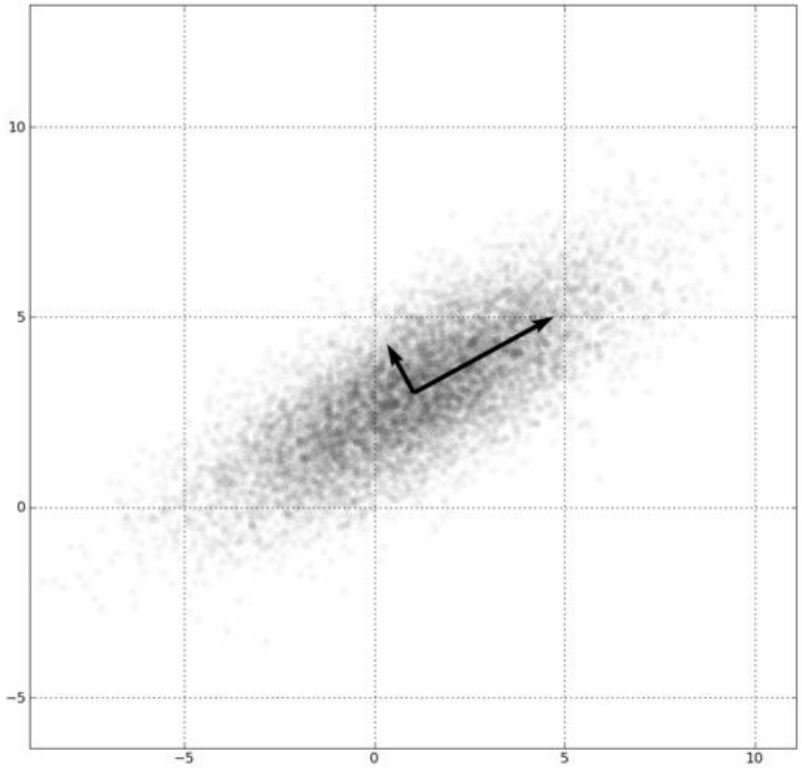
Introduction to Applied Multivariate Analysis with R

Principal Components Analysis

Principal Components Analysis

Basic goal is to describe variation in a set of **correlated** variables in terms of a new set of **uncorrelated** variables.

- ***Orthogonal transformation*** converts set of possibly correlated variables into set of linearly uncorrelated values called **principal components**.
- First principal component has ***largest possible variance*** and each succeeding principal component has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components.



Principal Components Analysis

Basic goal is to describe variation in a set of **correlated** variables in terms of a new set of **uncorrelated** variables.

The basic goal of principal components analysis is to describe variation in a set of correlated variables, $\mathbf{x}^\top = (x_1, \ldots, x_q)$, in terms of a new set of uncorrelated variables, $\mathbf{y}^\top = (y_1, \ldots, y_q)$, each of which is a linear combination of the $\mathbf{x}$ variables. The new variables are derived in decreasing order of "importance" in the sense that y_1 accounts for as much as possible of the variation in the original data amongst all linear combinations of $\mathbf{x}$. Then y_2 is chosen to account for as much as possible of the remaining variation, subject to being uncorrelated with y_1, and so on. The new variables defined by this process, $y_1, \ldots, y_q$, are the principal components.

Principal Components Analysis

Basic goal is to describe variation in a set of **correlated** variables in terms of a new set of **uncorrelated** variables.

The principal components are most commonly (and properly) used as a means of constructing an informative graphical representation of the data (see later in the chapter) or as input to some other analysis. One example of the latter is provided by regression analysis; principal components may be useful here when:

- There are too many explanatory variables relative to the number of observations.
- The explanatory variables are highly correlated.

Both situations lead to problems when applying regression techniques, problems that may be overcome by replacing the original explanatory variables with the first few principal component variables derived from them.

Rescaling the Principal Components

Coefficients defining the principal components are often rescaled.

The coefficients defining the principal components derived as described in the previous section are often rescaled so that they are correlations or covariances between the original variables and the derived components. The rescaled coefficients are often useful in interpreting a principal components analysis. The covariance of variable i with component j is given by

$$\mathsf{Cov}(x_i, y_j) = \lambda_j a_{ji}.$$

Rescaling the Principal Components

Rescaled coefficients from **principal components analysis** of a correlation matrix are analogous to ***factor loadings***.

The correlation of variable x_i with component y_j is therefore

$$r_{x_i,y_j} = \frac{\lambda_j a_{ji}}{\sqrt{\mathsf{Var}(x_i)\mathsf{Var}(y_j)}} = \frac{\lambda_j a_{ji}}{s_i\sqrt{\lambda_j}} = \frac{a_{ji}\sqrt{\lambda_j}}{s_i}.$$

If the components are extracted from the correlation matrix rather than the covariance matrix, the correlation between variable and component becomes

$$r_{x_i,y_j} = a_{ji}\sqrt{\lambda_j}$$

Example: `headsize` data

Table 3.1: `headsize` data. Head Size Data.

head1	breadth1	head2	breadth2	head1	breadth1	head2	breadth2
191	155	179	145	190	159	195	157
195	149	201	152	188	151	187	158
181	148	185	149	163	137	161	130
183	153	188	149	195	155	183	158
176	144	171	142	186	153	173	148
208	157	192	152	181	145	182	146
189	150	190	149	175	140	165	137
197	159	189	152	192	154	185	152
188	152	197	159	174	143	178	147
192	150	187	151	176	139	176	143
179	158	186	148	197	167	200	158
183	147	174	147	190	163	187	150
174	150	185	152				

Example: `heptathlon` data

Table 3.2: `heptathlon` data. Results of Olympic heptathlon, Seoul, 1988.

	hurdles	highjump	shot	run200m	longjump	javelin	run800m	score
Joyner-Kersee (USA)	12.69	1.86	15.80	22.56	7.27	45.66	128.51	7291
John (GDR)	12.85	1.80	16.23	23.65	6.71	42.56	126.12	6897
Behmer (GDR)	13.20	1.83	14.20	23.10	6.68	44.54	124.20	6858
Sablovskaite (URS)	13.61	1.80	15.23	23.92	6.25	42.78	132.24	6540
Choubenkova (URS)	13.51	1.74	14.76	23.93	6.32	47.46	127.90	6540
Schulz (GDR)	13.75	1.83	13.50	24.65	6.33	42.82	125.79	6411
Fleming (AUS)	13.38	1.80	12.88	23.59	6.37	40.28	132.54	6351
Greiner (USA)	13.55	1.80	14.13	24.48	6.47	38.00	133.65	6297
Lajbnerova (CZE)	13.63	1.83	14.28	24.86	6.11	42.20	136.05	6252
Bouraga (URS)	13.25	1.77	12.62	23.59	6.28	39.06	134.74	6252
Wijnsma (HOL)	13.75	1.86	13.01	25.03	6.34	37.86	131.49	6205
Dimitrova (BUL)	13.24	1.80	12.88	23.59	6.37	40.28	132.54	6171
Scheider (SWI)	13.85	1.86	11.58	24.87	6.05	47.50	134.93	6137
Braun (FRG)	13.71	1.83	13.16	24.78	6.12	44.58	142.82	6109
Ruotsalainen (FIN)	13.79	1.80	12.32	24.61	6.08	45.44	137.06	6101
Yuping (CHN)	13.93	1.86	14.21	25.00	6.40	38.60	146.67	6087
Hagger (GB)	13.47	1.80	12.75	25.47	6.34	35.76	138.48	5975
Brown (USA)	14.07	1.83	12.69	24.83	6.13	44.34	146.43	5972
Mulliner (GB)	14.39	1.71	12.68	24.92	6.10	37.76	138.02	5746
Hautenauve (BEL)	14.04	1.77	11.81	25.61	5.99	35.68	133.90	5734
Kytola (FIN)	14.31	1.77	11.66	25.69	5.75	39.48	133.35	5686
Geremias (BRA)	14.23	1.71	12.95	25.50	5.50	39.64	144.02	5508
Hui-Ing (TAI)	14.85	1.68	10.00	25.23	5.47	39.14	137.30	5290
Jeong-Mi (KOR)	14.53	1.71	10.83	26.61	5.50	39.26	139.17	5289
Launa (PNG)	16.42	1.50	11.78	26.16	4.88	46.38	163.43	4566

Example: `USairpollution` data

`USairpollution` data (partial). Air pollution in 41 US cities:

	SO2	temp	manu	popul	wind	precip	predays
Albany	46	47.6	44	116	8.8	33.36	135
Albuquerque	11	56.8	46	244	8.9	7.77	58
Atlanta	24	61.5	368	497	9.1	48.34	115
Baltimore	47	55.0	625	905	9.6	41.31	111
Buffalo	11	47.1	391	463	12.4	36.11	166
Charleston	31	55.2	35	71	6.5	40.75	148
Chicago	110	50.6	3344	3369	10.4	34.44	122
Cincinnati	23	54.0	462	453	7.1	39.04	132
Cleveland	65	49.7	1007	751	10.9	34.99	155
Columbus	26	51.5	266	540	8.6	37.01	134

Canonical Correlation Analysis

- **Canonical correlation analysis** is a way of making sense of **cross-covariance matrices**.
  - ***cross-covariance*** is **cov(*X*, *Y*)** between two random vectors ***X*** and ***Y***, to distinguish from the "covariance" of a random vector ***X***, which is the matrix of covariances between the scalar components of ***X***.
- If we have two sets of variables, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$, and there are correlations among the variables, then canonical correlation analysis will enable us to find linear combinations of the ***x***'s and the ***y***'s which have maximum correlation with each other.