STA 5106
Computational Methods in Statistics I

Department of Statistics
Florida State University

Class 5
September 10, 2019

2.5 Multiple Linear Regression Algorithms

Minimization Problem

- **Minimization problem:** For a given $y \in \mathbf{R}^m$ and $X \in \mathbf{R}^{m \times n}$ find
$$\hat{b} = \arg\min_b \| y - Xb \|^2$$
- Using orthogonal transformations the problem is equivalent to
$$\hat{b} = \arg\min_b \| y^* - X^* b \|^2$$
where $y^* = Oy$ and $X^* = OX$ for an $m \times m$ orthogonal matrix O.
- O can be obtained through Householder transformations so that $X^* = OX$ is upper triangular.
- Once O is found, then we can solve the minimization through backward substitution.

Backward Substitution

- Let y_j be the result of j Householder multiplications to the original y.

- At the j-th iterative stage, H_j is multiplied to the vector y_{j-1} to obtain y_j.

$$y_j(1:j-1) = y_{j-1}(1:j-1)$$

$$y_j(j:m) = \tilde{H}_j y_{j-1}(j:m) = (I - 2\frac{vv^T}{v^T v})[y_{j-1}(j:m)].$$

- Therefore,

$$y_j(j:m) = y_{j-1}(j:m) + \beta v, \quad \text{where } \beta = \frac{-2v^T[y_{j-1}(j:m)]}{v^T v}.$$

Multiple Linear Regression Algorithm

- **Algorithm 18 (Multiple Linear Regression)**

```
function b = multilinreg(X,y)
[m,n] = size(X);
for j=1:n
        v(1:m,1) = zeros(m,1);
        v(j:m,1) = house(X(j:m,j));
        X(j:m,j:n) = rowhouse(X(j:m,j:n),v(j:m));
        beta = -2*(v(j:m)^T*y(j:m))/(v(j:m)^T*v(j:m));
        y(j:m) = y(j:m) + beta*v(j:m);
end
b = backsub(X,y);
```

2.8 Singular Value Decomposition

Singular Value Decomposition

- **Theorem 2** For any $X \in \mathbf{R}^{m \times n}$ (assuming $m \geq n$) there exist orthogonal matrices $U \in \mathbf{R}^{m \times m}$ and $V \in \mathbf{R}^{n \times n}$ such that

  $$U^T X V = \Sigma, \text{ where } \Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n) \in \mathbf{R}^{m \times n},$$

  and where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$.

  σ_i' s are called the **singular values** of X and the columns of U and V are called the **singular vectors** of X.

- As $U^T X V = \Sigma$, we have $XV = U\Sigma$.

- Let $U = (u_1, u_2, \ldots, u_m)$ and $V = (v_1, v_2, \ldots, v_n)$. Then

  $$X(v_1, v_2, \ldots, v_n) = (u_1, u_2, \ldots, u_m)\Sigma.$$

Singular Vectors

- That is,

$$X(v_1, v_2, \cdots, v_n) = (u_1, u_2, \cdots, u_m)\begin{pmatrix} \sigma_1 & 0 & & 0 \\ 0 & \sigma_2 & & 0 \\ & & \ddots & \\ 0 & 0 & & \sigma_n \\ 0 & 0 & 0 & 0 \\ & \cdots & \cdots & \end{pmatrix}$$

$$= (\sigma_1 u_1, \sigma_2 u_2, \cdots, \sigma_n u_n)$$

- Therefore,

$$Xv_i = \sigma_i u_i, \text{ for } i = 1, \ldots, n.$$

Singular Values

- There must be an r between 0 and n such that

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0.$$

- The (Euclidean) norm of a matrix is given by the largest singular value, i.e.

$$\|X\| = \|X\|_2 = \sigma_1.$$

- If X happens to be a **symmetric, semi-positive definite matrix** then the orthogonal matrices U and V are the same. i.e.

$$U^T X U = \Sigma.$$

Dimension Reduction

- In analysis of systems with large number of components, it is common to reduce dimensions before any statistical analysis.
- The basic idea is to treat observations as elements of a large vector space, and use projections to transform them into smaller vectors.
- In view of their convenience, linear projections have become popular.
- If the original observation space is $\mathbf{R}^D$, then the problem reduces to finding an appropriate projection that takes elements of $\mathbf{R}^D$ to elements of $\mathbf{R}^M$ ($M < D$) in a linear fashion.

Principal Component Analysis

- The principal component analysis (PCA) rotates the coordinate axes to have the new coordinates with un-correlated properties.

- Let x be a vector of n random variables with mean zero and covariance Σ. If the original x has nonzero mean, then use $x - E[x]$ instead to make it *zero mean*.

- Let B be an $D \times D$ orthogonal matrix such that the elements of vector Bx are uncorrelated.

- **Definition 9** The uncorrelated elements of the vector $z = Bx$ are called the principal components of x.

Diagonal Covariance Matrix

- Denote the covariance matrix of z by Λ.
- Then Λ is a diagonal matrix (because all elements in z are uncorrelated).
- By definition, $\Lambda = \mathrm{E}(zz^T) = \mathrm{E}(Bx(Bx)^T) = B\mathrm{E}(xx^T)B^T = B\Sigma B^T$.
- This implies that $\Sigma = B^T \Lambda B$ for a diagonal matrix.
- For a singular value decomposition of Σ ($= U\Lambda U^T$), we can substitute the matrix of singular vectors in place of B^T.
- Hence the principal components of x are given by $z = U^T x$ where $U\Lambda U^T$ is the SVD of the covariance.

Important Properties

- Let the elements of Λ be non-increasing from top-left to bottom-right. i.e. $\boldsymbol{\lambda}_1 \geq \boldsymbol{\lambda}_2 \geq \ldots \geq \boldsymbol{\lambda}_n$.
- Then $\boldsymbol{\lambda}_1$ is the largest singular value of and $z_1 = u_1^T x$ is the first principal component of x. Here, u_1 is the first column of U ($U = (u_1, u_2, \ldots, u_D)$).
- z_1 has the largest variance, namely $\boldsymbol{\lambda}_1$, that can be obtained by projecting x on a unit length vector.
- In general, z_i is uncorrelated to the first $i-1$ principal components, and has the next highest variance.
- u_i's denote the directions of principal variations of x.

Important Properties

- For M significantly less than D, there will be an important speed up in the computational procedures analyzing x.
- To determine such a M, compute the ratio $\sum_{j=1}^{M} \lambda_j \Big/ \sum_{j=1}^{D} \lambda_j$ for increasing M and stop when this ratio exceeds some prescribed cutoff, say 0.95.

Sample Principal Components

- In practice, Σ is estimated using the samples of x, and the principal components are based on estimated valued of Σ.
- Let $x_1, x_2, \ldots, x_n$ be the sampled values of the vector x and

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T.$$

where $\bar{x}$ is the sample mean of x.

- Using SVD, we have $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^T$.
- $\hat{U}^T x$ contains the **sample principal components** of x.
- The observations of sample components are $\hat{U}^T x_i$.

PCA Algorithm

- **Algorithm 20 (PCA of Given Data)** Let X be the $D \times n$ matrix where **each column** denotes an independent observation vector for the random vector x.

  1. Find the sample covariance matrix $C \in \mathbf{R}^{D \times D}$ of the elements of X,

  2. Compute the singular value decomposition (SVD) of C to obtain the orthogonal matrix $U \in \mathbf{R}^{D \times D}$,

  3. Set U_1 to be the first M columns of U, and,

  4. define $Z = U_1^T X \in \mathbf{R}^{M \times n}$.