Principal Curves Revisited

Robert Tibshirani
Department of Preventive Medicine and Biostatistics
and
Department of Statistics
University of Toronto

Abstract

A principal curve (Hastie and Stuetzle, 1989) is a smooth curve passing through the "middle" of a distribution or data cloud, and is a generalization of linear principal components. We give an alternative definition of a principal curve, based on a mixture model. Estimation is carried out through an EM algorithm. Some comparisons are made to the Hastie-Stuetzle definition.

1 Introduction

Suppose we have a random vector $\mathbf{Y} = (Y_1, Y_2, \ldots Y_p)$ with density $g_{\mathbf{Y}}(y)$. How can we draw a smooth curve $\mathbf{f}(s)$ through the "middle" of the distribution of $\mathbf{Y}$? Hastie (1984) and Hastie and Stuetzle (1989) (hereafter HS) proposed a generalization of linear principal components known as *principal curves*. Let $\mathbf{f}(s) = (f_1(s), \ldots f_p(s))$ be a curve in R^p parametrized by a real argument s and define the *projection index* $s_{\mathbf{f}}(\mathbf{y})$ to be the value of s corresponding to the point on $\mathbf{f}(s)$ that is closest to $\mathbf{y}$. Then HS define a *principal curve* to be a curve satisfying the *self-consistency* property

$$\mathbf{f}(s) = \mathrm{E}(\mathbf{Y}|s_{\mathbf{f}}(\mathbf{y}) = s) \tag{1}$$

If we think of projecting each point $\mathbf{Y}$ to the curve $\mathbf{f}(s)$, this says that $\mathbf{f}(s)$ is the average of all points that project to it. In a sense, $\mathbf{f}(s)$ passes through the "middle" of the distribution of $\mathbf{Y}$. This is illustrated in Figure 1.

HS showed that a principal curve is a critical point of the squared distance function $\mathrm{E}\sum_1^p (Y_j - \mathbf{f}_j(s))^2$, and in this sense, it generalizes the minimum distance property of linear principal components.

HS proposed the following alternating algorithm for determining $\mathbf{f}$ and s:

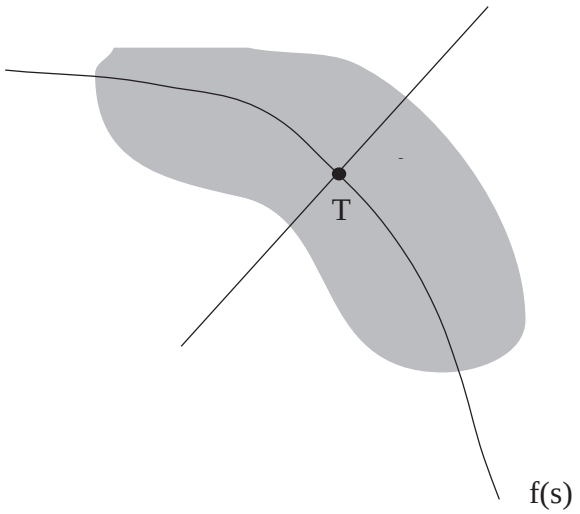

Figure 1: Principal curve schematic. The grey area is the distribution of $\mathbf{Y}$ and the solid curve is a principal curve of this density. Given any point on the curve (e.g. "T"), the principal curve is such that the average of all points that project to T (solid line) is T itself.

HS principal curve algorithm

a) Start with $\mathbf{f}(s) = \mathrm{E}(\mathbf{Y}) + \mathbf{d}s$ where $\mathbf{d}$ is the first eigenvector of the covariance matrix of $\mathbf{Y}$ and $s = s_{\mathbf{f}}(\mathbf{y})$ for each $\mathbf{y}$.

b) Fix s and minimize $\mathrm{E}\|\mathbf{Y} - \mathbf{f}(s)\|^2$ by setting $\mathbf{f}_j(s) = \mathrm{E}(Y_j | s_{\mathbf{f}}(\mathbf{y}) = s)$ for each j.

c) Fix $\mathbf{f}$ and set $s = s_{\mathbf{f}}(\mathbf{y})$ for each $\mathbf{y}$.

d) Iterate steps b and c until the change in $\mathrm{E}\|\mathbf{Y} - \mathbf{f}(s)\|^2$ is less than some threshold.

In the data case, the conditional expectations in step (b) are replaced by a *smoother* or *nonparametric regression estimate*. HS use locally weighted running lines or cubic smoothing splines. In step (c) $s_{\mathbf{f}}(\mathbf{y})$ is found by projecting $\mathbf{y}$ (numerically) to the curve $\mathbf{f}$.

While this definition seems appealing, HS note the following (somewhat unsettling) property. Suppose $\mathbf{Y}$ satisfies

$$Y_j = f_j(S) + \epsilon_j; \quad j = 1, 2, \ldots p$$

where S and $\epsilon_j, j = 1, 2 \ldots p$ are independent with $\mathrm{E}(\epsilon_j) = 0$ for all j. Then $\mathbf{f} = (f_1, f_2, \ldots f_p)$ is not in general a principal curve of the distribution of $\mathbf{Y}$.

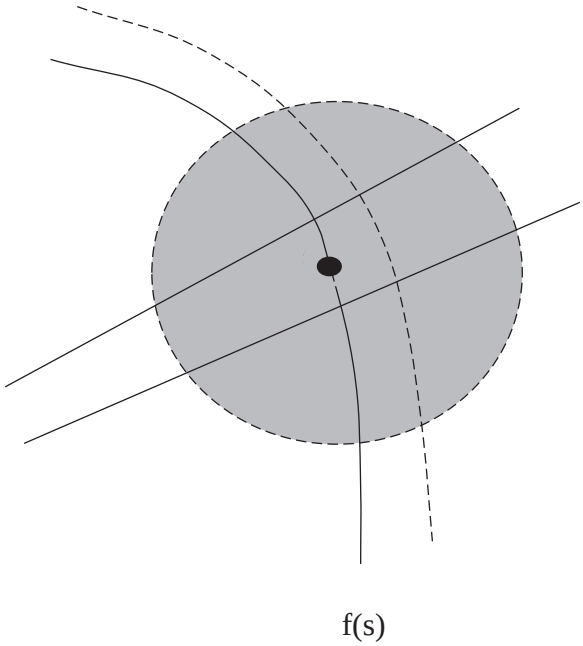

Figure 2: Bias in principal curve: the solid curve is the generating function, and the grey circle indicates a bivariate normal distribution centred at the black dot. The two straight lines indicate the part of the distribution that projects to the target point (black dot). Since there is more mass (between the lines) outside of the arc than inside, the principal curve (dashed curve) will fall outside of the generating curve. The resulting principal curve (broken curve) falls outside of the generating curve.

They give as an illustration the situation in Figure 2. S is uniform on the arc of a circle (solid curve), and the errors are circular normal. The two straight lines indicate the part of the distribution that projects to the target point (black dot). Since there is more mass (between the lines) outside of the arc than inside, the principal curve (dashed curve) will fall outside of the generating curve. This continues to hold in the limit, as the distance between the two straight lines goes to zero. As HS note, however, the estimation bias in the data case tends to cancel out this model bias, so that it is not clear whether this bias is a real problem in practice.

In this paper we view the principal curve problem in terms of a mixture model. This leads to a different definition of principal curves and a new algorithm for their estimation. The new definition does not share the difficulty mentioned above.

2 Principal curves for distributions

Let $\mathbf{Y} = (Y_1, Y_2, \ldots Y_p)$ be a random vector with density $g_{\mathbf{Y}}(y)$. In order to define a principal curve, we imagine that each $\mathbf{Y}$ value was generated in two stages: 1) a latent variable S was generated according to some distribution $g_S(s)$, and 2) $\mathbf{Y} = (Y_1, \ldots, Y_p)$ was generated from a conditional distribution $g_{\mathbf{Y}|s}$ having mean $\mathbf{f}(S)$, a point on a curve in R^p, with $Y_1, \ldots Y_p$ conditionally independent given s. Hence we define a *principal curve* of $g_{\mathbf{Y}}$ to be a triplet $\{g_S, g_{\mathbf{Y}|S}, \mathbf{f}\}$ satisfying the following conditions:

I. $g_S(s)$ and $g_{\mathbf{Y}|S}(y|s)$ are consistent with $g_{\mathbf{Y}}(y)$, that is, $g_{\mathbf{Y}}(y) = \int^y g_{\mathbf{Y}|s}(y|s) g_S(s) ds$

II. $Y_1, \ldots Y_p$ are conditionally independent given s.

III. $\mathbf{f}(s)$ is a curve in R^p parametrized over $s \in ?$, a closed interval in R^1, satisfying $\mathbf{f}(s) = \mathrm{E}(\mathbf{Y}|S = s)$

Notice that this definition involves not only a curve $\mathbf{f}$ but a decomposition of $g_{\mathbf{Y}}$ into g_S and $g_{\mathbf{Y}|S}$. From a conceptual standpoint, assumption (II) is not really necessary; however it is an important simplifying assumption for the estimation procedures described later in the paper.

One obvious advantage of this new definition is that it does not suffer from the problem of Figure 2. Suppose we define a distribution $g_{\mathbf{Y}}$ by $g_{\mathbf{Y}}(y) = \int^y g_{\mathbf{Y}|s}(y|s) g_S(s) ds$ where a latent variable S has density g_s, and $\mathbf{Y} \sim g_{\mathbf{Y}|s}$ with mean $\mathbf{f}(s)$. Then by definition the generating triplet $\{g_S, g_{\mathbf{Y}|S}, \mathbf{f}\}$ satisfies properties I, II, and III and hence is a principal curve. Thus for example the solid curve in Figure 2 is a principal curve according to I, II, and III.

When do the HS definition (1) and the new definition agree? In general they are not the same except in special cases. Suppose $S \sim g_S$, $\mathbf{f}(s)$ is linear, and

the support of the distribution $g_{\mathbf{Y}|s}$ is only on the projection line orthogonal to $\mathbf{f}(s)$ at s. $(\star)$

Then the $\mathbf{Y}$ values generated from the point $S = s$ are exactly the $\mathbf{Y}$ values on the projection line orthogonal to $\mathbf{f}(s)$ at s, and therefore

$$\mathrm{E}(\mathbf{Y}|S = s) = \mathrm{E}(\mathbf{Y}|\mathbf{s_f}(\mathbf{y}) = s)$$

The assumption $(\star)$ is somewhat unnatural, however, and violates our earlier assumption (II) .

There are other special cases for which a curve $\mathbf{f}(s)$ is a principal curve under both definitions. One can check that for a multivariate normal distribution, the principal components are principal curves under the new definition. HS note that for a multivariate normal distribution, the principal components are principal curves under their definition as well.

3 Principal curves for datasets

Suppose we have observations $\mathbf{y}_i = (y_{i1}, \ldots y_{ip})$, $i = 1, 2, \ldots n$ and unobserved latent data $s_1, \ldots s_n$. We assume the model of the previous section

$$s_i \sim g_s(S); \;\; \mathbf{y}_i \sim g_{\mathbf{Y}|s_i}; \;\; \mathbf{f}(s) = \mathrm{E}(\mathbf{Y}|s) \tag{2}$$

with $(y_{i1}, \ldots y_{ip})$ conditionally independent given s_i.

First notice that if we consider the unobserved values $\mathbf{s} = (s_1, \ldots s_n)$ as fixed parameters rather than random variables, and assume that the conditional distributions $g_{Y|S}$ are normal with equal known variance, then is easily seen that the HS principal curve algorithm for datasets can be derived as a penalized maximum likelihood estimator. More specifically, maximization of the penalized least squares criterion

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(y_{ij} - f_j(s_i))^2 + \sum_{j=1}^{p}\lambda_j \int_0^1 [f_j(s)^{''}]^2 ds \tag{3}$$

leads to the HS algorithm with cubic smoothing splines for estimating each f_j. The parameter $\lambda > 0$ governs the tradeoff between fit and smoothness of the coordinate functions. Details are given in HS section 5.5. We note that in problems where the number of "nuisance" parameters goes to infinity with the sample size, maximization of the likelihood over all of the parameters can lead to inconsistent or inefficient estimates.

Our approach, on the other hand, is to work with the likelihood implied by model (2). The form of $g_S(s)$ is left completely unspecified but $g_{\mathbf{Y}|S}$ is assumed to be some parametric family. We allow additional parameters $\mathbf{\Sigma}(\mathbf{s})$ in the specification of $g_{\mathbf{Y}|S}$ and let the complete set of unknowns be $\theta = \theta(s) = (\mathbf{f}(s), \mathbf{\Sigma}(s))$.

Consider maximum likelihood estimation $\mathbf{f}(s)$ and $\mathbf{\Sigma}(s)$, and nonparametric maximum likelihood estimation of g_S. The log-likelihood has the form of a mixture:

$$\ell(\theta) = \sum_{1}^{n} \log \int g_{Y|S}(\mathbf{y}_i|\theta) g_S(\mathbf{s}) d\mathbf{s}, \tag{4}$$

A general theorem on mixtures given by Lindsay (1983) implies that for fixed $\mathbf{f}(s)$ and $\mathbf{\Sigma}(s)$, the nonparametric maximum likelihood estimate of the mixing density g_S is discrete with at most n support points. Denote these support points by $a_1, a_2, \ldots a_n$.

Our approach to maximization of ℓ is via the EM algorithm (Dempster et al. 1977, sect. 4.3). EM uses the complete data log-likelihood

$$\ell_0(\theta) = \sum_{1}^{n} \log g_{Y|\mathbf{s}}(\mathbf{y}_i|\theta(s_i)) + \sum_{1}^{n} \log g_S(s_i)$$

The "E step" starts with a value $\mathbf{f}^0$ and computes the function

$$Q(\theta|\theta^0) = E\{\ell_0(\theta)|y, \theta^0\}$$

where y denotes the observations $(\mathbf{y}_i), i = 1, \ldots n$. Q is considered a function of θ with θ^0 fixed. The M step maximizes $Q(\theta|\theta^0)$ over θ to give θ^1 and the process is iterated until convergence.

Let $w_{ik} = Pr(s_i = a_k|y, \theta^0)$, $v_k = g_S(a_k) = Pr(s = a_k)$.

We may write Q as

$$Q(\theta|\theta^0) = \sum_{i=1}^{n}\sum_{k=1}^{n} w_{ik} \log g_{Y|S}(\mathbf{y}_i|\theta(a_k)) + \sum_{i=1}^{n}\sum_{k=1}^{n} w_{ik} \log v_k$$

Now by Bayes theorem

$$w_{ik} \sim g_{Y|S}(\mathbf{y}_i|\theta(a_k)) \prod_{h\neq i} g_Y(\mathbf{y}_h) v_k \tag{5}$$

The quantity $g_Y(\mathbf{y}_h)$ is computed from $g_Y(\mathbf{y}_h) = \sum_{k=1}^{n} g_{\mathbf{Y}|S}(\mathbf{y}_h|a_k, \theta^0)v_k$

A Gaussian form for $g_{\mathbf{Y}|S}$ is most convenient and gives

$$\begin{aligned} g_{Y|S}(\mathbf{y}_i|\theta(s)) &= \prod_{j=1}^{p} \phi_{f_j(s),\sigma_j(s)}(y_{ij}) \text{ where} \\ \phi_{f_j(a_k),\sigma_j(a_k)}(y) &= \frac{1}{\sigma_j(a_k)\sqrt{2\pi}} \exp[-(y - f_j(a_k))^2/2\sigma_j^2(a_k)] \end{aligned} \tag{6}$$

Maximizing $Q(\theta|\theta^0)$ gives

$$\begin{aligned} \hat{f}_j(a_k) &= \frac{\sum_{i=1}^{n} w_{ik} y_{ij}}{\sum_{i=1}^{n} w_{ik}} \\ \hat{\sigma}_j^2(a_k) &= \frac{\sum_{i=1}^{n} w_{ik} (y_{ij} - \hat{f}_j(a_k))^2}{\sum_{i=1}^{n} w_{ik}}; \quad j = 1 \ldots p \\ \hat{v}_k &= \frac{1}{n}\sum_{i=1}^{n} w_{ik} \end{aligned} \tag{7}$$

The first expression in (7) says that $\hat{f}_j(a_k)$ is a weighted average of $\{y_{ij}, i = 1, 2, \ldots n\}$. The weights are the relative probability under the current model that $s = a_k$ gave rise to y_{ij}. If σ_js are equal, the weight is a function of the Euclidean distance from $\mathbf{y}$ to $\mathbf{f}(a_k)$.

These equations are certainly not new: they are EM steps for fitting a multivariate normal mixture model. See for example Titterington et al. (1985, page 86-89). Notice that the log-likelihood is not a function of the support values $a_1, a_2, \ldots a_n$ but only the means $f_j(a_k)$ and variances $\sigma_j^2(a_k)$ at each of these points.

As an example, we generated 100 observations from the circle model

$$\begin{aligned} Y_1 &= 3\sin(\lambda) + .5\epsilon_1 \\ Y_2 &= 3\cos(\lambda) + .5\epsilon_2 \end{aligned}$$

where $\lambda \sim U[\pi/4, 5\pi/4]$ and $\epsilon \sim N(0,1)$. The results of the first 18 iterations of the EM procedure (showing every second iteration) are in Figure 3.

The procedure does a fair job of approximating the structure of the data. Although the approximating points seem to follow a fairly smooth path, there is no guarantee that this will happen in general. We show how to rectify this in the next section.

4 Multiple minima and regularization

Consider the log-likelihood (4) in the Gaussian case (6). As mentioned in section 3, a theorem of Lindsay (1983) tells us that for fixed $\mathbf{f}(s)$ and $\mathbf{\Sigma}(\mathbf{s})$, the nonparametric maximum likelihood estimate of the mixing distribution g_S is discrete with at most n support points. We note the following:

a) The log-likelihood has a global maximum of $+\infty$ when $\hat{\mathbf{f}}(s)$ is any curve that interpolates the data, $\hat{\mathbf{f}}(a_k) = \mathbf{y}_k,\ k = 1, 2, \ldots n,\ Pr(S = a_k) > 0,\ k = 1, 2, \ldots n$ and $\hat{\sigma}_j^2(a_k) = 0,\ k = 1, 2, \ldots n$.

b) For any candidate solution $\tilde{\mathbf{f}}$, the log-likelihood is only a function of the values of $\tilde{\mathbf{f}}$ at the support points with positive probability. Hence any curve that agrees with $\tilde{\mathbf{f}}$ at these points gives the same value of the log-likelihood.

In view of (a), the global maximum is uninteresting and we might hope that the EM algorithm converges instead to an "interesting" local maximum. This happened in Example 1 above, but may or may not occur in general.

A remedy for these problems is add a regularization component to the log-likelihood. We seek to maximize the penalized log-likelihood

$$j(\theta) = \ell(\theta) - (c_2 - c_1)\sum_{j=1}^{p} \lambda_j \int_{c_1}^{c_2} [f_j(s)'']^2 ds \tag{8}$$

For sufficiently large λ_j, the smoothness penalty forces each f_j to be smooth so that the global maximum does not occur for $\sigma_j^2(a_k) \to 0$. We assume that $\mathbf{f}$ is a parametrized curve over a finite interval, and the values c_1 and c_2 are the endpoints of the smallest interval containing the support of $g_S(s)$. (See Remark A below for more details). The corresponding Q function for the application of the EM algorithm is

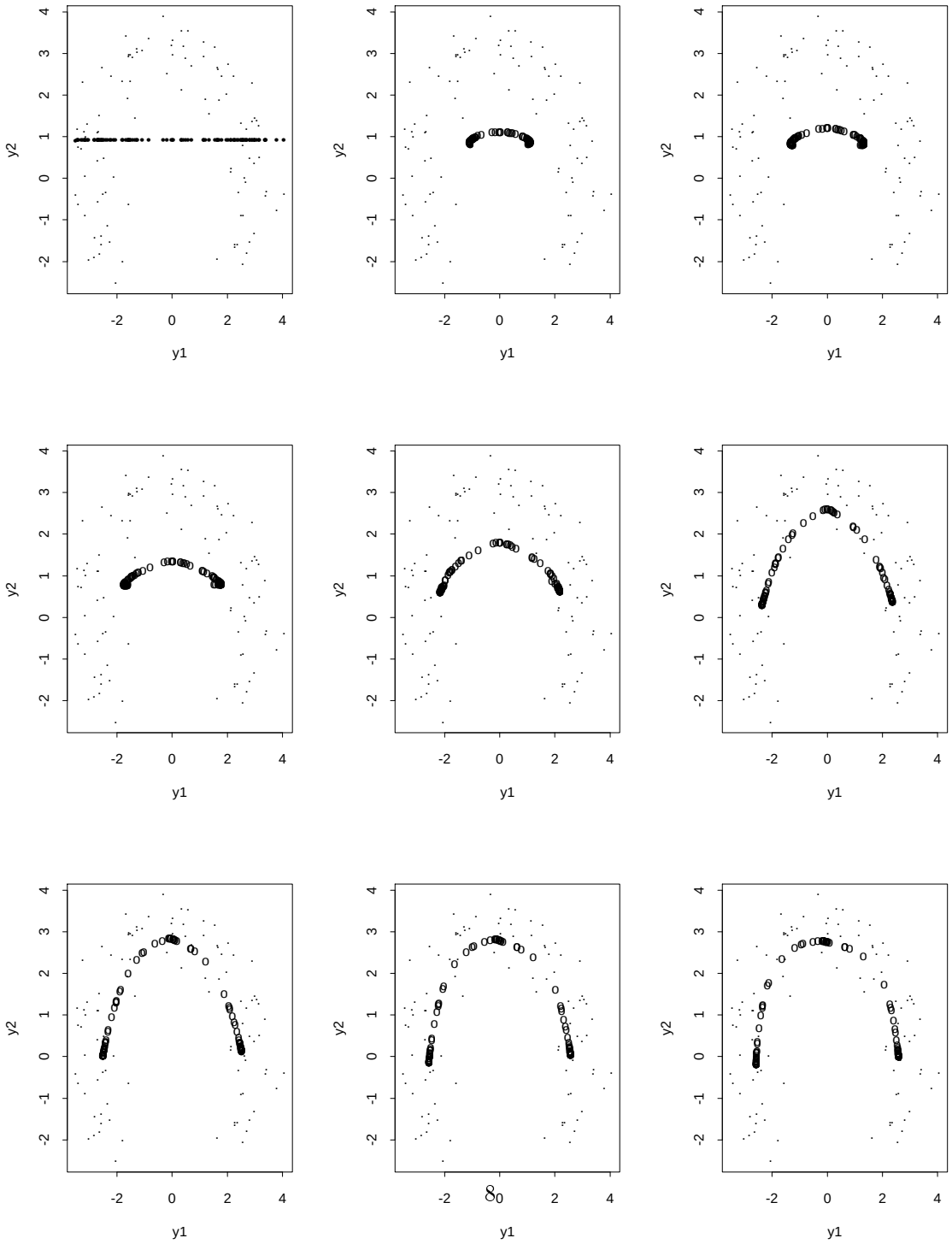

Figure 3: Circle approximations: dots are the data points, circles are the successive approximations using the EM procedure (7).

$$Q(\theta|\theta^0) = \sum_{i=1}^{n}\sum_{k=1}^{n} w_{ik} \log g_{Y|S}(\mathbf{y}_i|\theta(a_k)) + \sum_{i=1}^{n}\sum_{k=1}^{n} w_{ik} \log v_k - (c_2 - c_1)\sum_{j=1}^{p} \lambda_j \int_{c_1}^{c_2} [f_j(s)^{''}]^2 ds$$

The solutions are easily found in closed form. The last two equations in (7) give $\hat{\sigma}_j^2$ and $\hat{v}_k$; let $b_k = \sum_{i=1}^{n} w_{ik}$, D be a diagonal matrix with entries $b_1, b_2, \ldots b_n$ and $\bar{\mathbf{y}}_j$ be an n-vector with kth component $\sum_{i=1}^{n} w_{ik} y_{ij}$. Then

$$\hat{f}_j = (D + (c_2 - c_1)\lambda_j K_j)^{-1} D\{D^{-1}\bar{\mathbf{y}}_j\} \tag{9}$$

The matrix K_j is the usual quadratic penalty matrix associated with a cubic smoothing spline (see e.g. Hastie and Tibshirani, 1990 sec. 2.10). Equation (9) says that $\hat{f}_j$ is obtained by applying a weighted cubic smoothing to the quantity $D^{-1}\bar{\mathbf{y}}_j$; this latter quantity has elements $\sum_{i=1}^{n} w_{ik} y_{ij} / \sum_{i=1}^{n} w_{ik}$ which is exactly the estimate that appears in the (unregularized) version (7).

The only remaining obstacle is how to find the maximizing values $\hat{a}_1, \ldots \hat{a}_n$, the location of the support points. We use a Newton-Raphson procedure for this. Note however that full maximization of $Q(\theta|\theta^0)$ with respect to $\mathbf{f}$, σ_j^2, v_k and a_k would involve iteration between (9), the last two equations in (7) and the Newton-Raphson procedure for the a_k's, and this is computationally unattractive. We therefore seek only to increase $Q(\theta|\theta^0)$ at each iteration: this is called a generalized EM algorithm by Dempster el al. (1977) and the theory of that paper and Wu (1983) guarantees its convergence to a fixed point of the gradient equations. To ensure an increase in $Q(\theta|\theta^0)$, we apply (7) together with one Newton-Raphson step for the a_k's, using step size halving if necessary.

We summarize the algorithm below.

A new principal curve algorithm

a) Start with $\hat{\mathbf{f}}(s) = \mathrm{E}(\mathbf{Y}) + \mathbf{d}s$ where $\mathbf{d}$ is the first eigenvector of the covariance matrix of $\mathbf{Y}$, $\hat{v}_k = 1/n$. and $S = \hat{a}_1, \hat{a}_2, \ldots \hat{a}_n$ the projected values onto this line.

b) Compute

$$\hat{w}_{ik} \sim g_{Y|S}(\mathbf{y}_i|a_k, \theta) \prod_{h \neq i} g_Y(\mathbf{y}_h) v_k$$

under the Gaussian assumption (6) and normalize so that $\sum_{k=1}^{n} \hat{w}_{ik} = 1$.

c) Fix $\hat{\mathbf{f}}, \hat{\sigma}_j^2$, and $\hat{v}_k$ and apply a Newton-Raphson step to obtain a new set of support points $\hat{a}_1, \hat{a}_2, \ldots \hat{a}_n$. Reparametrize to arc-length parametrization.

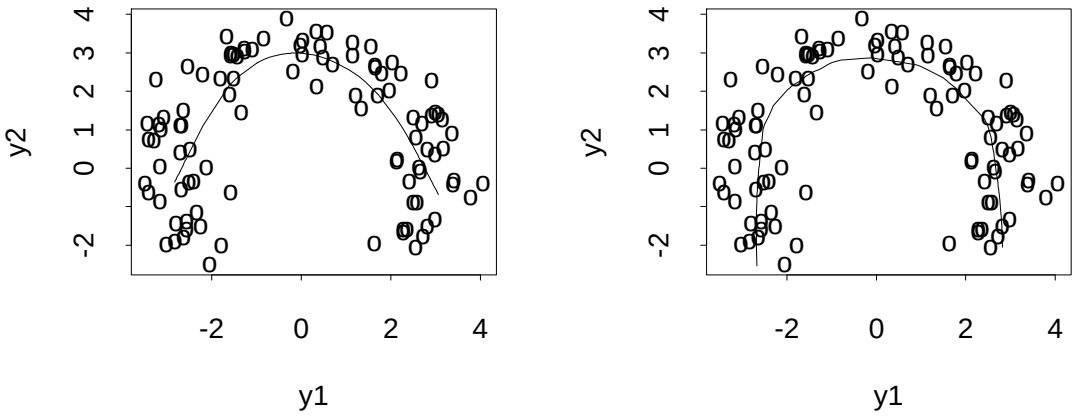

Figure 4: On the left is the result of the new principal curve procedure, applied to the circle example. On the right is the result of the HS principal curve procedure

d) For $k = 1, 2, \ldots n$ compute

$$\begin{aligned} \hat{f}_j &= (D + \lambda_j K_j)^{-1} D\{D^{-1}\bar{\mathbf{y}}_j\} \\ \hat{\sigma}_j^2(a_k) &= \frac{\sum_{i=1}^n \hat{w}_{ik}(y_{ij} - \hat{f}_j(a_k))^2}{\sum_{i=1}^n \hat{w}_{ik}}; \quad j = 1 \ldots p \\ \hat{v}_k &= \frac{1}{n}\sum_{i=1}^n \hat{w}_{ik} \end{aligned}$$

e) Iterate steps (b), (c), and (d) until the change in the log-likelihood is less than some threshold.

Step (b) is the E step of the EM algorithm, while (c) and (d) are the M step.

Figure 4 shows, for the earlier circle example, the result of this procedure (left panel) and the HS principal curve procedure (right panel). Notice that the HS estimate extends further in the bottom arms of the data cloud. This is to be expected: under model (2) specifying bivariate normal errors around the generating curve, there is insufficient data to determine the bottom arms of the generating curve.

REMARK A. The distribution $g_S(s)$ is only meaningful relative to some fixed parametrization of $\mathbf{f}(s)$. As HS (Sec. 5.5) point out, it is possible to find a curve defined over an arbitrarily large interval such that the curvature measure $\int [f_j''(s)]^2$ is arbitrarily small, and such a curve can visit every data point. To avoid this anomaly, HS restrict the curve to be defined over the unit interval, as in (3). However they do not say which parametrization they use over the unit interval. We include the additional multiplier $c_2 - c_1$ to account for the length of $\mathbf{f}(s)$; the effect of this is to make the penalty term invariant to scale changes $s \to cs$. After each iteration, the $s_1, s_2, \ldots s_n$ values are rescaled so that they correspond to arc-length along the current curve. As pointed out by Trevor Hastie (personal communication) a unit-speed parametrized curve ($\|\mathbf{f}'(s)\| = 1$), cannot be the minimizer of $Q(\theta|\theta_0)$ since a curve with cubic spline coordinate functions does not have unit speed. Thus we are not able to make explicit the parametrization that is being used here.

REMARK B. We choose the parameters λ_j via a fixed "degrees of freedom" strategy (see Hastie and Tibshirani, 1990 chap. 3). An automatic choice via cross-validation or generalized cross-validation might be useful in practice and is worthy of investigation. Most smoothers allow an automatic choice of smoothing parameter; one simple approach then would be to allow each smoother to choose its smoothing parameter in each coordinate function smoothing operation. However HS demonstrate that this results in a nearly interpolating curve after a number of iterations; they speculate that this may be due to autocorrelation in the errors. Altman (1990) studies this problem in the usual nonparametric regression situation.

REMARK C. The weights (5) and estimates (7) are dependent on the assumed Gaussian form of the distribution of $\mathbf{Y}$ given s. To make the estimates more robust to outliers one could use a resistant form for the weighted mean and variance in (7).

REMARK D. If the variables $Y_1, Y_2, \ldots Y_p$ are measured in different units it would be appropriate, as in principal components analysis, to standardize each variable first. If they are measured in the same units, no standardization is needed. In that instance, the procedure presented here allows for different variances in the different dimensions. The HS procedure seems to assume equal variances: however different variances could be facilitated by using a nonorthogonal projection in step (b) of their algorithm (see section 1)

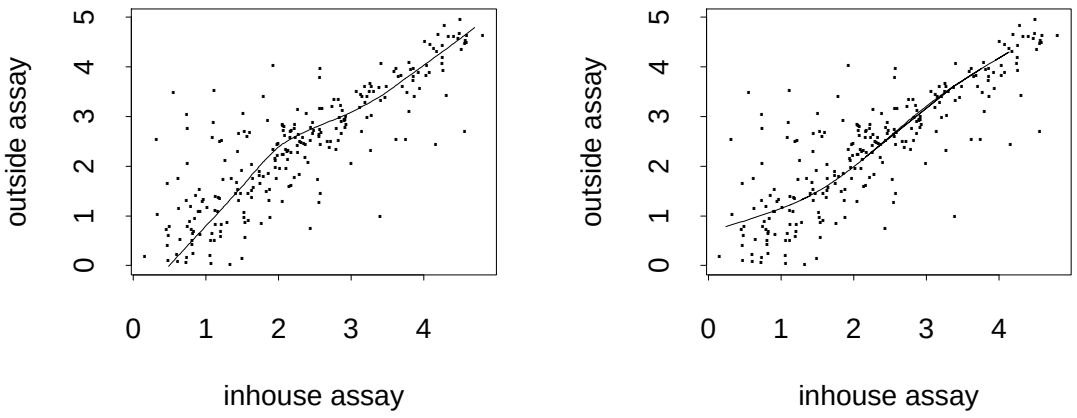

Figure 5: Left panel shows the gold pairs data and principal curve for the HS procedure. Right panel shows the gold pairs data and principal curve from the new principal curve procedure.

5 An Example

HS analyzed some data on the gold content of computer chip waste produced by two laboratories. Two hundred and forty nine samples were each split in two, and assays were carried out by an inhouse lab and an outside lab to determine the gold content. The objective was to investigate which laboratory produces lower or higher gold content measurements. A full description in given in HS (section 7.2). (HS report 250 pairs, but we received only 249 from Trevor Hastie). The assumed model has the form $Y_j = f_j(s) + \epsilon_j$, $j = 1, 2$, with $f_2(s)$ constrained to be the identity function. The only change needed in the estimation procedure is that the smooth of Y_2 versus s is not needed, and replaced by the identity function.

A scatterplot of the (log) measurements is shown in Figure (5) along with the estimated principal curve from the HS algorithm.

As HS report, the outside assay seems to be producing systematically higher gold levels than the inside assay, in the interval 1.4 to 4. However the curve estimated from the procedure of this paper (right panel) suggests that the outside lab gives higher gold content below 1. In this example, then, it seems that the choice of underlying model has an effect on the substantive conclusions.

6 Relationship between the HS and new principal curve algorithms

In this section we explore in detail the relationship between the HS and new algorithm for principal curves, in the data case.

The HS principal curve procedure, when applied to a dataset, uses a nonparametric regression estimate in place of the conditional expectation $\mathrm{E}(Y_j|s)$. One nonparametric regression estimate used by HS is a cubic smoothing spline, resulting from the penalized least squares criterion (3).

The penalized form of the mixture likelihood, on the other hand, leads to the estimate given by (9). This differs from a cubic smoothing spline applied to the observations y_{ij} in that a) the weights w_{ik} are first applied to the y_{ij}, and b) a weighted cubic smoothing spline is used, with weights $\sum_{i=1}^{n} w_{ik}$. Hence the relationship between the two approaches hinges on the form of the weights.

Consider then estimation of $\mathbf{f}(s)$ at a point $s = s_0$. We assume that s measures arc-length along the curve $\mathbf{f}(s)$. The weight in the new principal curve procedure is proportional to the Euclidean distance between $\mathbf{y}$ and $\mathbf{f}(s_0)$ (at least when the σ_js are equal). Now suppose $\mathbf{Y}$ projects to the curve $\mathbf{f}$ at $S = s$. The Euclidean distance between $\mathbf{y}$ and $\mathbf{f}(s_0)$ is not in general the same as $(s - s_0)^2$. However, when $\mathbf{f}(s)$ is a straight line, a simple application of the Pythagorean theorem shows that shows that the two distances are proportional to one another.

The left panel of Figure 6 illustrates this. It shows the contours of constant weight for estimation of $\mathbf{f}(s_0)$ (indicated by "T") and these are also the lines of constant s value. Hence the weighting operation averages the y_{ij}'s by their s value, and can be expressed as a kernel smoothing of the data. The cubic smoothing spline is then applied to the averaged values $\bar{y}_j$, and since a cubic smoothing spline is approximately a (variable) kernel smoother (Silverman, 1984), the result is the approximately the convolution of two kernel smoothers applied to the data, which is just another kernel smoother. Hence when $\mathbf{f}(s)$ is a straight line, the HS and new procedures will give similar results.

The story is different however when $\mathbf{f}(s)$ is curved, as in the right panel of Figure 6.

The contours are curved in the opposite direction to the curvature of $\mathbf{f}(s)$.

In both examples in Figure 6, we have assumed $\sigma_1(a_k) = \sigma_2(a_k)$= constant. If $\sigma_1 \neq \sigma_2$, the contours are tilted.

7 Discussion

We have presented an alternative definition of principal curves based on a mixture model. Although there are theoretical reasons for preferring this definition to that given by HS, we have no evidence that the resulting procedure estimation procedure works any better in practice.

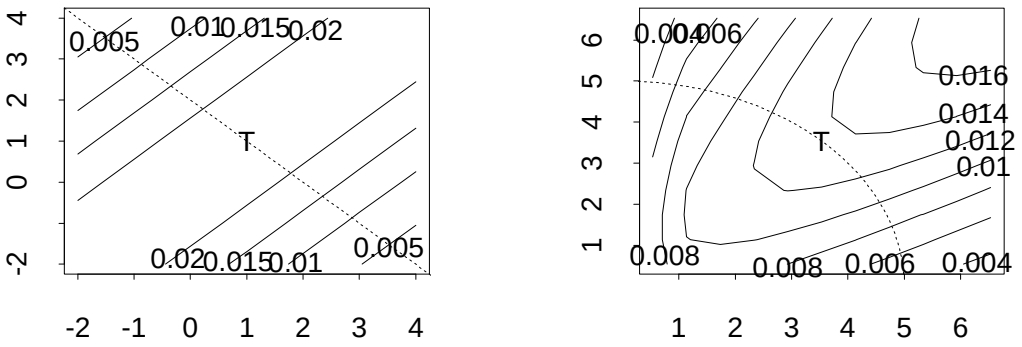

Figure 6: Contours of constant weight for computing the estimate at the target point "T". On the left the underlying function (broken line) is straight, while on the right it is curved.

We have focussed on one-dimensional curves, for which s is real-valued. Extensions to globally parametrized surfaces (vector-valued s) seem straightforward in principle but we have not studied the details. A Bayesian analysis of the model might also prove to be interesting. The Gibbs sampler (Geman and Geman 1984, Tanner and Wong 1987, Gelfond and Smith 1990) could be used to simulate from the posterior distribution. Verdinelli and Wasserman (1990) carry this out for a simpler normal mixture problem and show the structure of the mixture model lends itself naturally to the Gibbs sampler.

ACKNOWLEDGMENTS

This work was stimulated by questions raised by Geoff Hinton and Radford Neal, and I thank them for their questions and some fruitful discussions. I would also like to thank Mike Leblanc, Mary L'Esperance, Trevor Hastie and two referees for helpful comments, and Trevor for a copy of his program for fitting principal curves. Support by the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

References

Altman, N.S. (1990). Kernel smoothing of data with autocorrelated errors. *J. Amer. Statist. Assoc.*, 85:749–759.

Dempster, A.P., Laird, N.M., and Rubin, D.B (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, 39:1–38.

Gelfand, A. E. and Smith. A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc*, 85:398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Hastie, T. (1984). Principal curves and surfaces. Technical report, Stanford University.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84(406):502–516.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Annal of Statistics*, 11:86–94.

Silverman, B.W. (1984). Spline smoothing: the equivalent kernel method. *Annal of Statistics*, 12:898–9164.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.*, 82:528–550.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical analysis of finite mixture distributions.* Wiley, New York.

Verdinelli I, and Wasserman L. (1990). Bayesian analysis of outlier problems, using the gibbs sampler. Technical report, Carnegie-Mellon niversity.

Wu, C.F.J. (1983). On the convergence properties of the em algorithm. *Annal of Statistics*, 11:95–103.