smfl14.tex

Lecture 14. 18.2.2011

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

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

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

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 $S_y = AS_xA^T$. 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 $S^{-1/2}SS^{-1/2} = I$. 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 transformaton*

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

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 aareas. 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 the one-dimensional case).

Objectives.

These may vary widely.

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 intuitiion, 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 (with two types, this is called *discriminant analysis*).

Exploratory Data Analysis (EDA).

As in one dimension, one should begin by 'getting to know the data' by examining it visually. One should 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 eigenvalues 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 $a^Tx = \sum_1^p a_ix_i$ of x is a *standardised linear combination* (SLC) if $\sum_1^p a_i^2 = 1$ (i.e. $a^Ta = 1$).

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 $\alpha := a^Tx$ 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

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

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

But $\sum c_i^2 = 1$ and $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$, so $var\ \alpha = \sum \lambda_ic_i^2$ 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. //