smfl15.tex
Lecture 15. 18.2.2011

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 principal components. As a corollary: if Σ has rank $k < p$, *all* the variability is explained by the first k principal components (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, which typically have three parts –

(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

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

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, $L = diag(l_i)$, where $l_1 \geq \ldots \geq l_p \geq 0$ are the eigenvalues of S, $G = (g_1, \ldots, g_p)$ the orthogonal matrix of corresponding eigenvectors. As before,

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

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)

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

so as $Y = (y_{(1)}, \ldots, y_{(p)})$,

$$y_{(k)} = (X - \mathbf{1}\overline{x}^T)g_k$$

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, just as people generally do; most students have a definite preference for one or the other; 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 (though this 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 because of the smaller size of l_4, l_5, and have no clear interpretation. We would retain 3 principal components here.
Covariances v. correlations.

One of the main problems with PCA is that it is *scale-dependent*: the outcome depends on the numbers, and these depend on the units used to measure them. 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 we have now forced all p variables to have the same importance, which may or may not be sensible, again depending on context. Moral: think carefully whether to use S or R *before* doing PCA. For a thorough discussion, see e.g. [K] Section 2.2.5, esp. p.65-66.

Other topics.

1. *Analysis of variance (ANOVA).* This is closely related to regression. It tests the hypothesis that the *means* of p different samples are the same by analysing *variances* (= variability): if the means differ, variability *between* groups compared to *within* groups is higher than when the means are the same. For a full treatment, see e.g. [BF], Ch. 2.

2. *Chi-square distributions.* The *chi-square* distribution with n degrees of freedom (df), $\chi^2(n)$, is defined as the law of the sum of squares of n independent standard normals. The distribution theory in this area reduces to the distributions of quadratic forms in normal variates. As the relevant matrices are *projections* (satisfy $P^2 = P$), this can be reduced to *linear forms* in normal variates; see e.g. [BF], Ch. 3.

3. *Sum-of-squares decompositions.* Using sums of orthogonal projections from Linear Algebra, one can reduce the relevant distribution theory to decomposing sums of squares into independent sums of squares; all χ^2 distributed. Quotients of these have Fisher F-distributions. Hypotheses may be tested by suitable F-tests; for details, see e.g. [BF], Ch. 4, 6.

4. *Analysis of covariance (ANCOVA).*
This is a hybrid of regression and ANOVA, involving both qualitative and quantitative covariates. For details, see e.g. [BF], Ch. 5.

5. *Wishart distributions.* These are matrix analogues of chi-square distributions. They are important in multivariate hypothesis testing and multivariate ANOVA (MANOVA); see e.g. [MKB] 3.4, [K] 7.3. NHB