Principal curves revisited

ROBERT TIBSHIRANI

Department of Preventive Medicine and Biostatistics and Department of Statistics,
McMurrich Building, University of Toronto, Toronto M5S IA8, Canada

Received September 1991 and accepted March 1992

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.

Keywords: Principal curves, EM algorithm

1. Introduction

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

$$f(s) = \mathbb{E}(Y | s_f(y) = s)$$  \hspace{1cm} (1)

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

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

HS proposed the following alternating algorithm for determining $f$ and $s$.

**HS principal curve algorithm**

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

(b) Fix $s$ and minimize $\mathbb{E} \| Y - f(s) \|^2$ by setting $f_j(s) = \mathbb{E}(Y_j | s_f(y) = s)$ for each $j$.

(c) Fix $f$ and set $s = s_f(y)$ for each $y$.

(d) Iterate steps (b) and (c) until the change in $\mathbb{E} \| Y - f(s) \|^2$ 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) $s_f(y)$ is found by projecting $y$ (numerically) to the curve $f$.

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

$$Y_j = f_j(S) + \epsilon_j; \hspace{1cm} j = 1, 2, \ldots, p$$

where $S$ and $\epsilon_j, j = 1, 2, \ldots, p$ are independent with $\mathbb{E}(\epsilon_j) = 0$ for all $j$. Then $f = (f_1, f_2, \ldots, f_p)$ is not in general a principal curve of the distribution of $Y$. They give as an illustration the situation in Fig. 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 \( Y = (Y_1, Y_2, \ldots, Y_p) \) be a random vector with density \( g_Y(y) \). In order to define a principal curve, we imagine that each \( Y \) value was generated in two stages: (1) a latent variable \( S \) was generated according to some distribution \( g_S(s) \), and (2) \( Y = (Y_1, \ldots, Y_p) \) was generated from a conditional distribution \( g_{Y|s}(y|s) \) having mean \( f(s) \), a point on a curve in \( \mathbb{R}^p \), with \( Y_1, \ldots, Y_p \) conditionally independent given \( s \). Hence we define a principal curve of \( g_Y \) to be a triplet \( \{g_S, g_{Y|s}, f\} \) satisfying the following conditions:

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

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

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

Notice that this definition involves not only a curve \( f \) but a decomposition of \( g_Y \) into \( g_S \) and \( g_{Y|s} \). 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 Fig. 2. Suppose we define a distribution \( g_Y \) by

\[
g_Y(y) = \int g_{Y|s}(y|s)g_S(s) \, ds
\]

where a latent variable \( S \) has density \( g_S \), and \( Y \sim g_{Y|s} \) with mean \( f(s) \). Then by definition the generating triplet \( \{g_S, g_{Y|s}, f\} \) satisfies properties I, II, and III and hence is a principal curve. Thus for example the solid curve in Fig. 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 \( S \sim g_S, f(s) \) is linear, and

\[
\text{the support of the distribution } g_{Y|s} \text{ is only on the projection line orthogonal to } f(s) \text{ at } s \quad (*)
\]

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

\[
\mathbb{E}(Y | S = s) = \mathbb{E}(Y | S_Y = s)
\]

The assumption (*) is somewhat unnatural, however, and violates our earlier assumption (II).

There are other special cases for which a curve \( f(s) \) 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 \( y_i = (y_{i1}, \ldots, y_{ip}), \quad i = 1, 2, \ldots, n \) and unobserved latent data \( s_1, \ldots, s_n \). We assume the model of the previous section

\[
s_i \sim g_S(S); \quad y_i \sim g_{Y|s_i}; \quad f(s) = \mathbb{E}(Y | s) \quad (2)
\]

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

First notice that if we consider the unobserved values \( s = (s_1, \ldots, s_n) \) as fixed parameters rather than random variables, and assume that the conditional distributions