Statistics 202: Data Mining

Dimension reduction, PCA & eigenanalysis
Based in part on slides from textbook, slides of Susan Holmes

©Jonathan Taylor

October 3, 2012
Dimension reduction

**Combinations of features**

- Given a data matrix $\mathbf{X}_{n \times p}$ with $p$ fairly large, it can be difficult to visualize structure.
- Often useful to look at linear combinations of the features.
- Each $\beta \in \mathbb{R}^p$, determines a *linear* rule

$$f_\beta(\mathbf{x}) = \mathbf{x}^T \beta$$

- Evaluating this on each row $\mathbf{X}_i$ of $\mathbf{X}$ yields a vector

$$(f_\beta(\mathbf{X}_i))_{1 \leq i \leq n} = \mathbf{X}_\beta.$$
Dimension reduction

- By choosing $\beta$ appropriately, we may find “interesting” new features.
- Suppose we take $k$ much smaller than $p$ vectors of $\beta$ which we write as a matrix
  \[ B_{p \times k} \]
- The new data matrix $XB$ has fewer dimensions than $X$.
- This is dimension reduction …
Dimension reduction

**Principal Components**

- This method of dimension reduction seeks features that maximize sample variance . . .
- Specifically, for $\beta \in \mathbb{R}^p$, define the sample variance of $V_\beta = X\beta \in \mathbb{R}^n$

$$\hat{\text{Var}}(V_\beta) = \frac{1}{n-1} \sum_{i=1}^{n} (V_{\beta,i} - \bar{V}_\beta)^2$$

$$\bar{V}_\beta = \frac{1}{n} \sum_{i=1}^{n} V_{\beta,i}$$

- Note that $\hat{\text{Var}}(V_{C\beta}) = C^2 \cdot \hat{\text{Var}}(V_\beta)$ so we usually standardize $||\beta||_2 = \sqrt{\sum_{i=1}^{n} \beta_i^2} = 1$
Dimension reduction

**Principal Components**

- Define the $n \times n$ matrix

\[ H = I_{n \times n} - \frac{1}{n} \mathbf{1} \mathbf{1}^T \]

- This matrix removes means:

\[ (Hv)_i = v_i - \overline{v}. \]

- It is also a projection matrix:

\[ H^T = H \]
\[ H^2 = H \]
Dimension reduction

Principal Components with Matrices

- With this matrix,
  \[ \hat{\text{Var}}(V_\beta) = \frac{1}{n-1} \beta^T X^T H X \beta. \]

- So, maximizing sample variance, with \( \|\beta\|_2 = 1 \) is
  \[ \maximize_{\|\beta\|_2=1} \beta^T \left( X^T H X \right) \beta. \]

- This boils down to an eigenvalue problem ...
Dimension reduction

Eigenanalysis

- The matrix $X^THX$ is symmetric, so it can be written as

$$X^THX = VDV^T$$

where $D_{k \times k} = \text{diag}(d_1, \ldots, d_k)$, rank($X^THX$) = $k$ and $V_{p \times k}$ has orthonormal columns, i.e. $V^TV = I_{k \times k}$.

- We always have $d_i \geq 0$ and we take $d_1 \geq d_2 \geq \ldots d_k$. 
Dimension reduction

Eigenanalysis & PCA

- Suppose now that $\beta = \sum_{j=1}^{k} a_j v_j$ with $v_j$ the columns of $V$. Then,

$$\|\beta\|_2 = \sqrt{\sum_{i=1}^{k} a_i^2}$$

$$\beta^T \left(X^T H X\right) \beta = \sum_{i=1}^{k} a_i^2 d_i$$

- Choosing $a_1 = 1$, $a_j = 0$, $j \geq 2$ maximizes this quantity.
Dimension reduction

Eigenanalysis & PCA

- Therefore, \( \hat{\beta}_1 = v_1 \) the first column of \( V \) solves

\[
\text{maximize } \beta^T \left( X^T H X \right) \beta \quad \text{subject to } \|\beta\|_2 = 1
\]

- This yields scores \( HX\hat{\beta}_1 \in \mathbb{R}^n \).
- These are the 1st principal component scores.
Dimension reduction

Higher order components

- Having found the direction with “maximal sample variance” we might look for the “second most variable” direction by solving

\[
\begin{align*}
\text{maximize} & \quad \beta^T \left( X^T H X \right) \beta \\
\text{subject to} & \quad \beta^T v_1 = 0, \beta^T \beta = 1
\end{align*}
\]

- Note we restricted our search so we would not just recover \( v_1 \) again.

- Not hard to see that if \( \beta = \sum_{j=1}^{k} a_j v_j \) this is solved by taking \( a_2 = 1, a_j = 0, j \neq 2 \).
Dimension reduction

Higher order components

- In matrix terms, all the principal components scores are
  \[(HXV)_{n \times k}\]
  
  and the loadings are the columns of \(V\).

- This information can be summarized in a biplot.

- The loadings describe how each feature contributes to each principal component score.
Olympic data
Olympic data: screeplot

```
princomp(olympicStab)
```

Variances

Comp.1  Comp.3  Comp.5  Comp.7  Comp.9
Dimension reduction

The importance of scale

- The PCA scores are not invariant to scaling of the features: for $Q_{p \times p}$ diagonal the PCA scores of $XQ$ are not the same as $X$.
- Common to convert all variables to the same scale before applying PCA.
- Define the scalings to be the sample standard deviation of each feature. In matrix form, let

$$S^2 = \frac{1}{n-1} \text{diag} \left( X^T HX \right)$$

- Define $\tilde{X} = HXS^{-1}$. The normalized PCA loadings are given by an eigenanalysis of $\tilde{X}^T \tilde{X}$. 
Olympic data
Olympic data: screeplot

princomp(olympic$Tab, cor = TRUE)
Dimension reduction

PCA and the SVD
- The singular value decomposition of a matrix tells us we can write
  \[ \tilde{X} = U\Delta V^T \]
  with \( \Delta_{k \times k} = \text{diag}(\delta_1, \ldots, \delta_k) \), \( \delta_j \geq 0 \), \( k = \text{rank}(\tilde{X}) \), \( U^T U = V^T V = I_{k \times k} \).
- Recall that the scores were
  \[ \tilde{X} V = (U\Delta V^T) V = U\Delta \]
- Also,
  \[ \tilde{X}^T \tilde{X} = V\Delta^2 V^T \]
  so \( D = \Delta^2 \).
Dimension reduction

**PCA and the SVD**

- Recall $\tilde{X} = HXS^{-1/2}$.
- We saw that normalized PCA loadings are given by an eigenanalysis of $\tilde{X}^T \tilde{X}$.
- It turns out that, via the SVD, the normalized PCA scores are given by an eigenanalysis of $\tilde{X}\tilde{X}^T$ because

$$
\begin{pmatrix}
\tilde{X}V
\end{pmatrix}_{n \times k} = U\Delta
$$

where $U$ are eigenvectors of

$$
\tilde{X}\tilde{X}^T = U\Delta^2 U^T.
$$
Dimension reduction

Another characterization of SVD

- Given a data matrix $\mathbf{X}$ (or its scaled centered version $\mathbf{\tilde{X}}$) we might try solving

$$\minimize_{\mathbf{Z}: \text{rank}(\mathbf{Z})=k} \parallel \mathbf{X} - \mathbf{Z} \parallel_F^2$$

where $F$ stands for Frobenius norm on matrices

$$\parallel A \parallel_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} A_{ij}^2$$
Dimension reduction

Another characterization of SVD

- It can be proven that

\[ Z = S_{n \times k}(V^T)_{k \times p} \]

where \( S_{n \times k} \) is the matrix of the first \( k \) PCA scores and the columns of \( V \) are the first \( k \) PCA loadings.

- This approximation is related to the screeplot: the height of each bar describes the additional drop in Frobenius norm as the rank of the approximation increases.
Olympic data: screeplot

princomp(olympic$med, cor = TRUE)
Dimension reduction

Other types of dimension reduction

- Instead of maximizing sample variance, we might try maximizing some other quantity ...

- Independent Component Analysis tries to maximize “non-Gaussianity” of $V_\beta$. In practice, it uses skewness, kurtosis or other moments to quantify “non-Gaussianity.”

- These are both unsupervised approaches.

- Often, these are combined with supervised approaches into an algorithm like:

  Feature creation  Build some set of features using PCA.
  Validate      Use the derived features to see if they are helpful in the supervised problem.
Olympic data: 1st component vs. total score