmatrix} A_{1:r,1:s} & B_{1:r,(s+1):n} \\ C_{(r+1):m,1:s} & D_{(r+1):m,(s+1):n} \end{pmatrix}$$
- Fit a principal component regression of C on D: $\hat{\beta} = (\hat{D}^{(k)})^{-}C$
  - $D^{(k)}$ is the best rank-k approximation to D
  - "-" is the Moore-Penrose generalized inverse
- Get "estimate" of A by $B\hat{\beta}$
- Do this for different hold-out portion A

$$BCV(k) = \sum_{i=1}^{h}\sum_{j=1}^{l} \|A(i,j) - B(i,j)(\hat{D}(i,j))^{-}C(i,j)\|_F^2$$

- Counterintuitive: Use the best rank for D as the best rank for Y.

Properties

- Theoretical properties ("Model Selection Consistency"):
  - Self-consistency property
  - Pure Gaussian noise (true k=0)
    - Asymptotically: $E[BCV(1)] > E[BCV(0)]$
  - Rank 1 plus Gaussian noise (true k=1)
    - Asymptotically: $E[BCV(1)] < E[BCV(0)]$ under some conditions
- Empirical properties (next time...)
  - Cross-validation error is U-shape with respect to k

Model Selection Consistency

- Self-consistency property
  - $\mathbf{Y} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$
  - Conditions: rank(Y)=rank(D)=k
  - Conclusion: $A - B(\hat{D}^k)^- C = A - BD^- C = 0$
- Eastment and Krzanowski (1982) generally doesn't have such property

Model Selection Consistency

- True rank k=0: $Y_{m\times n} = 0_{m\times n} + Z_{m\times n}$
  - $Z_{ij} \overset{\text{iid}}{\sim} N(0,1)$
  - $c \approx m/n$: size of matrix
  - hold out proportion is constant: $r/m = s/n = \theta$
- True rank k=1: $Y_{m\times n} = \kappa u_{m\times 1} {v_{n\times 1}}^T + Z_{m\times n}$
  - $u_{m\times 1}$ and $v_{n\times 1}$ are unit vectors
  - root mean square of noise: $\left(E[Z_{ij}^2]\right)^{1/2} = 1$
  - root mean square of signal: $(E(\kappa uv^T)^2/mn)^{1/2} = \kappa\sqrt{mn}$
  - Assume $\kappa^2 = \delta\sqrt{mn}, \delta > 1$ represents the strength of signal.

Choice of hold out portion

TABLE 1

This table summarizes expected cross-validated squared errors from the text. The lower right entry is conservative as described in the text. The value $\eta \in (1/2, 1)$ represents a lower bound on the proportion not held out, for each singular vector u and v. The value θ is an assumed common value for r/m and s/n and $\delta > 1$ is a measure of signal strength

$\mathbb{E}(\mathbf{BCV}(k)) - mn$	**True $k = 0$**	**True $k = 1$**
Fitted $k = 0$	0	$\delta\sqrt{mn}$
Fitted $k = 1$	$\frac{1}{1-\theta}\frac{\sqrt{mn}}{\sqrt{c}+1/\sqrt{c}+2}$	$\sqrt{mn}(\delta(1-1/\eta)^2 + c^{3/2} + c^{-3/2} + 1/\delta)$

- Smaller aspect ratio ($c \approx m/n$ closer to 1) is advantageous.
- Larger hold out portion θ will favor selection of lower rank.
  - Small holdouts is more prone to overfitting.
  - Large holdouts is more prone to underfitting.
- In practice, the authors recommend a $(2 \times 2) -$ *fold* or $(3 \times 3) -$ *fold* BCV. (more about this next time...)

Summary

- Lower rank approximation to the observed data can be obtained via truncated SVD.
- Current practice of choice of k is arbitrary
- Bi-cross-validation(BCV) is a reasonable generalization of cross-validation to this unsupervised learning setting.
- Some theoretical justifications for BCV is presented, and more needs to be discovered.
- Choice of holdout size is still an open problem.

Coming next...

- Simulation results
- Real data application
- Discussion

Questions?