INDIAN INSTITUTE OF SCIENCE

STOCHASTIC HYDROLOGY

Lecture -32

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

Summary of the previous lecture

- Eigenvalues and eigenvectors
- Principal component analysis

$$Z = X\,A$$

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 independent variables

Z is nxp matrix of transformed data

A is pxp matrix consisting of eigenvectors

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 transformed data of p variables, and

B is a px1 vector of unknown parameters.

Regression on Principal Components

- The matrix B is estimated as

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

p×1   p×n   n×p   p×n   n×1

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{p \times 1}$$

Example – 1

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 19 years is given.

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

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

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 – 1 (Contd.)

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

$$Y_{(19\times 1)} = X_{(19\times 10)} \; \mathrm{B}_{(10\times 1)}$$

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

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 – 1 (Contd.)

Using this regression equation,

$$R^2 = \frac{B'X'Y - n\bar{y}^2}{Y'Y - n\bar{y}^2}$$

$$R^2 = 0.988$$

PCA is performed on the data to reduce the size of the problem and to account for correlations among the rainfall values at10 stations.

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

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

Example – 1 (Contd.)

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

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

Station	Mean	Std.dev.
x_1	1873.2	434.3
x_2	3839.9	927.5
x_3	5925.8	1250.1
x_4	4436.0	1846.3
x_5	6068.9	914.9
x_6	4441.0	1091.3
x_7	3254.0	608.8
x_8	3307.3	599.4
x_9	2969.7	621.6
x_{10}	2859.6	462.4
y	3569	-

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 – 1 (Contd.)

The covariance matrix for the standardized data matrix

$$\operatorname{cov}(X_1, X_2) = s_{X_1,X_2} = \frac{\sum_{i=1}^{n}\left(x_{1_i} - \bar{x}_1\right)\left(x_{2_i} - \bar{x}_2\right)}{n-1}$$

$S =$

1	0.79	0.81	0.64	0.78	0.71	0.18	0.38	-0.03	0.08
0.79	1	0.65	0.72	0.80	0.68	-0.02	0.40	0.03	0.12
0.81	0.65	1	0.38	0.84	0.64	0.44	0.61	0.17	0.19
0.64	0.72	0.38	1	0.55	0.71	-0.35	0.25	-0.15	0.04
0.78	0.80	0.84	0.55	1	0.70	0.21	0.51	0.21	0.13
0.71	0.68	0.64	0.71	0.70	1	-0.10	0.48	0.08	0.18
0.18	-0.02	0.44	-0.35	0.21	-0.10	1	0.52	0.59	0.37
0.38	0.40	0.61	0.25	0.51	0.48	0.52	1	0.64	0.64
-0.03	0.03	0.17	-0.15	0.21	0.08	0.59	0.64	1	0.79
0.08	0.12	0.19	0.04	0.13	0.18	0.37	0.64	0.79	1

10 x 10

Example – 1 (Contd.)

The eigenvalues and eigenvectors for the covariance matrix

Eigenvalues $\left|S - \lambda I\right| = 0$									
4.945	2.631	1.047	0.364	0.307	0.257	0.205	0.140	0.063	0.042
Eigenvectors $\left(S - \lambda I\right)X = 0$									
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 – 1 (Contd.)

The eigenvalues and % variance explained: $\frac{\lambda_j}{Trace(S)}$

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

$Trace(S) = \sum_j \lambda_j$

Example – 1 (Contd.)

First six components are considered in the analysis and the modified data is obtained as $Z = X\,A$

$Z =$

0.17	0.36	-0.34	-0.82	-0.39	-0.73	-0.08	-0.95	-1.29	-1.21
0.89	-0.18	1.50	-0.60	0.95	-0.02	0.72	0.24	-0.17	-0.45
0.27	0.55	1.17	-0.87	0.23	0.35	2.01	1.83	0.96	0.82
0.29	-0.57	0.91	-0.54	0.63	0.34	1.64	1.37	0.77	0.40
0.49	0.42	0.51	-0.78	0.66	-0.44	1.06	0.42	1.78	1.57
-0.36	-1.21	-0.01	-0.56	-0.51	-0.73	-0.28	-1.45	-0.54	-0.59
-1.13	-1.60	-0.70	-0.80	-1.01	-1.22	1.31	-0.50	0.13	0.07
-0.93	-0.90	-1.58	-0.95	-1.04	-0.81	-0.79	-2.33	-1.64	-2.46
-1.20	-1.50	-1.32	-1.10	-1.47	-0.81	-0.72	-0.61	-0.06	-0.27
-0.62	0.38	-0.01	-0.44	0.87	0.64	-0.11	0.59	1.60	0.53
-1.05	-0.23	-0.62	-0.63	-0.92	-0.59	-0.97	-0.48	-0.63	1.10
-0.49	0.91	0.37	1.05	1.36	0.22	-0.36	0.60	0.28	-0.52
0.19	0.77	-0.20	0.92	0.00	-0.72	-1.07	-0.86	-1.32	-0.89
0.23	0.00	0.08	0.66	-0.21	-0.38	-0.36	1.02	-0.62	-0.26
0.51	0.37	0.65	1.24	-0.03	1.95	-1.30	0.16	-0.90	-0.69
2.94	2.47	1.78	2.50	2.53	2.77	-0.60	0.36	-0.76	0.08
-0.93	-0.96	-1.42	0.29	-1.01	0.27	-1.07	0.31	0.69	1.50
-0.42	-0.14	-1.38	0.28	-0.78	-0.60	-0.11	-0.35	0.66	0.17
1.15	1.07	0.62	1.15	0.14	0.51	1.06	0.63	1.07	1.09

19 x 10

0.390	-0.165	0.211	-0.191	0.451	-0.304
0.381	-0.188	-0.053	-0.543	-0.127	0.215
0.393	0.029	0.382	0.235	0.074	-0.128
0.298	-0.321	-0.390	-0.111	0.246	0.400
0.404	-0.065	0.179	-0.121	-0.589	-0.056
0.371	-0.161	-0.229	0.546	-0.116	-0.394
0.122	0.462	0.521	-0.117	0.237	0.148
0.317	0.338	-0.122	0.444	0.069	0.603
0.136	0.529	-0.237	-0.201	-0.412	-0.110
0.160	0.443	-0.477	-0.192	0.351	-0.358

10 x 6

Example – 1 (Contd.)

$$Z = XA = \begin{bmatrix} -1.283 & -1.278 & 1.259 & -0.492 & 0.139 & 0.045 \\ 1.133 & 0.192 & 1.776 & 0.367 & -0.068 & -0.361 \\ 1.825 & 2.512 & 0.974 & 0.410 & 0.245 & 0.386 \\ 1.273 & 1.972 & 0.998 & 0.826 & 0.040 & 0.127 \\ 1.176 & 2.403 & 0.134 & -1.034 & -0.232 & -0.649 \\ -1.907 & -0.547 & 0.724 & 0.067 & 0.087 & -0.705 \\ -2.395 & 1.519 & 0.673 & 0.048 & 0.429 & 0.160 \\ -3.779 & -2.327 & 1.050 & -0.058 & -0.483 & -0.172 \\ -3.115 & 0.333 & -0.480 & 0.475 & -0.042 & -0.194 \\ 0.830 & 1.247 & -0.735 & 0.055 & -1.484 & -0.236 \\ -1.700 & 0.094 & -1.054 & -0.154 & 0.348 & -0.373 \\ 1.345 & -0.585 & -0.305 & -0.126 & -1.214 & 1.016 \\ -0.430 & -2.241 & 0.012 & -0.817 & 0.203 & 0.574 \\ 0.243 & -0.433 & -0.169 & 0.391 & 0.593 & 1.067 \\ 1.336 & -2.171 & -0.751 & 1.326 & 0.157 & -0.176 \\ 5.527 & -2.835 & -0.199 & -0.183 & 0.169 & -0.643 \\ -1.202 & 0.857 & -2.516 & 0.433 & 0.246 & -0.259 \\ -1.219 & 0.366 & -0.975 & -0.743 & 0.059 & 0.311 \\ 2.341 & 0.921 & -0.416 & -0.792 & 0.808 & 0.081 \end{bmatrix}_{19 \times 6}$$

Example – 1 (Contd.)

Regression analysis is performed on these components

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

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

$$Y = \begin{bmatrix} -314.2 \\ 113.3 \\ 352.5 \\ 339.9 \\ 199.5 \\ -463.0 \\ -500.0 \\ -629.2 \\ -554.2 \\ 383.7 \\ -397.0 \\ 221.6 \\ -224.6 \\ 80.9 \\ 309.3 \\ 1036.7 \\ -70.6 \\ -328.4 \\ 444.0 \end{bmatrix}$$

Example – 1 (Contd.)

$$(Z'Z)^{-1} = \begin{bmatrix} 89.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 47.36 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 18.85 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 6.55 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 5.52 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 4.62 \end{bmatrix}_{6 \times 6}$$

$$\hat{B} = (Z'Z)^{-1} Z'Y = \begin{bmatrix} 192.2 \\ 13.3 \\ -33.1 \\ 73.9 \\ -64.1 \\ -15.7 \end{bmatrix}_{6 \times 1}$$

β_1, β_2, ..., β_6

Example – 1 (Contd.)

The regression equation is

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

$$R^2 = 0.978$$

Example – 1 (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

> 85% variance explained by first 3 principal components

Example – 1 (Contd.)

First three components are considered in the analysis and the modified data is obtained as $Z = X\,A$

$$
Z = \begin{bmatrix}
0.17 & 0.36 & -0.34 & -0.82 & -0.39 & -0.73 & -0.08 & -0.95 & -1.29 & -1.21 \\
0.89 & -0.18 & 1.50 & -0.60 & 0.95 & -0.02 & 0.72 & 0.24 & -0.17 & -0.45 \\
0.27 & 0.55 & 1.17 & -0.87 & 0.23 & 0.35 & 2.01 & 1.83 & 0.96 & 0.82 \\
0.29 & -0.57 & 0.91 & -0.54 & 0.63 & 0.34 & 1.64 & 1.37 & 0.77 & 0.40 \\
0.49 & 0.42 & 0.51 & -0.78 & 0.66 & -0.44 & 1.06 & 0.42 & 1.78 & 1.57 \\
-0.36 & -1.21 & -0.01 & -0.56 & -0.51 & -0.73 & -0.28 & -1.45 & -0.54 & -0.59 \\
-1.13 & -1.60 & -0.70 & -0.80 & -1.01 & -1.22 & 1.31 & -0.50 & 0.13 & 0.07 \\
-0.93 & -0.90 & -1.58 & -0.95 & -1.04 & -0.81 & -0.79 & -2.33 & -1.64 & -2.46 \\
-1.20 & -1.50 & -1.32 & -1.10 & -1.47 & -0.81 & -0.72 & -0.61 & -0.06 & -0.27 \\
-0.62 & 0.38 & -0.01 & -0.44 & 0.87 & 0.64 & -0.11 & 0.59 & 1.60 & 0.53 \\
-1.05 & -0.23 & -0.62 & -0.63 & -0.92 & -0.59 & -0.97 & -0.48 & -0.63 & 1.10 \\
-0.49 & 0.91 & 0.37 & 1.05 & 1.36 & 0.22 & -0.36 & 0.60 & 0.28 & -0.52 \\
0.19 & 0.77 & -0.20 & 0.92 & 0.00 & -0.72 & -1.07 & -0.86 & -1.32 & -0.89 \\
0.23 & 0.00 & 0.08 & 0.66 & -0.21 & -0.38 & -0.36 & 1.02 & -0.62 & -0.26 \\
0.51 & 0.37 & 0.65 & 1.24 & -0.03 & 1.95 & -1.30 & 0.16 & -0.90 & -0.69 \\
2.94 & 2.47 & 1.78 & 2.50 & 2.53 & 2.77 & -0.60 & 0.36 & -0.76 & 0.08 \\
-0.93 & -0.96 & -1.42 & 0.29 & -1.01 & 0.27 & -1.07 & 0.31 & 0.69 & 1.50 \\
-0.42 & -0.14 & -1.38 & 0.28 & -0.78 & -0.60 & -0.11 & -0.35 & 0.66 & 0.17 \\
1.15 & 1.07 & 0.62 & 1.15 & 0.14 & 0.51 & 1.06 & 0.63 & 1.07 & 1.09
\end{bmatrix}_{19 \times 10}
\quad
\begin{bmatrix}
0.390 & -0.165 & 0.211 \\
0.381 & -0.188 & -0.053 \\
0.393 & 0.029 & 0.382 \\
0.298 & -0.321 & -0.390 \\
0.404 & -0.065 & 0.179 \\
0.371 & -0.161 & -0.229 \\
0.122 & 0.462 & 0.521 \\
0.317 & 0.338 & -0.122 \\
0.136 & 0.529 & -0.237 \\
0.160 & 0.443 & -0.477
\end{bmatrix}_{10 \times 3}
$$

Example – 1 (Contd.)

$$Z = X\,A = \begin{bmatrix} -1.283 & -1.278 & 1.259 \\ 1.133 & 0.192 & 1.776 \\ 1.825 & 2.512 & 0.974 \\ 1.273 & 1.972 & 0.998 \\ 1.176 & 2.403 & 0.134 \\ -1.907 & -0.547 & 0.724 \\ -2.395 & 1.519 & 0.673 \\ -3.779 & -2.327 & 1.050 \\ -3.115 & 0.333 & -0.480 \\ 0.830 & 1.247 & -0.735 \\ -1.700 & 0.094 & -1.054 \\ 1.345 & -0.585 & -0.305 \\ -0.430 & -2.241 & 0.012 \\ 0.243 & -0.433 & -0.169 \\ 1.336 & -2.171 & -0.751 \\ 5.527 & -2.835 & -0.199 \\ -1.202 & 0.857 & -2.516 \\ -1.219 & 0.366 & -0.975 \\ 2.341 & 0.921 & -0.416 \end{bmatrix}_{19 \times 3}$$

Example – 1 (Contd.)

Regression analysis is performed on these components

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

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

$$Y = \begin{bmatrix} -314.2 \\ 113.3 \\ 352.5 \\ 339.9 \\ 199.5 \\ -463.0 \\ -500.0 \\ -629.2 \\ -554.2 \\ 383.7 \\ -397.0 \\ 221.6 \\ -224.6 \\ 80.9 \\ 309.3 \\ 1036.7 \\ -70.6 \\ -328.4 \\ 444.0 \end{bmatrix} \qquad Z = \begin{bmatrix} -1.283 & -1.278 & 1.259 \\ 1.133 & 0.192 & 1.776 \\ 1.825 & 2.512 & 0.974 \\ 1.273 & 1.972 & 0.998 \\ 1.176 & 2.403 & 0.134 \\ -1.907 & -0.547 & 0.724 \\ -2.395 & 1.519 & 0.673 \\ -3.779 & -2.327 & 1.050 \\ -3.115 & 0.333 & -0.480 \\ 0.830 & 1.247 & -0.735 \\ -1.700 & 0.094 & -1.054 \\ 1.345 & -0.585 & -0.305 \\ -0.430 & -2.241 & 0.012 \\ 0.243 & -0.433 & -0.169 \\ 1.336 & -2.171 & -0.751 \\ 5.527 & -2.835 & -0.199 \\ -1.202 & 0.857 & -2.516 \\ -1.219 & 0.366 & -0.975 \\ 2.341 & 0.921 & -0.416 \end{bmatrix}$$

Example – 1 (Contd.)

Considering first three components

$$(Z'Z)^{-1} = \begin{bmatrix} 89.01 & 0.00 & 0.00 \\ 0.00 & 47.36 & 0.00 \\ 0.00 & 0.00 & 18.85 \end{bmatrix}$$

Considering first six components

$$\begin{bmatrix} 89.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 47.36 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 18.85 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 6.55 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 5.52 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 4.62 \end{bmatrix}$$

$$\hat{B} = (Z'Z)^{-1} Z'Y = \begin{bmatrix} 192.2 \\ 13.3 \\ -33.1 \end{bmatrix}$$

$$\begin{bmatrix} 192.2 \\ 13.3 \\ -33.1 \\ 73.9 \\ -64.1 \\ -15.7 \end{bmatrix}$$

Example – 1 (Contd.)

The regression equation is

$$y = 192.1569\ P_{c1} + 13.29536\ P_{c2} - 33.1304\ P_{c3}$$

$$R^2 = 0.961$$

Regression on Principal Components

- The numerical value for the β's retained in the regression will not be altered by reducing the size.
- 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.
- Disadvantage is that even if some of the principal components are eliminated, all of the original variables must be still measured.

MULTIVARIATE STOCHASTIC MODELS

Multivariate Stochastic models

- Stochastic models discussed for single site in relation with the auto correlations and auto covariance.
  - Thomas Fiering models
    - Stationary and non-stationary models
  - ARMA models
    - Box Jenkins models

Multivariate Stochastic models

First order Markov process:

Random component

$$X_{t+1} = \underbrace{\mu_x + \rho_1 (X_t - \mu_x)}_{\text{Deterministic component}} + \varepsilon_{t+1}$$

$\varepsilon \sim$ Mean 0 and variance σ_ε^2

$$X_{t+1} = \mu_x + \rho_1 \left(X_t - \mu_x \right) + u_{t+1} \sigma_x \sqrt{1 - \rho_1^2}$$

Multivariate Stochastic models

First order Markov model with non-stationarity, for stream flow generation:

$$X_{i,j+1} = \mu_{j+1} + \rho_j \frac{\sigma_{j+1}}{\sigma_j}\left(X_{ij} - \mu_j\right) + t_{i,j+1}\sigma_{j+1}\sqrt{1-\rho_j^2}$$

ρ_j is serial correlation between flows of j^{th} month and $j+1^{th}$ month.

$t_{i,\, j+1} \sim N(0, 1)$

ARIMA Models

ARMA (p, q)

Residuals of order 'q'

$$X_t = \underbrace{\phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p}}_{\text{AR of order 'p'}} + \overbrace{\theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q}}^{\text{Residuals of order 'q'}} + e_t$$

AR of order 'p'

$\{e_t\}$ is the residual series

Assumptions : $\{e_t\}$ has zero mean with uncorrelated terms

$$\hat{X}_{t+1} = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j e_{t-j}$$

Multivariate Stochastic models

- Data generation (or forecasting) on a random variate depending on two or more sites is usually required.
  - For example, in the design of a reservoir, the flow from all the streams fed to the reservoir must be considered.
- If the time series for the random variables are independent, then the generation techniques for the single site can be used.
- Important to consider the simultaneous behavior of the random variables.

Multivariate Stochastic models

- Correlation of a random variable between two sites is cross-correlation.
- Lag zero cross-correlation is the correlation of a random variable at two points in the same time period.
- Lag k cross-correlation is the correlation between one random variable at one point and the random variable k time points later at other point.
- Denoted by $r_{j,h}\ (k)$

Multivariate Stochastic models

$$r_{i,h}\left(k\right)=\frac{\sum_{i=1}^{n}\left(x_{j,i}-\overline{x}_{j}\right)\left(x_{h,i+k}-\overline{x}_{h}\right)}{\left(n-k\right)s_{x,j}s_{x,h}}$$

where

n is the total number of pairs of observations on X_j and X_k,

$X_{j,i}$ is the i^{th} observation on X_j

$\overline{x}_j, s_{x,j}$ are the mean and standard deviation of the observations on X_j

Multivariate Stochastic models

Multisite Markov model (Two sites):

- Fiering (1964) presented a two site generation model that preserves means, variances, skewness, lag one serial correlation and lag zero cross-correlations.
- One site is to be selected as key site.
- Selection may be based on the length of the data and the quality of the record.
- Consider j is the key site and h is a subordinate site to key site j.
- A sequence of observations are generated for site j using single site generation technique.

Ref.: Fiering, M.B. (1964) Multivariate technique for synthetic hydrology, Proceedings American Society of Civil Engineers 90(HY5):43-60

Multivariate Stochastic models

- A cross-correlation model is used to generate values of site h based on generated values at site j.

$$X_{h,i} = \bar{x}_h + r_{j,h}(0)\frac{s_h}{s_j}\left(X_{ij} - \bar{x}_j\right) + u_i s_h \sqrt{1 - r_{j,h}^2(0)}$$

where

u_i is a standardized random variate adjusted to incorporate the serial correlation at site h.

$$u_i = \zeta \frac{\left(X_{h,i-1} - \bar{x}_h\right)}{s_h} + t_i \sqrt{1 - \zeta^2}$$

Multivariate Stochastic models

t_i is a standardized random variate adjusted for skewness.

$$\zeta = \frac{r_h(1) - r_j(1)\, r_{j,h}^2(0)}{\sqrt{1 - r_{j,h}^2(0)}}$$

Multivariate Stochastic models

Multisite Markov model:

- Multisite generation requires simultaneous generation of data at several sites while preserving the correlation between the data at various sites.
- Consider $x_{j,i}$ is a standardized value of the data at site j during the period i

$$x_{j,i} = \frac{\left(X_{j,i} - \overline{x}_j\right)}{s_j}$$

Multivariate Stochastic models

- The first order Markov model for site h is

$$x_{h,i+1} = \rho_h\left(1\right)x_{h,i} + \varepsilon_{h,i+1}\sqrt{1-\rho_h^2\left(1\right)}$$

- The first order Markov model for site j is

$$x_{j,i+1} = \rho_j\left(1\right)x_{j,i} + \varepsilon_{j,i+1}\sqrt{1-\rho_j^2\left(1\right)}$$

- The equations are written in matrix form

Multivariate Stochastic models

$$X_{i+1} = EX_i + G\mathrm{E}$$

where

X_i is a p x 1 vector if standardized values of the variable generated at time i,

E is a p x p diagonal matrix whose jth diagonal element is $\rho_j(1)$,

G is a p x p diagonal matrix whose jth diagonal element is $\sqrt{1-\rho_j^2(1)}$

E is a p x 1 vector of random elements

Multivariate Stochastic models

- E is defined to preserve the first order serial correlation of the x_j ‘s and the lag zero cross-correlation between x_j and x_h .
- E is made of elements that are $\varepsilon_{j,i+1}$; each $\varepsilon_{j,i+1}$ is independent of $x_{j,i}$; ε_j is N(0,1)
- The correlation between ε_j and ε_h is $\rho^*_{j,h}(0)$,

$$\rho^*_{j,h}\left(0\right) = \frac{\left\{1-\rho_j\left(1\right)\rho_h\left(1\right)\right\}\rho_{j,h}\left(0\right)}{\sqrt{\left\{1-\rho_j^2\left(1\right)\right\}\left\{1-\rho_h^2\left(1\right)\right\}}}$$

Multivariate Stochastic models

- Matalas (1967) given a multisite normal generation model that preserves the means, variances, lag one serial correlation, lag one cross-correlations and lag zero cross-correlations.

$$X_{i+1} = AX_i + G\mathrm{E}_{i+1}$$

where

X_i is a m x 1 vector whose p^th^ element is $X_i^p - \mu_X^p$ with X_i^p the i^th^ value of X at station p and μ_X^p the mean value at station p.

Ref.: Matalas, N.C. (1967) Mathematical assessment of synthetic hydrology, Water Resources Research 3(4):937-945

Multivariate Stochastic models

E_{i+1} is a form of N(0,1) with E_{i+1} independent of X_i.

The m x m matrices A and B are defined in terms of two new m x m matrices M_1 and M_0.