Chapter 6

Examples

This chapter contains six examples that demonstrate the procedures on real and simulated data. We also introduce some ideas such as bootstrapping, robustness, and outlier detection.

Example 6.1. Gold assay pairs.

This real data example illustrates:

- A principal curve in 2-space,
- non-linear errors in variables regression,
- co-ordinate function plots, and
- bootstrapping principal curves.

A California based company collects computer chip waste in order to sell it for its content of gold and other precious metals. Before bidding for a particular cargo, the company takes a sample in order to estimate the gold content of the the whole lot. The sample is split in two. One sub-sample is assayed by an outside laboratory, the other by their own inhouse laboratory. (The names of the company and laboratory are withheld by request). The company wishes to eventually use only one of the assays. It is in their interest to know which laboratory produces on average lower gold content assays for a given sample.

The data in figure 6.1a consists of 250 pairs of gold assays. Each point is represented by

$$\boldsymbol{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix}$$

where $x_{ji} = \log(1 + \text{assay yield for } i\text{th assay pair for lab } j)$ and where $j = 1$ corresponds to the inhouse lab and $j = 2$ the outside lab. The log transformation tends to stabilize the variance and produce a more even scatter of points than in the untransformed data. (There were many more small assays (1 oz per ton) than larger ones (> 10 oz per ton)).

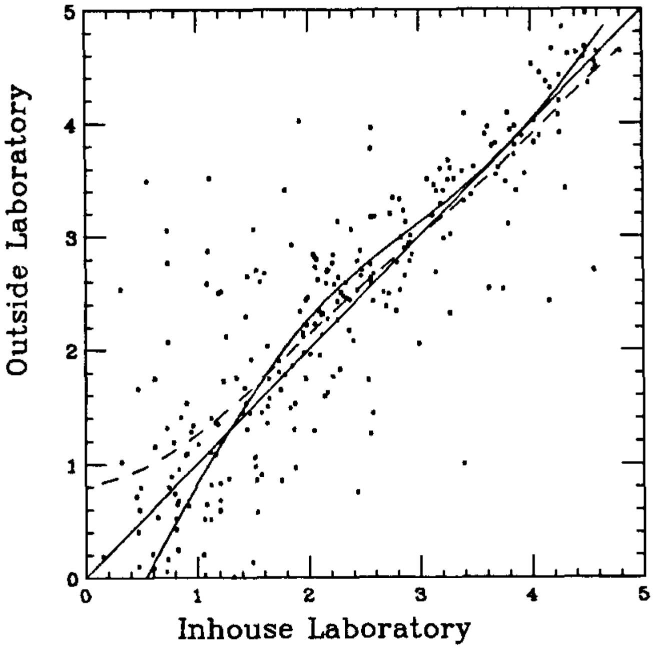

Figure 6.1a Plot of the log assays for the inhouse and outside labs. The solid curve is the principal curve, the dashed curve the scatterplot smooth.

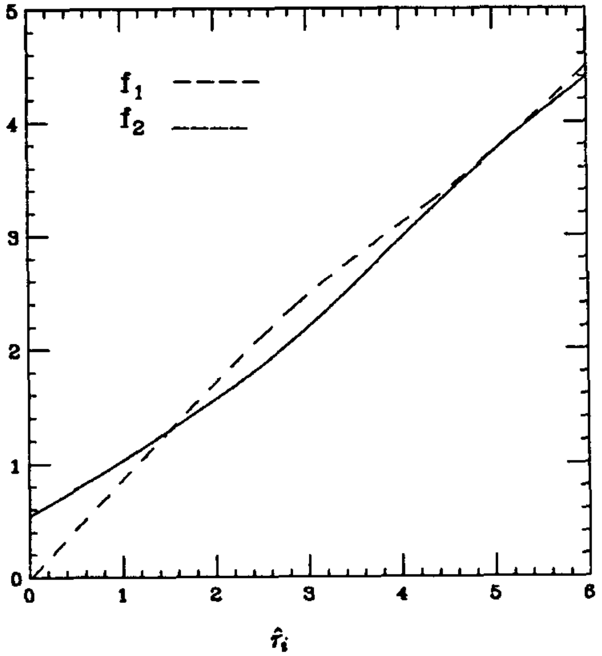

Figure 6.1b Estimated coordinate functions. The dashed curve is the outside lab, the solid curve the inhouse lab.

A standard analysis might be a paired t-test for an overall difference in assays. This would not reflect local differences which can be of great importance since the higher the level of gold the more important the difference.

The data was actually analyzed by smoothing the differences in log assays against the average of the two assays. This can be considered a form of symmetric smoothing and was suggested by Cleveland (1983). We discuss the method further in chapter 7.

The model presented here for the above data is

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} f_1(\tau_i) \\ f_2(\tau_i) \end{pmatrix} + \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \tag{6.1}$$

where τ_i is the unknown true gold content for sample i (or any monotone function thereof), $f_j(\tau_i)$ is the expected assay result for lab j, and e_{ji} is measurement error. We wish to analyze the relationship between f_1 and f_2 for different true gold contents.

This is a generalization of the errors in variables model or the structural model (if we regard the τ_i themselves as unobservable random variables), or the functional model (if the τ_i are considered fixed). This model is traditionally expressed as a linear model:

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} \alpha + \beta z_i \\ z_i \end{pmatrix} + \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \tag{6.2}$$

where $f_2(\tau_i) = z_i$ and

$$\begin{aligned} f_1(\tau_i) &= f_1 \circ f_2^{-1}(z_i) \qquad \text{(assuming } f_2 \text{ is monotone)} \\ &= \alpha + \beta z_i \end{aligned}$$

It suffers, however, from the same drawback as the t-test in that only global inference is possible.

We assume that the e_{ji} are pairwise independent and that *

$$\mathbf{Var}(e_{1i}) = \mathbf{Var}(e_{2i}) \quad \forall\, i.$$

The model is estimated using the principal curve estimate for the data and is represented by the solid curve in figure 6.1a. The dashed curve is the usual scatterplot smooth of x_2 against x_1 and is clearly misleading as a scatterplot summary. The curve lies above the 45° line in the interval 1.4 to 4 which represents an untransformed assay interval of 3 to 15 oz/ton. In this interval the inhouse average assay is lower than that of the outside lab. The difference is reversed at lower levels, but this is of less practical importance since at these levels the cargo is less valuable. This is more clearly seen by examining the estimated coordinate function plots in figure 6.1b.

A natural question arising at this point is wether the kink in the curve is real or not. If we had access to more data from the same population we could simply calculate the principal curves for each and see how often the kink is reproduced. We could then perhaps construct a 95% confidence tube for the true curve.

In the absence of such repeated samples, we use the bootstrap (Efron 1981, 1982) to simulate them. We would like to, but cannot, generate samples of size n from F, the true distribution of $\boldsymbol{x}$. Instead we generate samples of size n from $\hat{F}$, the empirical or estimated distribution function, which puts mass $1/n$ on each of the sample points $\boldsymbol{x}_i$. Each such sample, which samples the points $\boldsymbol{x}_i$ with replacement, is called a bootstrap sample.

* In the linear model one usually requires that $Var(e_{ji}) = constant_j$. This assumption can be relaxed here.

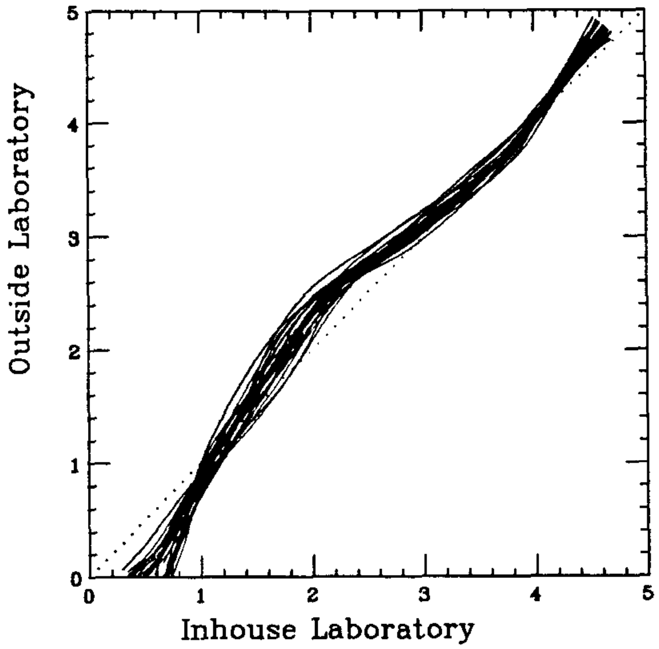

Figure 6.1*c* 25 bootstrap curves. The data X is sampled 25 times with replacement, each time yielding a bootstrap sample X^*. Each curve is the principal curve of such a sample.

Figure 6.1c shows the principal curves obtained for 25 such bootstrap samples. The 45° line is included in the figure, and we see that none of the curves cross the line in the region of interest. This provides strong evidence that the kink is indeed real.

When we compute a particular bootstrap curve, we use the principal curve of the original sample as a starting value. Usually one or two iterations are all that is required for the procedure to converge. Also, since each of the bootstrap points occurs at one of the sample sites, we know where they project onto this initial curve.

It is tempting to extract from the procedure estimates of $\hat{\tau}_i$, the true gold level for sample i. However, $\hat{\tau}_i$ need not be the true gold level at all. It may be any variable that orders the pairs $f(\hat{\tau}_i)$ along the curve, and is probably some monotone function of the true gold level. It is clear that both labs could consistently produce biased estimates of the true gold level and there is thus no information at all in the data about the true level.

Estimates of τ_i do provide us with a good summary variable for each of the pairs, if that is required:

$$\hat{r}_i = h(\boldsymbol{x}_i)$$

since we obtain $\hat{r}_i$ by projecting the point $\boldsymbol{x}_i$ onto the curve. Finally we observe that the above analysis could be extended in a straightforward way to include 3 or more laboratories. It is hard to imagine how to tackle the problem using standard regression techniques.

Example 6.2. The helix in three-space.

This is a simulated example illustrating:

- A principal curve in 3-space,
- co-ordinate plots, and
- cross-validation and span selection.

We looked at the bias of the principal curve procedure in estimating the helix in chapter 4. We now demonstrate the procedure by generating data from that model. We have

$$\boldsymbol{f}(\lambda) = \begin{pmatrix} \sin(4\pi\lambda) \\ \cos(4\pi\lambda) \\ 4\lambda \end{pmatrix} + \boldsymbol{e},$$

where $\lambda \sim U[0,1]$ and $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, .3I)$. This situation does not present the principal curve procedure with any real problems. The reason is that the starting vector passes down the middle of the helix and the data projects onto it in nearly the correct order. Table 6.1shows the steps in the iterations as the procedure converges at each of the ***procedural spans*** shown. At a span of $s = .2$ we use cross-validation to find the minimum ***mse span***.

Figure 6.2c shows the $CVRSS$ curve used to select the span, which is 0.1 with a value of $CVRSS$ of 0.1944. One more step is performed and the procedure is terminated. Figure 6.2d shows the estimated co-ordinate functions for this choice of span. We see that the estimate of the linear co-ordinate is rather wiggly. It is clear that a small span was required to estimate the sinusoidal co-ordinates, but a large span would suffice for the linear co-ordinate. This suggests a different scheme for cross-validation—choosing the spans separately for each co-ordinate. The results are shown in figures 6.2e and 6.2f. As predicted, a larger span is chosen for the linear co-ordinate, and its estimate is no longer wiggly. This is the final model referred to in the table and represented in figure 6.2.

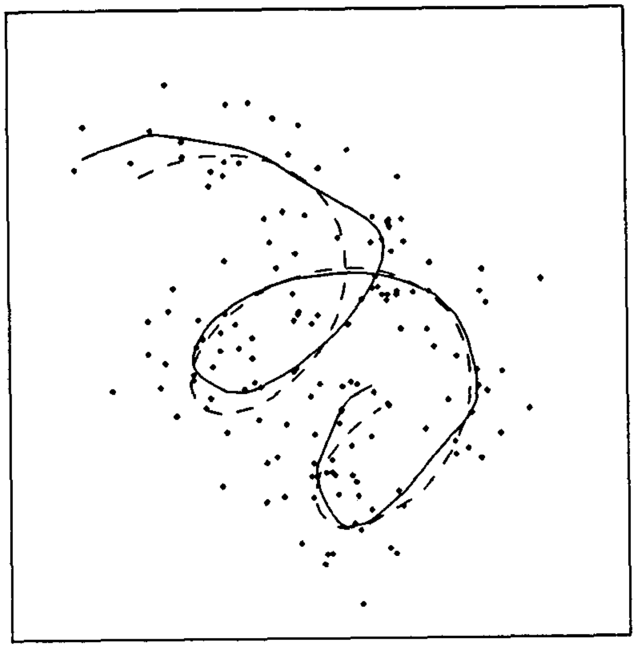

Figure 6.2a Data generated from a helix with independent errors on each coordinate. The dashed curve is the original helix, the solid curve is the principal curve estimate.

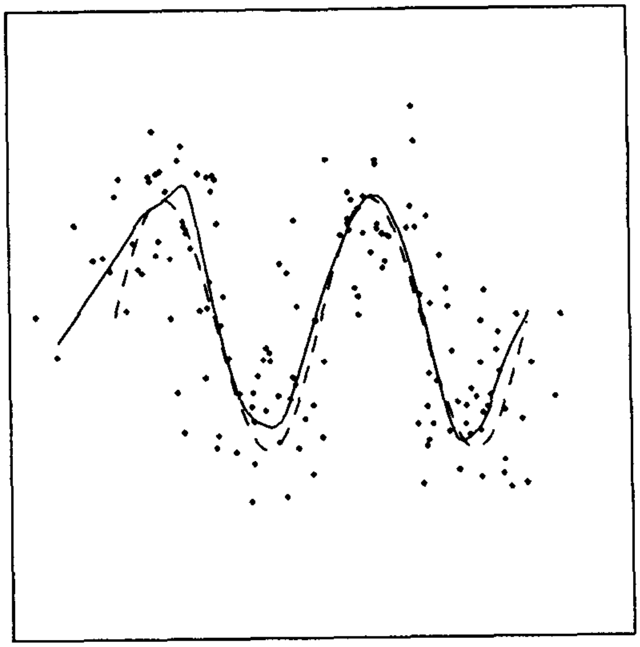

Figure 6.2b Another view of the helix, the data and the principal curve.

Table 6.1. The steps in the iterations. Initially the procedure goes through a regimen of procedural spans. Then the final span is found by cross-validation.

Iteration #	Span	D^2	d.o.f.	Comments
	procedural spans			
start	1.0	1.110	2.0	principal component line
1	0.4	0.740	4.2	initial span
2	0.4	0.565	4.6	
3	0.4	0.550	4.7	
4	0.4	0.549	4.7	converged
5	0.3	0.376	5.7	reduce span
6	0.3	0.361	5.4	
7	0.3	0.360	5.4	converged
8	0.2	0.222	7.3	reduce span
9	0.2	0.217	6.9	
10	0.2	0.217	6.9	converged
	mse spans			
final	0.07, 0.09, 0.35	0.162	9.7	
		0.189		cross-validated

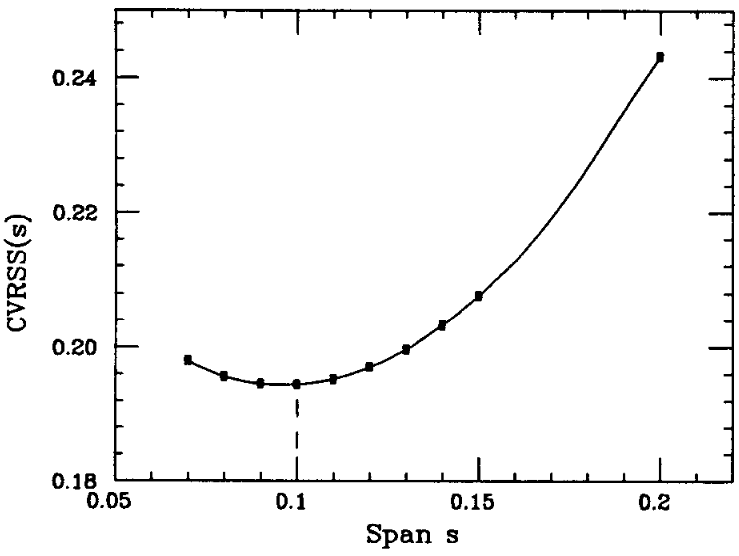

Figure 6.2c The cross-validation curve shows $CVRSS(s)$ as a function of the span s. One span is used for all 3 co-ordinates.

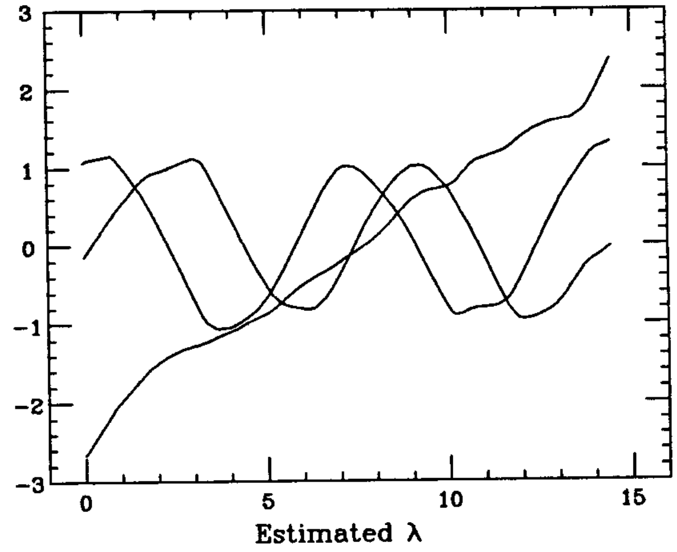

Figure 6.2d The estimated co-ordinate functions for the helix, using the span found in figure 6.2c.

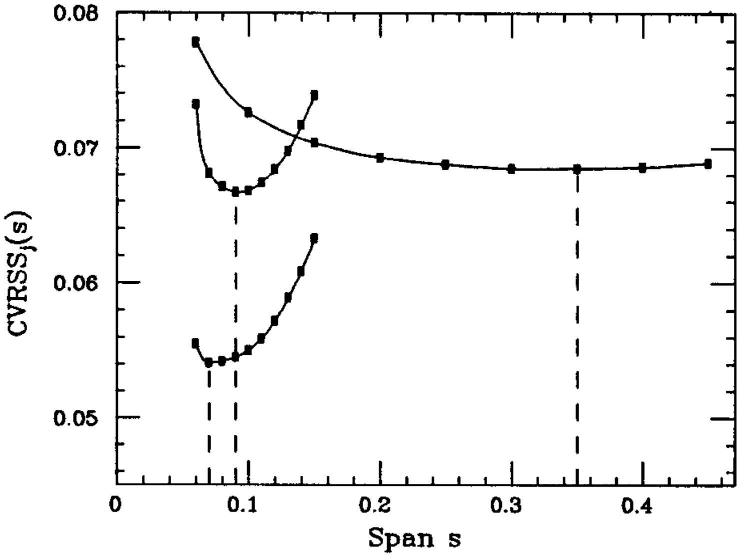

Figure 6.2e The cross-validation curve shows $CVRSS_j(s)$ as a function of the span s. A separate span is found for each co-ordinate.

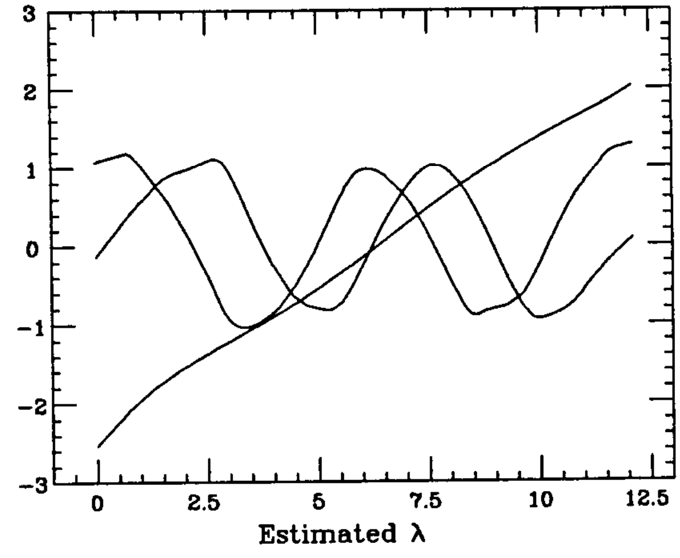

Figure 6.2f The estimated co-ordinate functions for the helix, using the spans found in figure 6.2f.

The entry labelled d.o.f. in table 6.1is an abbreviation for degrees of freedom. In linear regression the number of parameters used in the fit is given by $\mathrm{tr}\,(H)$ where H is the projection or ***hat*** matrix. If the response variables y_i are iid with variance σ^2, then

$$\begin{aligned}\sum_{i=1}^{n} \mathbf{Var}\,(\hat{y}_i) &= \sum_{i=1}^{n} \mathbf{Var}\,(\boldsymbol{h}_i'\boldsymbol{y}) \\ &= \sigma^2 \mathrm{tr}\,(H'H) \\ &= \sigma^2 \mathrm{tr}\,(H)\end{aligned}$$

We can do the same calculation for a linear smoother matrix C, and in fact for the local straight lines smoother we even have $\mathrm{tr}\,(C'C) = \mathrm{tr}\,(C)$. As the span decreases, the diagonal entries of C get larger, and thus the variance of the estimates increases, as we would expect. One can also approach this from the other side by looking at the residual sum of squares. In the absence of bias we have

$$\begin{aligned}\mathbf{E}RSS &= \mathbf{E}\,\|(I-C)\boldsymbol{y}\|^2 \\ &= \mathbf{E}\boldsymbol{y}'(I-C)'(I-C)\boldsymbol{y} \\ &= \mathrm{tr}\,[(I-C)'(I-C)\;\mathbf{Cov}\,(\boldsymbol{y})] \\ &= (n-\mathrm{tr}\,(C))\sigma^2\end{aligned} \tag{6.3}$$

if $\mathrm{tr}\,(C'C) = \mathrm{tr}\,(C)$. * More motivation for regarding $\mathrm{tr}\,(C)$ as the number of parameters or d.o.f. can be found in Cleveland (1979) and Tibshirani (1984). Some calculations similar to those in 3.5.1 show that the expected squared distance of X from the true $\boldsymbol{f}$ is $D^2 \approx 2\sigma^2$, or more precisely $D^2 \approx 2\sigma^2 - \sigma^4/(4\rho^2)$ where ρ is the radius of curvature, which in our example is $1 + 1/\pi^2$. Thus $D^2 \approx 0.18$. The cross validated residual estimate $\sum CVRSS_j$ was found to be 0.189. The orthogonal distance from the final curve is $D^{2(11)} = 0.162$. This is deflated due to overfitting. The average value of d.o.f for the final curve is (one for each co-ordinate) 9.7, or a total of 29.1. Some simple heuristics show that the we should scale this value up by by $2n/(2n - d.o.f) = 300/(300 - 29.1) = 1.11$. We then get $2n/(2n - d.o.f)D^{2(11)} = 0.179$ which is back in the correct ballpark.

It is more convenient to view the 3 dimensional examples on a color graphics system (such as the Chromatics system of the Orion group, Stanford University). This allows one to rotate the points in real time and thus see the 3rd dimension.

* For our smoothers, each row of C is the row of a projection matrix, and hence $\boldsymbol{c}_i'\boldsymbol{c}_i = c_{ii}$.

Example 6.3. Geological data.

This real data example illustrates:

- Data modelling in 3 dimensions,
- non-linear factor analysis, and
- outlier detection and robust fitting.

The data in this example consists of measurements of the mineral content of 64 core samples, each taken at different depths (Chernoff, 1973). Measurements were made of 10 minerals in each sample. We simply label the minerals $X_1, \cdots, X_{10}$, and analyze the first three.

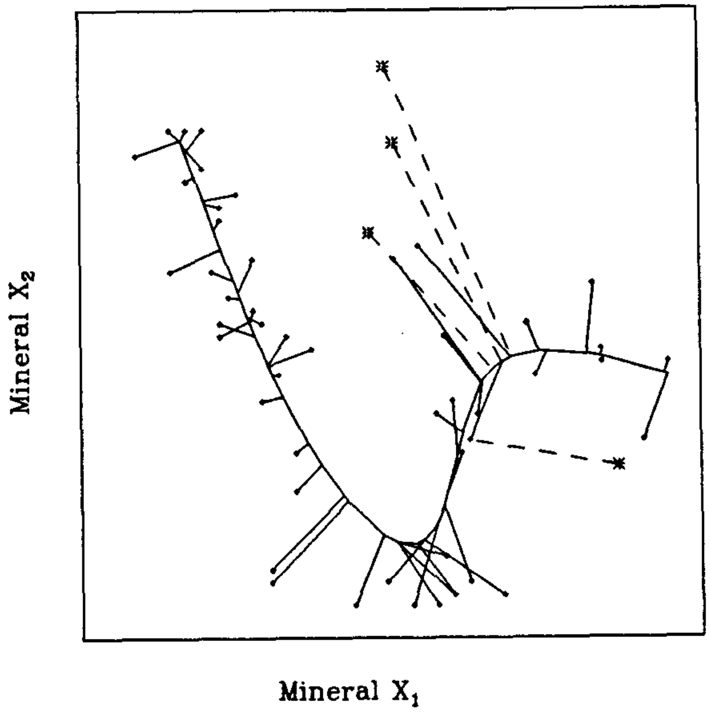

Figure 6.3a The principal curve for the mineral data. (Variable X_3 is into the page). The spikes join the points to their projection on the curve. The 4 outliers are joined to the curve with the broken lines.

Figure 6.3*a* shows the data and the solution curve. (A final span of 0.35 was manually selected.) In 3-D the picture looks like a dragon with its tail pointing to the left and the

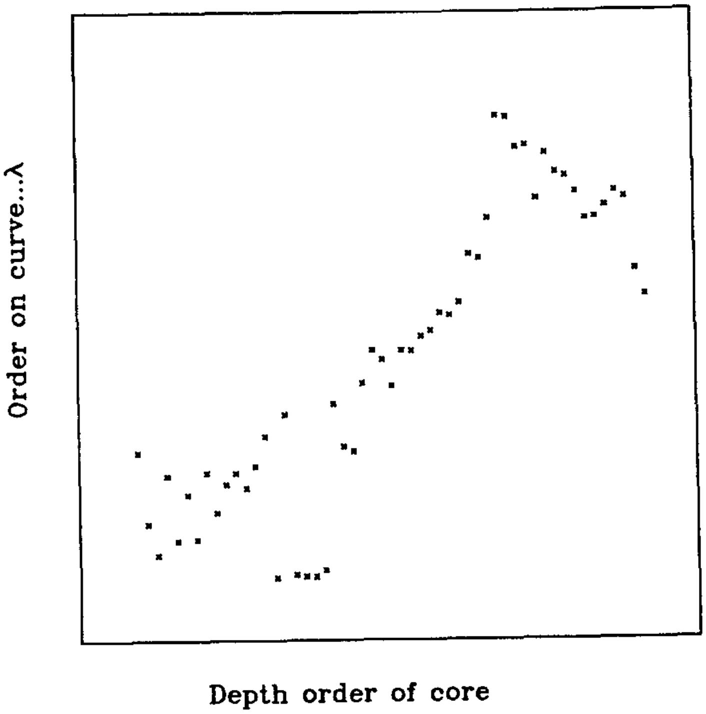

Figure 6.3b The values $\lambda_{\hat{f}}(x_i)$ are plotted against the depth order of the core samples.

long (outlier) spikes could be a mane. The linear principal component explains 55% of the variance, whereas this solution explains 82%.

The spikes join the observations to their closest projections on the curve. This is a useful device for spotting outliers. A robust version of the principal curve procedure was used in this example. After the first iteration, points receive a weight which is inversly proportional to their distance from the curve. In the smoothing step, a weighted smooth is used, and if the weight is below a certain threshhold, it is set to 0. Four points were identified as outliers, and are labelled differently in figure 6.3*a* . We would really consider them model outliers, since in that region of the curve the model does not appear to fit very well.

Figure 6.3*b* shows the relationship between the order of the points on the curve, and the depth order of the core samples. The curve appears to recover this variable for the most part. The area where it does not recover the order is where the curve appears to fit the data badly anyway. So here we have uncovered a hidden variable or factor that we are able to validate with the additional information we have about the ordering. The co-ordinate

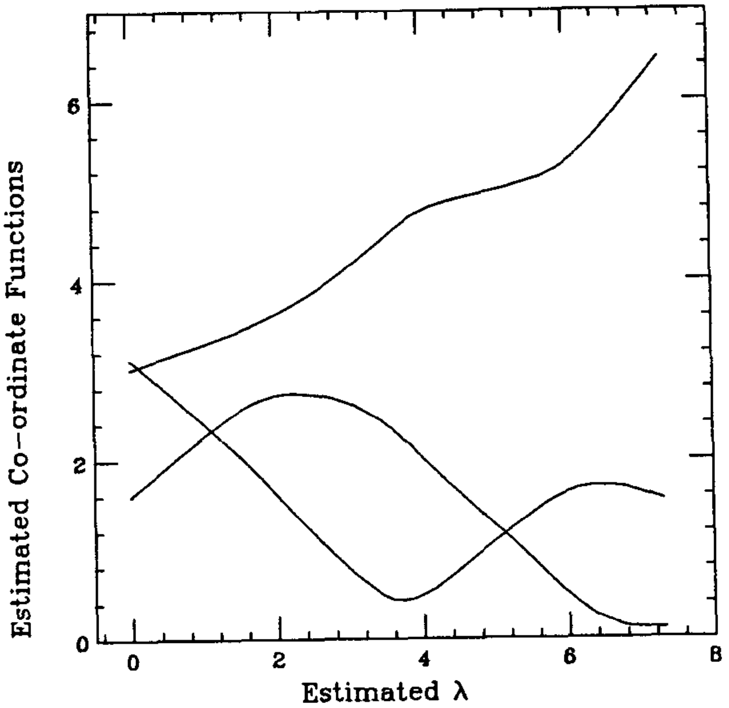

Figure 6.3c The estimated co-ordinate functions or *factor loading curves* for the three minerals.

plots would then represent the mean level of the particular mineral at different depths (see figure 6.3*c*). Usually one would have to use these co-ordinate plots to identify the factors, just as one uses the factor loadings in the linear case.

Example 6.4. The uniform ball.

This example illustrates:

- A principal surface in 3 space, and
- a connection to multidimensional scaling.

The data is artificially constructed, with no noise, by generating points uniformly from the surface of a sphere. It is the same data used by Shepard and Carroll (1966) to demonstrate their parametric mapping algorithm. (see reference and chapter 7). We simply use it here to demonstrate the ability of the principal surface algorithm to produce surfaces that are not a function of the starting plane (in analogy to the circle example in chapter 3).

There are 61 data points, as shown in figure 6.4a. One point is placed at each intersection of 5 equally spaced parallels and 12 equally spaced meridians. The extra point

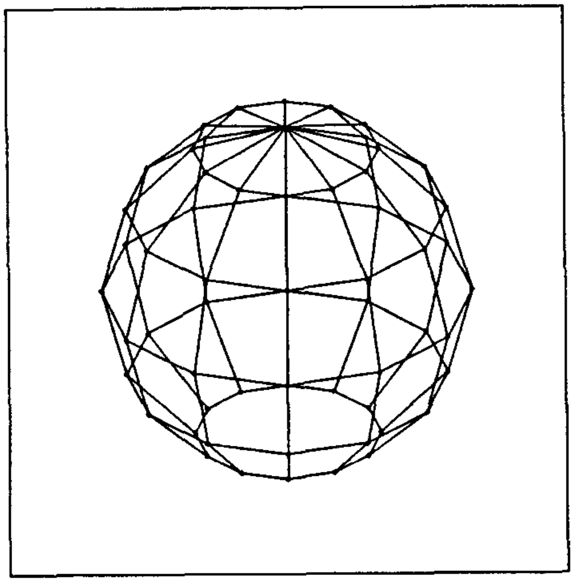

Figure 6.4a The data points are placed in a uniform pattern on the surface of a sphere. The south pole is missing.

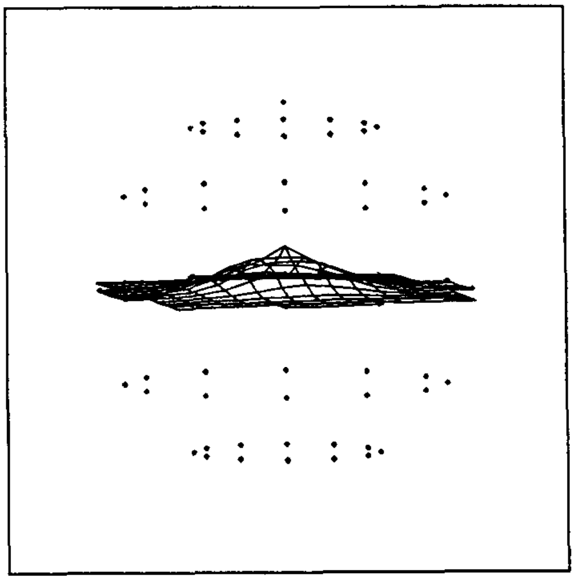

Figure 6.4b The second iteration of the principal surface procedure finds a surface that is a function of the first iteration.

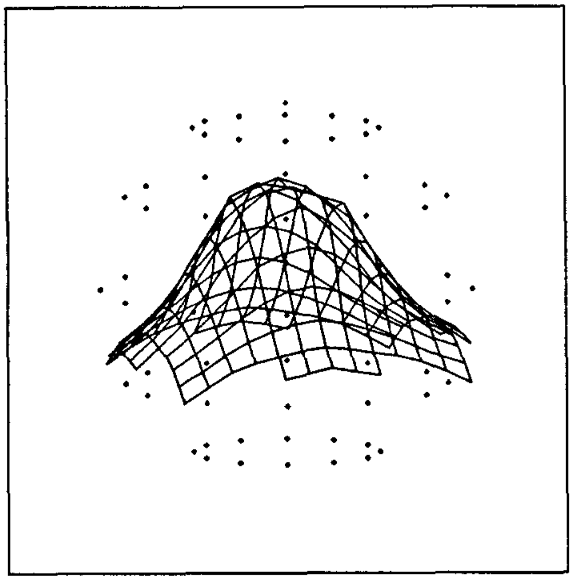

Figure 6.4c An intermediate stage in the iterations.

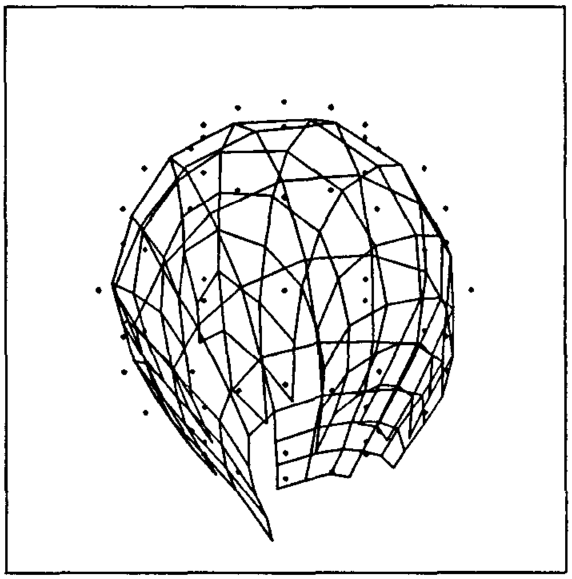

Figure 6.4d The final surface produced by the principal surface routine.

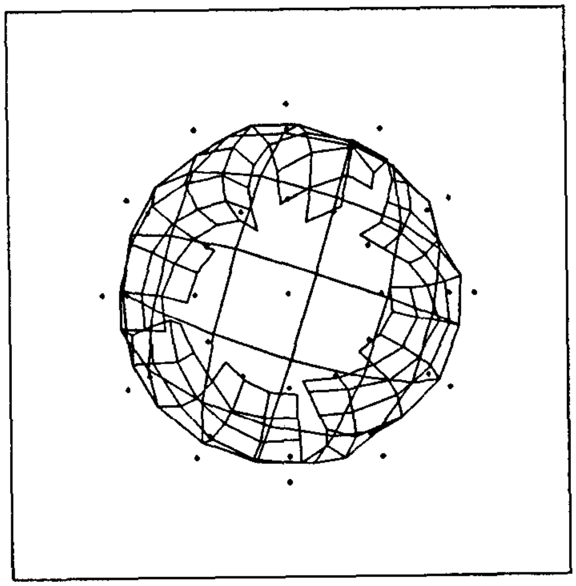

Figure 6.4e Another view of the final principal surface.

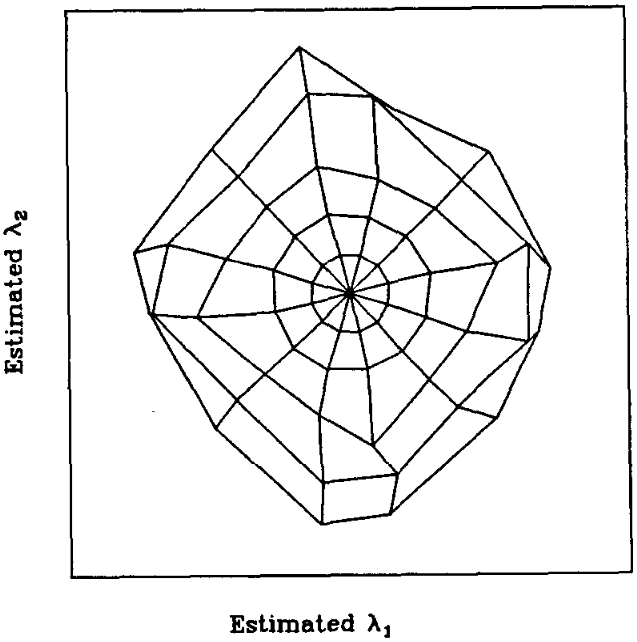

Figure 6.4f The λ map is a two dimensional summary of the data. It resembles a stereographic map of the world.

is placed at the north pole. (If we placed a point at the south pole the principal surface procedure would never move from the starting plane, which is in fact a principal surface.) Figures 6.4b to 6.4d show various stages in the iterative procedure, and figure 6.4e shows another view of the final surface. Figure 6.4f is a parameter map of the two dimensional $\hat{\lambda}$. It resembles a stereographic map of the earth. (A stereographic map is obtained by placing the earth, or a model thereof, on a piece of paper. Each point on the surface is mapped onto the paper by extrapolating the line segment joining the north pole to the point until it reaches the paper.) Points in the southern hemisphere are mapped on the inside of a circle, points in the northern hemisphere on the outside, and there is a discontinuity at the north pole. Points close together on this map are close together in the original space, but the converse is not necessarily true. This map provides a two dimensional summary of the original data. If we are presented with any new observations, we can easily locate them on the map by finding their closest position on the surface.

Example 6.5. One dimensional color data.

This almost real data example illustrates:

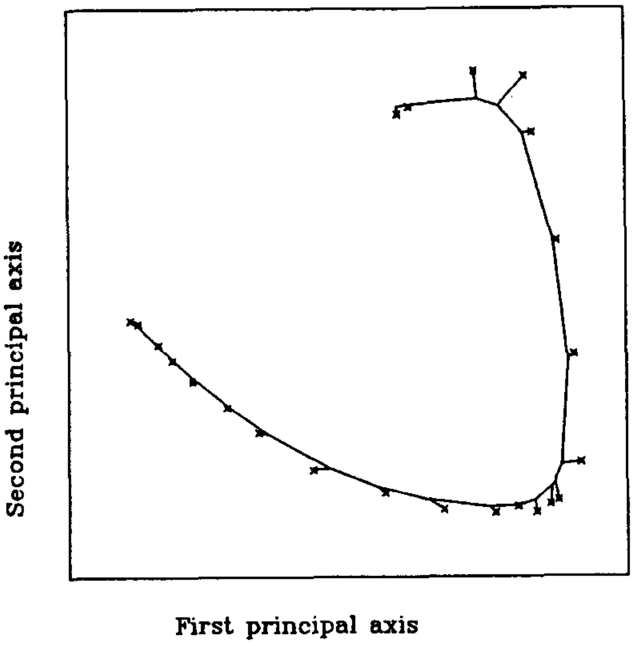

Figure 6.5a The 4 dimensional color data projected onto the first principal component plane. The principal curve, found in the original four-space, is also projected onto this plane.

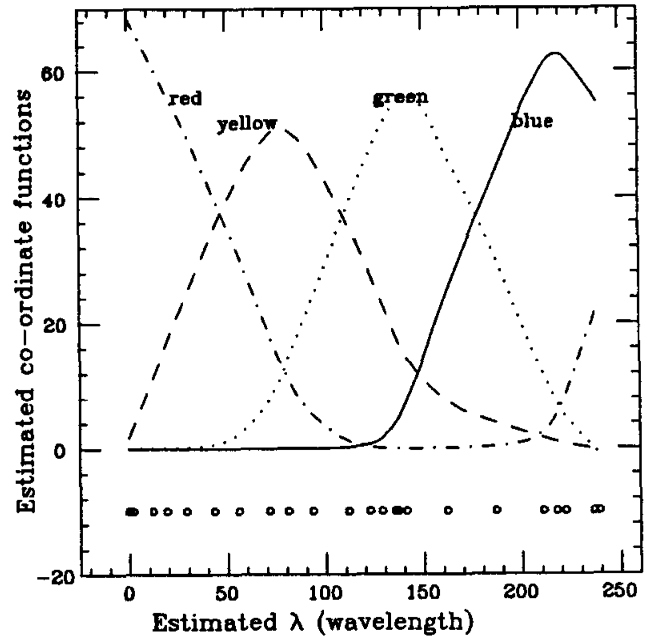

Figure 6.5b The estimated co-ordinate functions plotted against the arc length of the principal curve. This $\hat{\lambda}$ will be monotone with the true wavelength.

- A principal curves in 4-space, and
- a one dimensional MDS example.

These data were used by Shepard and Carroll (1966) (who cite the original source as Boynton and Gordon (1965)) to illustrate a version of their parametric data representation techniques called proximity analysis. We give more details of this technique in chapter 7.

Each of the 23 observations represents a spectral color at a specific wavelength. Each observation has 4 psychological variables associated with it. They are the relative frequencies with which 100 observers named the color *blue, green, yellow* and *red*. As can be seen in figure 6.5a, there is very little error in this data, and it is one dimensional by construction. Since the color changes slowly with wavelength, so should these relative frequencies, and they should thus fall on a one dimensional curve, as they do. The data, by construction lies in a 3 dimensional simplex since the four variables add up to 1. The pictures we show are projections of this simplex onto the 2-D subspace spanned by the first two linear principal components. Figure 6.5a shows the solution curve and figure 6.5b shows the recovered parameters and co-ordinate functions. This solution is in qualitative agreement with the data and with the solution produced by Shepard and Carroll.

Example 6.6. Lipoprotein data.

This real data example illustrates:

- A principal surface in 3 space with some interpretations,
- a principal curve suggested by the surface, and
- co-ordinate plots for surfaces.

Williams and Krauss (1982) conducted a study to investigate the inter-relationships between the serum concentrations of lipoproteins at varying densities in sedentry men. We focus on a subset of the data, and consider the serum concentrations of LDL 3-4 (Low Density Lipoprotein with flotation rates between $S_f3 - 4$), LDL 7-8, and HDL 3 (High Density Lipoprotein) in the sample of 81 men. Figures 6.6a-d are different views of the principal surface found for the data. Quantitively this surface explains 97.4% of the variability in the data, and accounts for 80% of the residual variance unexplained by the principal component plane. Qualitatively, we see that the surface has interesting structure in only two of the co-ordinates, namely LDL 3-4 and LDL 7-8. We can infer from the the surface that the bow shaped relationship between these two variables does not change for varying levels of HDL 3. It exhibits an independent behaviour. We have included a co-ordinate plot (figure 6.6e) of the estimated co-ordinate function for the variables LDL 7-8 which helps confirm this claim. The relationship between LDL 7-8 and $(\hat{\lambda}_1, \hat{\lambda}_2)$ depends mainly on the level of $\hat{\lambda}_1$. Similar information is conveyed by the other co-ordinate plots, or can be seen from the estimated surface directly. This suggests a model of the form

$$\begin{pmatrix} \text{LDL 3-4} \\ \text{LDL 7-8} \\ \text{HDL 3} \end{pmatrix} = \begin{pmatrix} f_1(\lambda_1) \\ f_2(\lambda_1) \\ f_3(\lambda_2) \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}.$$

As specified λ_2 is confounded with HDL 3, and is thus unidentifiable. We need to estimate the first two components of the model. This is a principal curve model, and figure 6.6f shows the estimated curve. It exhibits the same dependence between LDL 7-8 and LDL 3-4 as did the surface. The curve explains 92.6% of the variance in the two variables, whereas the principal component line explains only 80%.

Williams and Krauss performed a similar analysis looking at pairs of variables at a time. We discuss their techniques in chapter 7. Their results are qualitatively the same as ours for the LDL pair.

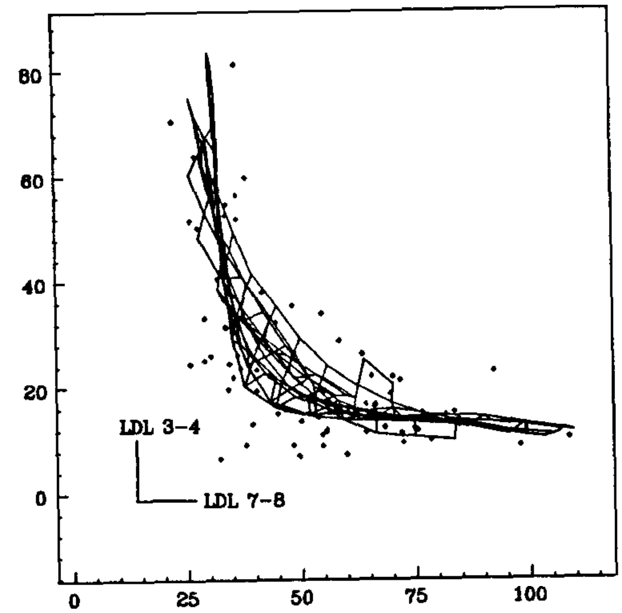

Figure 6.6a The principal surface for the serum concentrations LDL 7-8, LDL 3-4 and HDL 3 in a sample of 81 sedentary men. Variable HDL 3 is into the page.

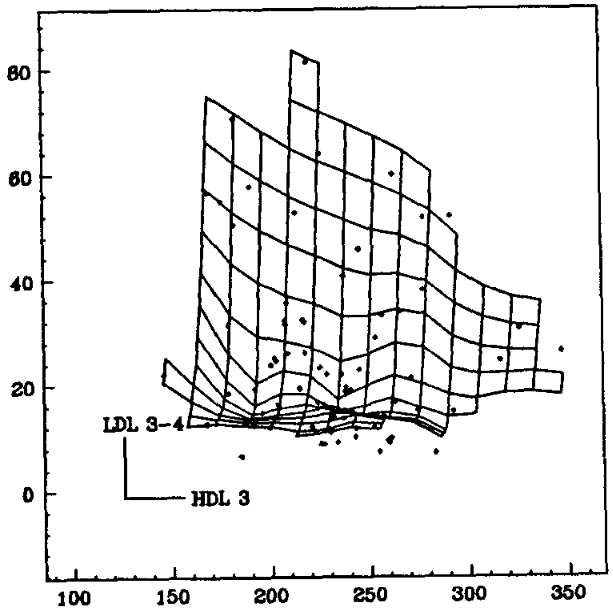

Figure 6.6b The principal surface as in figure 6.6a from a different viewpoint. Variable LDL 7-8 is into the page.

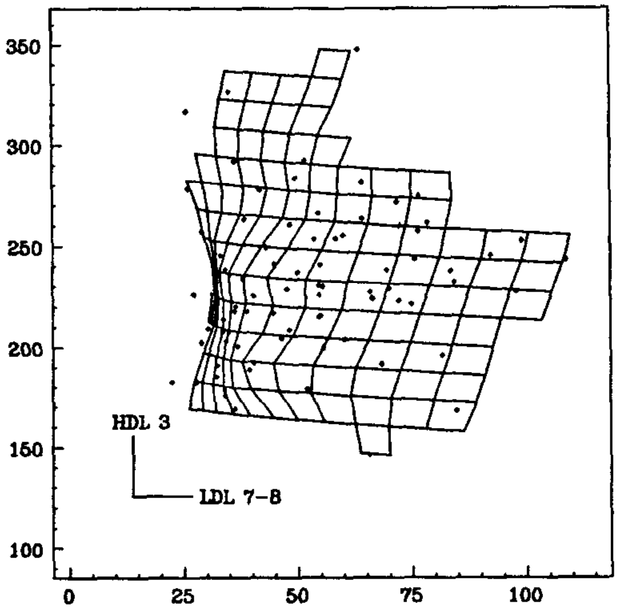

Figure 6.6c The principal surface as in figure 6.6a from a different viewpoint. Variable LDL 3-4 is into the page.

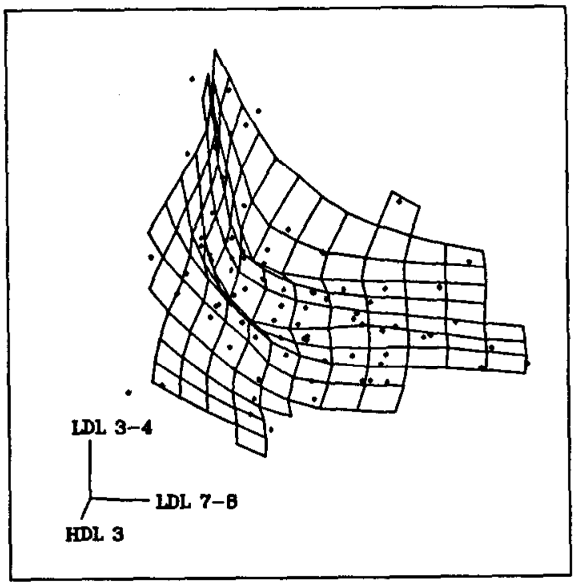

Figure 6.6d The principal surface as in figure 6.6a from a slightly oblique perspective.

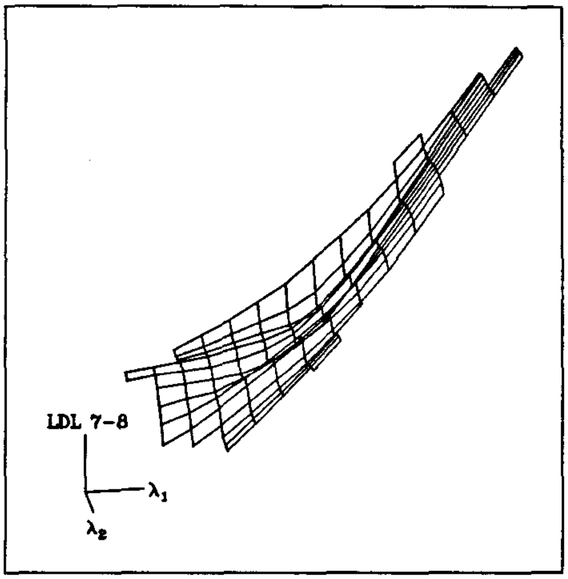

Figure 6.6e The estimated co-ordinate function for LDL 7-8 versus $\hat{\lambda}$. $\hat{\lambda}_2$ has little effect.

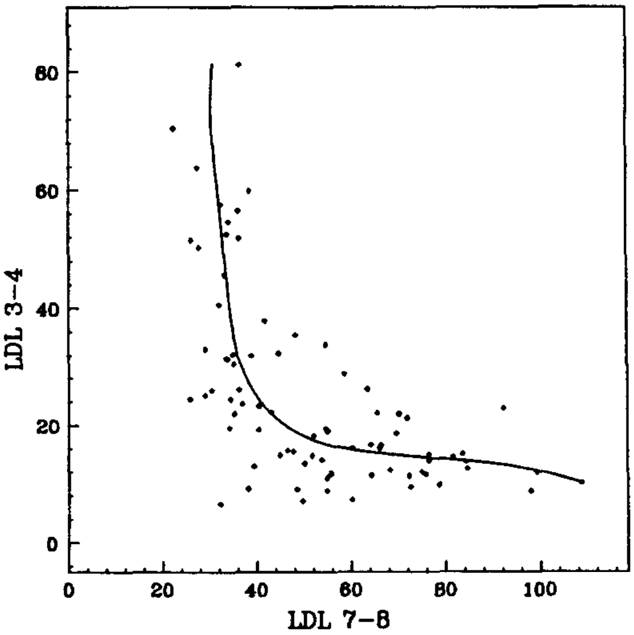

Figure 6.6f The principal curve for the serum concentrations LDL 7-8 and LDL 3-4 in a sample of 81 sedentry men.