smfd6.tex

Day 6. 5.11.2014.

As always, n may be large – the larger the better, as large samples are more informative than small ones. The size of p varies with the problem. But typically p might be of the order of 10 or 12, say. A 12-dimensional 'variable space' is unwieldy for many purposes, and we might want a lower-dimensional representation of the data, with as little loss of information as possible. Background: [MKB] Ch. 1, [K] Ch. 1.

Notation.

$$X = (x_{(1)}, \dots, x_{(p)}) = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}.$$

So the *column-vectors* x_i, $x_{(j)}$ relate to the ith *object* and the jth *variable*.

Mean vector. $\overline{x}_i := \frac{1}{n}\sum_{r=1}^n x_{ri}$ is the sample mean of the ith variable; the *sample mean vector* is

$$\overline{x} := \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{pmatrix}.$$

The sample variance s_{ij} between the ith and jth variables is

$$s_{ij} := \frac{1}{n}\sum_{r=1}^n (x_{ri} - \overline{x}_i)(x_{rj} - \overline{x}_j) = \frac{1}{n}\sum_{r=1}^n x_{ri}x_{rj} - \overline{x}_i\overline{x}_j.$$

Form these into a matrix, the *sample covariance matrix* $S := (s_{ij})$:

$$S = \frac{1}{n}\sum_{r=1}^n (x_r - \overline{x})(x_r - \overline{x})^T = \frac{1}{n}\sum_{r=1}^n x_r x_r^T - \overline{x}\ \overline{x}^T.$$

Now $X^T = (x_1, \dots, x_n)$ (row of columns corresponding to *objects*), so

$$XX^T = (x_1, \dots, x_n)\begin{pmatrix} x_1^T \\ \vdots \\ x_n \end{pmatrix} = \sum x_r x_r^T.$$

Write $\mathbf{1}$ for a column-vector of n 1s. Then (check) $\mathbf{1}\mathbf{1}^T$ is the $n \times n$ matrix with each element 1, and (check) $X^T\mathbf{1}\mathbf{1}^T X = n^2\overline{x}\ \overline{x}^T$. So

$$S = \frac{1}{n}X^T X - \frac{1}{n^2}X^T\mathbf{1}\mathbf{1}^T X = \frac{1}{n}X^T H X, \quad \text{where} \quad H := I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$$

is the $n \times n$ *centring matrix*. We call $M := X^T X = \sum_1^n x_r x_r^T$ the *matrix of sums of squares and products*.

Since s_{ii} is the sample variance of the ith variable, $s_i := \sqrt{s_{ii}}$ is its sample SD. Form the *sample correlation matrix* $R := (r_{ij})$, where

$$r_{ij} := s_{ij}/s_i s_j$$

is the sample correlation coefficient between the ith and jth variables (so $|r_{ij}| \leq 1$). If

$$D := diag(s_i) = diag(\sqrt{s_{ii}}),$$

$$R = D^{-1} S D^{-1}, \qquad S = DRD.$$

One can check:
(i) H is symmetric and idempotent (i.e. $H^2 = H$);
(ii) S is symmetric and non-negative definite;
(iii) R is symmetric and non-negative definite.
Scaling.

If our data is subjected to an affine transformation (change of location and scale) $x \mapsto y := Ax + b$, then (check) $\overline{y} = A\overline{x} + b$, and $S_y = AS_x A^T$. In particular, if

$$y_r := D^{-1}(x_r - \overline{x}) \qquad (*)$$

then Y has mean vector 0 and covariance matrix $D^{-1}S(D^{-1})^T = D^{-1}SD^{-1} = R$, the correlation matrix of X. So the affine transformation $(*)$ *scales* the data X to new data Y, with *zero means* and *unit variances* (1s on the diagonal of S_y – and correlations = covariances r_{ij} of modulus ≤ 1 off the diagonal). This eliminates dependence of the data on arbitrary choices of location and scale in the units, and makes the data *dimensionless*.
Mahalanobis transformation.

Recall that S is non-negative definite, and is positive definite in the typical, or generic, case. Then S^{-1} exists, and hence so do $S^{\pm 1/2}$. If

$$z_r := S^{-1/2}(x_r - \overline{x}) \quad (r = 1, \ldots, n), \qquad (**)$$

then Z has mean vector 0 and covariance matrix $S^{-1/2} S S^{-1/2} = I$. The map $X \mapsto Z$ is the *Mahalonobis transformation*, which not only centres (means to 0) and scales (variances to 1) as above, but also makes the variables *uncorrelated*.
Principal component transformation.

By the Spectral Decomposition Theorem, we can write $S = GLG^T$, where G is an orthogonal matrix and L is a diagonal matrix of eigenvalues of S. Since S is non-negative definite, its eigenvalues ℓ_i are non-negative, and w.l.o.g. we can re-order the variables so that they decrease in size:

$$\ell_1 \geq \ell_2 \geq \ldots \geq \ell_p \geq 0.$$

The *principal component transformation*

$$y_r := G^T(x_r - \overline{x}) \quad (r = 1, \ldots, n) \qquad (***)$$

takes data X to new data Y, with zero mean and covariance matrix $S_y = G^T S_x G = G^T G L G^T G = L$, as G is orthogonal: $S_y = L$ is diagonal. So the y_r are *uncorrelated* linear combinations of the data, called *principal components*.
R-techniques and Q-techniques.

Multivariate Analysis splits into two broad areas. In the first, we are interested in the p *variables*, that is, in the p *columns* of our data matrix. Methods used here are called *R-techniques*, because they depend on the correlation matrix R. In the second, we are interested in the n *objects*, that is, in the n *rows* of our data matrix. Methods used here are called *Q-techniques*, because they deal directly with the source data (Quelle = source, German).
R-techniques include:
principal components analysis (PCA) [MKB Ch. 8, K 2.3];
factor analysis [MKB Ch. 9, K 16.2];
canonical correlation analysis [MKB Ch. 10, K 14.5].
Q-techniques include:
discriminant analysis [MKB Ch. 11, K 12.3];
cluster analysis [MKB Ch. 13, K 3.1, 9.4];
multidimensional scaling [MKB Ch. 14, K 3.2, 3.3, 9.3].

4. Sample and Population

To describe the population in the p-dimensional case, we need a *population mean (vector)* and a *population covariance (matrix)*:

$$\mu := Ex; \qquad \Sigma := var\ x = E[(x - \mu)(x - \mu)^T].$$

Then (check)

$$E[\overline{x}] = \mu, \qquad var(\overline{x}) = \frac{1}{n}\Sigma, \qquad E[S] = \frac{n-1}{n}.\Sigma.$$

The *unbiased sample covariance matrix* is

$$S_u := \frac{n}{n-1} S;$$

then $E[S_u] = \Sigma$, so S_u is unbiased as an estimator for Σ (as in one dim.).
Objectives.
R-techniques. Here we are interested in the p *variables* (columns of X). If $p = 2$ we can use plots in two dimensions (paper, whiteboard, computer screen); if $p = 3$, we can use our 3-dimensional geometric intuition, and then use computer graphics (based on projective geometry) to represent 3-dimensional reality in 2 dimensions. But if p is 10 or 12, say, it is hard to visualise the data in 10 or 12 dimensions, and so we seek some *lower-dimensional representation* of the data. This will entail some loss of information, which we seek to minimise. We also seek a *parsimonious summarisation* of the data (Principle of Parsimony; Occam's Razor; Einstein's Dictum). One useful technique here is PCA (below). Another is *projection pursuit*.
Q-techniques. Here we are interested in the *objects*. We might want to
(i) represent them as points in space, with closeness corresponding to similarity (multidimensional scaling);
(ii) subdivide or classify into types (cluster analysis);
(iii) assign objects to types (e.g. two types – *discriminant analysis*).
Exploratory Data Analysis (EDA).

As in one dimension, one should begin by 'getting to know the data' by examining it visually. Check for unusual readings (which may be errors – or may be valid and highly informative!), or *outliers*, and decide what to do about any missing readings (e.g. fill in from existing readings – 'imputation').

5. Principal Components Analysis (PCA)

PCA is due to Harold Hotelling (1895-1978) in 1933, following Karl Pearson (1857-1936) in 1901.

We met PCA above in its sample form (see $(***)$); we now turn to the population counterpart of this. We take a random p-vector x, with mean μ and covariance matrix Σ (no distributional assumptions yet). By spectral decomposition of Σ,

$$\Sigma = \Gamma \Lambda \Gamma^T, \qquad \Lambda = \Gamma^T \Sigma \Gamma \qquad (\Sigma = \sum_1^p \lambda_i \gamma_i \gamma_i^T),$$

with $\Lambda = diag(\lambda_i)$, $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ the e-values of Σ, w.l.o.g. in decreasing order, $\Gamma = (\gamma_1, \ldots, \gamma_p)$ the orthogonal matrix of eigenvectors. Write

$$y := \Gamma^T(x - \mu): \qquad y_i = \gamma_i^T(x - \mu),$$

is called the ith *principal component* of x. Then (check)

$$Ey = 0, \qquad var\ y = \Lambda,$$

a diagonal matrix, so the y_i are uncorrelated. Also the $var\ y_i = \lambda_i$ are in decreasing order; their sum and product are the trace and determinant of Σ. *Definition.* A linear combination $a^T x = \sum_1^p a_i x_i$ of x is a *standardised linear combination* (SLC) if $\sum_1^p a_i^2 = 1$ (i.e. $a^T a = 1$).

Theorem. The first principal component

$$y_1 = \gamma_1^T(x - \mu)$$

is the SLC of x with the largest variance, λ_1.

Proof. Since $\gamma_i^T \gamma_i = 1$ (the eigenvectors are normalised to have length 1), y_1 is a SLC, and has variance λ_1 by above. If $\alpha := a^T x$ is any other SLC, write

$$a = c_1\gamma_1 + \ldots + c_p\gamma_p$$

(any p-vector can be written like this, as the columns γ_i are linearly independent, so form a basis). Then

$$var\ \alpha = var(a^T a) = a^T \Sigma a = (\sum_i c_i \gamma_i^T)(\sum_j \lambda_j \gamma_j \gamma_j^T)(\sum_k c_k \gamma_k)$$

$$= \sum_{ijk} c_i \lambda_j c_k \gamma_i^T \gamma_j \gamma_j^T \gamma_k = \sum_{ijk} c_i \lambda_j c_k \delta_{ij} \delta_{jk} = \sum_1^p \lambda_i c_i^2.$$

But $\sum c_i^2 = 1$ and $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$, so $var\ \alpha = \sum \lambda_i c_i^2$ is maximised for $c_1 = 1$, $c_i = 0$ for $i = 2, \ldots, p$, when $a = \gamma_1$, and its maximum value is λ_1. //

Note. This choice of $a^T x = \gamma_1^T x$ differs from the first principal component $y_1 = \gamma_1^T(x - \mu)$ only by a constant $\gamma_1^T \mu$, so has the same variance.

Theorem. For each $k = 0, 1, \ldots, p-1$, if $\lambda_k > 0$ the $(k+1)$th principal component

$$y_{k+1} = \gamma_{k+1}^T (x - \mu)$$

is the SLC of x with largest variance uncorrelated with the first k principal components, and this variance is λ_{k+1}.

Proof. If the SLC is $a^T x$ as above, then in the notation above

$$\begin{aligned} cov(a^T x, y_k) &= cov(a^T x, \gamma_k^T (x - \mu)) \\ &= E[(a^T x - E(a^T x)).\gamma_k^T (x - \mu)] \\ &= E[a^T (x - \mu)(x - \mu)^T \gamma_k] \quad (\gamma_k^T (x - \mu) \text{ a scalar, so its own transpose}) \\ &= a^T \Sigma a \qquad (E[(x - \mu)(x - \mu)^T] = \Sigma) \\ &= \sum_1^p c_i \gamma_i \Sigma \gamma_k = \sum_1^p c_i (\Gamma^T \Sigma \Gamma)_{ik}, \end{aligned}$$

which is $\sum c_i \lambda_{ik}$ by spectral decomposition, or $\sum c_i \lambda_i \delta_{ik}$ as Λ is diagonal, which is $c_k \lambda_k$. This is 0 if $a^T x$ is uncorrelated with y_k, but by assumption, $\lambda_k > 0$ (and so $\lambda_1 \geq \ldots \geq \lambda_k > 0$). So $c_k = 0$. Similarly, $c_1 = \ldots = c_{k-1} = 0$. So $a = \sum_{k+1}^p c_i \gamma_i$. As before, $var(a^T x) = \sum_{k+1}^p \lambda_i c_i^2$; as the λ_i are decreasing this is maximised for $c_{k+1} = 1$ and the rest 0, with maximum λ_{k+1}. //

Interpretation. We think of

$$\sum_1^p var\ y_i = \sum_1^p \lambda_i = trace(\Lambda) = trace(\Sigma)$$

as the 'total variability' in the distribution, and $var\ y_1 = \lambda_1$ the 'contribution' of the 1st principal component y_1 to 'explaining' this variability, $var\ y_2 = \lambda_2$ the contribution of y_2, etc. So $\lambda_i / (\lambda_1 + \ldots + \lambda_p)$ is the *proportion* of the total variability explained by the ith principal component, and $(\lambda_1 + \ldots + \lambda_i)/(\lambda_1 + \ldots + \lambda_p)$ is the proportion of the variability explained by the first k PCs. So: if Σ has rank $k < p$, *all* the variability is explained by the first k PCs (the remaining eigenvalues are 0).

How many components to retain?

If we retain k components, there is a trade-off between k large (to explain more variability) and k small (to give a parsimonious representation). We should choose k bearing in mind the *purpose* of our study.

To assist in choice of k, a diagram is often drawn. Plot the points (k, λ_k), or equivalently $(k, \lambda_k/(\sum \lambda_i))$, and join adjacent points by straight-line segments.As the λ_i decrease, the resulting 'broken line' (continuous piecewise-linear function) decreases. We hope to see it decrease steeply at first, then more slowly, then level off. By analogy with mountain-sides, typically with
(i) the steepest, rocky or cliff, part at the top, then
(ii) a less steep, scree slope in the middle, then
(iii) a gently sloping grassy part below –
such a diagram is called a *scree diagram* (R. B. Cattell (1905-1998) in 1966). Generally we will retain components until somewhere on the scree slope – where depending on how we value parsimony v. accuracy. We may look for an 'elbow', where the gradient flattens out.

Sample principal components

Return to our data matrix X. Let a be a unit p-vector. Then

$$Xa = \begin{pmatrix} x_1^T a \\ \vdots \\ x_n^T a \end{pmatrix}$$

gives n observations of a new variable $x^T a$. The sample variance is $a^T S a$, where S is the sample variance matrix of X; we look for SLCs with maximum variance. Let

$$S = GLG^T$$

be the spectral decomposition of S, $L = diag(l_i)$, with $l_1 \geq \ldots \geq l_p \geq 0$ the e-values of S, $G = (g_1, \ldots, g_p)$ the orthogonal matrix of e-vectors:

$$y_r := G^T(x_r - \overline{x}) \qquad (r = 1, \ldots, n)$$

takes the data matrix X to Y, with mean 0 and covariance matrix L, which is diagonal, so the y_r are *uncorrelated*. Now (check)

$$Y = (X - \mathbf{1}\overline{x}^T)G = (X - \mathbf{1}\overline{x}^T)(g_1, \ldots, g_p), \qquad y_{(k)} = (X - \mathbf{1}\overline{x}^T)g_k$$

gives the SLC of maximal variance, l_k, uncorrelated with $y_{(1)}, \ldots, y_{(k-1)}$. Taking the rth row,

$$y_{rk} = (x_r^T - \overline{x}^T)g_k = g_k^T(x_r - \overline{x}).$$

If the subscript r is unimportant, we can drop it: $y_i = g_i^T(x - \overline{x})$.

Example: Examination scores ([MKB], 1.2.3, Table 1.2.1). This gives data on 88 students' scores on each of 5 Mathematics exams (Mechanics, Vectors, Algebra, Analysis, Statistics); the first two are closed book (C), the last three open book (O). So here $n = 88, p = 5$. The eigenvalues of S are

$$l_1 = 679.2, \quad l_2 = 199.8, \quad l_3 = 102.6, \quad l_4 = 83.7, \quad l_5 = 31.8.$$

The five principal components are found.
1. y_1 gives positive (and comparable) weighting to all 5 marks. This is thus a *weighted average* of the marks, and reflects overall ability (or studiousness – it is difficult to tell these apart from exam performances alone!).
2. y_2 gives positive weight to C and negative weight to O. This is thus a contrast between open-book and closed-book exams. (Students differ greatly, like people generally; most students have a definite preference here; this is often gender-linked).
3. y_3 gives positive weight to Vectors, Algebra and Aalysis, and negative weight to Mechanics and Statistics. This is thus a pure-applied contrast (but would also depend on who taught what!). Again, most students have a definite preference for one or the other.
The last two are less important, as l_4, l_5 are smaller and lack a clear interpretation. We would retain 3 principal components here. We could also use three factors (see above for references to factor analysis).

Similarly for financial stock prices, where the three main factors may be: state of the economy; industrial sector; quality of management.
Covariances v. correlations.

One of the main problems with PCA is that it is *scale-dependent*: the outcome depends on the numbers, hence on the units used. The choice of units is often arbitrary, and then PCA does not have any *intrinsic* meaning. Also PCA looks for SLCs of maximum variability, and the variability can be increased arbitrarily by blowing up the scale in which some variable is measured. So we need to look at and choose the scale of each variable, and this depends on context.

If we use the covariance matrix S, we allow different variables to have differing importance. If we standardise each variance to 1, we pass from S to the correlation matrix R. This is independent of scale and intrinsically meaningful, but now all p variables have the same importance, which may/may not be sensible, depending on context. Moral: think carefully whether to use S or R *before* doing PCA. For more here, see e.g. [K] 2.2.5, esp. p.65-66.