An Algorithm for Data-Driven Bandwidth Selection

Dorin Comaniciu, *Member*, *IEEE*

Abstract—The analysis of a feature space that exhibits multiscale patterns often requires kernel estimation techniques with locally adaptive bandwidths, such as the variable-bandwidth mean shift. Proper selection of the kernel bandwidth is, however, a critical step for superior space analysis and partitioning. This paper presents a mean shift-based approach for local bandwidth selection in the multimodal, multivariate case. Our method is based on a fundamental property of normal distributions regarding the bias of the normalized density gradient. We demonstrate that, within the large sample approximation, the local covariance is estimated by the matrix that maximizes the magnitude of the normalized mean shift vector. Using this property, we develop a reliable algorithm which takes into account the stability of local bandwidth estimates across scales. The validity of our theoretical results is proven in various space partitioning experiments involving the variable-bandwidth mean shift.

Index Terms—Variable-bandwidth mean shift, bandwidth selection, multiscale analysis, Jensen-Shannon divergence, feature space.

1 Introduction

THE objective of variable-bandwidth kernel estimation is to improve the performance of kernel estimators by adapting the kernel bandwidth to the local data statistics. It can be shown that the estimation bias of sample point density estimators [11] decreases in comparison to the fixed-bandwidth estimators, while the covariance remains the same. Only recently, these estimators have been used in computer vision applications, such as histogram construction from color invariants [9]. We have introduced the variable-bandwidth mean shift as an adaptive estimator of the density's normalized gradient and applied it for mode detection in complex feature spaces [5]. Although theoretically promising, variable-bandwidth methods rely heavily on the selection of local bandwidth. In the case when the bandwidth is not properly selected, the performance is suboptimal and often worse than that of fixed-bandwidth methods.

Data-driven bandwidth[1] selection for multivariate data is a complex problem, largely unanswered by the current techniques [28, p. 109], [11]. Depending on the prior knowledge on input data, we distinguish two classes of problems. If the data statistics is homogeneous, then one *global bandwidth* suffices for the analysis. If the data statistics are, however, changing across the feature space, *local bandwidths* should be computed. Unfortunately, most of the tasks encountered in autonomous vision reduce to the latter class of problems, i.e., the input is represented by multdimensional features, whose properties (scales) are variable in space and might change in time. Examples of such tasks are background modeling, tracking, or segmentation.

One can identify two general approaches to bandwidth selection: *statistical analysis-based* and *task-oriented* methods. Statistical methods compute the global bandwidth by balancing between the bias and variance of the density estimate obtained with that bandwidth, over the entire space. Asymptotic approximations are used to express the quality of the density estimate. A reliable method for univariate data is the plug-in rule [24], shown superior to least-squares cross-validation and biased cross-validation [14], [21], [26, p. 46]. The global bandwidth, however, is not effective when data exhibits multiscale patterns. In addition, for the multivariate case the optimal bandwidth formula [25, p. 85], [28, p. 99] is of little practical use since it depends on the Laplacian of the unknown density being estimated. The most often used method for local bandwidth adaptation follows Arbamson's rule, which takes the bandwidth proportional to the inverse of the square root of a first approximation of the local density [1]. The proportionality constant is an important choice of the method [26, p. 46].

Task-oriented methods for bandwidth selection typically rely on the stability of feature space partitioning. The bandwidth is taken as the center of the largest operating range over which the same number of partitions are obtained for the given data [8, p. 541]. This strategy is also implemented within the framework of scale-space theory [17]. Nevertheless, it assumes that the space is homogeneous, i.e., all the partitions should have roughly the same scale, which is not always true. In a related class of techniques, the best bandwidth maximizes an objective function, which expresses the quality of space partitioning and is called index of cluster validity. The objective function compares inter- versus intra-cluster variability [13], [15], or evaluates the isolation and connectivity of the delineated clusters [22]. See [20], for an evaluation of a large set of such indices.

This paper presents a new and effective approach to local bandwidth selection for multimodal and multivariate data. The method estimates for each data point the covariance matrix which is *the most stable across scales*. The analysis is unsupervised and the only assumption is that the range of scales at which structures appear in the data is known. In almost all vision scenarios, this information is available from prior geometric, camera, or dynamical constraints. The selected bandwidth matrices are employed in the variable-bandwidth mean shift for adaptive mode detection and feature space partitioning, as shown in Fig. 1.

The paper is organized as follows: A more general form of the variable-bandwidth mean shift, including fully parameterized bandwidth matrices, is introduced in Section 2. Section 3 presents the theoretical criterion for bandwidth selection based on the normalized mean shift vector. Section 4 details the proposed algorithm and shows bandwidth selection experiments. In Section 5, we apply the variable-bandwidth mean shift to partition feature spaces. Discussions are presented in Section 6.

2 Variable-Bandwidth Mean Shift

Let $\mathbf{x}_i$, $i = 1 \ldots n$ be a set of d-dimensional points in the space R^d and assume that a symmetric positive definite $d \times d$ bandwidth matrix $\mathbf{H}_i$ is defined for each data point $\mathbf{x}_i$. The matrix $\mathbf{H}_i$ quantifies the uncertainty associated with $\mathbf{x}_i$ [12]. The sample point density estimator with d-variate normal kernel, computed at the point $\mathbf{x}$ is given by

$$\hat{f}_v(\mathbf{x}) = \frac{1}{n(2\pi)^{d/2}} \sum_{i=1}^{n} \frac{1}{|\mathbf{H}_i|^{1/2}} \exp\left(-\frac{1}{2} D^2(\mathbf{x}, \mathbf{x}_i, \mathbf{H}_i)\right), \quad (1)$$

where

$$D^2(\mathbf{x}, \mathbf{x}_i, \mathbf{H}_i) \equiv (\mathbf{x} - \mathbf{x}_i)^\top \mathbf{H}_i^{-1} (\mathbf{x} - \mathbf{x}_i) \quad (2)$$

is the Mahalanobis distance from $\mathbf{x}$ to $\mathbf{x}_i$. Let $\mathbf{H}_h$ be the data-weighted harmonic mean of the bandwidth matrices computed at $\mathbf{x}$

$$\mathbf{H}_h^{-1}(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) \mathbf{H}_i^{-1}, \quad (3)$$

where the weights

$$w_i(\mathbf{x}) = \frac{\frac{1}{|\mathbf{H}_i|^{1/2}} \exp\left(-\frac{1}{2} D^2(\mathbf{x}, \mathbf{x}_i, \mathbf{H}_i)\right)}{\sum_{i=1}^{n} \frac{1}{|\mathbf{H}_i|^{1/2}} \exp\left(-\frac{1}{2} D^2(\mathbf{x}, \mathbf{x}_i, \mathbf{H}_i)\right)} \quad (4)$$

1. The terms *bandwidth* and *scale* will be considered equivalent in this paper. Bandwidth will be preferred when used in conjunction with a kernel, while scale will be employed to underline the idea of size.

• *D. Comaniciu is with the Real-Time Vision and Modeling Department, Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540. E-mail: comanici@scr.siemens.com.*

Manuscript received 18 Mar. 2002; revised 19 July 2002; accepted 25 July 2002. Recommended for acceptance by S. Sclaroff.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 116109.

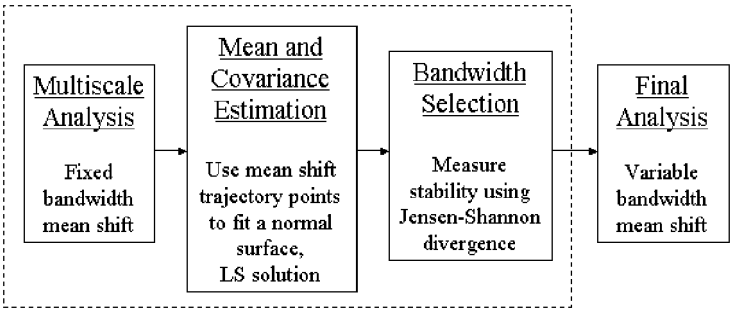

Fig. 1. Feature space analysis with variable-bandwidth. For the initial step, the fixed-bandwidth mean shift procedure [5] is applied with different analysis scales and, at each scale, each data point is classified into a local mode. The trajectory points and mean shift vectors are then used to fit a normal surface to the density surrounding each mode. For each data point, the most stable covariance matrix across scales is then selected using a specialized version of the Jensen-Shannon divergence. Finally, the covariance matrices are used in the variable-bandwidth mean shift.

satisfy $\sum_{i=1}^{n} w_i(\mathbf{x}) = 1$. An estimator of the gradient of the true density is the gradient of $\hat{f}_v$

$$\hat{\nabla} f_v(\mathbf{x}) \equiv \nabla \hat{f}_v(\mathbf{x}) = \frac{1}{n(2\pi)^{d/2}} \sum_{i=1}^{n} \frac{\mathbf{H}_i^{-1}(\mathbf{x}_i - \mathbf{x})}{\mid \mathbf{H}_i \mid^{1/2}} \exp\left(-\frac{1}{2} D^2(\mathbf{x}, \mathbf{x}_i, \mathbf{H}_i)\right). \quad (5)$$

By multiplying (5) to the left with $\mathbf{H}_h(\mathbf{x})$ and using (1), it results that

$$\mathbf{H}_h(\mathbf{x}) \hat{\nabla} f_v(\mathbf{x}) = \hat{f}_v(\mathbf{x}) \mathbf{m}_v(\mathbf{x}), \quad (6)$$

where

$$\mathbf{m}_v(\mathbf{x}) \equiv \mathbf{H}_h(\mathbf{x}) \sum_{i=1}^{n} w_i(\mathbf{x}) \mathbf{H}_i^{-1} \mathbf{x}_i - \mathbf{x} \quad (7)$$

is the variable-bandwidth mean shift vector. From (6), we also have

$$\mathbf{m}_v(\mathbf{x}) = \mathbf{H}_h(\mathbf{x}) \frac{\hat{\nabla} f_v(\mathbf{x})}{\hat{f}_v}, \quad (8)$$

which shows that the variable-bandwidth mean shift vector is an adaptive estimator of the normalized gradient of the underlying density.

If the bandwidth matrices $\mathbf{H}_i$ are all equal to a fixed matrix $\mathbf{H}$, called *analysis bandwidth*, the sample point estimator (1) reduces to the simple multivariate density estimator with normal kernel

$$\hat{f}(\mathbf{x}) = \frac{1}{n \mid 2\pi \mathbf{H} \mid^{1/2}} \sum_{i=1}^{n} \exp\left(-\frac{1}{2} D^2(\mathbf{x}, \mathbf{x}_i, \mathbf{H})\right). \quad (9)$$

Equation (8) becomes

$$\mathbf{m}(\mathbf{x}) = \mathbf{H} \frac{\hat{\nabla} f(\mathbf{x})}{\hat{f}(\mathbf{x})}, \quad (10)$$

where

$$\mathbf{m}(\mathbf{x}) \equiv \frac{\sum_{i=1}^{n} \mathbf{x}_i \exp\left(-\frac{1}{2} D^2(\mathbf{x}, \mathbf{x}_i, \mathbf{H})\right)}{\sum_{i=1}^{n} \exp\left(-\frac{1}{2} D^2(\mathbf{x}, \mathbf{x}_i, \mathbf{H})\right)} - \mathbf{x} \quad (11)$$

is the fixed-bandwidth mean shift vector.

A mode seeking algorithm can be derived by iteratively computing the fixed- or variable-bandwidth mean shift vector [4], [5]. The partition of the feature space is obtained by grouping together all the data points that converged to the same mode. Theoretically, the partition quality in the variable-bandwidth case is better, however, it depends on the selected bandwidth matrices $\mathbf{H}_i$. The next sections are devoted to the proper computation of these matrices.

3 Bandwidth Selection Theorem

This section exploits a fundamental property of the normalized gradient of normal distributions, whose estimate is proportionally downward biased [27]. The direct consequence of this property is that, within the large sample approximation, the estimation bias can be canceled, allowing the estimation of the true local covariance of the underlying distribution.

Our assumption is that, in the neighborhood of location $\mathbf{x}$, the data is a distributed multivariate normal with unknown mean $\boldsymbol{\mu}$ and covariance matrix Σ. The direct estimation of Σ is generally difficult since, to locally fit a normal one needs a priori knowledge of the neighborhood size in which the fitting parameters are to be estimated. If the estimation is performed for several neighborhood sizes, a scale invariant measure of the goodness of fit is needed. The following theorem, however, presents an elegant solution to the problem. It is valid when the number of available samples is large.

Theorem 1. *Assume that the true distribution f is $N(\boldsymbol{\mu}, \Sigma)$ and the fixed-bandwidth mean shift is computed with a normal kernel $K_{\mathbf{H}}$. The bandwidth normalized norm of the mean shift vector is maximized when the analysis bandwidth $\mathbf{H}$ is equal to Σ.*

Proof. Since the true distribution f is normal with covariance matrix Σ, it follows that the mean of $\hat{f}(\mathbf{x})$, $E\left[\hat{f}(\mathbf{x})\right] \equiv \phi(\mathbf{x}; \Sigma + \mathbf{H})$ is also a normal surface with covariance $\Sigma + \mathbf{H}$ [27]. Likewise, since the gradient is a linear operator, we have $E\left[\nabla \hat{f}(\mathbf{x})\right] = \nabla \phi(\mathbf{x}; \Sigma + \mathbf{H})$. When the large sample approximation is valid, the variances of the means are relatively small. By employing (10), this implies that

$$\begin{aligned} \text{plim } \mathbf{m}(\mathbf{x}) &= \mathbf{H} \frac{E\left[\nabla \hat{f}(\mathbf{x})\right]}{E\left[\hat{f}(\mathbf{x})\right]} = \mathbf{H} \frac{\nabla \phi(\mathbf{x}; \Sigma + \mathbf{H})}{\phi(\mathbf{x}; \Sigma + \mathbf{H})} \\ &= -\mathbf{H}(\Sigma + \mathbf{H})^{-1}(\mathbf{x} - \boldsymbol{\mu}), \end{aligned} \quad (12)$$

where plim denotes the probability limit with $\mathbf{H}$ held constant. The norm of the bandwidth normalized mean shift is given by

$$m(\mathbf{x}; \mathbf{H}) \equiv \left\| \mathbf{H}^{-1/2} \text{plim } \mathbf{m}(\mathbf{x}) \right\| = \left\| \mathbf{H}^{1/2} (\Sigma + \mathbf{H})^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\|. \quad (13)$$

It is shown in Appendix A that $m(\mathbf{x}; \mathbf{H})$ is maximized iff $\mathbf{H} = \Sigma$.□

Theorem 1 leads to an interesting scale selection criterion: The underlying distribution has the local covariance equal to the analysis bandwidth that maximizes the magnitude of the normalized mean shift vector. The main idea of this property is underlined in Fig. 2 for a unidimensional case. Given the input data drawn from $N(12, 4)$, we computed the magnitude of the normalized mean shift for different locations and using different bandwidths. Each curve in Fig. 2b represents the results for one location. Since the locations were chosen on both sides of the mean, the curves appear in pairs. The upper curves are for the points located far from the mean. Observe

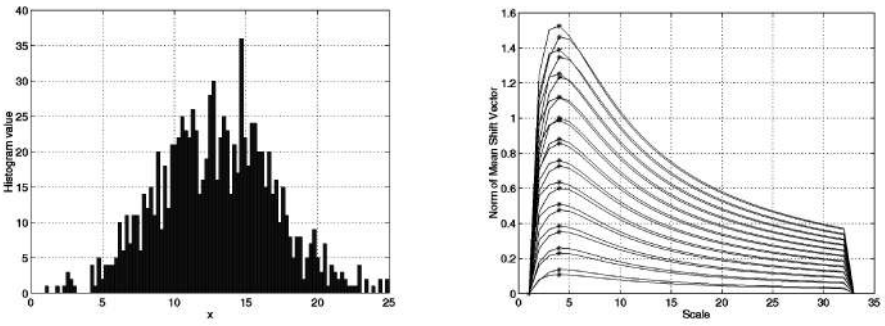

Fig. 2. Local mean shift-based scale selection. (a) Input data. $N(12,4)$, $n = 2,000$. (b) Each curve represents the magnitude of the normalized mean shift computed at one location, but with different analysis bandwidths. The maxima of the curves correctly indicate that the standard deviation of the input data is equal to 4.

that each curve is maximum when the analysis bandwith is $h_0 = 4$, indicating, according to the Theorem 1, that the standard deviation of the input data is equal to 4.

Since the theorem is valid in the neighborhood of each mode, a more global solution (least squares) can be obtained by using multiple measurements represented by the mean shift trajectories of all data points converging to the same mode. Note also that the input data might be multimodal with asymmetric structures, while neighboring structures might contaminate each other. In this case, the normality assumption of Theorem 1 is not valid and the result will depend on the analysis bandwidth $\mathbf{H}$. To solve this problem, we propose a procedure which selects the most stable bandwidth across scales. These ideas are discussed in the next section.

4 Algorithm for Bandwidth Selection

We derive in the sequel a least-squares solution for covariance matrix estimation and show how to choose the most stable result across scales. Then, the bandwidth selection algorithm is summarized and experiments are presented.

4.1 Least-Squares Solution

Let us denote by $\mathbf{x}_i$, $i = 1 \dots n_u$ all the data points associated with the $u - th$ mode and by $\mathbf{y}_i$, $i = 1 \dots t_u$ the location of all trajectory points associated with the same mode. The partition is obtained using the mean shift procedure with analysis bandwidth $\mathbf{H}$. Assume that $(\boldsymbol{\mu}, \Sigma)$ are the mean and covariance of the underlying structure.

We note that the mean and covariance of the points $\mathbf{x}_i$, $i = 1 \dots n_u$ are not reliable estimates of $(\boldsymbol{\mu}, \Sigma)$. The reason is that the data partitioning is nonparametric, based on the peaks and valleys of the density probability function of the entire data set. As a result, the set $\mathbf{x}_i$, $i = 1 \dots n_u$ is an incomplete sample from the local underlying distribution. It can be asymmetric (depending on the neighboring structures) and it might not contain the tail. Hence, the sample mean and variance differ from $(\boldsymbol{\mu}, \Sigma)$.

The solution is to fit a normal surface to the density values computed in the trajectory points associated with the mode. The fitting problem is easily solved by using the mean shift vector. For each trajectory point $\mathbf{y}_i$, we apply (12) to obtain

$$\mathbf{m}(\mathbf{y}_i) = -\mathbf{H}(\Sigma + \mathbf{H})^{-1}(\mathbf{y}_i - \boldsymbol{\mu}), \tag{14}$$

where $(\boldsymbol{\mu}, \Sigma)$ are the mean and covariance of the true distribution. By fixing the mean $\boldsymbol{\mu}$ as the the local peak in the density surface (see Fig. 3), we can derive a least-squares solution for the covariance matrix. If $\mathbf{H} = h^2\mathbf{I}$ and $\Sigma = \sigma^2\mathbf{I}$, the least-squares solution for σ^2 is

$$\sigma^2 = h^2\left[\frac{\sum_{i=1}^{t_u} \mathbf{m}_i^\top(\boldsymbol{\mu} - \mathbf{y}_i)}{\sum_{i=1}^{t_u} \|\mathbf{m}_i\|^2} - 1\right]. \tag{15}$$

If $\mathbf{H} = \mathbf{diag}\left[h_1^2 \dots h_d^2\right]$ and $\Sigma = \mathbf{diag}\left[\sigma_1^2 \dots \sigma_d^2\right]$, then

$$\sigma_v^2 = h_v^2\left[\frac{\sum_{i=1}^{t_u} m_{iv}^\top(\mu_v - \mathbf{y}_{iv})}{\sum_{i=1}^{t_u} m_{iv}^2} - 1\right], \tag{16}$$

where the subindex $v = 1 \dots d$ denotes the vth component of a vector.

Although a fully parameterized covariance matrix can be computed using (14), this is not necessarily advantageous [28, p. 107] and, for dimensions $d > 2$, the number of parameters introduced are too large to make reliable decisions. We will therefore use in the sequel only (15) and (16).

4.2 Multiscale Analysis

When the underlying data distribution is normal, the analysis bandwidth $\mathbf{H}$ does not influence the computation of $(\boldsymbol{\mu}, \Sigma)$. When the underlying structure deviates from normality, $\mathbf{H}$ affects the estimation. Therefore, in the final step of the algorithm, we test the stability of $(\boldsymbol{\mu}, \Sigma)$ against the variation of the analysis bandwidth. The simplest test is to take $\mathbf{H} = h^2\mathbf{I}$ and vary h on a logarithmic scale with constant step.

Let $\mathbf{H}_1 = h_1^2\mathbf{I}, \dots, \mathbf{H}_b = h_b^2\mathbf{I}$ be a set of analysis bandwidths generated as above. The range of these bandwidths is assumed known a priori. Denote by $(\boldsymbol{\mu}_1, \Sigma_1), \dots, (\boldsymbol{\mu}_b, \Sigma_b)$ the corresponding set of estimates and by $p_1 \dots p_b$ the associated normal distributions. The stability test for distribution p_j involves the computation of the overall dissimilarity between p_j and its neighbors across scale $p_{j-w} \dots p_{j-1}, p_{j+1} \dots p_{j+w}$. The simplest choice is $w = 1$.

The dissimilarity is measured using a specialized version of the Jensen-Shannon divergence, which is defined for the d-variate normal distributions p_j, $j = 1 \dots r$ as

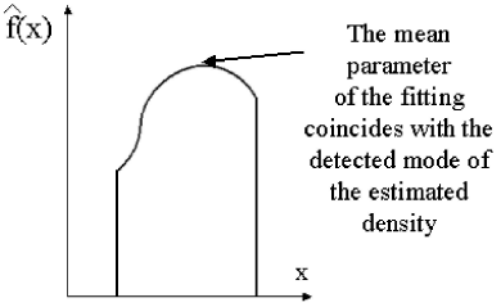

Fig. 3. Fitting a normal surface to the density values computed in the trajectory points. Observe that, even for asymmetric regions, the mean $\boldsymbol{\mu}$ of the normal surface should be taken equal to the mode of the density.

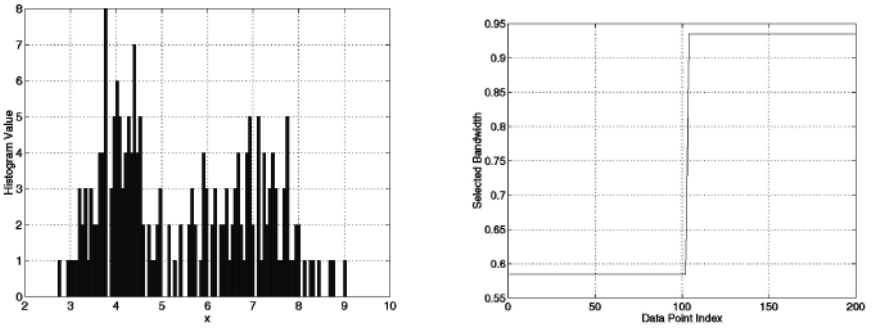

Fig. 4. Bandwidth selection example. (a) Histogram of input data drawn with equal probability from two normals $N(4, 0.5)$ and $N(7, 1)$ with total $n = 200$. (b) Bandwidth selection for each data point using the proposed algorithm. For presentation, the data point index increases with location.

$$JS(p_1 \dots p_r) = \frac{1}{2}\log\frac{\left|\frac{1}{r}\sum_{j=1}^{r}\Sigma_j\right|}{\sqrt{\prod_{j=1}^{r}|\Sigma_j|}} + \frac{1}{2}\sum_{j=1}^{r}\left(\mu_j - \mu\right)^{\top}\left(\sum_{j=1}^{r}\Sigma_j\right)^{-1}\left(\mu_j - \mu\right) \quad (17)$$

with $\mu = \frac{1}{r}\sum_{j=1}^{r}\mu_j$. This formula is derived in Appendix B. Observe that, for $r = 2$, the specialized Jensen-Shannon divergence reduces to the well-known Bhattacharyya distance [8, p. 99].

4.3 Bandwidth Selection Summary

The proposed algorithm solves the bandwidth selection problem in two stages. The first stage is defined at the partition level and determines a mean and covariance matrix for each mode detected through multiscale analysis. The second stage is defined at the data level and selects for each data point the most stable mean and covariance across the analysis scale. The algorithm is presented below.

Bandwidth Matrix Selection
Given n data points $\mathbf{x}_i$, $i = 1 \dots n$ and a set of analysis matrices $\mathbf{H}_1 = h_1^2\mathbf{I}, \dots, \mathbf{H}_b = h_b^2\mathbf{I}$ constructed on a logarithmic scale:

A. Evaluate the bandwidth at the partition level. For each $\mathbf{H}_j$, $j = 1 \dots b$

1. Partition the data using the mean shift procedure.
2. Compute (μ_{ju}, Σ_{ju}) for each mode u of the partition using the location of the mode for the mean and (15) or (16) for the covariance.
3. Associate to each data point $\mathbf{x}_i$ the mean and covariance of its mode.

B. Evaluate the bandwidth at the data level. For each data point $\mathbf{x}_i$

1. Based on the set of estimates $(\mu_1, \Sigma_1) \dots (\mu_b, \Sigma_b)$, define the normal distributions $p_1 \dots p_b$.
2. Select the most stable pair (μ, Σ) by minimizing the Jensen-Shannon divergence between neighboring distributions across scales. Σ represents the selected bandwidth for $\mathbf{x}_i$.

The complexity of the algorithm is b times larger than the complexity of data partitioning using mean shift analysis with one scale. Direct implementation of mean shift analysis with one scale has a complexity of $O(n^2)$, where n is the number of data points. However, by selecting a set of q representative data points using irregular tessellation of the space and only computing trajectories of those points, the complexity of mean shift analysis can be decreased to $O(qn)$, with $q \ll n$ [3].

4.4 Sample Size

While the large sample approximation is not critical for (12), the sparse data needs attention. The local sample size should be sufficiently large for inference. The approach we take is based on the *Effective Sample Size* [10] which computes the kernel weighted count of the number of points in each window

$$ESS(\mathbf{x}; \mathbf{H}) = \frac{\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)}{K_{\mathbf{H}}(\mathbf{0} - \mathbf{0})} = \frac{\sum_{i=1}^{n} \exp\left(-\frac{1}{2}D^2(\mathbf{x}, \mathbf{x}_i, \mathbf{H})\right)}{\exp\left(-\frac{1}{2}D^2(\mathbf{0}, \mathbf{0}, \mathbf{H})\right)}. \quad (18)$$

Using the binomial rule of thumb, we cancel the inference when $ESS(\mathbf{x}; \mathbf{H}) < 5$.

4.5 Bandwidth Selection Examples

A first example for a bimodal data set generated with equal probability from $N(4, 0.5)$ and $N(7, 1)$ is presented in Fig. 4. The standard deviation for each distribution (measured before amalgamating the data) is 0.53 and 0.92. Our algorithm resulted in 0.58 and 0.93, respectively. We used eight analysis bandwidths in the range of 0.3-1.42 with a ratio of 1.25 between two consecutive bandwidths. For all the experiments presented henceforth, we will use the same ratio of 1.25 between two consecutive bandwidths. The specialized Jensen-Shannon divergence was computed with $r = 3$ (three consecutive bandwidths). No other additional information was used.

For the next example, the data is drawn with equal probability from $N(8, 2)$, $N(25, 4)$, $N(50, 8)$, and $N(100, 16)$. The data histogram is shown in Fig. 5a, while our bandwidth selection is shown in Fig. 5b. We used 12 analysis bandwidths in the range of 1.5-17.46.

Another example is shown in Fig. 6 for bivariate data. We run the algorithm with six analysis bandwidths in the range 0.5-1.5. The algorithm detected three classes of bandwidths: 0.96, 1.04, and 1.08. In Fig. 6b, the bandwidth associated with each data point is indicated by the bullet (smallest bullets for 0.96, largest bullets for 1.08). The allocated bandwidths are very close to the true data scale, which is equal to 1.

5 Feature Space Partitioning

This section presents results for feature space partitioning using the variable-bandwidth mean shift with bandwidth selection. Only the range of analysis scales is provided for each experiment.

5.1 Nonlinear Structures with Multiple Scales

For the data shown in Fig. 7a, the algorithm was run with six analysis bandwidths in the range of 0.1-0.3. This time, we used expression (16) to estimate a diagonal form for the covariance matrix associated with each data point. The results are presented in Fig. 7c for the scales associated with the coordinate x and Fig. 7d for the scales associated with the coordinate y of each data point.

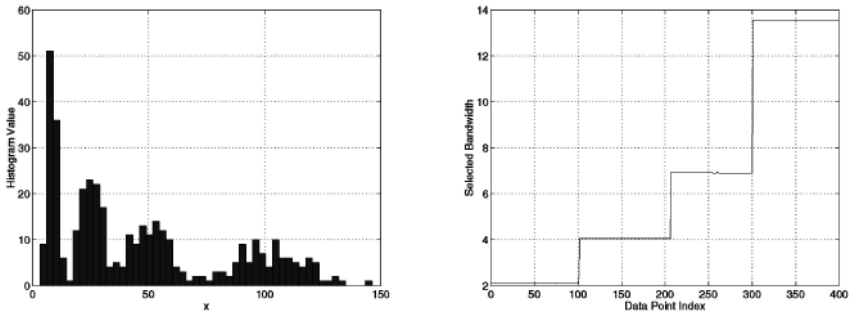

Fig. 5. Bandwidth selection example. (a) Histogram of input data drawn with equal probability from four normals $N(8,2)$, $N(25,4)$, $N(50,8)$, and $N(100,16)$ with total $n = 400$. (b) Bandwidth selection for each data point using the proposed algorithm. For presentation, the data point index increases with location.

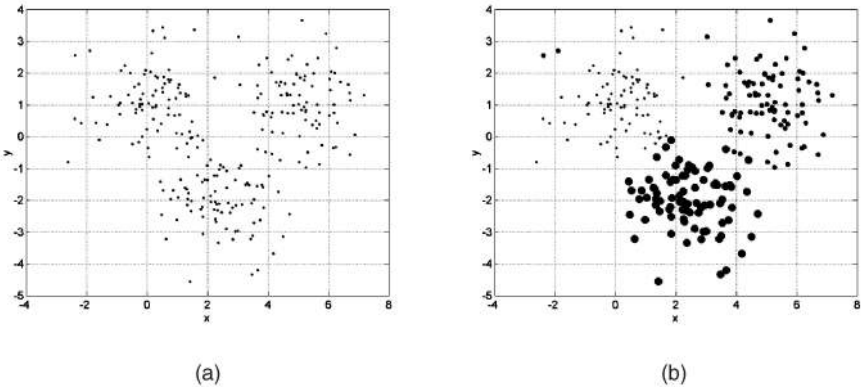

Fig. 6. Bandwidth selection example. (a) Bivariate data drawn with equal probability from $N([0,1],\mathbf{I})$, $N([2.5,-2],\mathbf{I})$, $N([5,1],\mathbf{I})$ with total $n = 250$. (b) Bandwidth selection for each data point using the proposed algorithm. Three classes of bandwidth were detected. See text for details.

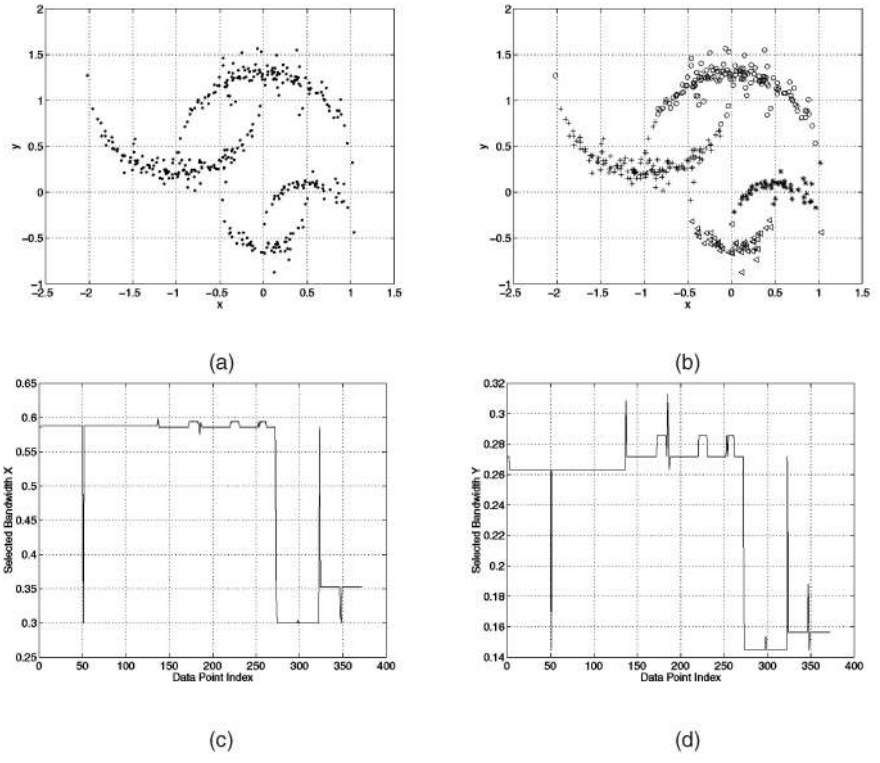

Fig. 7. Nonlinear data analysis. (a) Input containing structures at different scales ($n = 400$). (b) Final decomposition obtained through variable-bandwidth mean shift. Each structure is marked differently. (c) Scale selection for the x coordinate of each data point. (d) Scale selection for the y coordinate of each data point.

Observe that the elongated structure of the data is reflected in a larger bandwidth for the coordinate x. Also, each graph contains two distinct groups of scale values corresponding to the two scales in the data. The spurious peaks represent points located on the

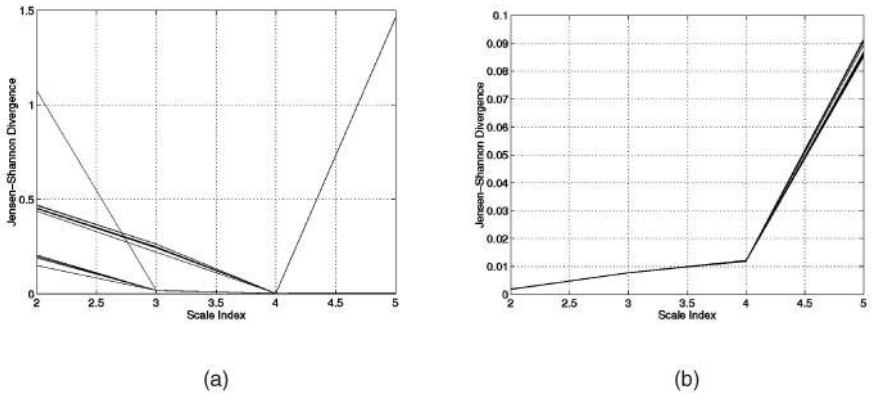

Fig. 8. Jensen-Shannon divergence for data from Fig. 7. (a) Points from large structures. (b) Points from small structures.

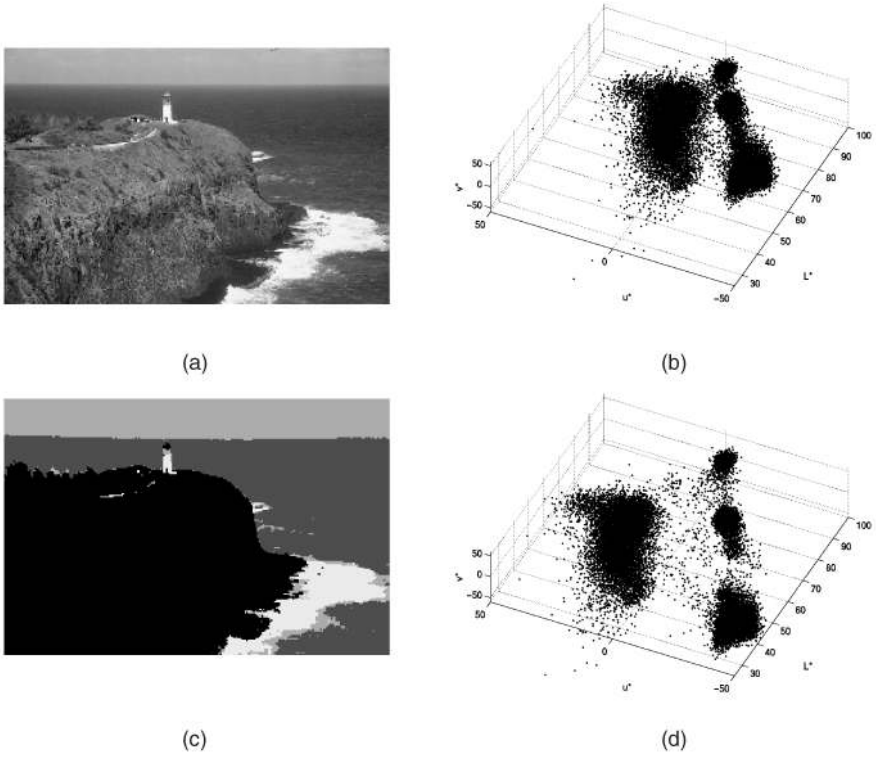

Fig. 9. Color clustering experiment 1. (a) Original image, 500×333 pixels. (b) $L^*u^*v^*$ color space containing 166,500 points. (c) Segmented image in pseudogray levels. (d) Obtained clusters. The position of each cluster is shifted to show the delineation.

border between two structures. Note also that, for both coordinates, the smaller scale is approximately half of the larger scale, similar to the data characteristics.

Fig. 8 shows of the specialized Jensen-Shannon divergence for points from the large structures (Fig. 8a) and small structures (Fig. 8b). As one can observe, in the case of large structures, the estimation is most stable (small divergence) for the analysis scales from the middle. On the contrary, in the case of small structures, the estimation is the most stable for the smallest analysis scale.

The last step involves the application of the variable-bandwidth mean shift with the bandwidths shown in Fig. 7c and Fig. 7d. The algorithm detected four modes and the resulting partitioning is shown in Fig. 7b. Note that most of the algorithms using one analysis bandwidth are prone to fail for this type of data. If the bandwidth is large, the two small structures will be joined together. If the bandwidth is small, each of the two large structures will be divided.

5.2 Color Clustering

We tested the new algorithm for the task of color clustering in the three-dimensional $L^*u^*v^*$ space. The selected examples contain large and elongated clusters in the vicinity of small clusters, a difficult scenario for fixed-bandwidth analysis. A first test image is shown in Fig. 9a. The sky, ocean, and the waves generate compact and small clusters, while the texture from the land generates a large cluster (Fig. 9b). Using six analysis bandwidths in the range of 3-9, our algorithm correctly obtained the four clusters shown in Fig. 9d which corresponds to the segmentation from Fig. 9c.

The same analysis bandwidths have been employed for processing the color data coming from the test image shown in Fig. 10a. Observe again, the presence of a large cluster in the vicinity of small clusters (Fig. 10b). The algorithm identified three clusters (Fig. 10d) which are associated to the main structures in the image, as can be seen in the corresponding segmented image (Fig. 10c).

6 Discussion

It is useful to contrast the proposed algorithm against some classical alternatives. The EM algorithm [23] also assumes a mixture of normal structures and finds iteratively the maximum-likelihood estimates of the a priori probabilities, means, and

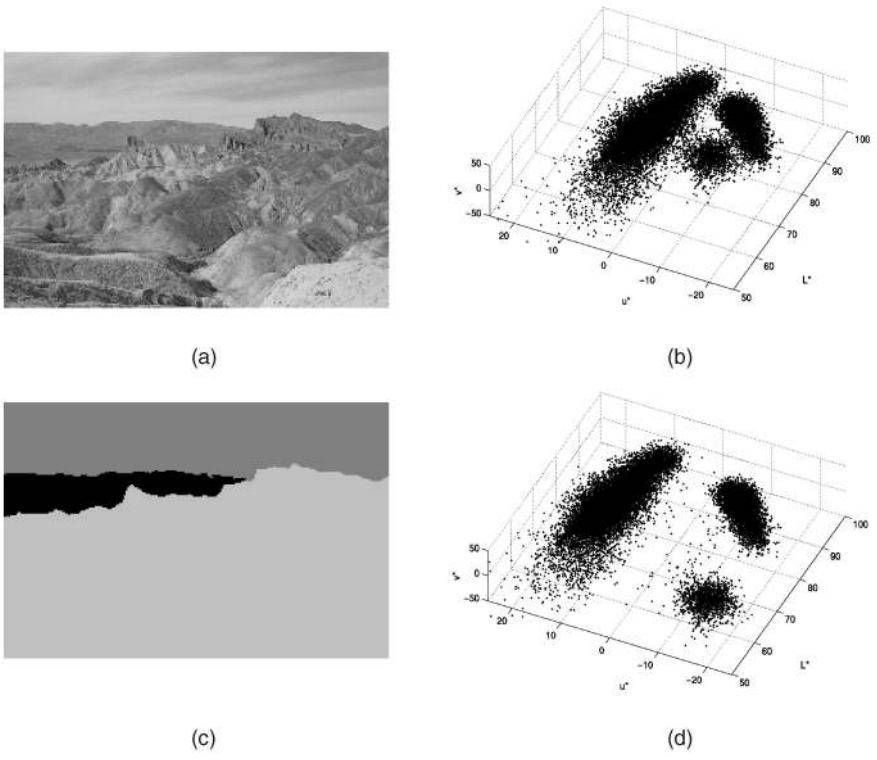

Fig. 10. Color clustering experiment 2. (a) Original image, 500×333 pixels. (b) $\mathrm{L^*u^*v^*}$ color space containing 166,500 points. (c) Segmented image in pseudogray levels. (d) Obtained clusters. The position of each cluster is shifted to show the delineation.

covariances. However, the EM needs the specification of the number of clusters, needs a good initialization, and does not deal with non-normal structures. In addition, its convergence is difficult when the number of clusters is large, determining the increase of the number of parameters. See [7] for a discussion and changes of the EM to overcome some of these limitations.

Our algorithm is not affected by the number of clusters since it does not employ a global criterion that should be optimized. We only require the a priori knowledge of a range of viable scales, which is a very practical criterion. In almost all situations, the user has this knowledge. In addition, the normality assumption is only for bandwidth selection, while the overall algorithm maintains the ability of analyzing complex, non-normal structures. The only limitation of our method comes with the dimensionality of the data. It is known that nonparameteric techniques are not reliable in high-dimensional spaces.

Let us also contrast the proposed algorithm with methods based on multiscale analysis. From this point of view and according to our knowledge, this is the first method which tests the stability of the second order statistics derived from the data. Up to now, the stability testing was limited to the first order statistics such as the mean, the mode, or direction vectors (see, for example, [2]). By checking the stability of the covariance matrix through the specialized Jensen-Shannon divergence, we increase the amount of information involved in the test. Finally, the method can be improved by replacing the least-square estimation with a robust method. This work mostly presented the theory related to the new algorithm. The algorithm is useful for scenarios involving multiscale patterns, such as feature space partitioning in tracking, background modeling, and segmentation. An interesting subject of future research is to analyze the relation between the proposed method and scale selection techniques for image features [19].

APPENDIX A

THE MAGNITUDE OF THE BANDWIDTH NORMALIZED MEAN SHIFT VECTOR $m(\mathbf{x};\mathbf{H})$ IS MAXIMIZED WHEN $\mathbf{H} = \Sigma$

Recall that the magnitude of the bandwidth normalized mean shift vector is given by

$$m(\mathbf{x};\mathbf{H}) = \left\|\mathbf{H}^{1/2}(\Sigma+\mathbf{H})^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\|. \tag{A.19}$$

We assume that $\mathbf{H}$ and Σ are symmetric, positive definite matrices, and the magnitude of $\mathbf{x}-\boldsymbol{\mu}$ is strictly positive. We will show that

$$m(\mathbf{x};\Sigma)^2 - m(\mathbf{x};\mathbf{H})^2 \geq 0 \tag{A.20}$$

with equality iff $\mathbf{H} = \Sigma$.

The left side of (A.20) becomes

$$\begin{aligned} & m(\mathbf{x};\Sigma)^2 - m(\mathbf{x};\mathbf{H})^2 \\ &= \frac{1}{4}\left[\left\|\Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})\right\|^2 - 4\left\|\mathbf{H}^{1/2}(\Sigma+\mathbf{H})^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\|^2\right] \\ &= \frac{1}{4}(\mathbf{x}-\boldsymbol{\mu})^\top\left[\Sigma^{-1} - 4(\Sigma+\mathbf{H})^{-1}\mathbf{H}(\Sigma+\mathbf{H})^{-1}\right](\mathbf{x}-\boldsymbol{\mu}) \\ &= \frac{1}{4}(\mathbf{x}-\boldsymbol{\mu})^\top(\Sigma+\mathbf{H})^{-1}(\mathbf{H}\Sigma^{-1}-\mathbf{I})^2\Sigma(\Sigma+\mathbf{H})^{-1}(\mathbf{x}-\boldsymbol{\mu}), \end{aligned} \tag{A.21}$$

where $\mathbf{I}$ is the $d \times d$ identity matrix. Within the conditions stated, all the matrices in the last term of (A.21) are positive definite, excepting $(\mathbf{H}\Sigma^{-1}-\mathbf{I})^2$ which is equal to $\mathbf{0}$ iff $\mathbf{H} = \Sigma$.

APPENDIX B

OVERALL DISSIMILARITY OF A SET OF MULTIVARIATE NORMAL DISTRIBUTIONS

One of the few measures of the overall difference of more than two distributions is the generalized Jensen-Shannon divergence [18].

Given r probability distributions p_j, $j = 1 \dots r$, their Jensen-Shannon divergence is defined as

$$JS(p_1 \dots p_r) = H\left(\frac{1}{r}\sum_{j=1}^{r} p_j\right) - \frac{1}{r}\sum_{j=1}^{r} H(p_j), \tag{B.22}$$

where

$$H(p(\mathbf{x})) = -\int p(\mathbf{x}) \log p(\mathbf{x}) \mathbf{dx} \tag{B.23}$$

is the entropy of $p(\mathbf{x})$. This divergence is positive and equal to zero iff all p_j are equal. Using (B.23) in (B.22), we obtain

$$JS(p_1 \dots p_r) = \frac{1}{r}\sum_{j=1}^{r} \int p_j(\mathbf{x}) \log \frac{p_j(\mathbf{x})}{q(\mathbf{x})} \mathbf{dx} \;\; \text{width} \;\; q(\mathbf{x}) = \frac{1}{r}\sum_{j=1}^{r} p_j. \tag{B.24}$$

For the d-variate normal case, the distributions p_j are defined by

$$p_j(\mathbf{x}) = \frac{1}{|2\pi\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right). \tag{B.25}$$

A specialized version of the Jensen-Shannon divergence can be obtained by taking $q(\mathbf{x})$ as the most likely normal source for the homogeneous model $\frac{1}{r}\sum_{j=1}^{r} p_j$, having the mean $\boldsymbol{\mu} = \frac{1}{r}\sum_{j=1}^{r} \boldsymbol{\mu}_j$ and covariance $\Sigma = \frac{1}{r}\sum_{j=1}^{r} \Sigma_j$ [6]. The new measure is equivalent to a goodness-of-fit test between the empirical distributions p_j, $j = 1 \dots r$ and the homogeneous model $\frac{1}{r}\sum_{j=1}^{r} p_j$.

To derive a closed form expression, we use (B.25) and the identity $\mathbf{x}^\top \mathbf{\Sigma}^{-1}\mathbf{x} = \operatorname{tr} \mathbf{\Sigma}^{-1}\mathbf{x}\mathbf{x}^\top$ to obtain [16, p.189]

$$\begin{aligned}\log \frac{p_i(\mathbf{x})}{q(\mathbf{x})} =& \frac{1}{2}\log\frac{|\Sigma|}{|\Sigma_i|} \\ &- \frac{1}{2}\operatorname{tr} \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^\top + \frac{1}{2}\operatorname{tr} \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\end{aligned} \tag{B.26}$$

for $i = 1 \dots r$, where tr denotes the trace of a matrix. Performing the integration yields

$$\begin{aligned}&\int p_i(\mathbf{x}) \log \frac{p_i(\mathbf{x})}{q(\mathbf{x})} \mathbf{dx} \\ &= \frac{1}{2}\log\frac{|\Sigma|}{|\Sigma_i|} + \frac{1}{2}\operatorname{tr} \Sigma_i\Sigma^{-1} - \frac{d}{2} + \frac{1}{2}\operatorname{tr} \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top.\end{aligned} \tag{B.27}$$

Summing (B.27), for $i = 1 \dots r$ and substituting $\Sigma = \frac{1}{r}\sum_{j=1}^{r} \Sigma_j$, we have

$$\begin{aligned}&JS(p_1 \dots p_r) \\ &= \frac{1}{2}\log\frac{\left|\frac{1}{r}\sum_{j=1}^{r} \Sigma_j\right|}{\sqrt[r]{\prod_{j=1}^{r}|\Sigma_j|}} + \frac{1}{2r}\operatorname{tr}\left(\sum_{j=1}^{r}\Sigma_j\right)\left(\frac{1}{r}\sum_{j=1}^{r}\Sigma_j\right)^{-1} - \frac{r}{2} \\ &\quad + \frac{1}{2r}\operatorname{tr}\left(\frac{1}{r}\sum_{j=1}^{r}\Sigma_j\right)^{-1}\sum_{j=1}^{r}\Big(\boldsymbol{\mu}_j - \boldsymbol{\mu}\Big)\Big(\boldsymbol{\mu}_j - \boldsymbol{\mu}\Big)^\top \\ &= \frac{1}{2}\log\frac{\left|\frac{1}{r}\sum_{j=1}^{r} \Sigma_j\right|}{\sqrt[r]{\prod_{j=1}^{r}|\Sigma_j|}} + \frac{1}{2}\sum_{j=1}^{r}\Big(\boldsymbol{\mu}_j - \boldsymbol{\mu}\Big)^\top\left(\sum_{j=1}^{r}\Sigma_j\right)^{-1}\Big(\boldsymbol{\mu}_j - \boldsymbol{\mu}\Big),\end{aligned} \tag{B.28}$$

where $\boldsymbol{\mu} = \frac{1}{r}\sum_{j=1}^{r} \boldsymbol{\mu}_j$.

ACKNOWLEDGMENTS

The author would like to thank Yakup Genc from Siemens Corporate Research for valuable discussions on this work.

REFERENCES

[1] I. Abramson, "On Bandwidth Variation in Kernel Estimates—A Square Root Law," *The Ann. Statistics*, vol. 10, no. 4, pp. 1217-1223, 1982.
[2] N. Ahuja, "A Transform for Multiscale Image Segmentation by Integrated Edge and Region Detection," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 18, pp. 1211-1235, 1996.
[3] D. Comaniciu and P. Meer, "Distribution Free Decomposition of Multivariate Data," *Pattern Analysis and Applications*, vol. 2, pp. 22-30, 1999.
[4] D. Comaniciu and P. Meer, "Mean shift: A Robust Approach toward Feature Space Analysis," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 24, no. 5, pp. 603-619, May 2002.
[5] D. Comaniciu, V. Ramesh, and P. Meer, "The Variable Bandwidth Mean Shift and Data-Driven Scale Selection," *Proc. Eighth Int'l Conf. Computer Vision*, vol. I, pp. 438-445, July 2001.
[6] R. El-Yaniv, S. Fine, and N. Tishby, "Agnostic Classification of Markovian Sequences," *Proc. Advances in Neural Information Processing Systems*, vol. 10, pp. 465-471, 1997.
[7] M. Figueiredo and A. Jain, "Unsupervised Learning of Finite Mixture Models," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 24, no. 3, pp. 381-396, Mar. 2000.
[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second ed. Academic Press, 1990.
[9] T. Gevers, "Robust Histogram Construction from Color Invariants," *Proc. Int'l Conf. Computer Vision*, vol. I, pp. 615-620, July 2001.
[10] F. Godtliebsen, J. Marron, and P. Chaudhuri, "Significance in Scale Space for Density Estimation," Unpublished manuscript, Available at www.stat.unc.edu/faculty/marron/marron_papers.html 1999.
[11] P. Hall, T. Hui, and J. Marron, "Improved Variable Window Kernel Estimates of Probability Densities," *The Ann. Statistics*, vol. 23, no. 1, pp. 1-10, 1995.
[12] M. Irani and P. Anandan, "Factorization with Uncertainty," *Proc. Sixth European Conf. Computer Vision*, pp. 539-553, 2000.
[13] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*. Prentice Hall, 1988.
[14] M. Jones, J. Marron, and S. Sheather, "A Brief Survey of Bandwidth Selection for Density Estimation," *J. Am. Statistical Assoc.*, vol. 91, pp. 401-407, 1996.
[15] L. Kauffman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. J. Wiley & Sons, 1990.
[16] S. Kullback, *Information Theory and Statistics*. Dover, 1997.
[17] Y. Leung, J. Zhang, and Z. Xu, "Clustering by Scale-Space Filtering," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 22, no. 12, pp. 1396-1410, Dec. 2000.
[18] J. Lin, "Divergence Measures Based on the Shannon Entropy," *IEEE Trans. Information Theory*, vol. 37, pp. 145-151, 1991.
[19] T. Lindeberg, "Feature Detection with Automatic Scale Selection," *Int'l J. Computer Vision*, vol. 30, no. 2, pp. 79-116, 1998.
[20] G. Milligan and M. Cooper, "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, vol. 50, pp. 159-179, 1985.
[21] B. Park and J. Marron, "Comparison of Data-Driven Bandwidth Selectors," *J. Am. Statistical Assoc.*, vol. 85, pp. 66-72, 1990.
[22] E.J. Pauwels and G. Frederix, "Finding Salient Regions in Images," *Computer Vision and Image Understanding*, vol. 75, pp. 73-85, 1999.
[23] R. Redner and H. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Rev.*, vol. 26, pp. 195-239, 1984.
[24] S. Sheather and M. Jones, "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *J. Royal Statistical Soc. B*, vol. 53, pp. 683-690 1991.
[25] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.
[26] J. Simonoff, *Smoothing Methods in Statistics*. Springer-Verlag, 1996.
[27] T. Stoker, "Smoothing Bias in Density Derivative Estimation," *J. Am. Statistical Assoc.*, vol. 88, no. 423, pp. 855-863, 1993.
[28] M.P. Wand and M. Jones, *Kernel Smoothing*. Chapman & Hall, 1995.