Chapter 7

Discussion and conclusions

In this chapter we discuss some of the existing techniques for symmetric smoothing, as well as the various generalizations of principal components and factor analysis. We compare these techniques with the methodology developed here. The chapter concludes with a summary of the uses of principal curves and surfaces.

7.1. Alternative techniques.

Other non-linear generalizations of principal components exist in the literature. They can be broadly classified according to two dichotomies.

- We can estimate either the non-linear manifold or the non-linear constraint that defines the manifold. In linear principal components the approaches are equivalent.
- The non-linearity can be achieved by transforming the space or by transforming the model.

The principal curve and surface procedures model the non-linear manifold by transforming the model.

7.1.1. Generalized linear principal components.

This approach corresponds to modeling either the nonlinear constraint or the manifold by transforming the space. The idea here is to introduce some extra variables, where each new variable is some non-linear transformation of the existing co-ordinates. One then seeks a subspace of this non linear co-ordinate system that models the data well. The subspace is found by using the usual linear eigenvector solution in the new *enlarged* space. This technique was first suggested by Gnanadesikan & Wilk (1966, 1968), and a good description can be found in Gnanadesikan (1977). They suggested using polynomial functions of the original p co-ordinates. The resulting *linear* combinations are then of the form (for $p = 2$ and quadratic polynomials)

$$\lambda_j = a_{1j}x_1 + a_{2j}x_2 + a_{3j}x_1x_2 + a_{4j}x_1^2 + a_{5j}x_2^2 \tag{7.1}$$

and the a_j will be eigenvectors of the appropriate covariance matrix.

This model has appeal mainly as a dimension reducing tool. Typically the linear combination with the smallest variance is set to zero. This results in an implicit non-linear constraint equation as in (7.1) where we set $\lambda = 0$. We then have a rank one reduction that tells us that the data lies close to a quadratic manifold in the original co-ordinates.

The model has been generalized further to include more general transformations of the co-ordinates other than quadratic, but the idea is essentially the same as the above; a linear solution is found in a transformed space. Young, Takane & de Leeuw (1978) and later Friedman (1983) suggested different forms of this generalization to include non-parametric transformations of the co-ordinates. The problem can be formulated as follows: Find $\boldsymbol{a}$ and $\boldsymbol{s}'(\boldsymbol{x}) = (s_1(x_1), \cdots, s_p(x_p))$ such that

$$\mathbf{E} \left\| \boldsymbol{s}(\boldsymbol{x}) - \boldsymbol{a}\boldsymbol{a}'\boldsymbol{s}(\boldsymbol{x}) \right\|^2 = \min! \tag{7.2}$$

or alternatively such that

$$\mathbf{Var}\left[\boldsymbol{a}'\boldsymbol{s}(\boldsymbol{x})\right] = \max! \tag{7.3}$$

where $\mathbf{E}\, s_j(x_j) = 0$, $\boldsymbol{a}'\boldsymbol{a} = 1$ and $\mathbf{E}\, s_j^2(x_j) = 1$. The idea is to transform the coordinates suitably and then find the linear principal components. If in (7.3) we replaced *max* by *min* then we would be estimating the constraint in the transformed space.

The estimation procedure alternates between finding the $s_j(\cdot)$ and finding the linear principal components in the transformed space.

- For a fixed vector of functions $\boldsymbol{s}(\cdot)$, we chose $\boldsymbol{a}$ to be the first principal component of the covariance matrix $\mathbf{E}\,\boldsymbol{s}(\boldsymbol{x})\boldsymbol{s}(\boldsymbol{x})'$.

- For $\boldsymbol{a}$ known, (7.2) can be written in the form

$$k\ \mathbf{E}\left[s_1(x_1) - \sum_{j=2}^{p} b_{1j} s_j(x_j)\right]^2 + \quad \text{terms in } s_2(\cdot), \cdots, s_p(\cdot), \tag{7.4}$$

  and b_{ji} are functions of $\boldsymbol{a}$ above. If $s_2, \ldots, s_p$ are known, equation (7.4) is minimized by

$$s_1(x_1) = \mathbf{E}\left(\sum_{j=2}^{p} b_{1j} s_j(x_j) \mid x_1\right)$$

  This is true for any s_j, and suggests an inner iterative loop. This inner loop is very similar to the ACE algorithm (Breiman and Friedman, 1982), except the normalization

is slightly different. Breiman and Friedman proved that the ACE algorithm converges under certain regularity conditions in the distributional case.

The disadvantages of this technique are:

- The space is transformed, and in order to understand the resultant fit, we usually would need to transform back to the original space. This can only be achieved if the transformations are restricted to monotone functions. In the transformed space the estimated manifold is given by

$$\begin{pmatrix} \hat{s}_1(x_1) \\ \vdots \\ \hat{s}_p(x_p) \end{pmatrix} = aa's(x).$$

  Thus if the $s_j(\cdot)$ are monotone, we get untransformed estimates of the form

$$\begin{pmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_p \end{pmatrix} = \begin{pmatrix} s_1^{-1}(a_1 z) \\ \vdots \\ s_p^{-1}(a_p z) \end{pmatrix} \tag{7.5}$$

  where $z = a's(x)$. Equation (7.5) defines a parametrized curve. The curve is not completely general since the co-ordinate functions are monotone. For the same reason, Gnanadesikan (1978) expressed the desirability of having procedures for estimating models of the type proposed in this dissertation.

- We are estimating manifolds that are close to the data in the transformed co-ordinates. When the transformations are non-linear this can result in distortion of the error variances for individual variables. What we really require is a method for estimating manifolds that are close to the data in the original p co-ordinates. Of course, if the functions are linear, both approaches are identical.

An advantage of the technique is that it can easily be generalized to take care of higher dimensional manifolds, although not in an entirely general fashion. This is achieved by replacing a with A where A is $p \times q$. We then get a q dimensional hyperplane in the transformed space given by $AA's(x_i)$. However, we end up with a *number* of implicit constraint equations which are hard to deal with and interpret. Despite the problems associated with generalized principal components, it remains a useful tool for performing rank 1 dimensionality reductions.

7.1.2. Multi-dimensional scaling.

This is a technique for finding a low dimensional representation of high dimensional data. The original proposal was for data that consists of $\binom{n}{2}$ dissimilarities or distances between n objects. The idea is to find a m (m small, 1, 2 or 3) dimensional euclidean representation for the objects such that the inter-object distances are preserved as well as possible. The idea was introduced by Torgerson (1958), and followed up by Shepard (1962), Kruskal (1964a ,1964b), Shepard & Kruskal (1964) and Shepard & Carroll (1966). Gnanadesikan (1978) gives a concise description.

The procedures have also been suggested for situations where we simply want a lower dimensional representation of high dimensional euclidean data. The lower dimensional representation attempts to reproduce the interpoint distances in the original space. We fit a principal curve to the color data in example 6.5; these data were originally analyzed by Shepard and Carroll (1966) using MDS techniques. Although there have been some intriguing examples of the technique in the literature, a number of problems exist.

- The solution consists of a vector of m co-ordinates representing the location of points on the low dimensional manifold, but *only for the n data points.* What we don't get, and often desire is a mapping of the whole space. We are unable, for example, to find the location of new points in the reduced space.

- The procedures are computationally expensive and unfeasible for large n ($nm > 300$ is considered large). They are usually expressed as non-linear optimization problems in nm parameters, and differ in the choice of criterion.

The principal curve and surface procedures partially overcome both the problems listed above; they are unable to find structures as general as those that can be found by the MDS procedures due to the averaging nature of the scatterplot smoothers, but they do provide a mapping for the space. We have demonstrated their ability to model MDS type data in examples 6.4 and 6.5. They do not, however, provide a model for dissimilarities which was the original intention of multidimensional scaling.

7.1.3. Proximity models.

Shepard & Carroll (1966) suggested a functional model similar in form to the model we suggest. They required only to estimate the n vectors of m parameters for each point, and considered the data to be functions thereof. The parameters (nm altogether) are found

by direct search as in MDS, with a different criterion to be minimized. Their procedure, however, was geared towards data without error, as in the ball data in example 6.4. This becomes evident when one examines the criterion they used, which measures the continuity of the data as a function of the parameters. When the data is not smooth, as is usually the case, we need to estimate functions that vary smoothly with the parameters, and are close to the data.

7.1.4. Non-linear factor analysis.

More recently, Etezadi-Amoli and McDonald (1983) approached the problem of non-linear factor analysis using polynomial functions. They use a model of the form

$$X = f(\lambda) + e$$

where f is a polynomial in the unknown parameters or factors. Their procedure for estimating the unknown factors and coefficients is similar to ours in this restricted setting. * Their emphasis is on the factor analysis model, and once the appropriate polynomial terms have been found, the problem is treated as an enlarged factor analysis problem. They do not estimate the λ's as we do, using the geometry of the problem, but instead perform a search in nq parameter space, where q is the dimension of λ and n is the number of observations. Our emphasis is on providing one and two dimensional summaries of the data. In certain situations, these summaries can be used as estimates of the appropriate non-linear functional and factor models.

7.1.5. Axis interchangeable smoothing.

Cleveland (1983) describes a technique for symmetrically smoothing a scatterplot which he calls *axis interchangeable smoothing* (which we will refer to as AI smoothing). We briefly outline the idea:

- standardize each coordinate by some (robust) measure of scale.
- rotate the coordinate axes by 45°. (if the correlation is positive, else rotate through $-45°$).
- smooth the transformed y against the transformed x.

---

* Their paper was published in the September, 1983 issue of Psychometrika, whereas Hastie (1983) appeared in July.

- rotate the axes back.
- unstandardize.

If the standardization uses regular standard deviations, then the rotation is simply a change of basis to the principal component basis. The resulting curve minimizes the distance from the points orthogonal to this principal component. It has intuitive appeal since the principal component is the line that is closest in distance to the points. We then allow the points to *tug in* the principal component line. It is simple and fast to compute the AI Smooth, and for many scatterplots it produces curves that are very similar to the principal curve solution. This is not surprising when we consider the following theorem:

Theorem 7.1

If the two variables in a scatterplot are standardized to have unit standard deviations, and if the smoother used is linear and reproduces straight lines exactly, then the axis interchangeable smooth is identical to the curve of the first iteration of the principal curve procedure.

Proof

Let the variables x and y be standardized as above. The AI Smooth transforms to two new variables

$$\begin{aligned} x^* &= \frac{(x+y)}{\sqrt{2}} \\ y^* &= \frac{(x-y)}{\sqrt{2}} \end{aligned} \tag{7.6}$$

Then the AI Smooth replaces (x^*, y^*) by $(x^*, \ \mathbf{Smooth}(y^* \mid x^*))$. But $\mathbf{Smooth}(x^* \mid x^*) = x^*$ since the smoother reproduces straight lines exactly.* Thus the AI Smooth transforms back to

$$\begin{aligned} \hat{x} &= \frac{(\ \mathbf{Smooth}(x^* \mid x^*) + \ \mathbf{Smooth}(y^* \mid x^*)}{\sqrt{2}} \\ \hat{y} &= \frac{(\ \mathbf{Smooth}(x^* \mid x^*) - \ \mathbf{Smooth}(y^* \mid x^*))}{\sqrt{2}} \end{aligned} \tag{7.7}$$

Since the smoother is linear, and in view of (7.6) , (7.7) becomes

$$\begin{aligned} \hat{x} &= \ \mathbf{Smooth}(x \mid x^*) \\ \hat{y} &= \ \mathbf{Smooth}(y \mid x^*) \end{aligned} \tag{7.8}$$

---

* Any weighted local linear smoother has this property. Local averages, however, do not unless the predictors are evenly spaced.

This is exactly the curve found after the first iteration of the principal curve procedure, since $\hat{\boldsymbol{\lambda}}^{(0)} = \boldsymbol{x}^*$. ∎

Williams and Krauss (1982) extended the AI smooth by iterating the procedure. At the second step, the residuals are calculated locally by finding the tangent to the curve at each point and evaluating the residuals from these tangents. The new fit at that point is the smooth of these residuals against their projection onto the tangent. This procedure would probably get closer to the principal curve solution than the AI smooth (we have not implemented the Williams and Krauss smooth). Analytically one can see that the procedures differ from the second step on.

This particular approach to symmetric smoothing (in terms of residuals) suffers from several deficiencies :

- the type of curves that can be found are not as general as those found by the principal curve procedure.
- they are designed for scatterplots and do not generalize to curves in higher dimensions.
- they lack the interpretation of principal curves as a form of conditional expectation.

7.2. Conclusions.

In conclusion we summarize the role of principal curves and surfaces in statistics and data analysis.

- They generalize the one and two dimensional summaries of multivariate data usually provided by the principal components.
- When the principal curves and surface are linear, they are the principal component summaries.
- Locally they are the critical points of the usual distance function for such summaries; this gives an indication that there are not too many of them.
- They are defined in terms of conditional expectations which satisfies our mental image of a summary.
- They provide the least squares estimate for generalized versions of factor analysis, functional models and the errors in variables regression models. The non-linear errors

in variables model has been used successfully a number of times in practical data analysis problems (notably calibration problems).

- In some situations they are a useful alternative to MDS techniques, in that they provide a lower dimensional summary of the *space* as opposed to the *data set*.
- In some situations they can be effective in identifying outliers in higher dimensional space.
- They are a useful data exploratory tool. Motion graphics techniques have become popular for looking at 3 dimensional point clouds. Experience shows that it is often impossible to identify certain structures in the data by simply rotating the points. A summary such as that given by the principal curve and surfaces can identify structures that would otherwise be transparent, even if the data could be viewed in a real three dimensional model.

Acknowledgements

My great appreciation goes to my advisor Werner Stuetzle, who guided me through all stages of this project. I also thank Werner and Andreas Buja for suggesting the problem, and Andreas for many helpful discussions. Rob Tibshirani helped me a great deal, and some of the original ideas emerged whilst we were suntanning alongside a river in the Californian mountains. Brad Efron, as usual, provided many insightful comments. Thanks to Jerome Friedman for his ideas and constant support. In addition I thank Persi Diaconis and Iain Johnstone for their help and comments, and Roger Chaffee and Dave Parker for their computer assistence. Finally I thank the trustees of the Queen Victoria, the Sir Robert Kotze and the Sir Harry Crossley scholarships for their generous assistence.