INDIAN INSTITUTE OF SCIENCE

STOCHASTIC HYDROLOGY

Lecture -31

Course Instructor : Prof. P. P. MUJUMDAR
Department of Civil Engg., IISc.

Summary of the previous lecture

- Design precipitation Hyetographs from IDF relationships.
- Multiple linear regression

PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis

- PCA is a way of identifying patterns in the data; data is expressed in such a way that the similarities and differences are highlighted.

- Once the patterns are found in the data, it can be compressed (reduce the number of dimensions) without losing much information.

Matrix Algebra

Eigenvectors and Eigenvalues:

- Let A be a square matrix. If λ is a scalar and X is a non–zero column vector satisfying

$$AX = \lambda X$$

X is an eigenvector of A; λ is an eigenvalue of A.

- Eigenvectors are possible only for square matrices.
- Eigenvectors of a matrix are orthogonal.

Matrix Algebra

- λ is an eigenvalue of an n x n matrix A, with corresponding eigenvector X.

$(A - \lambda I)X = 0$, with $X \neq 0$ leads to

$$\left|A - \lambda I\right| = 0$$

- There are at most n distinct eigenvalues of A.

Example – 1

Obtain the eigenvalues and eigenvectors for the matrix,

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

The eigenvalues are obtained as

$$\left| A - \lambda I \right| = 0$$

$$\begin{vmatrix} 1-\lambda & 2 \\ 2 & 1-\lambda \end{vmatrix} = 0$$

Example – 1 (Contd.)

$$(1-\lambda)(1-\lambda)-4=0$$

$$\lambda^2-2\lambda-3=0$$

Solving the equation,

$$\lambda=3,-1$$

The eigenvalues are 3 and -1 for matrix A.

The eigenvector is obtained by

$$(A-\lambda I)X=0$$

Example – 1 (Contd.)

For $\lambda_1 = 3$

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

$$\left(A - \lambda_1 I\right) X_1 = 0$$

$$\begin{bmatrix} 1-3 & 2 \\ 2 & 1-3 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$$

$$\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$$

$$-2x_1 + 2y_1 = 0$$

$$2x_1 - 2y_1 = 0$$

Example – 1 (Contd.)

which has solution $x_1 = y_1, \quad x_1$ arbitrary.

eigenvectors corresponding to $\lambda_1 = 3$ are the vectors $\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$, with $x_1 \neq 0$.

e.g., if we take $x_1 = 2$ then $y_1 = 2$

The eigenvector is $\begin{bmatrix} 2 \\ 2 \end{bmatrix}$

Example – 1 (Contd.)

For $\lambda_1 = -1$

$$\left(A - \lambda_2 I\right) X_2 = 0$$

$$\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = 0$$

$$2x_2 + 2y_2 = 0$$

$$2x_2 + 2y_2 = 0$$

which has solution $x_2 = -y_2$, x_2 arbitrary.

eigenvectors corresponding to $\lambda_2 = -1$ are the vectors, with $x_2 \neq 0$. $\begin{bmatrix} x_2 \\ -y_2 \end{bmatrix}$

Principal Component Analysis

Principal Component Analysis (PCA):

- Data on ‘p’ variables; these variables may be correlated.
- Correlation indicates information contained in one variable is also contained in some of the other p-1 variables.
- PCA transforms the ‘p’ original correlated variables into ‘p’ uncorrelated components (also called as orthogonal components or principal components)
- These components are linear functions of the original variables.

Principal Component Analysis

The transformation is written as

$$Z = X\,A$$

where

X is nxp matrix of n observations on p variables

Z is nxp matrix of n values for each of p components

A is pxp matrix of coefficients defining the linear transformation

All X are assumed to be deviations from their respective means, hence X is a matrix of deviations from mean

Principal Component Analysis

Steps for PCA:

- Start with the data for n observations on p variables.
- Form a matrix of size n x p with deviations from mean for each of the variables.
- Calculate the covariance matrix (p x p).
- Calculate the eigenvalues and eigenvectors of the covariance matrix.
- Choose principal components and form a feature vector.
- Derive the new data set.

Principal Component Analysis

The procedure is explained with a simple data set of the yearly rainfall and the yearly runoff of a catchment for 15 years.

Year	1	2	3	4	5	6	7	8	9	10
Rainfall (cm)	105	115	103	94	95	104	120	121	127	79
Runoff (cm)	42	46	26	39	29	33	48	58	45	20

Year	11	12	13	14	15
Rainfall (cm)	133	111	127	108	85
Runoff (cm)	54	37	39	34	25

Mean of Rainfall = 108.5 cm
Mean of Runoff = 38.3 cm

Principal Component Analysis

Step 2: Form a matrix with deviations from mean

Original matrix

$$\begin{bmatrix} 105 & 42 \\ 115 & 46 \\ 103 & 26 \\ 94 & 39 \\ 95 & 29 \\ 104 & 33 \\ 120 & 48 \\ 121 & 58 \\ 127 & 45 \\ 79 & 20 \end{bmatrix}$$

Matrix with deviations from mean

$$X = \begin{bmatrix} -1.3 & 3.4 \\ 8.7 & 7.4 \\ -3.3 & -12.6 \\ -12.3 & 0.4 \\ -11.3 & -9.3 \\ -2.3 & -5.6 \\ 13.7 & 9.4 \\ 14.7 & 19.4 \\ 20.7 & 6.4 \\ -27.3 & -18.6 \end{bmatrix}$$

Principal Component Analysis

Step 3: Calculate the covariance matrix

$$\operatorname{cov}(X,Y) = s_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\begin{bmatrix} \operatorname{cov}(X,X) & \operatorname{cov}(X,Y) \\ \operatorname{cov}(Y,X) & \operatorname{cov}(Y,Y) \end{bmatrix} = \begin{bmatrix} 216.67 & 141.35 \\ 141.35 & 133.38 \end{bmatrix}$$

Principal Component Analysis

Step 4: Calculate the eigenvalues and eigenvectors of the covariance matrix

$$A = \begin{bmatrix} 216.67 & 141.35 \\ 141.35 & 133.38 \end{bmatrix}$$

Eigenvalues:

$$|A - \lambda I| = 0$$

λ_1 = 322.4 and

λ_2 = 27.7

Eigenvectors:

$$(A - \lambda I)X = 0$$

$$X = \begin{bmatrix} 0.801 & -0.599 \\ 0.599 & 0.801 \end{bmatrix}$$

$\sqrt{(0.801)^2 + (0.599)^2} = 1$.

Principal Component Analysis

Step 5: Choose components and form a feature vector

The fraction of the total variance accounted for by the j^{th} principal component is

$$\frac{\lambda_j}{Trace(S)}$$

where

$$Trace(S) = \sum \lambda_j$$

Trace (S) = 322.4 + 27.7 = 350.1

Principal Component Analysis

The total variance accounted for by the first principal component is

$$\frac{\lambda_1}{Trace(S)} = \frac{322.3}{350.1} = 0.92$$

i.e., 92% of total system variance is represented by the first principal component and the remaining 8% is represented by the second component.

Hence the second principal component can be neglected and only the first one considered.

Principal Component Analysis

From the two eigenvectors, the feature vector is selected

$$A = \begin{bmatrix} 0.801 \\ 0.599 \end{bmatrix}$$

Principal Component Analysis

Step 6: Derive the new data set

$$Z = X\,A$$

$$\begin{bmatrix} -1.3 & 3.4 \\ 8.7 & 7.4 \\ -3.3 & -12.6 \\ -12.3 & 0.4 \\ -11.3 & -9.3 \\ -2.3 & -5.6 \\ 13.7 & 9.4 \\ 14.7 & 19.4 \\ 20.7 & 6.4 \\ -27.3 & -18.6 \end{bmatrix} \begin{bmatrix} 0.801 \\ 0.599 \end{bmatrix} = \begin{bmatrix} 0.995 \\ 11.39 \\ -10.2 \\ -9.61 \\ -14.8 \\ -5.20 \\ 16.60 \\ 23.39 \\ 20.41 \\ -33.0 \end{bmatrix}$$

$Y = f(x_1, x_2 \cdots x_p)$

p - principal components

$q < p$

Principal Component Analysis

Using both the eigenvalues, the new data set is

$$\begin{bmatrix} -1.3 & 3.4 \\ 8.7 & 7.4 \\ -3.3 & -12.6 \\ -12.3 & 0.4 \\ -11.3 & -9.3 \\ -2.3 & -5.6 \\ 13.7 & 9.4 \\ 14.7 & 19.4 \\ 20.7 & 6.4 \\ -27.3 & -18.6 \end{bmatrix} \begin{bmatrix} 0.801 & -0.599 \\ 0.599 & 0.801 \end{bmatrix} = \begin{bmatrix} 0.995 & 3.510 \\ 11.39 & 0.716 \\ -10.2 & -8.11 \\ -9.61 & 7.687 \\ -14.8 & -0.92 \\ -5.20 & -3.11 \\ 16.60 & -0.68 \\ 23.39 & 6.732 \\ 20.41 & -7.27 \\ -33.0 & 1.455 \end{bmatrix}$$

Regression on Principal Components

Regression on Principal components:

- In the development of a stochastic model for a dependent variable Y, the first step usually is to do PCA on the independent variables.
- The derived principal components are used as independent variables in a multiple regression analysis with the dependent variable Y.

Regression on Principal Components

Procedure:

- Independent variables are standardized.

$$x_{ij} = \frac{\left(X_{ij} - \bar{x}_j \right)}{s_j}$$

where $X_{i,j}$ is the i^{th} observation on the j^{th} variable, $\bar{x}_j$ and s_j are the mean and standard deviation of the j^{th} variable.

- Dependant variables are centered.

$$y_i = Y_i - \bar{y}$$

where Y_i is the i^{th} observation on y, $\bar{y}$ is the mean of y.

Regression on Principal Components

- The matrix Z is

$$Z = X\,A$$

where

X is nxp matrix of n observations on p variables

Z is nxp matrix of n values for each of p components

A is pxp matrix of coefficients defining the linear transformation

Regression on Principal Components

- The regression model is

$$Y = Z\,\mathrm{B} \quad \text{or} \quad y_i = \sum_{j=1}^{p} \beta_j z_{ij}$$

Where

Y is nx1 vector of ‘n’ observations of the centered dependent variable,

Z is nxp matrix of ‘n’ values for each of p components with z_{ij} elements representing i^{th} value of the j^{th} principal component, and

B is a px1 vector of unknown parameters.

Regression on Principal Components

- The matrix B is estimated as

$$\hat{\mathrm{B}} = \left(Z'Z\right)^{-1} Z'Y$$

Example – 2

The annual yield of a basin is to be obtained from annual rainfall of 10 stations in and around the basin. The annual rainfall in mm for the 10 stations (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9 and x_{10}) and the observed annual yield (Y) in mm for 19years is given.

Obtain the prediction model for calculating annual basin yield (Y) from annual rainfall using PCA.

The annual rainfall for 10 stations and observed basin annual yield (Y).

Year	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1979	1948	4177	5496	2922	5713	3640	3203	2739	2167	2299	3255.2
1980	2261	3670	7797	3327	6934	4424	3692	3451	2866	2653	3682.7
1981	1989	4353	7392	2837	6275	4827	4476	4403	3568	3241	3921.9
1982	1999	3307	7061	3439	6641	4815	4256	4129	3447	3046	3909.3
1983	2086	4230	6564	2987	6675	3959	3900	3559	4078	3583	3768.9
1984	1717	2714	5919	3394	5605	3648	3085	2440	2631	2587	3106.4
1985	1383	2357	5053	2958	5144	3106	4052	3006	3049	2890	3069.4
1986	1470	3004	3951	2691	5116	3557	2775	1909	1952	1723	2940.2
1987	1350	2446	4280	2397	4722	3556	2818	2945	2931	2733	3015.3
1988	1602	4188	5910	3619	6869	5142	3190	3660	3964	3107	3953.2
1989	1417	3631	5145	3282	5226	3793	2663	3017	2579	3367	3172.4
1990	1662	4683	6384	6376	7313	4679	3037	3666	3142	2621	3791.0
1991	1955	4553	5679	6141	6068	3651	2601	2791	2148	2448	3344.8
1992	1974	3836	6021	5646	5876	4026	3037	3920	2583	2742	3650.3
1993	2094	4183	6733	6720	6044	6573	2465	3406	2410	2539	3878.7
1994	3149	6128	8151	9048	8384	7467	2888	3522	2496	2895	4606.2
1995	1471	2952	4151	4975	5149	4733	2603	3493	3396	3554	3498.8
1996	1691	3711	4200	4962	5359	3782	3185	3099	3381	2938	3241.0
1997	2373	4836	6704	6563	6197	5001	3902	3685	3636	3365	4013.5

Example – 2 (Contd.)

A regression equation is obtained using all the 10 stations annual rainfall data is as follows

$$Y_{(19\times1)} = X_{(19\times10)} \; \mathrm{B}_{(10\times1)}$$

$$\hat{\mathrm{B}} = \left(X'X\right)^{-1} X'Y$$

The multiple linear regression equation is as follows.

$$Y = 782.4 + 0.1861\,x_1 + 0.0484\,x_2 - 0.0198\,x_3 + 0.0019\,x_4 + 0.1196\,x_5 + 0.1555\,x_6 + 0.0232\,x_7 + 0.1948\,x_8 + 0.0799\,x_9 - 0.0041\,x_{10}$$

Example – 2 (Contd.)

Using this regression equation,

$$R^2 = 0.988$$

PCA is performed on the data before obtaining the prediction model.

The annual rainfall data is standardized and the observed basin annual yield is centered.

$$x_{ij} = \frac{\left(X_{ij} - \bar{x}_j\right)}{s_j} \qquad y_i = Y_i - \bar{y}$$

Standardized annual rainfall and centered observed basin annual yield (Y).

Year	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1979	0.17	0.36	-0.34	-0.82	-0.39	-0.73	-0.08	-0.95	-1.29	-1.21	-314.2
1980	0.89	-0.18	1.50	-0.60	0.95	-0.02	0.72	0.24	-0.17	-0.45	113.3
1981	0.27	0.55	1.17	-0.87	0.23	0.35	2.01	1.83	0.96	0.82	352.5
1982	0.29	-0.57	0.91	-0.54	0.63	0.34	1.64	1.37	0.77	0.40	339.9
1983	0.49	0.42	0.51	-0.78	0.66	-0.44	1.06	0.42	1.78	1.57	199.5
1984	-0.36	-1.21	-0.01	-0.56	-0.51	-0.73	-0.28	-1.45	-0.54	-0.59	-463.0
1985	-1.13	-1.60	-0.70	-0.80	-1.01	-1.22	1.31	-0.50	0.13	0.07	-500.0
1986	-0.93	-0.90	-1.58	-0.95	-1.04	-0.81	-0.79	-2.33	-1.64	-2.46	-629.2
1987	-1.20	-1.50	-1.32	-1.10	-1.47	-0.81	-0.72	-0.61	-0.06	-0.27	-554.2
1988	-0.62	0.38	-0.01	-0.44	0.87	0.64	-0.11	0.59	1.60	0.53	383.7
1989	-1.05	-0.23	-0.62	-0.63	-0.92	-0.59	-0.97	-0.48	-0.63	1.10	-397.0
1990	-0.49	0.91	0.37	1.05	1.36	0.22	-0.36	0.60	0.28	-0.52	221.6
1991	0.19	0.77	-0.20	0.92	0.00	-0.72	-1.07	-0.86	-1.32	-0.89	-224.6
1992	0.23	0.00	0.08	0.66	-0.21	-0.38	-0.36	1.02	-0.62	-0.26	80.9
1993	0.51	0.37	0.65	1.24	-0.03	1.95	-1.30	0.16	-0.90	-0.69	309.3
1994	2.94	2.47	1.78	2.50	2.53	2.77	-0.60	0.36	-0.76	0.08	1036.7
1995	-0.93	-0.96	-1.42	0.29	-1.01	0.27	-1.07	0.31	0.69	1.50	-70.6
1996	-0.42	-0.14	-1.38	0.28	-0.78	-0.60	-0.11	-0.35	0.66	0.17	-328.4
1997	1.15	1.07	0.62	1.15	0.14	0.51	1.06	0.63	1.07	1.09	444.0

Example – 2 (Contd.)

The eigenvalues and eigenvectors for the standardized data matrix

Eigenvalues									
4.945	2.631	1.047	0.364	0.307	0.257	0.205	0.140	0.063	0.042
Eigenvectors									
0.390	-0.165	0.211	-0.191	0.451	-0.304	0.149	-0.043	-0.644	-0.079
0.381	-0.188	-0.053	-0.543	-0.127	0.215	-0.265	-0.574	0.189	0.157
0.393	0.029	0.382	0.235	0.074	-0.128	-0.328	0.319	0.227	0.600
0.298	-0.321	-0.390	-0.111	0.246	0.400	0.425	0.437	0.210	0.089
0.404	-0.065	0.179	-0.121	-0.589	-0.056	-0.093	0.393	-0.013	-0.522
0.371	-0.161	-0.229	0.546	-0.116	-0.394	0.301	-0.402	0.241	-0.087
0.122	0.462	0.521	-0.117	0.237	0.148	0.428	-0.136	0.393	-0.229
0.317	0.338	-0.122	0.444	0.069	0.603	-0.241	-0.134	-0.333	-0.135
0.136	0.529	-0.237	-0.201	-0.412	-0.110	0.388	0.031	-0.275	0.443
0.160	0.443	-0.477	-0.192	0.351	-0.358	-0.358	0.155	0.235	-0.234

Example – 2 (Contd.)

The eigenvalues and % variance explained is

Eigenvalues	% variance explained
4.945	49.447
2.631	26.310
1.047	10.470
0.364	3.641
0.307	3.069
0.257	2.565
0.205	2.047
0.140	1.399
0.063	0.629
0.042	0.423

> 95% variance explained by first 6 principal components

Example – 2 (Contd.)

First six components are considered in the analysis and the modified data is obtained as

$$Z = X\,A =$$

-1.283	-1.278	1.259	-0.492	0.139	0.045
1.133	0.192	1.776	0.367	-0.068	-0.361
1.825	2.512	0.974	0.410	0.245	0.386
1.273	1.972	0.998	0.826	0.040	0.127
1.176	2.403	0.134	-1.034	-0.232	-0.649
-1.907	-0.547	0.724	0.067	0.087	-0.705
-2.395	1.519	0.673	0.048	0.429	0.160
-3.779	-2.327	1.050	-0.058	-0.483	-0.172
-3.115	0.333	-0.480	0.475	-0.042	-0.194
0.830	1.247	-0.735	0.055	-1.484	-0.236
-1.700	0.094	-1.054	-0.154	0.348	-0.373
1.345	-0.585	-0.305	-0.126	-1.214	1.016
-0.430	-2.241	0.012	-0.817	0.203	0.574
0.243	-0.433	-0.169	0.391	0.593	1.067
1.336	-2.171	-0.751	1.326	0.157	-0.176
5.527	-2.835	-0.199	-0.183	0.169	-0.643
-1.202	0.857	-2.516	0.433	0.246	-0.259
-1.219	0.366	-0.975	-0.743	0.059	0.311
2.341	0.921	-0.416	-0.792	0.808	0.081

Example – 2 (Contd.)

Regression analysis is performed on these components

$$Y = Z\,\mathrm{B}$$

$$\hat{\mathrm{B}} = \left(Z'Z\right)^{-1} Z'Y = \begin{bmatrix} 192.2 \\ 13.3 \\ -33.1 \\ 73.9 \\ -64.1 \\ -15.7 \end{bmatrix}$$

Example – 2 (Contd.)

The regression equation is

$$y = 192.1569\ P_{c1} + 13.29536\ P_{c2} - 33.1304\ P_{c3}$$
$$73.92323\ P_{c4} - 64.0569\ P_{c5} - 15.6921\ P_{c6}$$

$$R^2 = 0.978$$

Regression on Principal Components

Advantages of Regression on Principal components:

- The numerical value for the β's retained in the regression will not be altered by eliminating any number of β's. (this is because of the orthogonal matrix of independent variables)
- Interpretation of β's in terms of the independent variables is simplified.
- The resulting regression coefficients are more stable when applied to a new set of data.

Regression on Principal Components

- Disadvantage is that even some of the principal components are eliminated, all of the original variables must be still measured.
- This disadvantage can be eliminated by examining the factor loadings and eliminating the variables that are not highly correlated with any of the components.

MULTIVARIATE STOCHASTIC MODELS