Chapter 2

Background and Motivation

Consider a data matrix X with n rows and p columns. The matrix consists of n points or vectors with p coordinates. In many situations the matrix will have arisen as n observations of a vector random variable.

2.1. Linear Principal Components.

The first (linear) principal component is the normalized linear combination of the p variables with the largest sample variance. It is convenient to think of X as a cloud of n points in p-space. The principal component is then the length of the projection of the n points onto a direction vector. The vector is chosen so that the variance of the projected points along it is largest. Any line parallel to this vector will have the same property. To tie it down we insist that it passes through the mean vector. This line then has the appealing property of being the line in p-space that is closest to the data. Closest is in terms of average squared euclidian distance. We think of the projection as being the best linear one dimensional summary of the data X. Of course this linear summary might be totally inadequate locally but it attempts to provide a reasonable global summary.

The theory and practical issues involved in linear principal components analysis are well known (Barnett 1981, Gnanadesikan 1977), and the technique is originally due to Spearman (1904), and then later developed by Hotelling (1933). We can find the the second component, orthogonal to the first, that has the next highest variance. The plane spanned by the two vectors and including the mean vector is the plane closest to the data. In general we can find the $m < p$ dimensional hyperplane that contains the most variance, and is closest to the data.

The solution to the problem is obtained by computing the singular value decomposition or basic structure of X, (centered with respect to the sample mean vector), or equivalently the eigen decomposition of the sample covariance matrix (Golub and Reinsch 1970, Greenacre 1984). Without any loss in generality we assume from now on that X is centered. If this is not the case, we can center X, perform the analysis, and uncenter the results by

adding back the mean vector.

In particular, the first principal component direction vector $\boldsymbol{a}$ is the largest normalized eigenvector of S, the sample covariance matrix. The principal component itself is $X\boldsymbol{a}$, an n vector with elements $\lambda_i = \boldsymbol{x}_i'\boldsymbol{a}$ where $\boldsymbol{x}_i'$ is the ith row of X and λ_i is the one dimensional summary variable for the ith observation. The coordinates in p-space of the projection of the ith observation on $\boldsymbol{a}$ are given by *

$$\boldsymbol{a}\lambda_i = \boldsymbol{a}\boldsymbol{a}'\boldsymbol{x}_i \tag{2.1}$$

There is no underlying model in the above. We merely regard the first component as a good summary of the original variables if it accounts for a large fraction of the total variance.

2.2. A linear model formulation.

In this section we describe a linear model formulation for the p variables. This formulation includes many familiar models such as linear regression and factor analysis. We end up showing in 2.2.2 that the estimation of the systematic component of some of these models is once again the principal component procedure.

2.2.1. Outline of the linear model.

Consider a model for the observed data

$$\boldsymbol{x}_i = \boldsymbol{u}_i + \boldsymbol{e}_i \tag{2.2}$$

where $\boldsymbol{u}_i$ is an unobservable systematic component and $\boldsymbol{e}_i$ an unobservable random component (We only get to see their sum). We usually impose some linear structure on $\boldsymbol{u}_i$, namely

$$\boldsymbol{u}_i = \boldsymbol{u}_0 + A\boldsymbol{\lambda}_i \tag{2.3}$$

where $\boldsymbol{u}_0$ is constant location vector, A is a $p \times m$ matrix and $\boldsymbol{\lambda}_i$ is an m-vector. For the procedures considered $\boldsymbol{u}_0$ is always estimated by the sample mean vector $\bar{\boldsymbol{x}}$; without loss of generality we will simply assume that X has been centered and ignore the term $\boldsymbol{u}_0$. We also

---

* If X is not centered we center it by forming $\tilde{X} = X - \mathbf{1}\bar{\boldsymbol{x}}$. Then the principal component is $\boldsymbol{\lambda} = \tilde{X}\boldsymbol{a}$ and the estimate in p space for the projection of the ith observation onto the principal component line $\bar{\boldsymbol{x}} + \boldsymbol{a}\gamma$ is $\bar{\boldsymbol{x}} + \boldsymbol{a}\lambda_i = \bar{\boldsymbol{x}} + \boldsymbol{a}\boldsymbol{a}'(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$

assume that $\boldsymbol{e}_i$ are mutually independent and identically distributed random vectors with mean $\boldsymbol{0}$ and covariance matrix Ψ and are independent of the $\boldsymbol{\lambda}_i$.

If the $\boldsymbol{\lambda}_i$ are considered to be random as well, the model is referred to as the linear structural model, or more commonly as the factor analysis model. If the $\boldsymbol{\lambda}_i$ are fixed it is referred as the linear functional model. The model (2.3) includes some familiar models as special cases:

- Let A be $p \times (p-1)$ with rank $(p-1)$. We can write A as

$$\begin{pmatrix} \boldsymbol{a}' \\ I \end{pmatrix}$$

where $\boldsymbol{a}$ is a $(p-1)$ vector and I is $(p-1) \times (p-1)$ since we can post-multiply A by an arbitrary non-singular $(p-1) \times (p-1)$ matrix and pre-multiply $\boldsymbol{\lambda}_i$ by its inverse. Thus we can write the model (2.3) as

$$\begin{pmatrix} x_{1i} \\ \boldsymbol{x}_{2i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}' \\ I \end{pmatrix} \boldsymbol{\lambda}_i + \begin{pmatrix} e_{1i} \\ \boldsymbol{e}_{2i} \end{pmatrix} \tag{2.4}$$

where $E(\boldsymbol{e}_i) = \boldsymbol{0}$ and assume $Cov(\boldsymbol{e}_i) = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$. If $\sigma_2^2 = \sigma_3^2 = \ldots = \sigma_p^2 = 0$ then we have the usual linear regression model with response x_{1i} and regressor variables $\boldsymbol{x}_{2i}$.

- If the variances are not zero we have the errors in variables regression model. The idea is to find a $(p-1)$ dimensional hyperplane in p-space that approximates the data well. The model takes care of errors in all the variables, whereas the usual linear regression model considers errors only in the response variable. This is a form of linear functional analysis.

- When the $\boldsymbol{\lambda}_i$ are random we have the usual factor analysis model, which includes the random effects Anova. This is also referred to as the linear structural model.

- If all the variances are zero and the $\boldsymbol{\lambda}_i$ are random and A is $p \times p$ the model represents the principal component change of basis. In this situation it is clear that the λ_{ji} are each functions of the $\boldsymbol{x}_i$.

For a full treatment of the above models see Anderson (1982).

2.2.2. Estimation.

We return for simplicity to the case where $m = 1$. Thus

$$\boldsymbol{x}_i = \boldsymbol{a}\lambda_i + \boldsymbol{e}_i \tag{2.5}$$

The systematic components $\boldsymbol{a}\lambda_i$ are points in p-space confined to the line defined by a multiple λ_i of the vector $\boldsymbol{a}$. We need to estimate λ_i for each observation, and the direction vector.

We now state some results which can be found in Anderson (1982).

If either

- the $\boldsymbol{e}_i$ are jointly Normal with a scalar covariance cI, where c is possibly unknown, and if $\boldsymbol{\lambda}_i$ are random or fixed, and we estimate by maximum likelihood

or

- as above but we drop the Normal assumption and estimate by least squares,

then the estimate of λ_i is once again the first principal component and that of $\boldsymbol{a}$ the principal component direction vector. In both cases the quantity we wish to minimize is

$$RSS(\boldsymbol{\lambda}, \boldsymbol{a}) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{a}\lambda_i\|^2 . \tag{2.6}$$

It is easy to see that for any $\boldsymbol{a}$ the appropriate value for λ_i is obtained by projecting the point $\boldsymbol{x}_i$ onto $\boldsymbol{a}$. Thus equation (2.6) reduces to

$$\begin{aligned} RSS(\boldsymbol{a}) &= \sum_{i=1}^{n} \left\|\boldsymbol{x}_i - \boldsymbol{a}\boldsymbol{a}'\boldsymbol{x}_i\right\|^2 \\ &= \mathrm{tr} XX' - \boldsymbol{a}'X'X\boldsymbol{a} \end{aligned} \tag{2.7}$$

The normalized solution to (2.7) is the largest eigenvector of $X'X$.

If the error covariance Ψ is general but known, we can transform the problem to the previous case. This is the same as using the Mahalanobis distance defined in terms of Ψ. In particular when Ψ is diagonal the procedure amounts to finding the line that minimizes the weighted distance to the points and is depicted in figure (2.1) below.

If the error covariance is unknown and not scalar then we require replicate observations in order to estimate it.

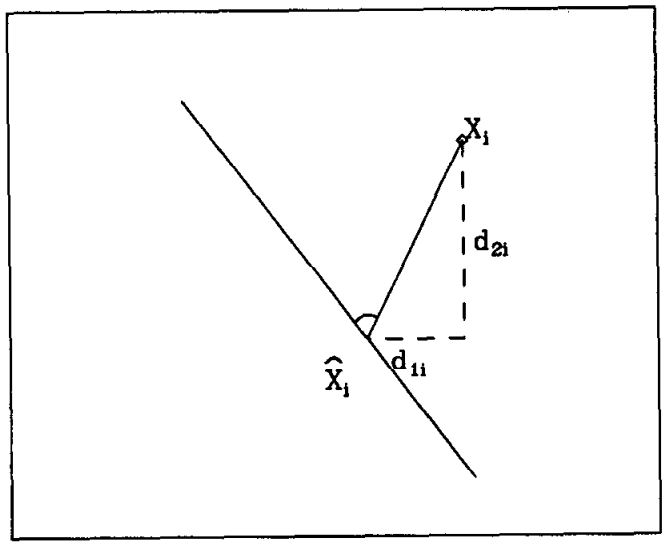

Figure 2.1 If $\Psi = \text{diag}(\sigma_1^2, \sigma_2^2)$ then we minimize the weighted distance $\sum_i (d_{1i}^2/\sigma_1^2 + d_{2i}^2/\sigma_2^2)$ from the points to the line.

2.2.3. Units of measurement.

It is often a problem in multivariate data analysis that variables have different error variances, even though they are measured in the same units. A worse situation is that often the variables are measured in completely different and incommensurable units. When we use least squares to estimate a lower dimensional summary, we explicitly combine the errors on each variable using the usual sum of components loss function, as in (2.6) . This then gives equal weight to each of the components. The solution is thus not invariant to changes in the scale of any of the variables. This is easily demonstrated by considering a spherical point cloud. If we scale up one of the co-ordinates an arbitrary amount, we can create as much linear structure as we like. In this situation we would really like to weigh the errors in the estimation of our model according to the variance of the measurement errors, which is seldom known. The safest procedure in this situation is to standardize each of the coordinates to have unit variance. This could destroy some of the structure that exists but without further knowledge about the scale of the components this yields a procedure that is invariant to coordinate scale transformations.

If, on the other hand, it is known that the variables are measured in the same units, we should not do any scaling at all. An apparent counter-example occurs if we make

measurements of the same quantities in different situations, with different measurement devices. An example might be taking seismic readings at different sights at the same instances with different recording devices. If the error variances of the two devices are different, we would want to scale the components differently.

To sum up so far, the principal component summary, besides being a convenient data reduction technique, provides us with the estimate of a formal parametric linear model which covers a wide variety of situations. An original example of the one factor model given here is that of Spearman (1904). The x_i are scores on psychological tests and the λ_i is some underlying unobservable general intelligence factor.

The estimation in all the cases amounts to finding a m-dimensional hyperplane in p-space that is closest to the points in some metric.

2.3. A non-linear generalization of the linear model.

The above formulation is often very restrictive in that it assumes that the systematic component in (2.2) is linear, as in (2.3). It is true in some cases that we can approximate a nonlinear surface by its first order linear component. In other cases we do not have sufficient data to estimate any more than a linear component. Apart from these cases, it is more reasonable to assume a model of the form

$$x_i = f(\lambda_i) + e_i \tag{2.8}$$

where λ_i is a m-vector as before and f is a p-vector of functions, each with m arguments. The functions are required to be smooth relative to the errors. This is a natural generalization of the linear model.

This dissertation deals with a generalization of the linear principal components. Instead of finding lines and planes that come close to the data, we find curves and surfaces. Just as the linear principal components are estimates for the variety of linear models listed above, so will our non- linear versions be estimates for models of the form (2.8) . So in addition to having a more general summary of multidimensional data, we provide a means of estimating the systematic component in a large class of models suitably generalized to include non-linearities. We refer to these summaries as principal curves and surfaces.

So far the discussion has concentrated on data sets. We can just as well formulate the above models for p dimensional probability distributions. We would then regard the data set

as a sample from this distribution and the functions derived for the data set will be regarded as estimates of the corresponding functions defined for the distribution. These models then define one and two dimensional surfaces that summarize the p dimensional distribution. The point $f(\lambda)$ on the surface that corresponds to a general point x from the distribution is a p dimensional random variable that can be summarized by a two dimensional random variable λ.

2.4. Other generalizations.

There have been a number of generalizations of the principal component model suggested in the literature.

- "Generalized principal components" usually refers to the adaptation of the linear model in which the coordinates are first transformed, and then the standard principal component analysis is carried out on the transformed coordinates.

- Multidimensional scaling (MDS) finds a low dimensional representation for the high dimensional point cloud, such that the sum of squared interpoint distances are preserved. This constraint has been modified in certain cases to cater only for points that are *close* in the original space.

- Proximity analysis provides parametric representations for data without noise.

- Non-linear factor analysis is a generalization similar to ours, except parametric coordinate functions are used.

We have been deliberately brief in listing these alternatives. Chapter 7 contains a detailed discussion and comparison of each of the above with the principal curve and surface models.