Journal of Classification: CLAS-D-14-00037R1

Design of Blurring Mean-Shift Algorithms for Data Classification

Carlo Grillenzoni
IUAV University, Venice, Italy

Abstract: The mean-shift algorithm is an iterative method of mode seeking and data clustering based on the kernel density estimator. The blurring mean-shift is an accelerated version which uses the original data only in the first step, then re-smoothes previous estimates. It converges to local centroids, but may suffer from problems of asymptotic bias, which fundamentally depend on the design of its smoothing components. This paper develops nearest-neighbor implementations and data-driven techniques of bandwidth selection, which enhance the clustering performance of the blurring method. These solutions can be applied to the whole class of mean-shift algorithms, including the iterative local mean method. Extended simulation experiments and applications to well known data-sets show the goodness of the blurring estimator with respect to other algorithms.

Keywords: Bandwidth selection; Cluster stability; Kernel density; Image segmentation; Local means; Monotone convergence; Nearest neighbors.

Acknowledgments: A sincere thanks to the Reviewers for their useful remarks.

Author's Address: Institute of Architecture, University of Venice, S. Croce 1957, 30135 Venezia, Italy, e-mail: carlog@iuav.it

1. Introduction

The mean-shift (MS) algorithm, Fukunaga and Hostetler (1975), Silverman (1986, p.132), is a method for clustering spatial data and segmenting digital images. The algorithm iteratively moves points and pixels toward the modes of the kernel density function (KDF) of the data, by just exploiting the first-order conditions. Convergence properties of MS have been studied by Comaniciu and Meer (2002), Carreira-Perpiñán (2007), Li et al. (2007) and Aliyari Ghassabeh (2013); they have proved that its estimates, initialized with the observed data, converge monotonically at a linear rate, to the modal values of the kernel density.

Blurring mean-shift (BMS) is a smoothed version of MS, which increases the speed of convergence up to a cubic rate, see Carreira-Perpiñán (2006). It is particularly useful in processing large data sets, as those produced in video sequences and laser scanning (e.g. Grillenzoni, 2007). However, BMS was criticized by Rao et al. (2009), by showing that it is less accurate than MS and is *biased* with Gaussian kernels, as it converges to a single cluster. This issue was already discussed by Cheng (1995), who distinguished between BMS with broad and flat kernels, and investigated the properties of truncated functions. Now, since the simple MS also converges to a single point when its bandwidth is large, one may wonder if the bias of BMS also depends on the bandwidth selection. This issue has not been cleared by Chen (2015), who has recently improved the analysis of Cheng (1995).

Carreira-Perpiñán (2006) showed the advantages of using Gaussian kernels in BMS, and provided a stopping rule for iterations to prevent convergence to biased solutions. He also developed a sparse matrix implementation which further increases the speed of the method. In this work we discuss the conditions of convergence of BMS and suggest a nearest-neighbor implementation which avoids explicit kernel truncations. We also develop automatic, data-driven, techniques of bandwidth selection which are based on the statistical optimization of the number of groups in a classification. Simulation experiments and real case-studies show the efficacy of the proposed solutions, both to identify the right number of clusters and to converge to their centroids. The superiority of Gaussian BMS algorithms will be proven.

The paper is organized as follows: Section 2 discusses censored BMS algorithms and their conditions of convergence. Section 3 develops techniques of bandwidth selection based on clustering. Section 4 presents various numerical applications.

2. Mean-Shift Algorithms

Let $\{\mathbf{x}_i\}_{i=1}^N$ be a real data-set in $\Re^d$ space. In spatial data, one usually has d=5 because in laser scanning (or seismology) it contains 3 spatial coordinates, the measurement time and the infrared reflectance (or the magnitude): $\mathbf{x}_i' = [\mathrm{x}_i, \mathrm{y}_i, \mathrm{z}_i; \mathrm{t}_i, \mathrm{r}_i]$. Similarly, in continuous data as digital images, it contains 2 spatial coordinates and 3 color intensities (RGB). The goal of clustering is to group data in $m > 1$ homogeneous sets, having minimum inner variance: $\left\{\{\boldsymbol{x}_{i,k}\}_{i=1}^{N_k}\right\}_{k=1}^{m}$, where $\sum_{k=1}^m N_k = N$. We have denoted the target groups in *italics*, because they are unknown quantities to be estimated, together with their centroids $\bar{\boldsymbol{x}}_k = N_k^{-1} \sum_{i=1}^{N_k} \boldsymbol{x}_{i,k}$.

Statistical properties of the data are entirely described by their probability density function $f(\boldsymbol{x})$. We assume that it is differentiable with $m_o > 1$ modal points $\boldsymbol{\mu}_k$, where first derivative vanishes: $f'(\boldsymbol{\mu}_k)$=$\mathbf{0}$. The function can be estimated with nonparametric methods, such as the kernel density (Silverman, 1986):

$$\hat{f}_N(\boldsymbol{x}; \beta) = \frac{1}{\beta^d N} \sum_{i=1}^{N} K\Big[(\boldsymbol{x} - \mathbf{x}_i)/\beta\Big], \qquad \boldsymbol{x} \in \Re^d,$$

where $K(\boldsymbol{z}) \propto \exp(-.5\|\boldsymbol{z}\|^2)$ is the Gaussian kernel and $0 < \beta < \infty$ is the bandwidth. The shape of $\hat{f}_N$ depends on the size of β; this is usually selected by minimizing the mean integrated squared error (MISE): $\mathrm{E}\{\int[\hat{f}_N(\boldsymbol{x}) - f(\boldsymbol{x})]^2 \mathrm{d}\boldsymbol{x}\}$. As shown by Silverman (1986), if $f(x)$ is univariate Gaussian, then the optimal bandwidth is given by $\beta_o \approx \sigma_x / N^{0.2}$, where σ_x^2 is the variance of data. For this reason, we do *not* equate the coefficients σ_x, β, where the latter tends to 0 as $N \to \infty$.

The clustering approach followed by MS is to shift each data point $\mathbf{x}_i$ toward the regions of $\hat{f}_N$ with higher density. This is obtained with the first-order condition $\hat{f}_N'(\boldsymbol{x}) = \sum_j K[(\boldsymbol{x} - \mathbf{x}_j)/\beta]\,(\boldsymbol{x} - \mathbf{x}_j) = \mathbf{0}$, and solving iteratively for $\boldsymbol{x}$:

$$\text{MS}: \quad \hat{\boldsymbol{x}}_i^{(t+1)} = \sum_{j=1}^{N} \frac{K\Big[(\hat{\boldsymbol{x}}_i^{(t)} - \mathbf{x}_j)/\beta\Big]}{\sum_{l=1}^{N} K[(\hat{\boldsymbol{x}}_i^{(t)} - \mathbf{x}_l)/\beta]}\, \mathbf{x}_j\,, \qquad \hat{\boldsymbol{x}}_i^{(0)} = \mathbf{x}_i\,, \tag{1}$$

where (t) is the iteration counter and $\mathbf{x}_i$ is the starting value. Since MS is hill climbing and $K(\cdot)$ is monotonically decreasing, the estimates (1) converge to the modes $\hat{\boldsymbol{\mu}}_k$ of the kernel density; see Li et al. (2007) and Aliyari Ghassabeh (2013):

$$\lim_{t\to\infty} \hat{\boldsymbol{x}}_i^{(t)} = \hat{\boldsymbol{\mu}}_k\,, \quad k = 1, \ldots, \hat{m}\,, \quad \forall\, i\,,$$
$$\hat{f}'_N(\hat{\boldsymbol{\mu}}_k) = \mathbf{0}\,, \quad \|\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k\| \preceq \|\mathbf{x}_i - \hat{\boldsymbol{\mu}}_h\|\,,$$

for any N, β. Notice that each $\hat{\boldsymbol{x}}_i^{(t)}$ converges to the nearest mode in its basin of attraction, and the number of estimated modes $\hat{m}$ depends on the size of β. In general, the optimal bandwidth (that minimizes MISE) does not provide unbiased identification of the number of modes m_o of f; hence, one cannot state $\mathrm{E}(\hat{\boldsymbol{\mu}}_k) = \boldsymbol{\mu}_k$. Also the selection techniques of β used in the kernel estimation of derivatives f' (e.g. Chaćon and Duong, 2013) are not suitable when m_o is large. In Section 3 we develop specific methods for selecting β which may approach m_o.

The basic idea of *blurring* MS is to treat the estimates (1) as new data in the subsequent iteration; in practice, it smoothes previously smoothed (i.e. blurred) data. This leads to the nested algorithm

$$\text{BMS}: \quad \tilde{\boldsymbol{x}}_i^{(t+1)} = \sum_{j=1}^{N} \frac{K\big[(\tilde{\boldsymbol{x}}_i^{(t)} - \tilde{\boldsymbol{x}}_j^{(t)})/\beta\big]}{\sum_{l=1}^{N} K[(\tilde{\boldsymbol{x}}_i^{(t)} - \tilde{\boldsymbol{x}}_l^{(t)})/\beta]}\, \tilde{\boldsymbol{x}}_j^{(t)}\,, \qquad \tilde{\boldsymbol{x}}_i^{(0)} = \mathbf{x}_i\,, \tag{2}$$

where the original data are used only in the first iteration. The convergence of (2) is allowed by the fact that its first iteration coincides with that of (1): $\tilde{\boldsymbol{x}}_i^{(1)} = \hat{\boldsymbol{x}}_i^{(1)}$ for all i. Subsequent iterations shrink previously shrinked data, so that the convergence of (2) is guaranteed (see Chen, 2015).

As a formal argument, let $\mathcal{H}_N$ be the convex-hull of $\{\boldsymbol{x}_i\}_{i=1}^N$, i.e. the minimal convex set which contains $\boldsymbol{x}_i$. For the estimates $\tilde{\boldsymbol{x}}_i^{(t)}$ we have $\tilde{\mathcal{H}}_N^{(t)} \supseteq \tilde{\mathcal{H}}_N^{(t+1)}$ for all t, because $\tilde{\boldsymbol{x}}_i^{(t+1)} = \sum_{j=1}^N \tilde{w}_{ij}^{(t)}\, \tilde{\boldsymbol{x}}_j^{(t)}$ is a convex combination of all points. Given the nested structure of sets $\tilde{\mathcal{H}}_N^{(t)}$, they converge to the bounded limit $\mathcal{H}_N = \bigcap_{t=0}^{\infty} \tilde{\mathcal{H}}_N^{(t)}$. Carreira-Perpiñán (2006) showed that the convergence speed of BMS increases up to a cubic rate over (1); however, the side-effect is the asymptotic bias of the algorithm. This means the tendency of estimates $\tilde{\boldsymbol{x}}_i^{(t)}$ to converge to a single point $\mathcal{H}_N = \boldsymbol{\mu}_0$, after having reached local centroids (not the modes of $\hat{f}_N$).

The asymptotic bias of the blurring estimator (2) stems from the fact that it implicitly maximizes the measure of concentration

$$C_N\Big(\{\boldsymbol{x}_l\}_{l=1}^N;\beta\Big)=\sum_{i=1}^N\sum_{j=1}^N K\Big[(\boldsymbol{x}_i-\boldsymbol{x}_j)/\beta\Big]\,, \tag{3}$$

(see Cheng, 1995). Indeed, if $K(\cdot)$ is Gaussian, then (2) arises from the condition

$$\partial\, C_N/\partial\boldsymbol{x}_i=\sum_j K\Big[(\boldsymbol{x}_i-\boldsymbol{x}_j)/\beta\Big]\,(\boldsymbol{x}_i-\boldsymbol{x}_j)=\mathbf{0}\,,$$

and solving iteratively for $\boldsymbol{x}_i$. Since the function (3) is maximized by $\boldsymbol{x}_i=\boldsymbol{x}_j$ for all i,j, it follows that BMS estimates always converge to a single point $\boldsymbol{\mu}_0$. Obviously, this would make useless the clustering results provided by (2).

2.1 Censored Estimators. Last remarks can be extended to all infinite-support kernels, as they virtually cover the entire data range (see Cheng, 1995). However, in numerical applications, the computer precision actually cuts off infinite supports (this is clear in the random number generation). Hence, it is useful to investigate the behavior of algorithms with *truncated* kernels, such as

$$t\text{BMS}: \quad K^*_\alpha(z)=K_\beta(z)\cdot I(|z|\le\alpha)\,,\qquad 0<\alpha<\infty\,, \tag{4}$$

where $I(\cdot)$ is the indicator function. For suitable α,β, the kernel (4) does not cover the entire data set, and one can define conditions of convergence to multiple centers. In the Appendix it is shown that for $\alpha=3\beta$ (i.e. K^*_α covers 99.8% of K_β Gaussian), and local centroids $\{\boldsymbol{\mu}_k\}_1^m$, the blurring estimator based on (4)

$$\begin{aligned} &t\text{BMS converges to 1 center if} && \beta>\max_{k,h}\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_h\|/6\,,\\ &t\text{BMS converges to } m \text{ centers if} && \beta<\min_{h,k}\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_h\|/6\,, \end{aligned} \tag{5}$$

Condition (5) means that the kernel diameter $2\alpha=6\beta$ must be less than the minimum distance between $\boldsymbol{\mu}_k$. It allows point estimates to have no reciprocal influence, once they are converged to their nearest local centroids (see Appendix).

Instead of cutting off the kernel support, it may be preferable to compute BMS estimates only on their $n<N/m$ nearest neighbors (NN), as

$$n\text{BMS}: \quad \check{\boldsymbol{x}}_i^{(t+1)}=\sum_{j=1}^n \frac{K\Big[(\check{\boldsymbol{x}}_i^{(t)}-\check{\boldsymbol{x}}_{ji}^{(t)})/\beta\Big]}{\sum_{l=1}^n K[(\check{\boldsymbol{x}}_i^{(t)}-\check{\boldsymbol{x}}_{li}^{(t)})/\beta]}\;\check{\boldsymbol{x}}_{ji}^{(t)}\,,\qquad \check{\boldsymbol{x}}_i^{(0)}=\mathbf{x}_i\,, \tag{6}$$

where $\check{\boldsymbol{x}}_{ji}^{(t)}$ is the j-th NN term of $\check{\boldsymbol{x}}_i^{(t)}$. The algorithm (6) directly limits the number of observations that the estimates manage. It is equivalent to a truncated BMS (4) with *variable* bandwidth $\{\beta_i\}_1^N$ and $\alpha_i = 3\beta_i$; where these coefficients provide a fixed number n of observations to each estimate $\tilde{\boldsymbol{x}}_i^{(t)}$. The constraint $n < N/m$ involves a $\beta^* \leq \max_i(\beta_i)$ which satisfies the condition (5); hence, it enables the convergence to different centroids (see Appendix). Indeed as in (3), the objective function of (6) is given by $C_n = \sum_{i=1}^N \sum_{j=1}^n K\Big[(\boldsymbol{x}_i - \boldsymbol{x}_{ji})/\beta\Big]$; this is maximized by $\boldsymbol{x}_i = \boldsymbol{x}_{ji}\,,\ j = 1, 2 \ldots n$, and theoretically yield $\lfloor N/n \rceil$ groups.

It may be noted that the estimator (6) is related to the local mean (LM) algorithm, which computes arithmetic means on the neighbors of each point

$$\text{LM}: \quad \bar{\boldsymbol{x}}_i^{(t+1)} = \frac{1}{n+1} \sum_{j=0}^{n} \bar{\boldsymbol{x}}_{ji}^{(t)}\,, \qquad \bar{\boldsymbol{x}}_{ji}^{(0)} = \mathbf{x}_{ji}\,, \tag{7}$$

where $\mathbf{x}_{ji}$ is the j-th NN term of $\mathbf{x}_i = \mathbf{x}_{0i}$. The main difference between (7) and (6) is in the weights w_{ji}, which are uniform and decreasing respectively. Their difference should vanish as $t \to \infty$, because both methods shrink data while their bandwidths β, n remain constant. Unlike the K-means method (which randomly select m initial values, and does not shrink data), the algorithm (7) always converges to its global optimum, and for $n \leq \min_k(N_k) < N/m$ provides multiple centroids.

2.2 Stopping Rules. There are other ways to control the bias of the blurring estimator (2), without changing its structure. The first consists of letting the bandwidth $\beta \to 0$ as $t \to \infty$ at a cubic rate, which is the convergence rate of (2), see Carreira-Perpiñán (2006). Alternatively one may stop the iteration process when subsequent or neighboring estimates become sufficiently close, that is

$$\begin{aligned} \text{MS}: \quad \hat{t} &= \min\left\{t : \frac{1}{N}\sum_{i=1}^{N} \left\| \hat{\boldsymbol{x}}_i^{(t)} - \hat{\boldsymbol{x}}_i^{(t-1)} \right\| < \delta \right\}, \\ \text{BMS}: \quad \tilde{t} &= \min\left\{t : \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p} \left\| \tilde{\boldsymbol{x}}_i^{(t)} - \tilde{\boldsymbol{x}}_{ji}^{(t)} \right\| < \delta \right\}, \end{aligned} \tag{8}$$

for some $p \geq 1$, where $\tilde{\boldsymbol{x}}_{ji}$ is the j-th NN of $\tilde{\boldsymbol{x}}_i$, and $\delta > 0$ is a small constant. The first criterion does not work for BMS (2) with infinite support kernels, because the data points move together towards a single center (after having reached their

nearest local centroids). This problem is avoided by the solution (8), which considers simultaneous neighboring points. It represents a simple alternative to the *entropy* approach of Carreira-Perpiñán (2006), that focuses on the histogram of estimate variations. Obviously, stopping criteria for BMS (2) are also useful for algorithms (4)-(7), to reduce computations and the bias caused by the bad design of β.

3. Bandwidth Selection

As shown in Section 2, convergence and performance of MS-type algorithms fundamentally depend on the size of their bandwidth. The selection methods of β developed in kernel density estimation (e.g. Silverman, 1986) are not suitable for MS clustering. In this section we provide data-driven methods for the algorithms (1),(2),(6),(7); next we will check their validity with simulation experiments.

3.1 Clustering Statistics. We start with techniques based on the identification of the number of groups of classical cluster analysis. In this context, the optimal value of m is determined by maximizing information criteria (which balance likelihood function and cluster complexity), or F-type statistics, which compare explained and residual variances, weighted by their degrees of freedom. In the MS-algorithms m depends on β, thus one can select the bandwidth as

$$\hat{\beta} = \arg\max_{\beta} \left[F_N\big(m(\beta)\big) = \frac{B(m)/(m-1)}{W(m)/(N-m)} \right], \tag{9}$$

where B, W, are between-group and within-group total deviances, computed on the original data with the partition $\{\mathbf{x}_{i,k}\}_{k=1}^{m(\beta)}$ provided by MS-estimates. In practice,

$$B(m) = \sum_{k=1}^{m} N_k \|\bar{\mathbf{x}}_k - \bar{\bar{\mathbf{x}}}\|^2, \qquad W(m) = \sum_{k=1}^{m} \sum_{i=1}^{N_k} \|\mathbf{x}_{i,k} - \bar{\mathbf{x}}_k\|^2,$$

where $\bar{\mathbf{x}}_k = N_k^{-1} \sum_{i=1}^{N_k} \mathbf{x}_{i,k}$ and $\bar{\bar{\mathbf{x}}} = N^{-1} \sum_{i=1}^{N} \mathbf{x}_i$ are data centroids.

Algorithm. The detailed procedure for selecting the value of β with the F-statistic (9) in MS-type algorithms is as follows:

Step 1. Define a set $\{\beta_l\}_1^L$, such that the number of groups $m(\beta_1) \geq 2$ and $m(\beta_L) \leq M$, where $M \ll N$ is the expected maximum number.

Step 2. For each β_l run MS, BMS, LM algorithms until convergence to their centroids: $\hat{\boldsymbol{\mu}}_{i,k} = \lim_{t\to\infty} \hat{\boldsymbol{x}}_i^{(t)}(\beta_l)$, where $k = 1, \ldots, m(\beta_l)$.

Step 3. Perform data partition and labeling, i.e. find the groups $\{\mathbf{x}_{i,k}\}$ which correspond to every center $\hat{\boldsymbol{\mu}}_{i,k}$, where $i = 1, \ldots, N_k$.

Step 4. Use the clusters $\{\mathbf{x}_{i,k}\}$ to compute data-centroids $\bar{\mathbf{x}}_{i,k}$, between-group $B(m(\beta_l))$ and within-group $W(m(\beta_l))$ deviances.

Step 5. Compute the statistics $F_N(m(\beta_l))$ (9) and find the maximum value over the set $l = 1, \ldots, L$. This defines the optimal β.

Data partition at Step 3 usually requires the identification and removal of anomalous (isolated) observations. However, this problem does not arise in blurring algorithms, since their shrinking ability also allows for robustness.

Another approach for selecting the number of groups is based on the *silhouette* index S_N of Rouseeuw (1986). This measures the goodness of a classification by comparing the data distances within clusters, with the data distances between clusters. As in (9), MS-algorithms provide the partition $\{\mathbf{x}_{i,k}\}$, then the selection is

$$\hat{\beta} = \arg\max_{\beta} \left[S_N\Big(m(\beta)\Big) = \frac{1}{N} \sum_{k=1}^{m} \sum_{i=1}^{N_k} \frac{(\,\bar{B}_{i,k} - \bar{W}_{i,k}\,)}{\max\,(\bar{B}_{i,k}, \bar{W}_{i,k})} \right], \tag{10}$$

where $\bar{W}_{i,k}, \bar{B}_{i,k}$ are the mean distances of the i-th observation from the members of its class (k-th), and from the members of its nearest group (h-th), that is:

$$\begin{aligned} \bar{W}_{i,k} &= (N_k - 1)^{-1} \sum_j \|\mathbf{x}_{i,k} - \mathbf{x}_{j,k}\|\,, \\ \bar{B}_{i,k} &= \min_{h\neq k} \Big(N_h^{-1} \sum_j \|\mathbf{x}_{i,k} - \mathbf{x}_{j,h}\| \Big)\,. \end{aligned}$$

Notice that $S_N \in (-1, +1)$, where high values indicate good clustering, because $B > W$; the algorithm for computing (10) is similar to that of (9).

3.2 Specific Indicators. A heuristic solution for MS (1) is based of the difference between modal values and data-centroids. While the modes are the limit of the estimates $\hat{\boldsymbol{x}}_i^{(t)}$, the centroids $\hat{\mathbf{x}}_k$ are the corresponding means computed on original data. Since the two statistics are based on the same membership list, significant difference between them means inadequacy of fitting; hence, the selection of β is

$$\hat{\beta} = \arg\min_{\beta} \sum_{k=1}^{m} \Big\| \hat{\boldsymbol{\mu}}_k(\beta) - \hat{\mathbf{x}}_k(\beta) \Big\|\,, \tag{11}$$

where $\hat{\boldsymbol{\mu}}_k = \lim_{t\to\infty} \hat{\boldsymbol{x}}_i^{(t)}$ and $\hat{\mathbf{x}}_k = N_k^{-1} \sum_i \mathbf{x}_{i,k}$ use the same data-partition. The statistic (11) provides a fitting measure between data and MS estimates.

For BMS (2), we note that the index of concentration C_N (3) is monotonically increasing in β, whereas the number of clusters m is monotonically decreasing in β. The optimal bandwidth should then balance the two quantities, so as to find a compromise between variance reduction (C) and cluster complexity (m). Thus, β can be selected by minimizing the sum of the two rescaled functions

$$\begin{aligned} \tilde{\beta} &= \arg\min_{\beta} \left[J_N(\beta) = \tilde{C}_N^*(\beta) + \tilde{m}^*(\beta) \right], \qquad (12) \\ \tilde{C}_N(\beta) &= N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} K\left[(\tilde{\boldsymbol{x}}_i^{(t)} - \tilde{\boldsymbol{x}}_j^{(t)})/\beta \right], \end{aligned}$$

where $\tilde{m}$ is the number of centroids of $\tilde{\boldsymbol{x}}_i^{(t)}$. The indexes $\tilde{C}, \tilde{m}$ are computed over a grid of $\{\beta_l\}$, and then are standardized (denoted by *), to have unit scales. This approach is similar to the information criteria, but avoids inferential aspects.

In the presence of a-priori informations on the distance between centers, one can select the bandwidth by using the relationship (5), namely

$$\beta = \min_{h,k} \| \boldsymbol{\mu}_h - \boldsymbol{\mu}_k \| / 6 \,, \qquad (13)$$

Similarly, for LM algorithm (7) one can select the NN size as n=$\min_k(N_k)$ or n=N/m. The validity of these solutions can be checked by simulations.

Finally, the stopping rules of iterations may also be useful for the bandwidth selection itself, because the size of β determines the speed of the convergence process. Small bandwidths yield many clusters, whereas large β_s lead to few clusters; in both cases, many iterations are necessary to converge. Instead, the optimal bandwidth enables the right/easy clustering of the data; hence, it should involve the smallest number of iterations t. It follows that the bandwidth selection based on the stopping criterion (8) is simply given by:

$$\tilde{\beta} = \arg\min_{\beta} \tilde{t}_{\varepsilon}(\beta) \,. \qquad (14)$$

It should be noted that the use of stopping rules enlarges the set of admissible β_s (those which provide unbiased centroids), and in general reduces the bias of estimates caused by non-optimal bandwidths (see Grillenzoni, 2014).

The path of statistics (9)-(14) may not be strictly convex, but they have the global minima in correspondence of the optimal β. To avoid conflict between the methods, one can build a single index by summing their normalized values.

4. Simulations and Applications

As a simulation experiment, we consider the Gaussian mixture model of Rao et al. (2009), which consists of m=16 bivariate components (d=2), with spherical covariance matrix $\mathbf{I}_2\sigma^2$=0.01, and means placed on the unit circle ($\mathcal{C}$):

$$f(\boldsymbol{x}) = \sum_{k=1}^{16} \pi_k \, \mathrm{N}\Big(\boldsymbol{x}; \boldsymbol{\mu}_k, \mathbf{I}/100\Big) \,, \quad \boldsymbol{x} \in \Re^2, \;\; \boldsymbol{\mu}_k \in \mathcal{C}(0,1) \,. \tag{15}$$

The weighs are not uniform: $0.044 \leq \pi_k \leq 0.088$, but $\sum_k \pi_k$=1; the number of simulated data is N=1500, and a sample is displayed in Figure 1(a).

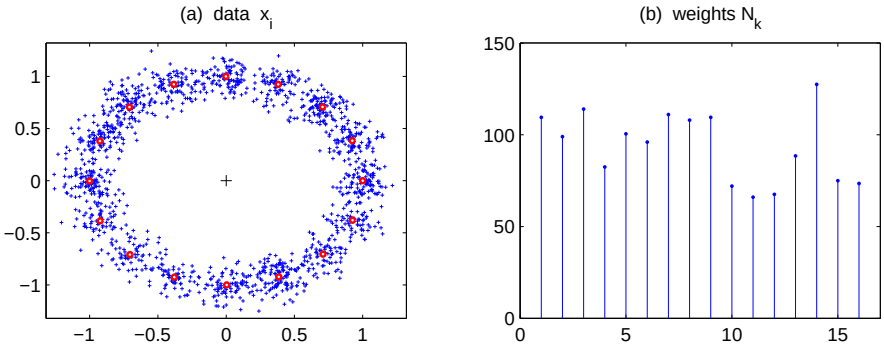

Figure 1. Simulated data: (a) Centers $\boldsymbol{\mu}_k$ (•, red) and N=1500 data $\mathbf{x}_i$ (+, blue) generated with the model (15); (b) Size of components $N_k = \pi_k N$.

Given the proximity of the centers, it is difficult to recover them by using classical clustering methods; however, the adaptive nature of MS-algorithms can tackle the problem. Rao et al. (2009) used the bandwidth β=σ=0.1 and two stopping criteria for the algorithms (1), (2), which yielded t=46, 20 iterations respectively. Their results showed the effectiveness of MS and the significant bias of BMS, which provided only m=14 centroids and 12 correct estimates. We have confirmed these results; however, by using the criterion (13), we note that the optimal bandwidth for BMS should be around β=0.4/6=0.067.

To shed light on this issue, we run BMS (2) with $\beta \in [0.05, 0.10]$, and t=100 iterations. Figure 2 displays the resulting functions $\tilde{C}_N$ (3) and $\tilde{m}$; they show that BMS provides 16 centroids for $\beta \in$[0.075, 0.08], and that C_N is constant for $\beta \leq$ 0.08. At a finer grid, we check that BMS is unbiased for $\beta \in \mathcal{S}_\beta = [0.072, 0.084]$, with centroids which are close to the modes of MS (see Figure 3). The width of this set is significantly different from zero and slightly decreases by letting t=1000, which means that BMS converges for suitable bandwidths. Notice that the identified set $\mathcal{S}_\beta$ does not contain β=0.067 of equation (13), because the value z=3 in (13) may be too large. In fact, using P($|z| \leq$2.57)=99% one can get β=0.4/(2*2.57)=0.077, where 0.4 is the minimum distance of centers.

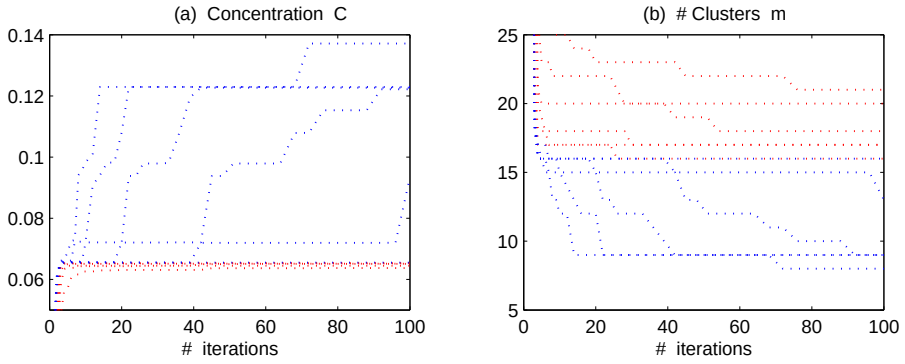

Figure 2. Path of BMS statistics: (a) $\tilde{C}_N(t|\beta)/N^2$; (b) $\tilde{m}(t|\beta)$; for $\beta \in$[0.05-0.075] ($\cdot$, red) and $\beta \in$[0.08-0.1] ($\cdot$, blue).

Table 1. Results of MS-methods applied to the data in Figure 1(a): $\mathcal{S}_\beta$ is the set of bandwidths which provide unbiased estimates (m=16). β_1^* is the *minimum* value which leads to a single centroid (m=1). The value of n of the algorithm (6) is 100.

Method	Eq.	$t_{\max}$	$\mathcal{S}_\beta$=$[\beta_{\min}, \beta_{\max}]$	$\dot{\beta}_0$	$\sum_k \|\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|/m$	$\max_k \|\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|$	β_1^*
MS	(1)	100	0.067 - 0.109	0.088	0.018	0.049	0.86
BMS	(2)	100	0.072 - 0.084	0.078	0.017	0.047	0.31
BMS	(2)	3000	0.068 - 0.078	0.073	0.017	0.045	0.28
nBMS	(6)	100	0.072 - 0.092	0.082	0.017	0.046	NA
sBMS	(8)	12	0.072 - 0.096	0.084	0.017	0.047	0.35
LM	(7)	100	61 - 65	63	0.019	0.053	876

We can also check the equation (5a) (which states that $\beta > 2/6$=0.33, where 2 is the range of data), by finding that the minimum β for which BMS converges to a single point is indeed β_1^*=0.31. Table 1 resumes the results of all MS algorithms applied to the data of Figure 1(a). The various methods perform similarly in terms of mean and maximum errors of the centroids estimated with $\dot{\beta}_0$ (the central value of $\mathcal{S}_\beta$). The only noticeable fact is that $\mathcal{S}_\beta$ of MS is wider than that of BMS; however, the latter can be enlarged by using the stopping rule (8) or the NN version (6) with n=100. In particular, the NN solution does not admit a finite β_1^*. Finally, the optimal value of n for LM (7) is close to $\min_k(N_k)$=66 mentioned in Section 2.

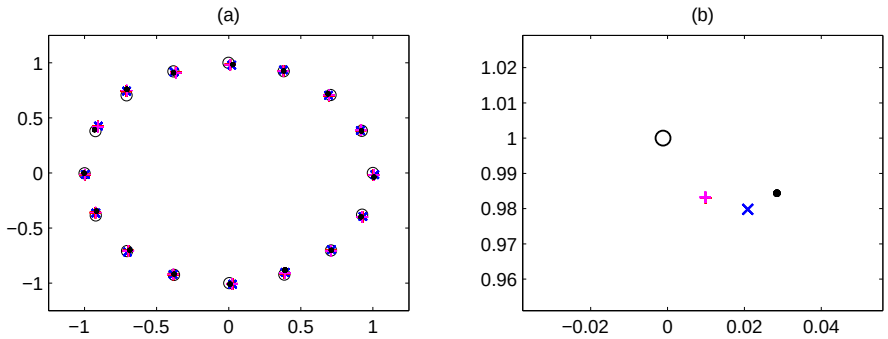

Figure 3. Final estimates $\hat{\boldsymbol{x}}_i^{(100)}$ obtained with the bandwidths $\dot{\beta}_0$ in Table 1: (a) Ground centers (o, black), MS ($\times$, blue), BMS (+, red), nBMS (+, magenta), LM ($\bullet$, black); (b) A particular of the first center.

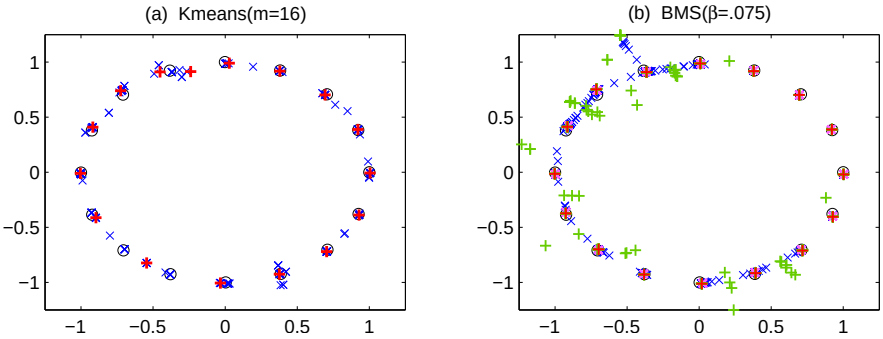

Figure 4. Comparison of clustering methods for the data in Figure 1(a): (a) Centroids $\bar{\mathbf{x}}_k$ provided by K-means with m=16 and r-replications: r=3-14 ($\times$, blue), r=15 (+, red); (b) Results of BMS with t=3-14 iterations: estimates $\tilde{\boldsymbol{x}}_i^{(t)}$ ($\times$, blue), their data-centroids $\tilde{\mathbf{x}}_k$ (+, green); both quantities at iteration t=15 ($*$, red).

Figure 3 displays the final estimates $\hat{\boldsymbol{x}}_i^{(100)}$ generated by the central value $\dot{\beta}_0$ in Table 1. All methods perform similarly and satisfactorily and, as said in the previous section, BMS and its NN version are almost identical. Figure 4 compares the performance of K-means and BMS methods; Cheng (1995) showed that the latter is a limiting case of the first. K-means was implemented with m=16 groups and r=3-15 replications (of starting points), and BMS with β=0.075 and t=3-15 iterations. It can be seen that K-means can be biased, whereas BMS reaches the target in few iterations. Figure 4(b) also shows the different path of BMS estimates $\tilde{\boldsymbol{x}}_i$ and their data-centroids $\tilde{\mathbf{x}}_k$, in converging to the same centers $\boldsymbol{\mu}_k$. This difference is exploited by the selection criterion (11).

4.1 Bandwidth Selection. The analyzes conducted so far have strongly relied on the use of the ground centers $\boldsymbol{\mu}_k$, which are unknown in real-life applications. We now check the ability of data-driven methods to select the MS bandwidths. The selection methods commonly used in kernel density estimation and its derivatives (e.g. Chacón and Duong, 2013) have provided very variable results, in general outside the admissible sets in Table 1. In particular, bandwidth values in Table 2 are external to $\mathcal{S}_\beta$=[0.067, 0.109] of the MS method; the sole exception is the univariate MISE plug-in method of Sheather and Jones (1991), implemented by Ripley and Wand (2014) in the program `KernSmooth`.

Table 2. Results of bandwidth selection performed with kernel smoothing packages of R-Cran (e.g. Duong, 2014), on the data of Figure 1(a). Methods are: CV=cross validation, BCV=biased cv, LSCV=least squares cv, NS=normal scale, PI=plug-in MISE, SCV=smoothed cv. The entries are the mean values $\hat{\beta} = (\hat{\beta}_1 + \hat{\beta}_2)/2$.

R-package	d	F	BCV	LSCV	NS	PI	SCV
`ks`	2	f	0.052	0.0024	0.045	0.008	0.010
`ks`	2	f'	.	0.148	**0.074**	0.015	0.025
`KernSmooth`	1	f	.	.	.	**0.075**	.
`sm`	2	f	0.048	.	0.210	0.063	.
`kedd`	1	f	0.046	0.041	.	0.376	0.052
`kedd`	1	f'	0.584	0.045	.	0.584	0.584

More useful results are provided by the methods described in Section 3. Figure 5 displays the estimates of the criteria (9)-(14) (over a grid of β) for the BMS algorithm with the stopping rule (8), and with 100 iterations (in red). It can be seen that all methods select the value β=0.08, which is close to the optimal one in Table 1. Notice that some paths are not strictly convex; however, they have well defined global minima/maxima.

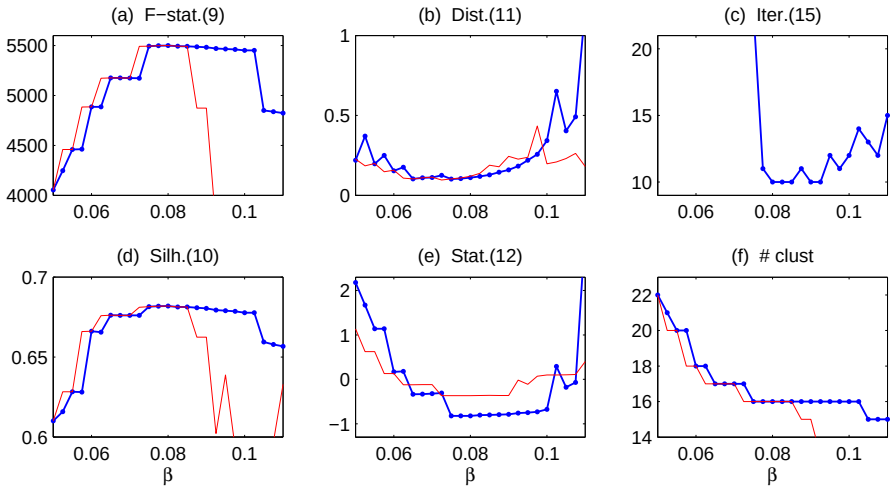

Figure 5. Graph of selection criteria of BMS bandwidth: (a) F-statistic (9), (b) Distance (11); (c) # Iterations (8); (d) Silhouette (10); (e) J-statistic (12); (f) # Clusters. Estimates with: stopping rule (8) ($\cdot-$, blue), and 100 iterations ($-$, red).

Figure 6 shows the combination of criteria (9)-(14) for MS, BMS, LM with the stopping rule (8) and with 100 iterations (in red). The various functions are standardized and changed in sign so as to make their scale and pattern homogeneous. It can be noted that all methods select values which agree with $\dot{\beta}_0$ in Table 1.

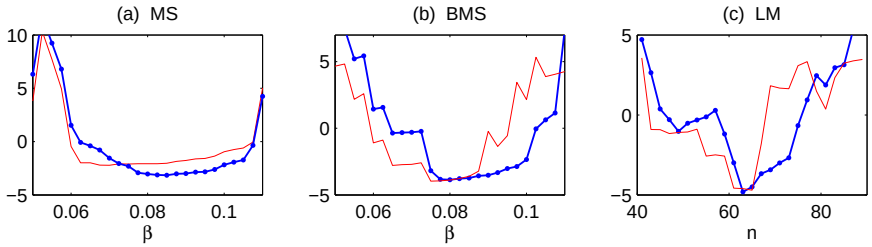

Figure 6. Sum of standardized criteria (9)-(14) for: (a) MS; (b) BMS; (c) LM. Estimates with stopping criteria (8) ($\cdot-$, blue), and with 100 iterations ($-$, red).

4.2 Monte Carlo. Previous results deal with a single sample from the model (15). We now perform a Monte Carlo simulation experiment, which consists of 100 replications. We focus on the BMS algorithm with stopping rule (8) and ε=1e-4. Figure 7(a) displays the selection functions of β, obtained by combining criteria (9)-(14), and their mean value (in white). Figure 7(b) shows the path of admissible sets $\mathcal{S}_\beta$ (which provide 16 centers) sorted by width, and the bandwidths selected with the functions in Figure 7(a). The results agree with those in Table 1, since the mean values (over 100 replicates) are $\bar{\mathcal{S}}_\beta$=[0.072, 0.098] and $\bar{\beta}$=0.079.

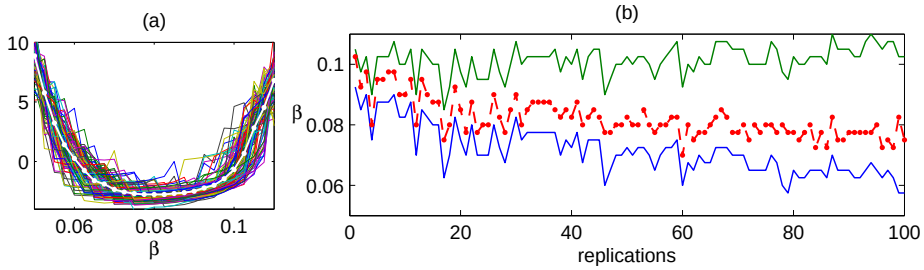

Figure 7. Results of BMS over 100 replications of (15): (a) Selection functions of β which combine (9)-(14), mean value ($\cdot-$, white); (b) Admissible sets $\mathcal{S}_\beta$ sorted by width (solid); Bandwidths $\hat{\beta}$ selected with the functions in panel (a) ($\cdot-$, red).

4.3 Image Analysis. We consider a problem of image segmentation which deals with the picture "hand" in Figure 8(a). As in Carreira-Perpiñán (2006), the original image is resized to a $40 \times 50 \times 3$ array, which provides a matrix with N=2000 rows; the resulting d=5 columns are standardized in order to use a single bandwidth. We apply MS-methods by selecting β through the combination of criteria (9)-(14); the results are shown in Figure 8. The K-means method in Panel (e) is run with the ideal value m=3 and 15 replicates; it can be noted that it is unable to detect the ring, and also fails to identify the background table as a single cluster. MS-methods rightly detect the 3 parts of the image, but the simple MS is not optimal in detecting the ring. Its problems also arise from the fact that the function in Panel (b) indicates two values β=0.6, 1; the first was excluded because yields 5 clusters. A similar problem is present in LM method (7) in Panels (d), where only n=600 provides 3 clusters; however, Panel (h) shows that its segmentation is similar to that

of K-means. Finally, the best method is BMS (2) with β=0.7, for three reasons: the convexity of the selection function 8(c), the convergence in 11 iterations with the stopping rule (8), and the right segmentation in 3 coherent groups.

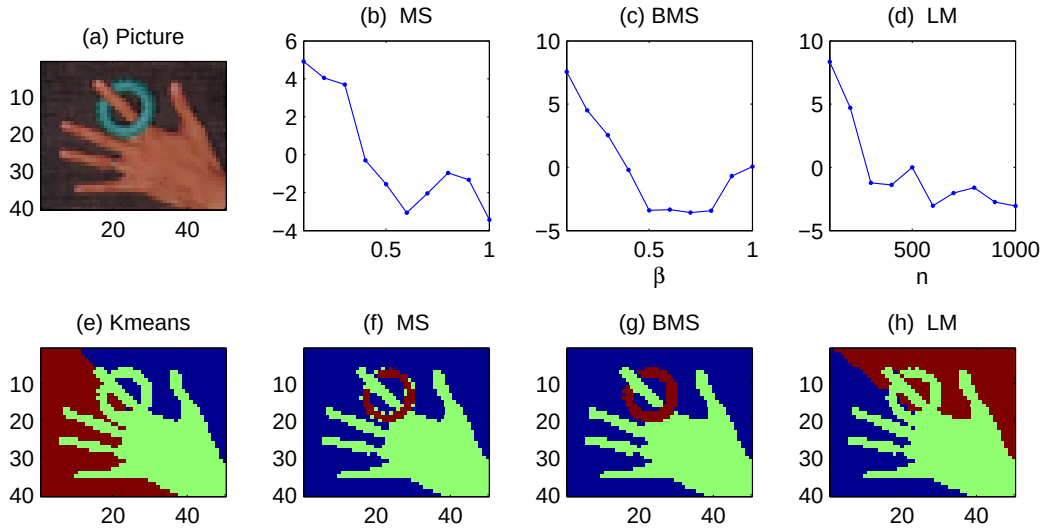

Figure 8. Results of image segmentation: (a) Resized test image; (b)-(d) Selection functions for β, n which combine (9)-(14); (e)-(h) Clustering results for K-means, MS, BMS and LM methods.

4.4 Complex data-sets. As suggested by a referee, MS methods may have problems when the clusters have very different dimensions and non-homogeneous dependence structure. We check this issue by simulating a 3D Gaussian mixture, which has m_o=8 centers placed on the vertexes of a cube, 2x4 proportions 0.05$\leq$ π_k $\leq$0.25 and 5 different covariance matrices. A sample of size N=1000 is displayed in Figure 9(a); Panels (c)-(g) show the bandwidth selection functions obtained by combining the statistics (9)-(14). Notice that only the LM method fails to identify its coefficient n; therefore, we adopt n^*=120 which allows for 8 clusters. This situation is a consequence of the fact that clusters have very different size, with N_k ranging from 50 to 200. Table 3 provides the MSE of modal estimates and data centroids (with respect to the ground centers $\boldsymbol{\mu}_k$); the best method is nBMS implemented with n=120, whereas the results of LM are very disappointing.

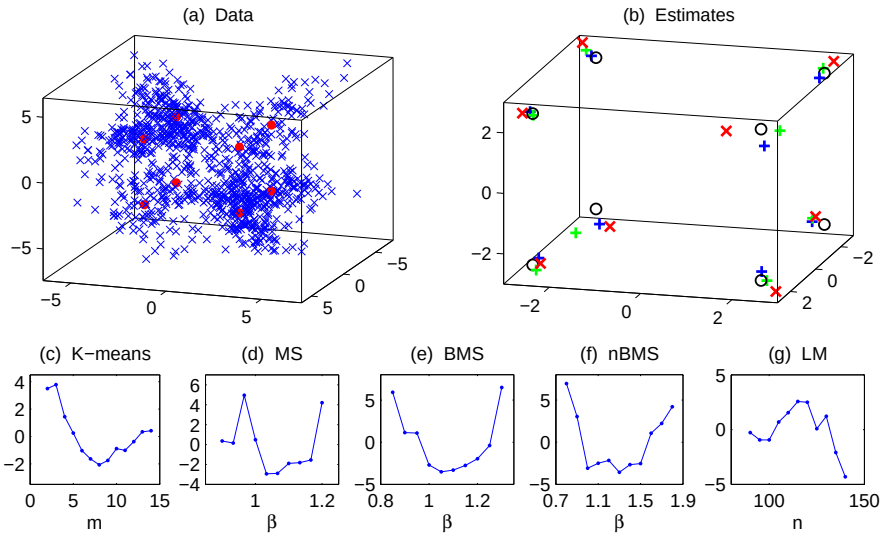

Figure 9. 3D Gaussian mixture simulation: (a) Data (x, blue), means (•, red); (b) Center estimates provided by K-means (×, red), and BMS: $\tilde{\boldsymbol{x}}_i^{(t)}$ (+, blue), its data-centroids $\tilde{\mathbf{x}}_k$ (+, green); (c)-(g) Bandwidth selection functions of various methods.

Table 3 also provides the results of MS methods applied to biological data, as Iris flowers (N=150, d=4, m_o=3), and Wine features (N=178, d=13, m_o=3), available from the UCI database (`http://archive.ics.uci.edu/ml/`). In both cases it is known the *true* cluster composition, so that wrong classifications can be directly detected. The two data-sets need different transformations in order to minimize the number of bad classifications of K-means: Iris prefers normalization of variables by their maximum value; whereas Wine wants classical standardization. Both methods enable to use a single bandwidth in MS methods. In K-means we select the number of groups m by combining the Silhouette index (10) with those of Davies-Bouldin, Calinski-Harabasz and Krzanowski-Lai (see Wang et al. 2009). In the Iris example, however, these are not able to detect the right value m_o=3; thus, in this and similar cases the coefficients (denoted by *) are assigned to obtain m_o groups. Summarizing Table 3, we can state that the worst method is the simple MS, whereas the best one is nBMS: the Gaussian blurring algorithm based on n nearest neighbors. It provides the best combination of fastness and accuracy and its bandwidth β can be selected with data-driven methods; also the value of n can be selected in this way, although nBMS is not very sensitive to it.

Table 3. Results of MS-methods applied to the data of Figure 9, and to Iris and Wine data-sets. The coefficients m, β, n with * are *not* selected with data-driven methods, and must satisfy the ground value m_o. The algorithm (6) uses $n = N/m$.

Data	Statistics	K-means	MS	BMS	nBMS	LM
Figure 9	m, β, n	8	1.05	1.05	1.15	120*
(original)	MSE($\hat{\boldsymbol{x}}_i^{(t)}, \boldsymbol{\mu}_k$)	.	0.671	0.415	**0.379**	2.54
·	MSE($\bar{\mathbf{x}}_k, \boldsymbol{\mu}_k$)	0.571	0.518	0.533	0.455	1.82
Iris	m, β, n	3*	0.073*	0.073	0.073	50
(normalized)	# wrong	6	26	**5**	**5**	6
Wine	m, β, n	3	1.42*	1.05*	2.3	50
(standardized)	# wrong	6	27	26	**5**	**5**

5. Conclusions

In this paper we have analyzed and compared clustering methods based on the mean-shift of kernel densities. The blurring version is a natural extension of the simple MS algorithm; it enjoys properties of fast convergence but also accuracy, as shown by many numerical applications. We have checked that its problems of bias (i.e. the tendency to converge to a single cluster), largely depend on the bandwidth selection. In this paper we have provided automatic (data-driven) criteria of bandwidth selection which lead to the correct detection of the number of clusters. The introduction of stopping criteria, and the nearest-neighbor implementation, can also improve the performance of the BMS method.

Appendix: Convergence of BMS Estimators

The analysis of the conditions of convergence of the algorithms (2),(4) and (6) requires some results of Cheng (1995) and Chen (2015). The latter considers BMS estimators in which the kernels may not be integrable. They must only be positive and decreasing with respect to the distance (PDD); that is $1 \geq K(u, v) = K(\|u - v\|) \to 0$ as $\|u - v\| \to 0$. Then, for a finite data set $\{\mathbf{x}_i\}_1^N$ one has:

Theorem (Chen, 2015). *If the function $K(x)$ in the BMS (2) is PDD, then there exist values $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M\}$, with $M \leq N$, such that $\lim_{t\to\infty} \tilde{\boldsymbol{x}}_i^{(t)} = \boldsymbol{c}_i$ for all i.*

The convergence here is of numerical type (as the data do not change), and does not mean that limit values $\boldsymbol{c}_i$ are different. The proof is elegant but long, and can be summarized as follows: One first considers the convex hulls (minimal convex sets) of the estimates; since they have a nested structure they converge to $\mathcal{H}_{\boldsymbol{C}}$. Next, for each vertex of $\mathcal{H}_{\boldsymbol{C}}$, there is at least one point which tends to it; finally, the influence between converged estimates and the other points tends to zero, i.e.

$$\lim_{t\to\infty} K\Big(\tilde{\boldsymbol{x}}_i^{(t)} - \tilde{\boldsymbol{x}}_j^{(t)}\Big) = 0 \quad \text{for all } j, \text{ such that} \quad \lim_{t\to\infty} \tilde{\boldsymbol{x}}_j^{(t)} \neq \lim_{t\to\infty} \tilde{\boldsymbol{x}}_i^{(t)} \tag{16}$$

The theorem only requires the PDD condition; however, there are kernels that produce trivial results, in which all data points converge to a single center $\boldsymbol{c}_0$. Theorem 3 of Cheng (1995) shows that this occurs when the kernel's support covers the entire data set; this result can be seen as a Corollary of the above.

Corollary (Chen, 2015). *Let the range $R_x = \max_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|$. If $K(x)$ is PDD with $K(R_x) > 0$, then there exists a point $\boldsymbol{c}_0$, such that $\lim_{t\to\infty} \tilde{\boldsymbol{x}}_i^{(t)} = \boldsymbol{c}_0$ for all i.*

As a proof note that $\left\|\tilde{\boldsymbol{x}}_i^{(t)} - \tilde{\boldsymbol{x}}_j^{(t)}\right\| \leq R_x$ for all t, i, j, by the convex hull property; and since K is decreasing, one has $K\Big(\tilde{\boldsymbol{x}}_i^{(t)} - \tilde{\boldsymbol{x}}_j^{(t)}\Big) \geq K(R_x) > 0$ for all t, i, j. This contradicts the condition (16); therefore all estimates must converge to $\boldsymbol{c}_0$.

It follows that a condition for non-trivial convergence of BMS, is that its function $K(x) = 0$ for $x > \mathrm{x}_0$, with $\mathrm{x}_0 < R_x$. For infinite support kernels, this may be approached by imposing an essential *truncation* on K; in the Gaussian case, the 3-sigma rule provides $K^*_{3\beta}(z) = K_\beta(z)\, I(|z| \leq 3\beta)$, where $I(\cdot)$ is the indicator function. In this case the kernel diameter is $\mathrm{x}_0 = 6\beta < R_x$, and the condition becomes $\beta < R_x/6$. However, this is only a necessary condition, the sufficient one replaces R_x with the range of centers R_c. We then have the following:

Proposition. *Let the data $\{\mathbf{x}_i\}_1^N$ have local centroids $\{\boldsymbol{c}_k\}_1^m$ with minimum distance r_c, and let $\boldsymbol{c}_{ki}$ be the nearest center to $\mathbf{x}_i$. Then BMS estimates with bandwidth $\beta < r_c/6$ and kernel $K^*_{3\beta}(z)$, are such that $\lim_{t\to\infty} \tilde{\boldsymbol{x}}_i^{(t)} = \boldsymbol{c}_{ki}$ for all i.*

The proof is a consequence of the above results, and of Theorem 4 in Cheng (1995). It is known that the convergence of BMS estimates proceeds in two phases: in the first, $\tilde{\boldsymbol{x}}_i^{(t)} \to \boldsymbol{c}_{ki}$ quickly, subsequently $\boldsymbol{c}_{ki} \to \boldsymbol{c}_0$ slowly (see Carreira-Perpiñán, 2006). However, owing to the above Corollary and to the kernel $K^*_{3\beta}$, the second phase takes place only if $\beta > R_c/6$, where $R_c = \max_{h,k} \|\boldsymbol{c}_k - \boldsymbol{c}_h\|$. On the contrary, if $\beta < r_c/6$, with $r_c = \min_{h,k} \|\boldsymbol{c}_k - \boldsymbol{c}_h\|$, then estimates remain on their nearest centroids by result (16). In fact, if two point data $\mathbf{x}_i$ on $\boldsymbol{c}_k$ and $\mathbf{x}_j$ on $\boldsymbol{c}_h$ have no actual influence, as $K^*_{3\beta}(\boldsymbol{c}_k - \boldsymbol{c}_h) = 0$, then they do not attract reciprocally.

As regards the nBMS algorithm (6) with $n < N/m$, notice that it is equivalent to a truncated BMS with variable β_i and $\alpha_i = 3\beta_i$; these coefficients have to allow the same number n of observations to each estimate. The resulting algorithm is encompassed by a truncated BMS with bandwidth $\beta^* \leq \max_i(\beta_i)$ and $\alpha^* = 3\beta^*$, which converges to multiple centroids. In fact, since n and β^* vary proportionally, then there exists a $n < N/m$ such that β^* satisfies the condition (5). In particular, as in (3) the objective function of (6) is given by $C_n = \sum_{i=1}^{N} \sum_{j=1}^{n} K\Big[(\boldsymbol{x}_i - \boldsymbol{x}_{ji})/\beta\Big]$; this is maximized by $\boldsymbol{x}_i = \boldsymbol{x}_{ji}\,,\ j = 1, 2 \ldots n$, and theoretically yield $m < N/n$ groups. Local centroids of nBMS then correspond to local maxima of $C_N(3)$.

Dynamic Analysis. A general framework for analyzing the behavior of BMS estimators (2)-(7) is to consider them as dynamical systems. Rewrite the algorithms as $\tilde{\boldsymbol{x}}_i^{(t+1)} = \sum_j \tilde{w}_{ij}^{(t)} \tilde{\boldsymbol{x}}_j^{(t)}$, and in matrix form one has

$$\tilde{\boldsymbol{X}}_\beta^{(t+1)} = \tilde{\boldsymbol{W}}_\beta^{(t)} \tilde{\boldsymbol{X}}_\beta^{(t)} = \prod_{s=0}^{t} \tilde{\boldsymbol{W}}_\beta^{(s)} \mathbf{X} = \tilde{\boldsymbol{P}}_\beta^{(t)} \mathbf{X} \tag{17}$$

where $\tilde{\boldsymbol{X}}_{N\times d}$ are the estimates, $\tilde{\boldsymbol{W}}_{N\times N}$ are the kernel weights, and $\mathbf{X}_{N\times d}$ are the data. Equation (17) is a large scale inhomogeneous Markov system, and $\tilde{\boldsymbol{W}}_\beta^{(t)}$ are row-probability matrices, because $\sum_j w_{ij}^{(t)} = 1$ for all i. Now, if $K(\cdot)$ has a bounded support as (4), or $\sum_j$ is limited to $n < N/m$ terms as in (6) and (7), then $\tilde{\boldsymbol{W}}_\beta^{(s)}$ are sparse and $\tilde{\boldsymbol{P}}_\beta^{(t)}$ converges to a block-diagonal matrix $\boldsymbol{P}_\beta$. The system (17) will then reach the steady-state $\tilde{\boldsymbol{X}} = \boldsymbol{P}_\beta \mathbf{X}$, with $\tilde{m} > 1$ different rows (centroids).

On the contrary, when the kernel support is broad as in (2), the weights are $\tilde{w}_{ij}^{(t)} > 0$ for all i, j, t, and $\tilde{\boldsymbol{P}}_\beta^{(t)}$ converges to a dense matrix with identical rows $\boldsymbol{\pi}'$

(the so-called equilibrium distribution). This result is due to the averaging affect of probability matrices, and is well known when $\tilde{\boldsymbol{W}}$ is constant, because $\tilde{\boldsymbol{P}}^{(t)} = \tilde{\boldsymbol{W}}^{t}$. This can be extended to variable matrices under the condition $\|\tilde{\boldsymbol{W}}_{\beta}^{(t)} - \tilde{\boldsymbol{W}}_{\beta}\| \to 0$ (see Isaacson and Madsen, 1976 p.170), which holds for (17) given the convergence to fixed points of BMS estimates for all β (see Theorem of Chen, 2015).

References

ALIYARI GHASSABEH, Y. (2013), "On the Convergence of the Mean Shift Algorithm in the One-Dimensional Space," *Pattern Recognition Letters*, *34*, 1423-1427.

CARREIRA-PERPIÑÁN, M.Á. (2006), "Fast Nonparametric Clustering with Gaussian Blurring Mean Shift," in *Proceedings of 23rd International Conference on Machine Learning, ICML 2006*, 153-160.

CARREIRA-PERPIÑÁN, M.Á. (2007), "Gaussian Mean Shift is an EM Algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*, 767-776.

CARREIRA-PERPIÑÁN, M.Á. (2008), "Generalized Blurring Mean Shift Algorithms for Nonparametric Clustering," *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008.*

CHACÓN, J.E., and DUONG, T. (2013), "Data-driven Density Derivative Estimation, with Applications to Nonparametric Clustering and Bump Hunting," *Electronic Journal of Statistics, 7*, 1-3169

CHEN, T.-L. (2015), "On the Convergence and Consistency of the Blurring Mean Shift Process," *Annals of the Institute of Statistical Mathematics, 67*, 157-176.

CHENG, Y. (1995), "Mean Shift, Mode Seeking and Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 17*, 790-799.

COMANICIU, D., and MEER, P. (2002), "Mean Shift: A Robust Approach Toward Feature Space Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 603619

DUONG, T. (2014), *Package 'ks'*, ver. 1.9.1., Cran R Project, available at: `http://cran.r-project.org/web/packages/ks/ks.pdf`

FUKUNAGA, K., and HOSTETLER, L.D. (1975), "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition," *IEEE Transactions on Information Theory, 21*, 32-40.

GRILLENZONI, C. (2007), "Pattern Recognition via Robust Smoothing, with Application to Laser Data," *Australian & N.Z. Journal of Statistics, 37*, 137-153.

GRILLENZONI, C. (2014), "Detection of Tectonic Faults by Spatial Clustering of Earthquake Hypocenters," *Spatial Statistics, 7*, 62-78.

ISAACSON, D.L., and MADSEN, R.W. (1976), *Markov Chains, Theory and Applications*, New York: Wiley.

LI, X., HU, Z., and WU F. (2007), "A Note on the Convergence of the Mean Shift," *Patter Recognition, 40*, 1756-1762.

RAO, S., DE MEDEIROS MARTINS A., and PRÍNCIPE, J. (2009), "Mean Shift: An Information Theoretic Perspective. *Patter Recognition Letters, 30*, 222-230.

RIPLEY, B., and WAND M. (2014), *Package 'KernSmooth'*, ver. 2.23-12, at link `http://cran.r-project.org/web/packages/KernSmooth/KernSmooth.pdf`

ROUSEEUW, P.J. (1986), "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computation and Applied Mathematics, 20*, 53-65.

SHEATHER, S.J., and JONES, M.C. (1991), "A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society, B, 53*, 683-690.

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

WANG, K., WANG B., and PENG L. (2009), "Validation for Cluster Analyses. *Data Science Journal, 8*, 88-93.