Bandwidth Selection for Mean-shift based Unsupervised Learning Techniques: a Unified Approach via Self-coverage

Jochen Einbeck *jochen.einbeck@durham.ac.uk*
Department of Mathematical Sciences, Durham University
Durham City, DH1 3LE, UK

Received November 24, 2010.

Abstract

The mean shift is a simple but powerful tool emerging from the computer science literature which shifts a point to the local center of mass around this point. It has been used as a building block for several nonparametric unsupervised learning techniques, such as density mode estimation, clustering, and the estimation of principal curves. Due to the localized way of averaging, it requires the specification of a window size in form of a bandwidth (matrix). This paper proposes to use a so-called self-coverage measure as a general device for bandwidth selection in this context. In short, a bandwidth h will be favorable if a high proportion of data points falls within circles or "hypertubes" of radius h centered at the fitted object. The method is illustrated through real data examples in the light of several unsupervised estimation problems.

Keywords: Mean shift clustering, local principal curves, coverage, goodness-of-fit.

1. Introduction

We are given a $p-$variate random vector $\boldsymbol{X}$ with density function $f(\cdot)$, mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. An unsupervised learning method attempts to gain some knowledge on structure, composition, or certain landmarks of $\boldsymbol{X}$. A special but important case is linear principal component analysis (PCA), which provides a sequence of best linear approximations to $\boldsymbol{X}$. More specifically, the linear $q-$variate subspace that minimizes the expected squared distance to $\boldsymbol{X}$ is given by $\boldsymbol{Z} = (\gamma_1, \ldots, \gamma_q)^T(\boldsymbol{X} - \mu)$, with γ_j, $j = 1, 2, \ldots, q$ being the first $q \leq p$ eigenvectors of Σ. Nonparametric versions of this concept have been developed, which in case of $q = 1$ lead to a principal curve [12]. A less known concept, which can be thought of as representing the case $q = 0$, are "principal points" [10], corresponding to the set $\{v_{(1)}, \ldots, v_{(k)}\} \in \mathbb{R}^p$ which minimizes the expected squared distance to $\boldsymbol{X}$, for (usually) predetermined k. As formalized by Tarpey & Flury [18], objects labelled "principal" share the joint property of self-consistency, where $\boldsymbol{Z}$ is defined to be self-consistent for $\boldsymbol{X}$ if $E(\boldsymbol{X}|\boldsymbol{Z}) = \boldsymbol{Z}$ almost surely. In the context of principal points, this means that each $v_{(j)} \in \boldsymbol{Z}$ is the expectation over all outcomes of $\boldsymbol{X}$ which are closest to $v_{(j)}$ in terms of Euclidean distance. For principal curves, it means that each point on the curve is the expectation over all points that project orthogonally onto that point.

Empirical algorithms for the estimation of principal points and curves have been developed. Flury [11] proposes to estimate principal points through Maximum Likelihood (assuming normality of $\boldsymbol{X}$) or via $k-$means. The latter approach is very intuitive in this framework as the final cluster centers are, by construction, self-consistent. Hastie & Stuetzle's (hereafter: HS) algorithm for the estimation of principal curves can be seen as a variant of $k-$means, in which, starting from the first principal component line, the data are

alternately projected onto the curve, and component–wise smoothed against the projection indices.

This paper deals with a family of unsupervised learning methods which are based on localization (rather than global optimization). The ground for methods of this type was laid by Cheng [2], who introduced the "mean shift" as the shift necessary to move a point $x \in \mathbb{R}^p$ towards the local mean around this point. Cheng showed that, when iterating the mean shift, the resulting sequence of points always converges to a local mode of a kernel density estimator $\hat{f}$ of f. Even more attractively, if one assigns each data point to the local mode to which its mean shift trajectory has converged, this turns into a clustering (or partitioning) technique which does not require pre-specification of the number of cluster centers. However, one still needs to define the size of the considered neighborhood through some bandwidth parameter(s). Some attempts to bandwidth selection in this context have been made. Comaniciu [3] discussed the possibility of using optimal bandwidths (based on asymptotic bias–variance trade–off) originally derived for the purpose of multivariate density estimation [16, 20], but he discarded this idea quickly for practical considerations. The suitability of such density–based optimal bandwidths is also questionable from a conceptual point of view, since the mean shift in general, and the mode selection problem in particular, are more related to the gradient of f than to f itself [5, 19]. An alternative family of methods, which is tailored towards the clustering problem, attempts to maximize the stability of the partitioning under variation of the bandwidth [1, 3, 5]. These methods have been found to work successfully, but here again, it is not clear whether a bandwidth which is optimal for the sake of clustering is necessarily optimal for the problem of finding principal points (in the form of local modes). In fact, in this article we will distinguish between these two cases. The necessity for bandwidth selection rules which are specifically adapted to the mode selection problem was also pointed out on a more theoretical level by Vieu [19].

Recently, the mean-shift has also been employed for the estimation of principal curves: the local principal curve algorithm [6] alternates between a mean shift and a local PCA step. In this context, it was suggested to choose a bandwidth which leads to a fitted principal curve such that a large proportion of data points are not further away from the curve than that very bandwidth. This leads to the idea of extracting bandwidths from a so-called self-coverage curve, but it is still open how to achieve this task algorithmically. This selection rule is provided in this paper, and its applicability is extended beyond the framework of local principal curves.

We invest Section 2 in clarifying the terminology, and reviewing the concepts of mean shift and its derived algorithms. In this course, we will also give a meaning to the yet undefined notion of a "local principal point", and argue that it makes sense to assign this attribute to the local modes of $\hat{f}$. In Section 3 we formulate a coverage-based goodness-of-fit criterion for principal points and curves, which can be interpreted similarly to the coefficient of determination used in regression analysis. This criterion is still valid for, and can be used for comparison with, methods which do not carry the label "local" (such as HS principal curves).

From Section 4 on we focus on *local* principal points and curves. The self-coverage measure is discussed in detail, and it is demonstrated how suitable bandwidths can be extracted from it. In Section 5 we will have a closer look at mean shift clustering, and we discuss how the self-coverage measure can be used to suit this purpose too. We close with the Discussion in Section 6, including some words on principal-curve based clustering.

2. Local principal points and curves

2.1 The mean shift

Assume that data $X = (x_1, \ldots, x_n)^T$, with $x_i \in \mathbb{R}^p$, have been sampled from $\boldsymbol{X}$. Let $x \in \mathbb{R}^p$ be an arbitrary point (which may or may not correspond to a data point). Let $K(\cdot)$ be a $p-$variate kernel function (for instance, a Gaussian density function, which will be used throughout this article), H a bandwidth matrix (which we assume to be of the shape $H = \mathrm{diag}(h_1^2, \ldots, h_p^2)$, with $h_j > 0$ being the component-wise bandwidth parameters), and $K_H(\cdot) = |H|^{-1/2} K(H^{-1/2}\cdot)$. Let $\mu_H(x)$ denote the local mean, or local center of mass, around x, i.e.

$$\mu_H(x) = \frac{\sum_{i=1}^n K_H(x_i - x) x_i}{\sum_{i=1}^n K_H(x_i - x)}.$$

Then the mean shift at x is given by

$$s_H(x) = \mu_H(x) - x = \frac{\sum_{i=1}^n K_H(x_i - x)(x_i - x)}{\sum_{i=1}^n K_H(x_i - x)}$$

i.e. $s_H(x)$ is a vector which shifts a point x to the local mean around x. The mean shift has several interesting properties [4], one of which being:

$$s_H(x) \propto H \frac{\nabla \hat{f}_H(x)}{\hat{f}_H(x)} \tag{1}$$

where $\hat{f}_H(x) = \frac{1}{n} \sum_{i=1}^n K_H(x_i - x)$ is a kernel density estimator of f. An asymptotic version of the result does also exist; one deduces easily from the proof of Theorem 2.1 in [15] that $s_H(x) = c(K) H \frac{\nabla f(x)}{f(x)} + o_P(H\mathbf{1})$, where $c(K)$ is defined through $c(K) I_p = \int uu^T K(u)\, du$, with identity matrix $I_p \in \mathbb{R}^{p \times p}$, and $\mathbf{1}$ is a vector only consisting of 1's. The significance of (1) is that the mean shift is 0 when the density gradient is 0, implying that, at a mode m_H of $\hat{f}_H$, one has $s_H(m_H) = 0$ and so

$$\mu_H(m_H) = m_H, \tag{2}$$

hence m_H being a fixed point of $\mu_H(\cdot)$.

We will from now on assume that the data are scaled. This is not strictly necessary for the estimation procedures that we present in the next two subsections, but will facilitate the bandwidth selection problem drastically. The scaling could be done by dividing each variable through its standard deviation or the interquartile range, but we choose in all our data examples the somewhat unusual convention to divide each variable by its range. *Global* variance is an irrelevant quantity for a local method; it is rather "spread" in a more colloquial sense which is more important, as the full data range needs to be represented by the fitted object. Another benefit of this way of scaling is the nice interpretability: A bandwidth h_j in $j-$th covariate direction covers $100h_j\%$ of the span of the $j-$th variable.

2.2 Local principal points as estimates of density modes

We can interpret the result (2) such that m_H is the local average of all points in a neighborhood of m_H. Recalling the definition of a self-consistent point as the average of all points which are closest to this point, it makes now sense to refer to property (2) as *local self-consistency*, and we may rebrand all such local modes as *local principal points* (LPPs), for bandwidth H.

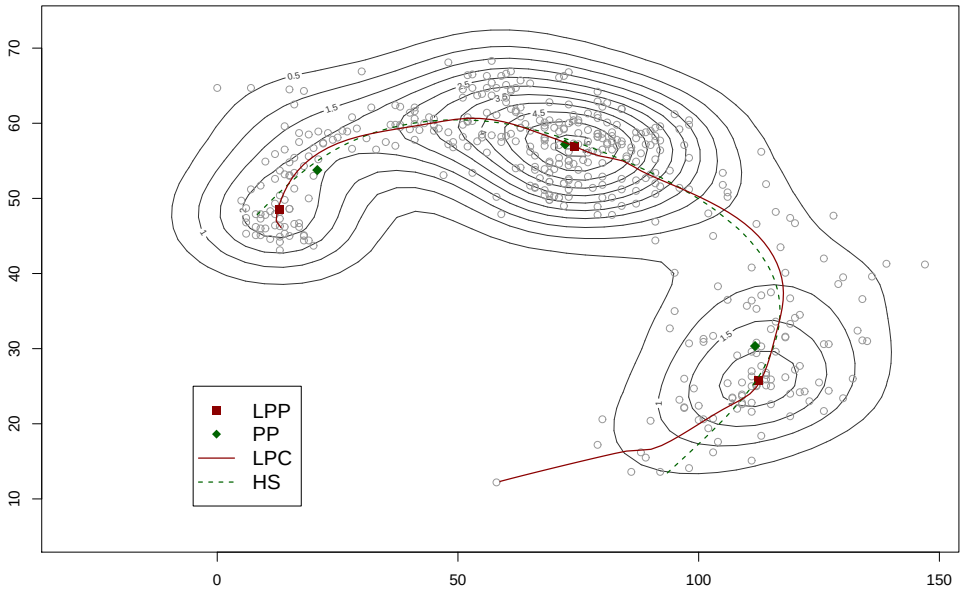

Fig. 1: Principal points and curves for Californian traffic data, with overlaid contour plot of a kernel density estimate. Note that all estimation is carried out in scaled data space; data and fitted objects were then re-scaled for convenience of plotting.

Comaniciu & Meer [4] showed that density modes may be estimated by iterating mean shift steps, i.e. for any x, the sequence $(m_\ell)_{\ell\geq 0}$ with $m_0 = x$, and $m_{\ell+1} = \mu_H(m_\ell)$ will eventually converge to a local principal point m_H, whereby the sequence of density estimates $(\hat{f}_H(m_\ell))_{\ell\geq 0}$ is monotonically increasing and achieves it maximum at m_H.

We may run such iterated mean shift procedures from every point $x_i, i = 1, \ldots, n$. This will eventually identify all local principal points, say $m_{(j)}, j = 1, \ldots, k$ (where the index H is omitted for notational convenience). These local principal points can be collected in a set $\boldsymbol{m}_k = \{m_{(1)}, \ldots, m_{(k)}\}$, which can be considered as an estimator of $\{v_{(1)}, \ldots, v_{(k)}\}$. Strictly speaking, $\boldsymbol{m}_k$ is a multiset, as several trajectories will converge to each local mode, but let us use the convention that we remove all multiple entries. In contrast to the $k-$means or the ML approach, the number k of principal points does not need to be known beforehand, and is only determined by the bandwidth H.

An example is provided in figure 1 for a simple data pattern involving vehicle speed and flow measurements on a Californian freeway (this type of data is sometimes referred to as a"fundamental diagram" in the traffic engineering literature). The diamond symbols correspond to the principal points (PPs) in Flury's sense, estimated with $k-$means, for $k = 3$. The solid squares show the local principal points, estimated using a constant bandwidth $h = h_1 = h_2 = 0.08$ on the scaled data. Additionally, we provide contours of a density estimate using the same bandwidth. It is evident that the LPPs occupy positions of higher (namely: locally maximal) density compared to the PPs estimated via $k-$means.

2.3 Local principal curves as estimates of density ridges

Local principal curves (LPCs) have been introduced by Einbeck et al. [6], as a mean-shift based principal curve algorithm. Given a starting point, say $x_{(0)}$, one alternates between a mean shift and a local PCA step, i.e. at $j-$th iteration one has

(A) $m_{(j)} = \mu_H(x_{(j)})$

(B) $x_{(j+1)} = m_{(j)} + t\gamma_{1,H}(x_{(j)})$

where $\gamma_{1,H}(x_{(j)})$ is the first eigenvector of the local covariance matrix

$$\Sigma_H(x_{(j)}) = \sum_{i=1}^{n} K_H(x_i - x_{(j)})(x_i - m_{(j)})(x_i - m_{(j)})^T / \sum_{i=1}^{n} K_H(x_i - x_{(j)})$$

If the end of the data cloud is reached (which is algorithmically determined by finding the point at which the distance between two neighboring $m_{(j)}$'s passes below a given threshold), then the same procedure is followed from $x_{(0)}$ into the direction of $-\gamma_{1,H}(x_{(0)})$. The local principal curve is given by the set of points, say $\boldsymbol{m}_d$ (d for discrete), which consists of all the $m_{(j)}$'s (on either side of $x_{(0)}$). If desired, a continuous and differentiable curve through $p-$variate space can be constructed from $\boldsymbol{m}_d$ by laying a natural cubic spline function $\boldsymbol{m}_s : \mathbb{R} \longrightarrow \mathbb{R}^p, \lambda \mapsto \boldsymbol{m}_s(\lambda)$ (s for smooth) through its elements. The parametrization λ is naturally defined through the arc-length of the curve. Each data point x_i can be projected onto the curve and represented ("compressed") through a univariate projection index λ_i. Details on these techniques are irrelevant for the presentation of this paper [7].

If $H = h^2 I_p$, one can show that, asymptotically [9],

$$m_{(j+1)} - m_{(j)} = \left[\frac{c(K)}{f(x_{(j)})} h^2 \pm \frac{1}{||\nabla f(x_{(j)})||} t \right] \nabla f(x_{(j)}) \tag{3}$$

where the first summand in the squared bracket is the mean shift contribution, while the second term is the contribution of the local PCA step. The sign is a "+"' when climbing uphill and a "–" when climbing downhill, implying that the curve stops at some point close to the boundary when equality of the two terms is met. The natural choice for the step size is $t = h$.

The starting point $x_{(0)}$ may be selected at random or by hand. Any of the principal points (in Flury's sense) or local principal points are sensible choices, and, in the example presented, any of these choices would lead to a practically identical curve. The local principal curve will only pass *exactly* through a local principal point, if this point is used as starting point for the LPC, and the same bandwidth is used for both. If the latter is true but not the former, then the curve will still pass *almost* exactly through those points, provided that the LPC covers that part of the data cloud at all (see Section 4 for tools which help checking this assumption). This is illustrated in figure 1, where the upper LPP is taken as starting point.

The interpretation of (3) is that local principal curves follow the gradient of the density, which means in practice that they will pass along the density ridges. This is illustrated in figure 2. The black "+" symbols are the local centers of mass, with the red triangle highlighting $x_{(0)}$. All local principal points, curves, and the density estimate used in figures 1 and 2 are computed using the same diagonal bandwidth matrix $H = \text{diag}(0.08^2, 0.08^2)$ after scaling.

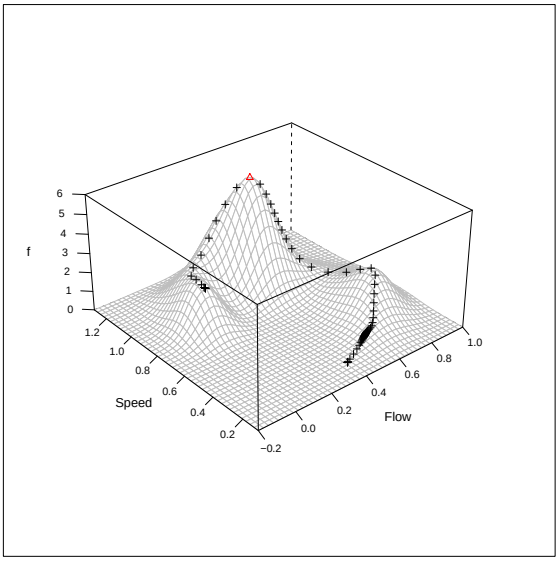

Fig. 2: Kernel density estimate of (scaled) speed-flow data, with LPC points $\boldsymbol{m}_d$ (+) superimposed. The LPP used as starting point is highlighted through a red triangle.

In summary, just as local principal points are estimators of density modes, local principal curves can be thought of as estimating density ridges. The relationship between LPCs and HS curves is just the same as that between LPPs and PPs: the local methods give biased estimates of the (self-consistent) principal points or curves, in the sense that they are stretched towards the high density regions. The essential common feature of local principal points and curves is that they are constructed through a sequence of local centers of mass. Their shape is only determined by the local topology of the data (and the bandwidth parameters), rather than global distance-minimizing criteria.

3. Coverage

We define the coverage $C_{\boldsymbol{m}}(\tau)$ of a principal object[1] $\boldsymbol{m}$ as the proportion of all data points whose distance to their nearest point on $\boldsymbol{m}$ is at most τ. For a set of principal points, this would be the proportion of points lying inside any circles ($p = 2$) or ("hyper"-)balls ($p \geq 3$) centered at $v_{(j)}, j = 1, \ldots, k$. For principal curves, one can think of the coverage as the proportion of points situated within a band ($p = 2$) or ("hyper"-)tube ($p \geq 3$) centered at the curve. Formally, for each $x_i \in \mathbb{R}^p$, define the "residual" ϵ_i as the shortest vector connecting x_i and (any point of) $\boldsymbol{m}$, with residual length $||\epsilon_i||$. Then we can write

$$C_{\boldsymbol{m}}(\tau) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{||\epsilon_i|| \leq \tau\}} \equiv F_n(\tau) \tag{4}$$

where F_n is just the empirical distribution function of the residual length. Note that $C_{\boldsymbol{m}}(\tau)$ is monotonically increasing with τ. Computing the coverage for all τ, one obtains the *coverage curve* $(\tau, C_{\boldsymbol{m}}(\tau))$.

1. We use this a general term encompassing any of principal components, points, curves, etc., with or without the attribute "local"

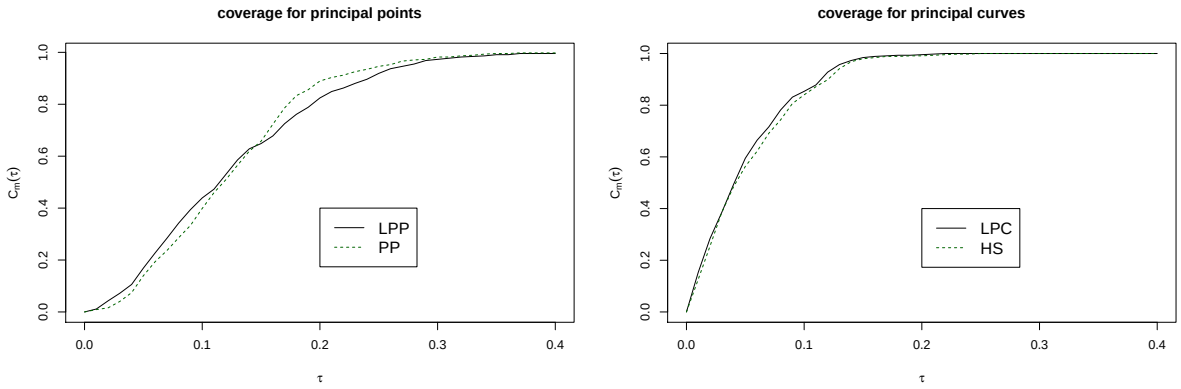

Fig. 3: Coverage curves for the fundamental diagram. Left: (local) principal points; right: (local) principal curves.

For the fitted (local) principal points/curves from figure 1, the coverage curves are provided in figure 3. In the left panel, we see that the coverage curve rises initially quicker for the LPPs than for the principal points estimated via $k-$means, but that the latter outperforms the former from about $\tau \approx 0.13$. For the principal curves, the LPC fit seems to be slightly superior to the HS curve throughout. One also notes that the coverage curves for principal curves rise more quickly than those for principal points. This is plausible, since it is harder to cover a data cloud through circles around a few points, compared to bands of the same size along a curve.

Using $C_{\boldsymbol{m}}(\tau)$ as a goodness-of-fit measure has an obvious drawback; it depends on τ and so no unambiguous conclusion can be drawn. So, the information provided by the coverage curve needs to be worked into a single summary statistics. Clearly, a "good" coverage curve will be concave and rise quickly. Hence, the immediate idea is to use the left top area, say A, between $\tau = 0$, $C_{\boldsymbol{m}}(\tau) = 1$, and the curve, as a measure of goodness-of-fit. Theoretically, this area has an appealing interpretation. Note that

$$A = \int_0^\infty (1 - F_n(\tau))d\tau = \frac{1}{n}\sum_{i=1}^n \int_0^\infty 1_{\{||\boldsymbol{\epsilon}_i|| > \tau\}} d\tau = \frac{1}{n}\sum_{i=1}^n ||\boldsymbol{\epsilon}_i||$$

is just the mean length of the residuals! Next, we set this area A in proportion to the area, say $A^{(\mathrm{P})}$, which would be obtained for the corresponding $q-$variate parametric benchmark, which in the case $q = 0$ is the overall mean ("1-means-estimator"), and in the case $q = 1$ is a linear principal component line. Computing "1 minus this ratio" yields the *coverage coefficient*, R_C

$$R_C \equiv 1 - \frac{A}{A^{(\mathrm{P})}} = 1 - \frac{\sum_{i=1}^n ||\boldsymbol{\epsilon}_i||}{\sum_{i=1}^n ||\boldsymbol{\epsilon}_i^{(\mathrm{P})}||} = \frac{\sum_{i=1}^n \left(||\boldsymbol{\epsilon}_i^{(\mathrm{P})}|| - ||\boldsymbol{\epsilon}_i||\right)}{\sum_{i=1}^n ||\boldsymbol{\epsilon}_i^{(\mathrm{P})}||},$$

which can be conveniently interpreted as the mean reduction in residual length. This double-interpretation, both in terms of residuals and in terms of coverage, makes R_C an attractive measure of goodness-of-fit in this context.

In the traffic example, the values of R_C for the LPP and the PP fits are 0.574 and 0.579, respectively, indicating that both methods fit almost equally well. Interpretationally, a

value of, say, $R_C = 0.574$ for the LPP fit means that the mean residual length reduces by 57.4% when using residuals to the next LPP rather than to the overall mean. For the principal curves, the value of R_C for the LPC fit is 0.627, while that for the HS fit is 0.606, implying that the LPC fits slightly better for these data.

Generally, R_C takes values in $(-\infty, 1]$, with 1 corresponding to the best possible fit, 0 corresponding to a 'bad' fit of the same quality as the parametric benchmark, and negative values corresponding to a fit being worse. In this sense, R_C behaves similar in spirit to the coefficient of determination (R^2) known from regression analysis. It also shares the property of R^2 that, the more complex the model, the higher values would be attained. That means that, if R^2 or R_C were used for model selection, or specifically bandwidth selection, then maximizing these coefficients would lead inevitably to overfitting. Flury [11] named this property of R^2 a "shortcoming", but one could argue that this depends on whether or not one takes it as what it is, namely as a *goodness–of–fit criterion*. The coefficient of determination checks, *given* some selected model, whether the attained fit is acceptable, but another criterion should be used for the actual selection of the model. This does not invalidate, in my opinion, its proper use as a measure of goodness–of–fit.

Precisely the same holds in our context. We may reasonably use the coverage curve, $C_{\boldsymbol{m}}(\tau)$, or its summary statistic, R_C, as measures of goodness–of–fit for any sort of principal objects. However, here it ends. We cannot use them, specifically, for bandwidth selection for local principal points or curves, since the solution maximizing these criteria would encompass all observations x_i, $i = 1, \ldots n$, which is obviously unacceptable. We will address this problem by manipulating the coverage curve appropriately.

4. Self-coverage

We wish to have a unifying procedure, which, for estimation problems of the type considered in Section 2, helps us to select a suitable bandwidth. Finding a full $p \times p$ bandwidth matrix H is a rather elusive task, as it would require to select $\frac{1}{2}p(p+1)$ bandwidths. Even the simpler problem of having to select the diagonal bandwidths in $H = \mathrm{diag}(h_1^2, \ldots, h_p^2)$ is challenging, but having scaled the variables as outlined in Section 2.1, there is very often no need to work with different degrees of smoothing in different coordinate directions.

Hence, we are working in this section in a setup in which

(A1) $H = h^2 I_p$;

(A2) in the case of principal curves: $t = h$;

implying that there is only one single smoothing or tuning parameter to select, namely the univariate bandwidth h.

The idea that we are going to convey is most intuitively explained in the context of principal curves, but still valid otherwise. Assume that, for some data cloud, a local principal curve of bandwidth h, say $\boldsymbol{m}(h)$, in either discrete or continuous representation, is being fitted. Then we would assume the curve to fit well, if the bandwidth h reflected the "thickness" of the data cloud around the curve. More statistically speaking, we would expect that, at each point along the curve, the data points to either side of it are scattered with a residual standard deviation of about h. If the width of the data band around the curve were much thinner than h, we would have chosen a bandwidth which is unnecessarily large, resulting in an oversmoothed principal curve. If the data cloud were much thicker, the chosen bandwidth would have been too small, implying that not all data would have been used for the construction of the curve. Hence, if a certain bandwidth h is good, then the coverage

at exactly this tube size should attain a good value as well. Similar considerations leading to the same conclusions apply if we argue in terms of a set of principal points $\boldsymbol{m}(h) = \boldsymbol{m}_k$: the detected cluster centers using bandwidth h should ideally correspond to the local means of all points of the respective cluster, so the coverage of $\boldsymbol{m}(h)$ at ball radius h should be large.

For either of local principal points or curves, computing the coverage *for the same radius, τ, which was used as bandwidth h*, and tracing this function over all values of h, leads to the definition of the *self-coverage*

$$S(h) = C_{\boldsymbol{m}(h)}(h). \tag{5}$$

The function $S(h)$ will eventually converge to 1, so we cannot simply take its maximum in order to detect the most suitable bandwidth. However, unlike $C(\tau)$, the self-coverage curve does not converge monotonically, but possesses distinctive features which we can exploit. There are essentially three sorts of features that we are interested in: (i) We may have the situation of an already reasonably fitting object, but using a smaller than ideal bandwidth. Then increasing h will increase $S(h)$ monotonically, until the full width of the data cloud is covered, from which point $S(h)$ will level off. (ii) There may be certain "threshold" bandwidths, from which on certain parts of the data cloud, which were previously inaccessible for smaller bandwidths, are now visited. In this case, there will be a very sudden jump in the self-coverage curve. (iii) At any stage, an increase of the bandwidth may blur previously well fitted parts of the fitted object, so that the self-coverage even decreases.

Bandwidths of type (i) would be the most desirable ones, but they are also the most difficult ones to detect reliably. Bandwidths of type (ii) and (iii) are very easy to detect, but have to be considered with care. They can be seen as the smallest or largest bandwidths, respectively, before the fit breaks down.

We are going to detect bandwidths according to scenarios (i)-(iii) through one single criterion, which targets points of negative curvature of $S(h)$. Assume we have evaluated $S(h)$ over a grid of bandwidths $h_1 < \ldots < h_L$ (a setting which works well is to use a grid with a spacing of 0.005; so if one investigates the span of, say, $h = 0.005$ to 0.4, one needs $L = 80$). The curvature, $S''(h_\ell)$, is then easily captured by considering second differences;

$$\triangle^2 S(h_\ell) = S(h_{\ell+1}) - 2S(h_\ell) + S(h_{\ell-1}). \tag{6}$$

Since we are interested in points of large *negative* curvature, we need to find the *minima* of (6). Let $h_{(j)}$ be the bandwidth yielding the $j-$th lowest value of (6) under the constraint

$$S(h_\ell) > \max\{S(h_1), \ldots, S(h_{\ell-1}), s\} \tag{7}$$

where $s \in (0, 1)$ is a pre–defined constant. The condition (7) will enforce that no bandwidth is selected which leads to a coverage which could already be achieved through a smaller bandwidth, or which falls below a threshold s. Based on experience with a wide range of real data sets, the setting $s = 2/3$ has been found to work generally well for local principal curves, while for local principal points, where it is more difficult to achieve high coverages, a lower value of $s = 1/3$ is recommended.

The optimal bandwidth under this criterion is $h_{(1)}$. Often it will be useful to consider also the next-best candidates, say $h_{(2)}$ and $h_{(3)}$, as different suitable solutions may exist at different degrees of resolution. This will be illustrated through real data examples below.

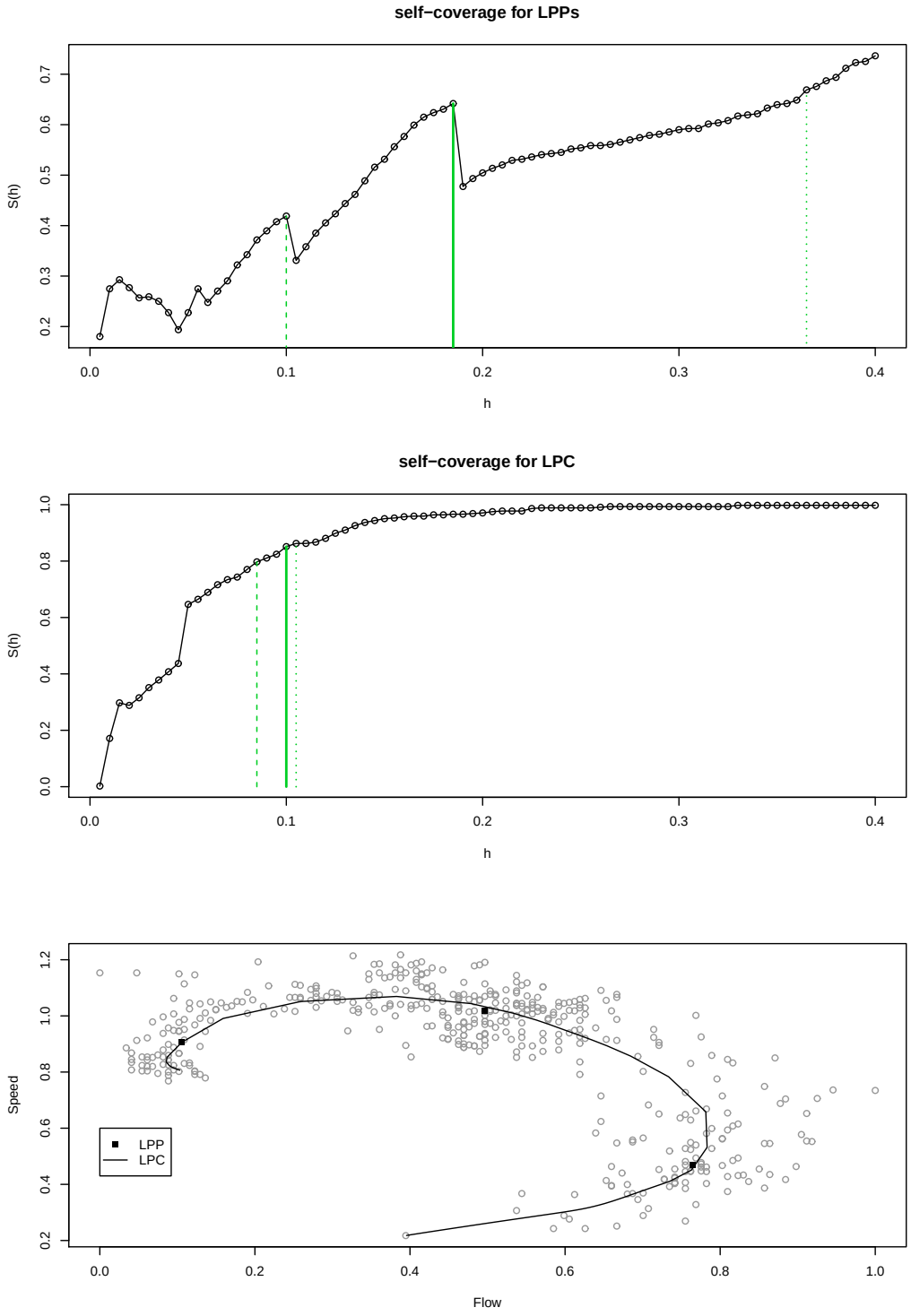

Fig. 4: Self-coverage curves for the fundamental diagram, with selected bandwidths $h_{(1)}$ (thick solid), $h_{(2)}$ (dashed), and $h_{(3)}$ (dotted). Top: selection for LPPs; middle: selection for LPCs; bottom: scaled data with LPPs and LPC, fitted using the selected bandwidths 0.100 each. The equality of the two bandwidths is coincidental.

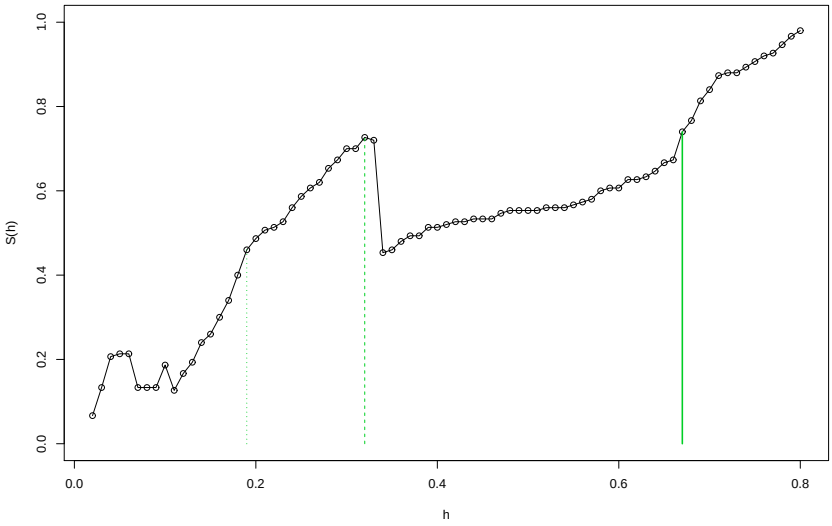

Fig. 5: Iris data: Self-coverage curve for LPPs, with selected candidate bandwidths at $h_{(1)} = 0.67, h_{(2)} = 0.32$, and $h_{(3)} = 0.19$.

Firstly, we consider the speed-flow data used previously. The self-coverage curve for local principal points, in the form as it is produced by default in R package LPCM [8], is provided in the first panel of figure 4, with the thick solid, dashed, and dotted lines corresponding to $h_{(1)}, h_{(2)}$ and $h_{(3)}$ respectively. The first and the second hump of the curve, at $h_{(2)} = 0.100$ and $h_{(1)} = 0.185$, correspond to the 3- and 2- cluster solutions, respectively. Both are bandwidths of type (iii). The 1-cluster solution is provided by $h_{(3)} = 0.365$, which is a type (i) bandwidth.

The self-coverage curve for the local principal curve (using the leftmost LPP as starting point) selects a bandwidth of type (i) at $h_{(1)} = 0.100$ (see the second panel of figure 4). Two further, but less significant, type (i) bandwidths are situated nearby. The resulting LPPs (using the 3–cluster solution) and the LPC are depicted in the bottom panel. We do not provide plots of the second differences $\triangle^2 S(h_\ell)$ as they have a quite erratic appearance and do not add much value.

In addition, we consider the well-known iris data, which are part of the standard R distribution [14]. The iris data feature 4-variate measurements (in cm) of petal and sepal length and width, respectively, of $n = 150$ flowers belonging to certain species of iris. We apply the methodology sequentially. Firstly, we apply the self-coverage technique on the problem of finding the density modes. The resulting self-coverage curve is provided in figure 5. From left to right, the bandwidths at $h_{(3)} = 0.19$ and $h_{(2)} = 0.32$ are both 2–cluster solutions of type (i) and (iii), respectively. The bandwidth at $h_{(1)} = 0.76$ is a 1–cluster solution of type (ii), which would probably not be of interest here, despite being the overall winner in terms of criterion (6). We choose to work with the bandwidth $h_{(3)} = 0.19$, yielding the fitted density modes in form of black triangles in figure 6. Taking now one of these two local principal points — we choose arbitrarily that one belonging to the green cluster — as the starting point for the local principal curve, one obtains for the latter the self-coverage curve provided in figure 7. Here we observe a type (ii) bandwidth at $h_{(1)} = 0.16$, which

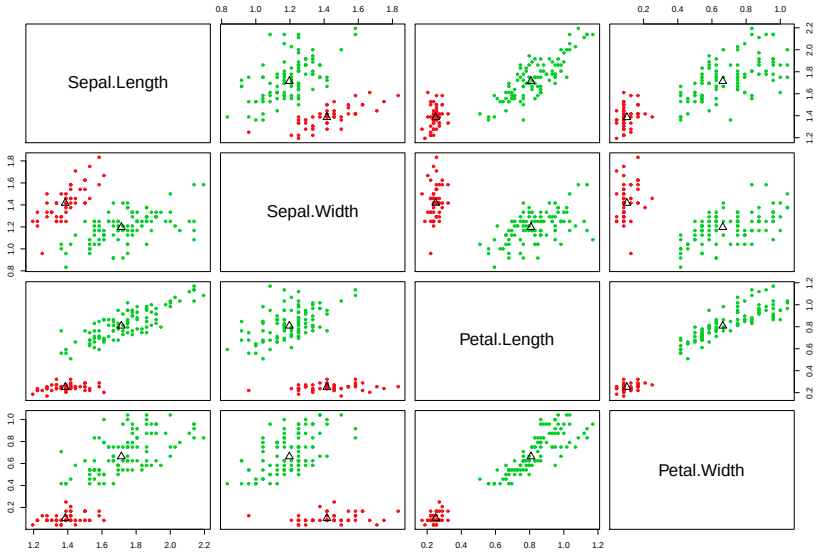

Fig. 6: Iris data (scaled) with local principal points (black triangles), estimated using $h_{(3)} = 0.19$. The colours correspond to the mean shift clusters, see Section 5.

is the smallest ("most wiggly") bandwidth such that the LPC is able to connect the two clusters. When increasing h further, two type (i) bandwidth candidates at $h_{(2)} = 0.17$ and $h_{(3)} = 0.20$ are found. From this point onwards the gain in coverage is rather due to the increase of the tube size, than due to the improvement of the fit. The fitted local principal curve using $h_{(1)} = 0.16$, which achieves a value of $R_C = 0.36$, is provided in figure 8.

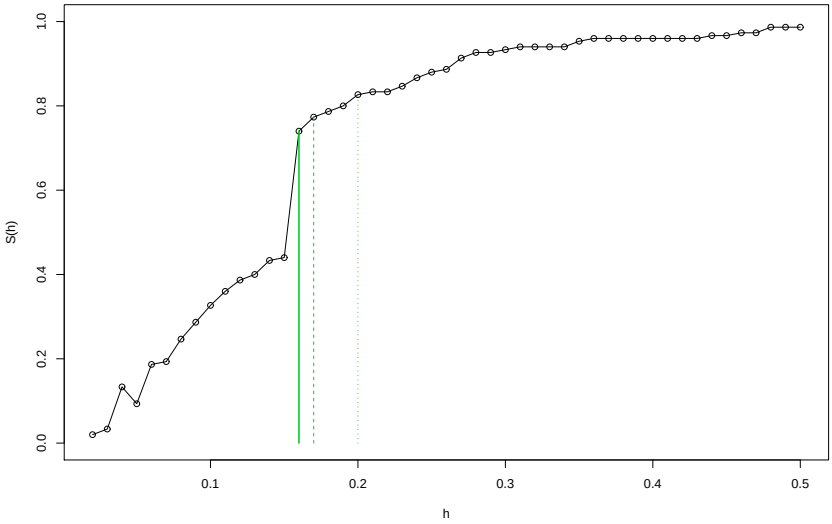

Fig. 7: Iris data: Self-coverage curve for LPCs, with selected bandwidths at $h_{(1)} = 0.16$, $h_{(2)} = 0.17$, and $h_{(3)} = 0.20$.

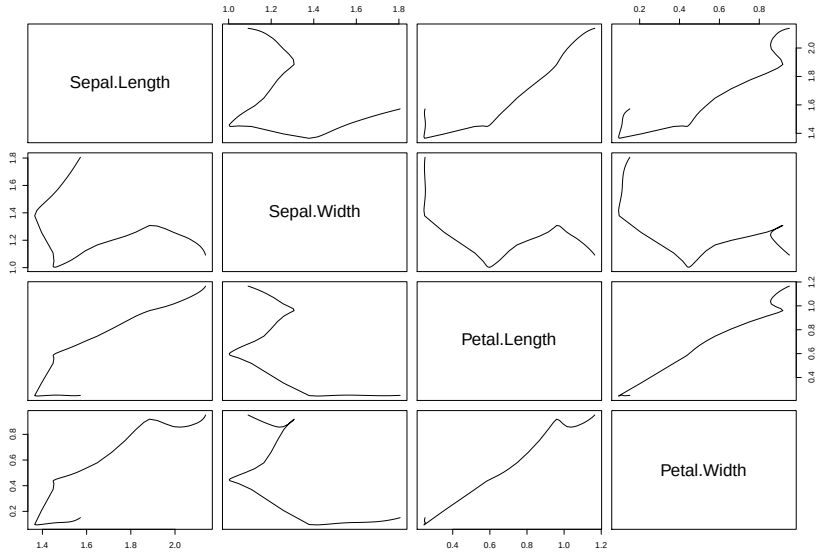

Fig. 8: LPC through iris data, fitted with $h_{(1)} = 0.16$.

We finish this section with some algorithmic technicalities. Firstly, if the principal curve is just provided in form of a set of points, such as $\boldsymbol{m}_d$, then the $||\epsilon_i||$ are computed through the distance to the nearest of these points (just as if they were principal points), so that the (self-) coverage curve is only approximate in this case [6]. The software provided in [8] uses numerical optimization in order to compute the nearest point on the smooth curve $\boldsymbol{m}_s$, yielding *exact* (self-) coverage curves. The differences between these two types of (self-) coverage curves are usually negligible, though the approximate self-coverage curves may result in little spurious humps at places, which may impact on the bandwidth selection task. All (self-)coverage curves shown in this paper are exact.

Secondly, one needs to meet a minor precaution to avoid overfitting at very small bandwidths. For density mode estimation, using $h \longrightarrow 0$ will lead to a local principal point at almost every data point, implying that $S(h) \approx 1$. Similarly, a very small bandwidth will lead to principal curve which performs some sort of random walk within the data cloud, which can lead to a very long but useless curve with a high coverage. In the former case, the precaution is to base the calculation of the self–coverage only on cluster centrers to which more than two mean shift trajectories converged. In the latter case, the precaution is to disallow principal curves to intersect themselves. Both measures only have an effect for very small bandwidths, but do not influence the self-coverage curve in bandwidth domains, say $h \gtrsim 0.05$, in which we are realistically interested.

5. Clustering

Clustering based on local principal points has been suggested already by Cheng [2]. The idea, as well as the implementation, is very simple: Each data point x_i is allocated to the density mode, say $c_i \in \{m_{(1)}, \ldots, m_{(k)}\}$, to which its mean shift trajectory has converged. This method, known as mean shift clustering, is illustrated in figure 9. Compared to other clustering algorithms, mean shift clustering has the massive advantage that it does not require to pre-determine the number of clusters. However, it requires the selection

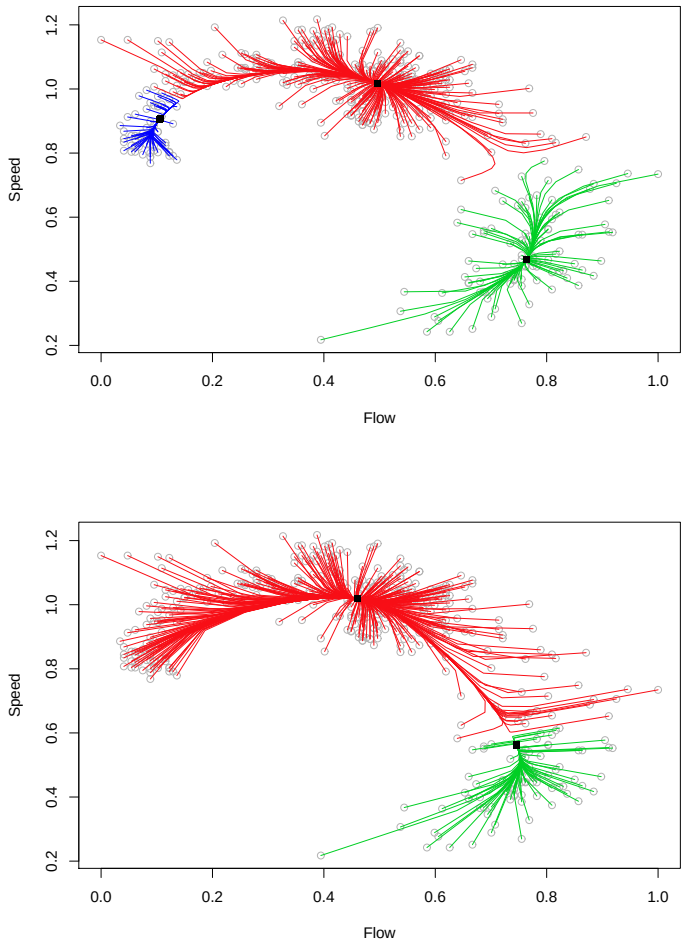

Fig. 9: Mean shift clustering for (scaled) speed-flow data. Top: $h = 0.010$, bottom: $h = 0.185$.

of the bandwidth(s), and the lack of automated procedures for this purpose has perhaps contributed to the fact that mean shift clustering has not yet become a widely accepted and applied tool.

Of course, we may just take the bandwidths selected by the usual self-coverage routine. These are the bandwidths used in figure 9, and often this will work well. However, conceptually there is a caveat in this approach: The coverage is estimated by considering circles, tubes, etc, around the fitted object. That is, for the coverage estimation, points are associated to the *nearest* cluster center, while the mean shift clustering itself does not necessarily assign points to the nearest cluster center. Hence, in (4), one should add the requirement that each x_i is attached to their respective cluster center c_i. The "cluster residual" would be defined as $\epsilon_i^c = x_i - c_i$ and using these adapted residuals in (4), and, eventually, in (5) and (6), will give bandwidths which are more tailored to the clustering problem. An illustration is given through the traffic data example. It is well known that the upper branch corresponds to free-flow traffic, while the lower branch corresponds to congested traffic. Figure 9 shows the clustering obtained using the two bandwidths $h = 0.100$ and $h = 0.185$ detected

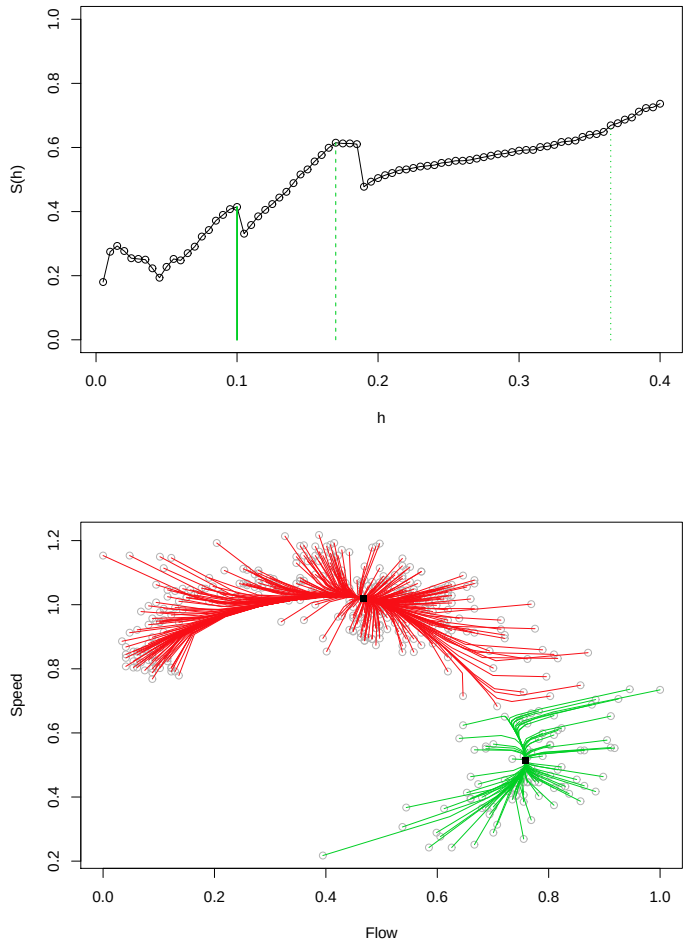

Fig. 10: Mean shift clustering for (scaled) speed-flow data. Top: self-coverage curve using cluster residuals; bottom: clustering with selected bandwidth $h_{(2)} = 0.170$.

previously. For the former bandwidth, a third (blue) cluster which can be attributed to night time driving is detected. The two cluster-solution does not appear satisfactory; the uncongested cluster seems too "greedy" and takes over a considerable amount of the congested data points. Using the suggested adaption to the self-coverage curve, this problem alleviates: The two-cluster bump in the self-coverage curve is now replaced by a plateau, of which, by criterion (6) with the usual default setting of s, $h_{(1)} = 0.100$ and $h_{(2)} = 0.170$ (notably lower than 0.185), are selected. The resulting 2-cluster solution, using $h_{(2)}$, is provided in figure 10. Note that this solution is already present, in weaker form, in figure 4: the little bend shortly before $h_{(1)}$ would correspond to the seventh–best bandwidth selected, $h_{(7)} = 0.170$.

6. Discussion

We have provided a semi-automatic tool for the selection of bandwidths for unsupervised multivariate mean-shift based learning techniques. The method is "semi–automatic" and

not "fully automatic", since it will often produce multiple feasible solutions, reflecting the fact that useful information may be present at different degrees of resolution.

An important simplification that has been made in this paper is the restriction (A1) to a diagonal bandwidth matrix, with equal entries in the diagonal. One may criticize this as being not flexible enough. However, if the data are scaled, then this problem is significantly alleviated. We note that the situation is here different to the regression context, where one investigates the *impact* of the components of $\boldsymbol{X}$ onto an external variable $\boldsymbol{Y}$, and the degree of localization needed to describe this impact may very well be very different with different components, regardless whether the data are scaled or not.

The techniques presented in this paper are, in principle, extendible and applicable beyond the framework considered here. For instance, the local principal surface algorithm [7] does make use of the mean shift as the essential tool of estimation too, so that the methods proposed here should straightforwardly extend to this case, and preliminary investigations brought encouraging results.

One may also consider extensions of this technique to "principal curve- based clustering" [17]. Local principal curves allow to fit separate branches to separate parts of the data cloud. Having selected a suitable starting point within each cluster, a principal curve may be launched from each of these. Each data point can then be projected to the nearest point on the nearest curve, and is accordingly classified. We illustrate this technique by considering, for ease of presentation, only those two variables of the iris data set which represent the "length" of the petals and sepals. A branched local principal curve using starting points selected via 2−means is shown in figure 11, along with the projections and the classification. In order to adapt the self-coverage techniques successfully to this context, one would need to disallow individual branches to creep into other clusters.

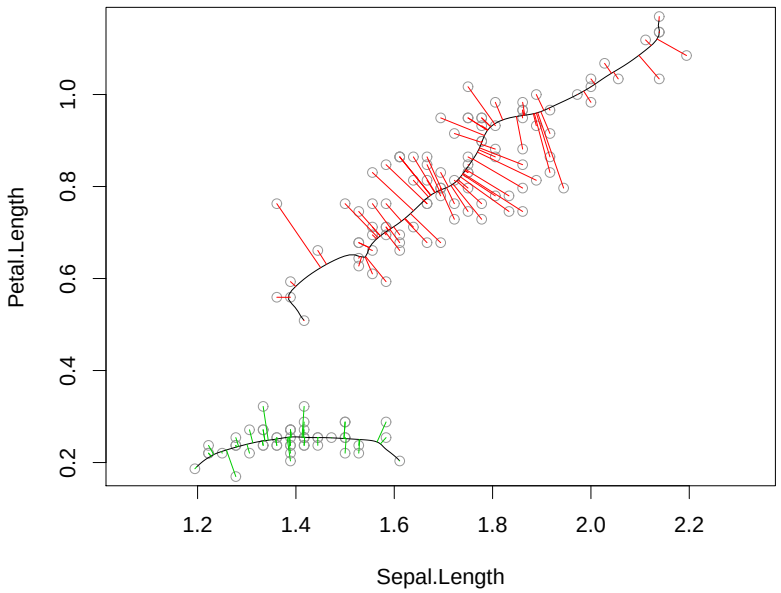

Fig. 11: Example for principal-curve based clustering.

Finally, one could investigate the extension to variable bandwidths, i.e. bandwidths $h = h(x)$ or even $H = H(x)$ which take different values in different regions of the covariate space. Methods of this type have been considered by computer scientists for applications such as image analysis [1] or object tracking [13]. Though the results were partially encouraging, care needs to be taken with this approach. In fact, it turns out that the corresponding density estimator $\hat{f}$ obtained via $H(x)$ is not a proper density, i.e. it does not integrate to 1 [1]. Secondly, such a technique may lead to undesired results at it could move the estimated modes [19].

It is finally noted that implementations of all methods in the statistical programming language R [14], as well as the traffic data set, are provided in [8].

Acknowledgments

The author is grateful to Prof. Simos Meintanis, University of Athens. His encouragement to present this material in his session at the EMS 2010 was the initial motivation to write up this paper. Support by the Durham Energy Institute (DEI) is gratefully acknowledged.

References

[1] Bugeau, A. and Pérez, P. (2007). Bandwidth selection for kernel estimation in mixed multi-dimensional spaces. Research Report RR-6286, INRIA.

[2] Cheng, Y. (1995). Mean shift, mode seeking and clustering. *IEEE Trans. Pattern Anal. Machine Intell.* **17**, 790–799.

[3] Comaniciu, D. (2003). An algorithm for data-driven bandwidth selection. *IEEE Trans. Pattern Anal. Machine Intell.* **25**, 281–288.

[4] Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell.* **24**, 603–619.

[5] Comaniciu, D., Ramesh, V., and Meer, P. (2001). The variable bandwidth mean shift and data–driven scale selection. *Proc. Eighth Int'l Conf. Computer Vision*, vol I, 438–445.

[6] Einbeck, J., Tutz, G., and Evers, L. (2005). Local principal curves. *Statistics and Computing* **15**, 301–313.

[7] Einbeck, J., Evers, L., and Powell, B. (2010). Data compression and regression through local principal curves and surfaces. *International Journal of Neural Systems* **20**, 177–192.

[8] Einbeck, J. and Evers, L. (2011). *LPCM: Local principal curve methods.* R package version 0.43. On CRAN, `http://cran.r-project.org/`.

[9] Einbeck, J. and Zayed, M. (2011). Some asymptotics for localized principal components and curves. Unpublished working paper, Durham University.

[10] Flury, B. D. (1990). Principal points. *Biometrika* **77**, 33–41.

[11] Flury, B. D. (1993). Estimation of principal points. *Applied Statistics (JRSSC)* **42**, 139–151.

[12] Hastie, T. and Stuetzle, W. (1989). Principal curves. *J. Amer. Statist. Assoc.* **84**, 502–516.

[13] Pu, J.-X. and Peng, N.-S. (2006). Adaptive kernel based tracking using mean-shift. *Image Analysis and Recognition: Lecture Notes in Computer Science* **4141**, 394–403.

[14] R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

[15] Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–1370.

[16] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

[17] Stanford, D. C. and Raftery, A. E. (2000). Finding curvilinear features in spatial point patterns: principal curve clustering with noise. *IEEE Trans. Pattern Anal. Machine Intell.* **22**, 601–609.

[18] Tarpey, T. and Flury, B. (1996). Self-consistency: A fundamental concept in statistics. *Statistical Science* **11**, 229–243.

[19] Vieu, P. (1996). A note on density mode estimation. *Statistics & Probability Letters* **26**, 297–307.

[20] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* Boca Raton: Chapman & Hall/CRC.