

Kernel Bandwidth Estimation for Nonparametric Modeling

Adrian G. Bors, *Senior Member, IEEE*, and Nikolaos Nasios

Abstract—Kernel density estimation is a nonparametric procedure for probability density modeling, which has found several applications in various fields. The smoothness and modeling ability of the functional approximation are controlled by the kernel bandwidth. In this paper, we describe a Bayesian estimation method for finding the bandwidth from a given data set. The proposed bandwidth estimation method is applied in three different computational-intelligence methods that rely on kernel density estimation: 1) scale space; 2) mean shift; and 3) quantum clustering. The third method is a novel approach that relies on the principles of quantum mechanics. This method is based on the analogy between data samples and quantum particles and uses the Schrödinger potential as a cost function. The proposed methodology is used for blind-source separation of modulated signals and for terrain segmentation based on topography information.

Index Terms—Image segmentation, kernel bandwidth estimation, kernel density estimation (KDE).

I. INTRODUCTION

STATISTICAL modeling and probability density function (pdf) estimation is required in various applications [1], [2]. In data analysis, in general, we can identify two main methodologies: 1) parametric and 2) nonparametric. The first methodology assumes the knowledge of a parametric model for data distribution. Parametric models have extensively been studied [2]–[6]. Such models rely on certain a priori assumptions about the statistical model. On the other hand, the nonparametric methodology can be used to model any pdf up to a certain approximation error and requires only minimal and very general assumptions. The nonparametric methods can be classified [7] into two approaches: 1) histogram based and 2) kernel based [8]–[10]. Nonparametric algorithms that use histograms require large data sets to guarantee convergence, and their data representation ability is affected by outliers. Kernel density estimation (KDE) produces smooth continuous differentiable functions while ensuring a good pdf approximation for a given data [6], [11]–[15]. KDE yields smaller bias than histograms when approximating the underlying pdf [16]. Kernel-based approaches are used in probabilistic neural networks [17], self-organizing maps, radial-basis functions [5], and other computational-intelligence methods. KDE has been employed in various applications, including image segmenta-

tion [13], [18], depth map segmentation [11], tracking in image sequences [12], blind-source separation [19], edge enhancement in images, and filtering [20].

In KDE, the kernel function is centered at each data sample location [9]. An influence region is defined with the maximum at the data sample location while decreasing in intensity with the distance from that location. A scale parameter, which is also called bandwidth or window width, controls the kernel function that smooths over the surrounding space. Comparisons among several different kernel functions have shown only small differences with respect to their efficiency in approximating the underlying data pdf [7], [21], [22]. Most studies choose the Gaussian as the kernel function due to its properties of approximation and for having the derivatives of all orders defined over the entire space [7], [16], [23], [24]. Mixtures of Gaussians have extensively been used in machine learning [3] and computer vision [5]. The mean shift is an updating algorithm that employs the local gradient to find the local pdf maxima [11], [13], [25], [26]. Other clustering algorithms are derived from the physics theory by associating ferromagnetic properties to data [27]. One algorithm that relies on the analogy between the pdf representation and the quantum potential of physical particles was proposed in [28] and was used in [29] for data segmentation.

According to several studies, the performance of KDE crucially depends on the value of the kernel bandwidth [7], [11], [16], [21], [23], [32]. The bandwidth influences the degree of smoothing for the resulting pdf function approximation and the location of its modes. Most KDE studies assume an identical bandwidth for all kernels and are applied to univariate data [7], [23]. Using a randomly generated bandwidth, as in [28], does not clearly provide an appropriate solution in most applications. The value of the bandwidth, which is taken to be equal to the Euclidean distance to the k th farthest data sample from the kernel center [10], is biased by outliers. The algorithms for finding the bandwidth in statistics can be classified into two categories: 1) quality-of-fit methods and 2) plug-in methods. The first category uses cross validation by leaving certain data samples out while approximating the pdf with the sum of kernels located at the remaining data [14], [20]. Usually, a least squares criterion is used in these methods [7]. Oftentimes, these methods produce a large variability in the estimated bandwidth, depending on the selection of specific data samples [23]. The plug-in methods calculate the bias in the pdf approximation such that it minimizes the mean integrated square error (MISE) between the real density and its kernel-based approximation [11], [22], [30]–[32]. However, plug-in algorithms require an initial pilot estimate of the bandwidth for an iterative estimation process [24]. Various comparative studies have been performed among bandwidth estimation algorithms [16], [23]. These

Manuscript received March 19, 2008; revised November 24, 2008 and February 7, 2009. First published June 19, 2009; current version published November 18, 2009. This paper was recommended by Associate Editor X. Jiang.

The authors are with the Department of Computer Science, University of York, York YO10 5DD, U.K. (e-mail: adrian.bors@cs.york.ac.uk).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TSMCB.2009.2020688

studies show that existing kernel bandwidth estimation methods lead to either spurious bumpiness or excessive smoothing in the resulting underlying density function. In many cases, existing bandwidth estimation methods fail to identify the modes of the underlying pdf [23], [24].

This paper proposes a new bandwidth estimation approach for KDE methods, which can be used in various computational-intelligence methods. Three different KDE approaches are studied in this paper. The first method corresponds to the classical KDE that corresponds to the scale space [6], [7], [12]. The second approach is the mean shift, which is an adaptive KDE procedure [13], [33]. The third approach is a new computational-intelligence method named the quantum clustering, which is adapted for use in data representation [28], [29]. A Bayesian approach is proposed to find the kernel bandwidth in the KDE methodology. Distributions of local variances are evaluated from the randomly sampled K -nearest neighbor (KNN) data sets. In this paper, we employ the Gamma distribution to model the distribution of local variances. Gamma distributions have successfully been used as priors for blind-source separation [39]. In this paper, a uniform prior is assumed for K , i.e., the number of data samples from a neighborhood. The proposed bandwidth estimation method is employed by all three KDE methods and applied to the blind detection of modulated signals and to segment vector fields. The vector fields represent surface normals that were estimated from synthetic-aperture radar (SAR) images of terrain by adapting shape from a shading methodology [34]. The KDE methodology is described in Section II. The estimation of the bandwidth is outlined in Section III. The experimental results are provided in Section IV, whereas the conclusion of this paper is drawn in Section V.

II. KERNEL DENSITY ESTIMATION BACKGROUND

KDE provides a smooth reliable estimation that asymptotically represents the true pdf for a given data set. Recently, KDE has been used in various pattern recognition and signal processing applications [6], [11]–[16], [23], [26]. In KDE, a kernel function is assigned to each data sample $\{\mathbf{X}_i, i = 1, \dots, N\}$, and the pdf that represents these data is approximated by

$$\psi(\mathbf{X}) = \frac{1}{N\sigma^d} \sum_{i=1}^N \mathcal{K}\left(\frac{\mathbf{X} - \mathbf{X}_i}{\sigma}\right) \quad (1)$$

where the kernel function $\mathcal{K}(\cdot)$ is defined in a d -dimensional space, and σ is the kernel bandwidth. A kernel function must satisfy the following conditions:

$$\int_{\mathbb{R}^d} \mathcal{K}(\mathbf{X}) d\mathbf{X} = 1 \quad (2)$$

$$\int_{\mathbb{R}^d} \mathbf{X} \mathcal{K}(\mathbf{X}) d\mathbf{X} = 0 \quad (3)$$

$$0 < \mu(\mathcal{K}) = \int_{\mathbb{R}^d} \mathbf{X}^2 \mathcal{K}(\mathbf{X}) d\mathbf{X} < \infty \quad (4)$$

where $\mu(\mathcal{K})$ is a constant that depends on the kernel function.

Various studies have been performed to choose the most appropriate kernel function for KDE [7]. The optimal efficiency

is provided by the Epanechnikov kernel [7], [22], [33], i.e.,

$$\mathcal{K}_E(\mathbf{X}_i) = \begin{cases} \frac{d+2}{2C_d} \left(1 - \frac{\|\mathbf{X} - \mathbf{X}_i\|^2}{\sigma^2}\right), & \text{if } \frac{\|\mathbf{X} - \mathbf{X}_i\|^2}{\sigma^2} < 1 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where \mathbf{X}_i are d -dimensional data samples, with $i = 1, \dots, N$, σ is the kernel bandwidth, C_d is the volume of the d -dimensional sphere with a radius of one, and $\|\cdot\|$ is the Euclidean distance. However, the Epanechnikov kernel lacks differentiability at $(\|\mathbf{X} - \mathbf{X}_i\|^2/\sigma^2) = 1$.

Most studies prefer to use the Gaussian kernel due to its well-known properties [6], [9], [15], [16], i.e.,

$$\mathcal{K}_G(\mathbf{X}_i) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{\|\mathbf{X} - \mathbf{X}_i\|^2}{2\sigma^2}\right) \quad (6)$$

where σ is the bandwidth parameter. The efficiency of the Gaussian kernel is 0.95, which is very close to that of the Epanechnikov kernel, which is 1, [7]. In the following discussion, we outline three KDE algorithms that are used in pattern recognition and computer vision.

A. Scale-Space Algorithm

The scale-space algorithm corresponds to representing the data as in (1) while aiming at identifying the modes of the resulting KDE function. The local modes can be used to segment and interpret the given data set. The pdf representation, which results after using either the Epanechnikov kernel in (5) or the Gaussian kernel in (6), depends on the bandwidth σ . The nonparametric segmentation procedure is described in detail in Section II-D.

B. Mean-Shift Algorithm

The mean-shift algorithm, which is used in several computer vision and image-processing applications, adapts the cluster centers by using the gradient of the density function $\psi(\mathbf{X})$ in (1) [11], [13], [25], [26]. The gradient of the density function $\psi_G(\mathbf{X}_i)$ is obtained by replacing $K_G(\mathbf{X})$ from (6) into (1). After differentiating the resulting function with respect to \mathbf{X} , we obtain

$$\nabla\psi_G(\mathbf{X}) = \frac{1}{N(2\pi)^{d/2}\sigma^{d+2}} \sum_{i=1}^N (\mathbf{X}_i - \mathbf{X}) \exp\left(-\frac{\|\mathbf{X} - \mathbf{X}_i\|^2}{2\sigma^2}\right). \quad (7)$$

The sample mean shift $M_\sigma(\mathbf{X})$ is defined as [13], [25]

$$M_\sigma(\mathbf{X}) = \frac{\sum_{i=1}^N \mathbf{X}_i \exp\left(-\frac{\|\mathbf{X} - \mathbf{X}_i\|^2}{2\sigma^2}\right)}{\sum_{i=1}^N \exp\left(-\frac{\|\mathbf{X} - \mathbf{X}_i\|^2}{2\sigma^2}\right)} - \mathbf{X}. \quad (8)$$

In the case of the Epanechnikov kernel in (5), the mean shift has a simpler expression and represents the difference between the average of the data samples \mathbf{X}_i , from inside a sphere that is

centered at \mathbf{X} of radius σ , and the location \mathbf{X} [11], [26], [33]. In general, we can express the mean shift as

$$M_\sigma(\mathbf{X}) = \frac{c\sigma^2 \nabla\psi(\mathbf{X})}{2 \psi(\mathbf{X})} \quad (9)$$

where c is a constant ($c = 1/((2\pi)^{d/2}\sigma^d)$ in the case of the Gaussian kernel). The modes are located among the zeros of the gradient $\nabla\psi(\mathbf{X})$. We can observe, based on (8), that the mean-shift vector $M_\sigma(\mathbf{X})$ points toward the direction of the steepest slope of the kernel density representation $\psi_G(\mathbf{X})$, leading to the modes of the underlying density without estimating the density [13]. The mean shift $M_\sigma(\mathbf{X})$ relies on the knowledge of the bandwidth σ . After choosing an appropriate bandwidth σ , an estimate of the local normalized gradient is calculated using the sample mean shift (9) by initially centering the kernel at each data sample. The mean shift is used to iteratively calculate the cluster center candidates, thus defining a path that leads to the local maximum of the density function. In the final step, two cluster candidates are considered distinct if there is a local minima on the line that joins them. Otherwise, the two clusters are merged. The local minima and maxima are defined according to the difference in the local gradient compared to a threshold [13].

C. Quantum Clustering

The kernel function can be assimilated with an energy field that manifests around particles [9]. By extending this analogy, the KDE that corresponds to a data set can be associated with a potential energy function. One new nonparametric algorithm called quantum clustering was proposed in [28]. This algorithm was derived from the analogy between the quantum potential as defined in quantum mechanics and data sample statistics. The state of a quantum mechanical system is completely specified at a location \mathbf{X} using $\psi(\mathbf{X})$, similar to the kernel function based on (1).

According to the fifth postulate of quantum mechanics, a quantum system evolves according to the Schrödinger differential equation. The time-independent Schrödinger equation is given by

$$\mathcal{H}\psi(\mathbf{X}) \equiv \left(-\frac{\sigma^2}{2} \nabla^2 + V(\mathbf{X}) \right) \psi(\mathbf{X}) = E \cdot \psi(\mathbf{X}) \quad (10)$$

where \mathcal{H} is the Hamiltonian operator, E is the eigenvalue energy level associated with a specific particle orbit, $\psi(\mathbf{X})$ corresponds to the state of the given quantum system, $V(\mathbf{X})$ is the Schrödinger potential, and ∇^2 is the Laplacian. In quantum mechanics, the potential $V(\mathbf{X})$ is given, and (10) is solved to find solutions $\psi(\mathbf{X})$. In this case, $\psi(\mathbf{X})$ describes the probability of locating a particle on a specific orbit. Based on the machine-learning perspective, we consider the inverse problem by assuming the location of data samples and their state, as given by (6), to be known. This location is considered a solution for (10), which is subject to the calculation of a set of constants. We want to evaluate the potential $V(\mathbf{X})$ that was produced by the quantum-mechanics system, which is assimilated with the given data set $\{\mathbf{X}_i, i = 1, \dots, N\}$.

In (10), the potential is always positive, $V(\mathbf{X}) > 0$. After replacing $\psi(\mathbf{X})$ from (6) into (10), we calculate the Schrödinger

potential for the given data set [28], [29] as

$$V(\mathbf{X}) = E - \frac{d}{2} + \frac{1}{2\sigma^2\psi(\mathbf{X})} \sum_{i=1}^N \|\mathbf{X} - \mathbf{X}_i\|^2 \exp\left(-\frac{\|\mathbf{X} - \mathbf{X}_i\|^2}{2\sigma^2}\right). \quad (11)$$

Based on the statistics point of view, the quantum-potential formulation can be written as

$$V(\mathbf{X}) = E - \frac{d}{2} + \frac{\sum_{i=1}^N \|\mathbf{X} - \mathbf{X}_i\|^2 P(\mathbf{X}|\mathbf{X}_i)}{2\sigma^2} \quad (12)$$

where $P(\mathbf{X}|\mathbf{X}_i)$ is the a posteriori probability for \mathbf{X} , given the data samples $\{\mathbf{X}_i, i = 1, \dots, N\}$. This expression represents the weighted Euclidean distance from \mathbf{X} to a set of given data samples, where the weights are represented by their a posteriori probabilities. The bandwidth is used in the calculation of the probabilities $P(\mathbf{X}|\mathbf{X}_i)$ and as a general weighting factor. This resulting function represents the hypersurface of the potential function that was produced by the quantum-clustering algorithm. One adaptive version of this algorithm can be implemented similar to that employed by the mean shift but based on the second derivative, thus leading to the detection of local minima instead of local maxima for characterizing pdf modes.

D. Nonparametric Segmentation

In all KDE algorithms, the precision and the smoothness of the pdf representation depend on the size of the kernel bandwidth σ [13], [15], [16], [21], [23], [24], [32]. Although the mean-shift algorithm was proven to lead to the local maxima [26], the other two algorithms provide an intermediate representation given by a potential function. To apply these algorithms in various applications, we should identify the modes in the resulting kernel density representation [4], [13]. These modes are represented by the local maxima when using $\psi(\mathbf{X})$ based on (6) or by the local minima in the case of the quantum potential $V(\mathbf{X})$ based on (11), [29], [37].

The local Hessian is known as a good indicator of the variability in a function and has been used in parametric models [4], [38]. We consider a regular rectangular grid that corresponds to the given data range along each data dimension. The grid facilitates the calculation of the local Hessian [37]. We separate the knots of the grid according to the signs of the eigenvalues that result from the singular value decomposition of the potential function that was evaluated at the grid locations. Local maxima are detected when all the eigenvalues are negative, whereas local minima correspond to the case when all eigenvalues are positive. Saddle knots are decided when one eigenvalue has its sign opposite the others. We want to detect the local maxima in the case of the scale-space algorithm and local minima when considering the quantum-clustering algorithm. The detection of the modes is performed by evaluating the mass energy that was locally contained in the potential function. The ratio of the local energy from that corresponding to the entire data set is used as the criterion for identifying the modes. A label is assigned to each mode, and the entire grid is split by using a region-growing process that was applied on labeled regions [29]. This algorithm does not require the calculation of distances to a set of centers and preserves the shape of clusters without employing any parametric assumptions.

III. STATISTICAL ESTIMATION OF THE BANDWIDTH

A. Classical Bandwidth Estimation

The choice of the specific kernel function does not significantly influence the KDE, as previous studies have suggested [7]. One important problem in computational-intelligence methods that employ KDE is to find the appropriate bandwidth (scale parameter) [11], [13], [16], [21]–[24], [28], [32]. In the following discussion, we outline various kernel bandwidth estimators when processing 1-D data. The bandwidth can be chosen from measuring the estimation error between the underlying function $f(\mathbf{X})$ and its estimate $\psi(\mathbf{X})$. One well-known error measure is the MISE [7], i.e.,

$$MISE = E \left[\int (f(\mathbf{X}) - \psi(\mathbf{X}))^2 d\mathbf{X} \right] \quad (13)$$

where $\psi(\mathbf{X})$ depends on σ and is provided in (1). This expression can be decomposed into the variance and squared bias. The asymptotic MISE (AMISE), as $\sigma \rightarrow 0$ and $N \rightarrow \infty$, is given by

$$AMISE = \frac{\nu(\mathcal{K})}{N\sigma} + \frac{\sigma^4}{4} \mu^2(\mathcal{K}) \int [f''(\mathbf{X})]^2 d\mathbf{X} \quad (14)$$

where $\mu(\mathcal{K})$ is provided in (4), $f''(\mathbf{X})$ represents the second derivative of the underlying pdf function, and

$$\nu(\mathcal{K}) = \int \mathcal{K}^2(\mathbf{X}) d\mathbf{X}. \quad (15)$$

The bandwidth that minimizes $AMISE$ based on (14) is estimated as [7], [16], [24]

$$\hat{\sigma}_{AMISE} = \left[\frac{\nu(\mathcal{K})}{N\mu^2(\mathcal{K}) \int [f''(\mathbf{X})]^2 d\mathbf{X}} \right]^{1/5}. \quad (16)$$

The main problem with estimates such as $\hat{\sigma}_{AMISE}$ is that they require the underlying function $f(\mathbf{X})$, which is our problem in the first place.

The rule-of-thumb bandwidth estimation methods replace the integral depending on $f''(\mathbf{X})$ with a parametric family, which is estimated from the data [7], [16]. Under the conditions of Gaussianity for the given data, we have the following estimate for the $AMISE$ bandwidth:

$$\hat{\sigma}_{ROT} = \frac{1.06}{N^{1/5}} S \quad (17)$$

where S is the standard deviation of the given data $\{\mathbf{X}_i, i = 1, \dots, N\}$. For practical reasons, however, Silverman recommended using

$$\hat{\sigma}_{SROT} = \frac{0.9}{N^{1/5}} \min\{S, Q/1.34\} \quad (18)$$

where Q denotes the data sample interquartile range.

An approximation of $\hat{\sigma}_{AMISE}$, by considering that the distribution $Beta(4, 4)$ of variance S^2 minimizes $f''(\mathbf{X})$, was provided by Terrell in [22] as

$$\hat{\sigma}_{TER} = \frac{1.114}{N^{1/5}} S. \quad (19)$$

Plug-in bandwidth estimators replace the integrals in (16) with approximations and require proper initial estimates for the pilot bandwidth. Sheather and Jones developed a plug-in

method that chooses the bandwidth such that it solves the equation that results from (16), where the functions from the right-hand side depend on the kernel bandwidth [30]. An iterative procedure using the Newton–Raphson algorithm is used for solving the resulting equation. Ruppert *et al.* employed local least squares kernel regression to approximate the variance of the error for the plug-in estimate of the bandwidth [31]. Three different approaches are proposed, which are adaptations of existing plug-in bandwidth selectors for KDE: 1) the rule-of-thumb method; 2) the direct plug-in method; and 3) the solve-the-equation method. The first two approaches based on [31] are implemented using linear fitting to binned data, whereas the third approach is an extension of the method in [30]. According to several comparative studies, existing bandwidth estimation algorithms are known to underperform [16], [23], [24]. In this paper, we propose a novel approach to kernel bandwidth estimation based on the Bayesian methodology. The stages of the proposed methodology are described in the following sections.

B. Defining Local Neighborhoods

The proposed approach considers that the bandwidth σ can be associated with the local data spread, which is statistically characterized by the variance. Let us consider that the bandwidth is modeled by the a posteriori pdf $p(s|\mathbf{X})$, where s is the statistical variable that is associated with the bandwidth, and \mathbf{X} represents a data sample. The bandwidth determines the local smoothness in the resulting pdf approximation; thus, it should be calculated from a data subset that was locally defined. Let us consider K , i.e., the number of nearest neighbors to a specific data sample \mathbf{X}_i . All the other data samples are ordered according to their Euclidean distance to \mathbf{X}_i as

$$\|\mathbf{X}_{i,(1)} - \mathbf{X}_i\| < \dots < \|\mathbf{X}_{i,(K)} - \mathbf{X}_i\| < \dots < \|\mathbf{X}_{i,(N-1)} - \mathbf{X}_i\| \quad (20)$$

where $\mathbf{X}_{i,(j)}$ is the j th ordered data sample according to the Euclidean distance from \mathbf{X}_i , and $\mathbf{X}_{i,(j)} \neq \mathbf{X}_i$ for $j = 1, \dots, N-1$. The proposed approach estimates the kernel bandwidth by using the statistics of the variances that correspond to KNN data sample populations [2], where $K < N$.

Let us assume n neighborhoods of various sizes $\{K_j, j = 1, \dots, n\}$. We define a pdf of the bandwidth $p(s|\mathbf{X})$ that is expressed as a pseudolikelihood of the form

$$P(s|\mathbf{X}) = \prod_{j=1}^n P(s|\mathbf{X}_{K_j}) \quad (21)$$

where $P(s|\mathbf{X}_{K_j})$ represents the probability of the bandwidth, depending on the K_j nearest neighborhood data samples to \mathbf{X}_{K_j} . These probabilities can be evaluated over an entire range of K_j , i.e.,

$$P(s|\mathbf{X}_{K_j}) = \int P(s|K_j, \mathbf{X}_{K_j}) P(K_j|\mathbf{X}_{K_j}) dK_j \quad (22)$$

and after using the Bayes rule, we obtain

$$P(K_j|\mathbf{X}_{K_j}) = \frac{P(\mathbf{X}_{K_j}|K_j)P(K_j)}{P(\mathbf{X}_{K_j})} \quad (23)$$

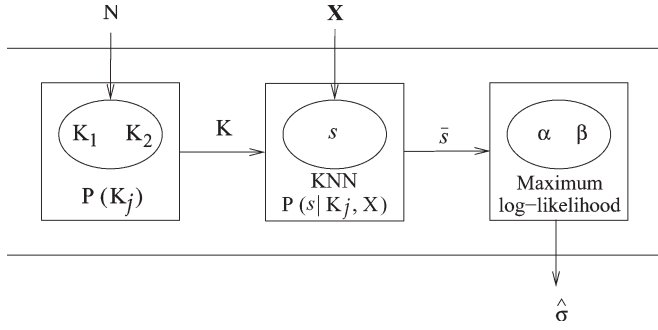


Fig. 1. Graphical model that describes the stages for estimating the bandwidth $\hat{\sigma}$.

where $P(\mathbf{X}_{K_j}|K_j)$ is the probability of the data sample population, depending on the specific neighborhood size. In the following discussion, we consider $P(K_j)$ as a uniform distribution limited to the range $[K_1, K_2]$. To find the bandwidth, we use the maximum log-likelihood estimation in the expression that results after replacing (22) into (21).

The bandwidth estimation approach consists of three steps, and it is outlined in the graphical model in Fig. 1. KNN is an estimation algorithm that considers a localized data subset that has been used in various pattern recognition applications [2]. In this case, the only necessary elements are the bounds of the uniform distribution that characterize $P(K_j)$, as given by $K_j \in [K_1, K_2], j = 1, \dots, n$. These bounds are chosen depending on N , i.e., the number of data samples.

C. Modeling the Local Variance

After sampling K_j from the uniform distribution $P(K_j)$, we randomly sample the given data set $\{\mathbf{X}_i, i = 1, \dots, N\}$ and consider these data samples $\mathbf{X}_{i'}$ as centers for evaluating their K_j -nearest neighbors (KNN). The K_j -nearest neighbors are selected according to their Euclidean distance to sampled $\mathbf{X}_{i'}$, as in (20). This step results in data sets \mathbf{X}_{K_j} for which we calculate the variance as

$$s_i = \frac{\sum_{k=1}^{K_j} \|\mathbf{X}_{i,(k)} - \mathbf{X}_{i'}\|^2}{K_j - 1} \quad (24)$$

where $\{\mathbf{X}_{i,(k)}, k = 1, \dots, K_j\}$ are the nearest neighbors to the sampled data $\mathbf{X}_{i'}$. The estimate s_i in (24) is considered a random variable, whose statistics can be used to infer the estimated bandwidth $\hat{\sigma}$. We consider that the bandwidth should depend on the local data variance distribution. In the following discussion, we fit the empirical distributions of variances, which were estimated from the given data sets, to their underlying pdf.

The distribution of variances for data samples that are generated independent of the normal distribution with mean 0 and variance 1 are modeled by the chi-square distribution. Gamma distribution is a generalization of the chi-square distribution and is suitable for modeling distributions of data sample variances in the case when we have no knowledge about the underlying data distribution. Due to its generality, Gamma distribution can model the bandwidth distribution for kernels that are not necessarily Gaussian. Gamma distributions have been used as priors on the distributions of both the source signals and the

mixing coefficients in blind-source separation applications [39]. The Gamma pdf is given by the following expression [1]:

$$P(s|\alpha, \beta) = \frac{\beta^\alpha s^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta s} \quad (25)$$

where $s \geq 0, \alpha > 0$ is the shape parameter, and $\beta > 0$ is the scale parameter of the Gamma distribution. When β is greater or smaller than 1, it allows the density function to stretch or compress, respectively. $\Gamma(\cdot)$ represents the Gamma function

$$\Gamma(t) = \int_0^\infty r^{t-1} e^{-r} dr. \quad (26)$$

After calculating the local variance for each of the data subsets that describe $P(s|\mathbf{X}_{K_j})$ and after taking into account $P(K_j)$, when uniformly varying $K_j \in [K_1, K_2]$, we form a data set that corresponds to the random variable s , which was calculated as in (24). This data set is used as an empirical a priori probability for the local variance. In the following section, we describe the maximum log-likelihood approach for inferring the Gamma distribution parameters α and β .

D. Estimating Gamma Distribution Parameters

Two methods have been proposed to estimate the parameters of a Gamma distribution: 1) data moments [35] and 2) the maximum likelihood estimation [36]. The data moments method lacks the efficiency required for estimating the shape parameter of the Gamma distribution. The maximum likelihood approach was proposed in [36] to overcome this drawback in numerical estimation. The likelihood function that corresponds to the distribution in (25) is

$$\mathcal{L}(\alpha, \beta) = \prod_{i=1}^M p(s_i|\alpha, \beta) = \frac{\beta^{\alpha M}}{\Gamma^M(\alpha)} \left[\prod_{i=1}^M s_i \right]^{\alpha-1} e^{-\beta \sum_{i=1}^M s_i} \quad (27)$$

where we consider M data samples $\{s_i|i = 1, \dots, M\}$, with each representing the variance of a local neighborhood, calculated according to (24). We estimate the parameters α and β by equating the likelihood derivatives to zero as follows:

$$\begin{cases} \frac{\partial \ln \mathcal{L}(\alpha, \beta)}{\partial \alpha} = 0 \\ \frac{\partial \ln \mathcal{L}(\alpha, \beta)}{\partial \beta} = 0. \end{cases} \quad (28)$$

This estimation results in the following system of equations:

$$\begin{cases} \ln(\hat{\alpha}) - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + \ln \left[\frac{\left(\prod_{i=1}^M s_i \right)^{1/M}}{\bar{s}} \right] = 0 \\ \hat{\beta} = \frac{\hat{\alpha}}{\bar{s}} \end{cases} \quad (29)$$

where \bar{s} is the sample mean for the variable s , i.e.,

$$\bar{s} = \frac{\sum_{i=1}^M s_i}{M}. \quad (30)$$

The third term in the first equation of the system in (29) is the logarithm of the ratio between the geometric and arithmetic means. The first nonlinear equation in (29) is solved using the

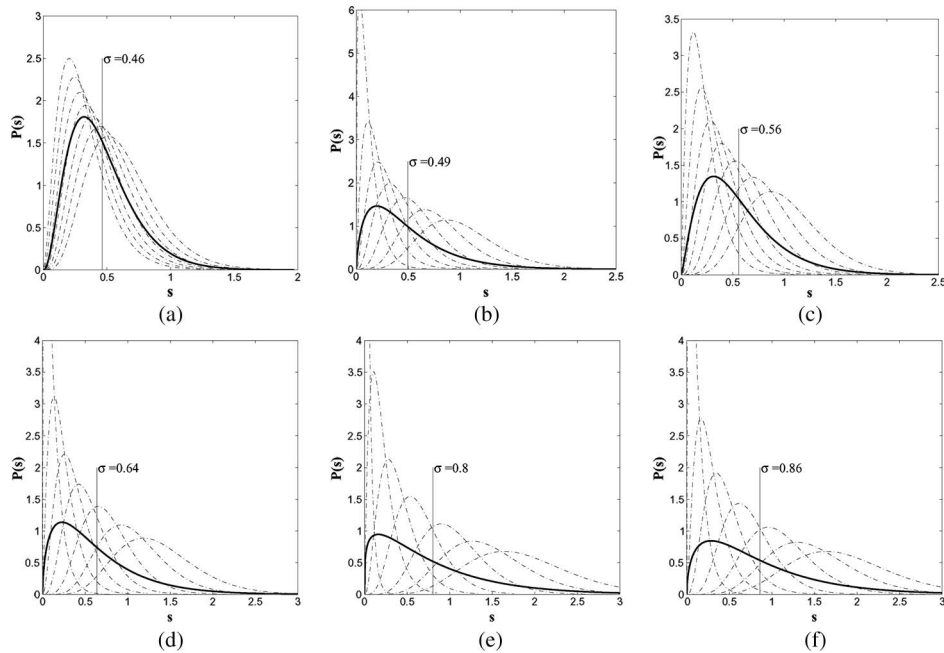


Fig. 2. Evaluation of the scaling parameter for 4-QAM modulated data when considering various intervals for the boundaries of the uniform distribution for K . (a) $[K_1, K_2] = [\frac{N}{7}, \frac{N}{4}]$. (b) $[K_1, K_2] = [\frac{N}{20}, \frac{N}{3}]$. (c) $[K_1, K_2] = [\frac{N}{10}, \frac{N}{3}]$. (d) $[K_1, K_2] = [\frac{N}{20}, \frac{2N}{5}]$. (e) $[K_1, K_2] = [\frac{N}{80}, \frac{N}{2}]$. (f) $[K_1, K_2] = [\frac{N}{20}, \frac{N}{2}]$.

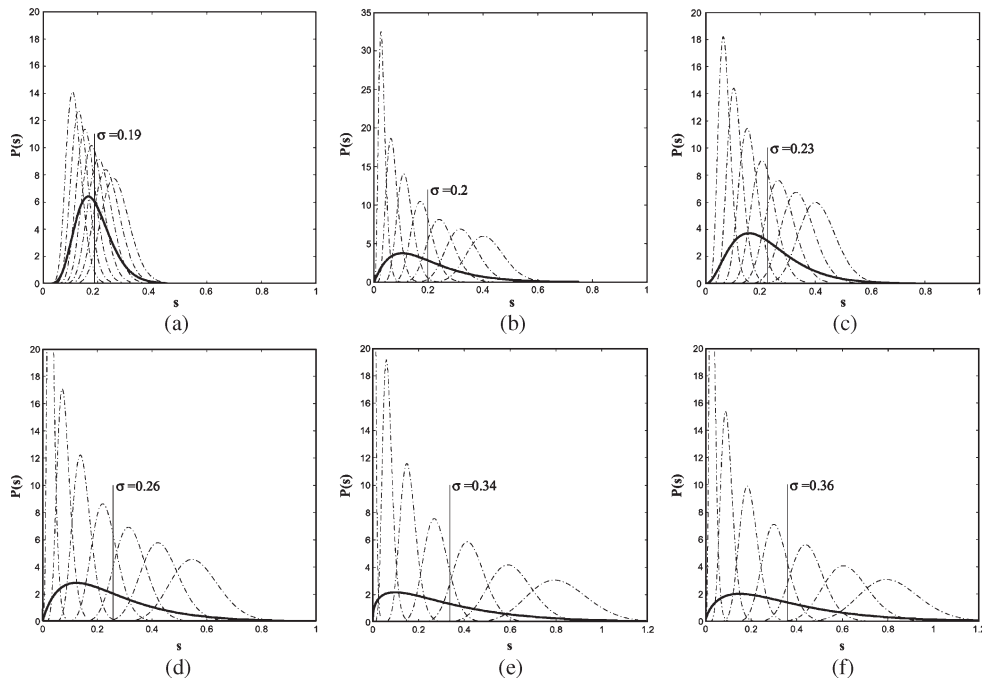


Fig. 3. Evaluation of the scaling parameter for 8-PSK modulated data when considering various intervals for the boundaries of the uniform distribution for K . (a) $[K_1, K_2] = [\frac{N}{7}, \frac{N}{4}]$. (b) $[K_1, K_2] = [\frac{N}{20}, \frac{N}{3}]$. (c) $[K_1, K_2] = [\frac{N}{10}, \frac{N}{3}]$. (d) $[K_1, K_2] = [\frac{N}{20}, \frac{2N}{5}]$. (e) $[K_1, K_2] = [\frac{N}{80}, \frac{N}{2}]$. (f) $[K_1, K_2] = [\frac{N}{20}, \frac{N}{2}]$.

Newton–Raphson iterative algorithm with respect to $\hat{\alpha}$ [36]. This step results in the following updating equation:

$$\hat{\alpha}_{t+1} = \hat{\alpha}_t - \frac{\ln(\hat{\alpha}_t) - \Psi(\hat{\alpha}_t) + \ln \left[\frac{(\prod_{i=1}^M s_i)^{1/M}}{\sum_{i=1}^M s_i} \right]}{1/\hat{\alpha}_t - \Psi(\hat{\alpha}_t)} \quad (31)$$

where $\Psi(\hat{\alpha}_t)$ is the Digamma function, which represents the logarithmic derivative of the Gamma function, i.e.,

$$\Psi(\hat{\alpha}) = \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} \quad (32)$$

where $\Gamma(\cdot)$ is provided in (26), and $\Gamma'(\cdot)$ represents its first derivative. After estimating $\hat{\alpha}$ at the convergence of (31), we

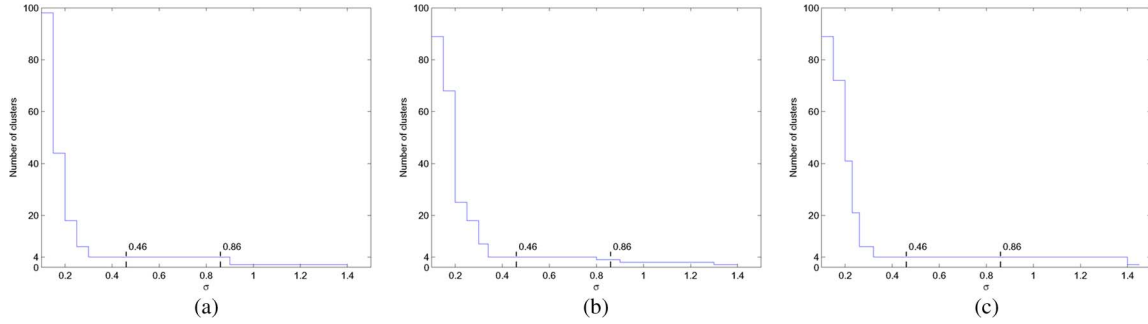


Fig. 4. Finding the number of modes in 4-QAM using nonparametric modeling. (a) Using the scale-space method. (b) Using the mean-shift method. (c) Using the quantum-clustering method.

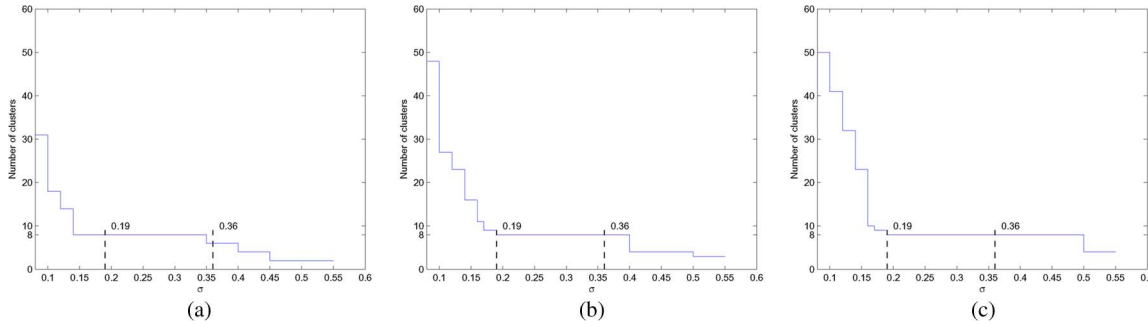


Fig. 5. Finding the number of modes in 8-PSK using nonparametric modeling. (a) Using the scale-space method. (b) Using the mean-shift method. (c) Using the quantum-clustering method.

replace it in the second equation in (29) and estimate $\hat{\beta}$. A good initialization for the Newton–Raphson optimization is achieved using the moments method estimate for $\hat{\alpha}$ [35]. The maximum likelihood estimation was shown to provide a minimal bias when estimating the Gamma distribution parameters in [36]. We estimate the bandwidth as the mean of the Gamma function that corresponds to distributions of KNN variances.

IV. EXPERIMENTAL RESULTS

The proposed KDE methodology was applied on various data sets. These include artificially generated blind-source separation data sets as well as vector fields representing surface normals estimated from radar images of terrain. For the blind-source separation data we consider two cases of modulated signals: quadratic-amplitude modulated (QAM) signals and phase-shift keying (PSK) modulated signals that have also been used in [3]. The perturbation channel equations, assuming interference and noise, that were considered for the 8-PSK data are

$$\begin{cases} x_I(t) = I(t) + 0.2I(t-1) - 0.2Q(t) \\ \quad - 0.04Q(t-1) + \mathcal{N}(0, 0.11) \\ x_Q(t) = Q(t) + 0.2Q(t-1) + 0.2I(t) \\ \quad + 0.04I(t-1) + \mathcal{N}(0, 0.11) \end{cases} \quad (33)$$

where $(x_I(t), x_Q(t))$ are the in-phase and in-quadrature signal components at time t on the communication channel, whereas $I(t)$ and $Q(t)$ correspond to the signal symbols that represent the ground-truth signals. For 8-PSK, there are eight signal symbols, which were equidistantly located on a circle, whereas 4-QAM has four signal symbols. We consider that the given signals are corrupted by additive Gaussian noise. In the case of 8-PSK, the noise statistics corresponds to a signal-to-noise

ratio (SNR) of 22 dB. We have generated $N = 960$ signals by assuming equal probabilities for all possible consecutive inter-symbol combinations. For 4-QAM signals, we assume additive noise, with SNR of 8 dB and no interference. A certain degree of overlap occurs among the components of the generated data clusters.

The proposed bandwidth estimation methodology was applied to the KDE algorithms in Section II: 1) scale-space; 2) mean-shift; and 3) quantum clustering. The blind separation of sources is achieved by data clustering according to the corresponding pdf modeling. The scale-space algorithm directly employs the mode-seeking procedure to the function $\psi(\mathbf{X})$ defined in (6) and assigns the signals to their corresponding clusters. The mean-shift algorithm defines clusters by updating their centers according to the local gradient using (9) by employing the Epanechnikov kernel in (5). Quantum clustering calculates the quantum potential $V(\mathbf{X})$ by using (11) and splits it into regions according to its modality, with each region corresponding to a symbol.

Following the proposed methodology for selecting the bandwidth, according to the graphical model in Fig. 1, we consider that K is defined by a uniform distribution bounded in the interval $K \in [K_1, K_2]$, which was chosen as a prior. After choosing the limits of the bounded interval, K_1 and K_2 , we select several values for K_j , i.e., the number of nearest neighbors, as follows:

$$K_j = K_1 + j \frac{K_2 - K_1}{n}. \quad (34)$$

We consider $n = 7$, and we generate the neighborhood sizes $\{K_j | j = 1, \dots, n\}$, which were equally spaced in the given interval. We apply the methodology in Sections III-B and C by subsampling the given data set. We fit the Gamma pdf $p(s|\alpha, \beta)$ in (25) to the empirical distributions of the local variance by

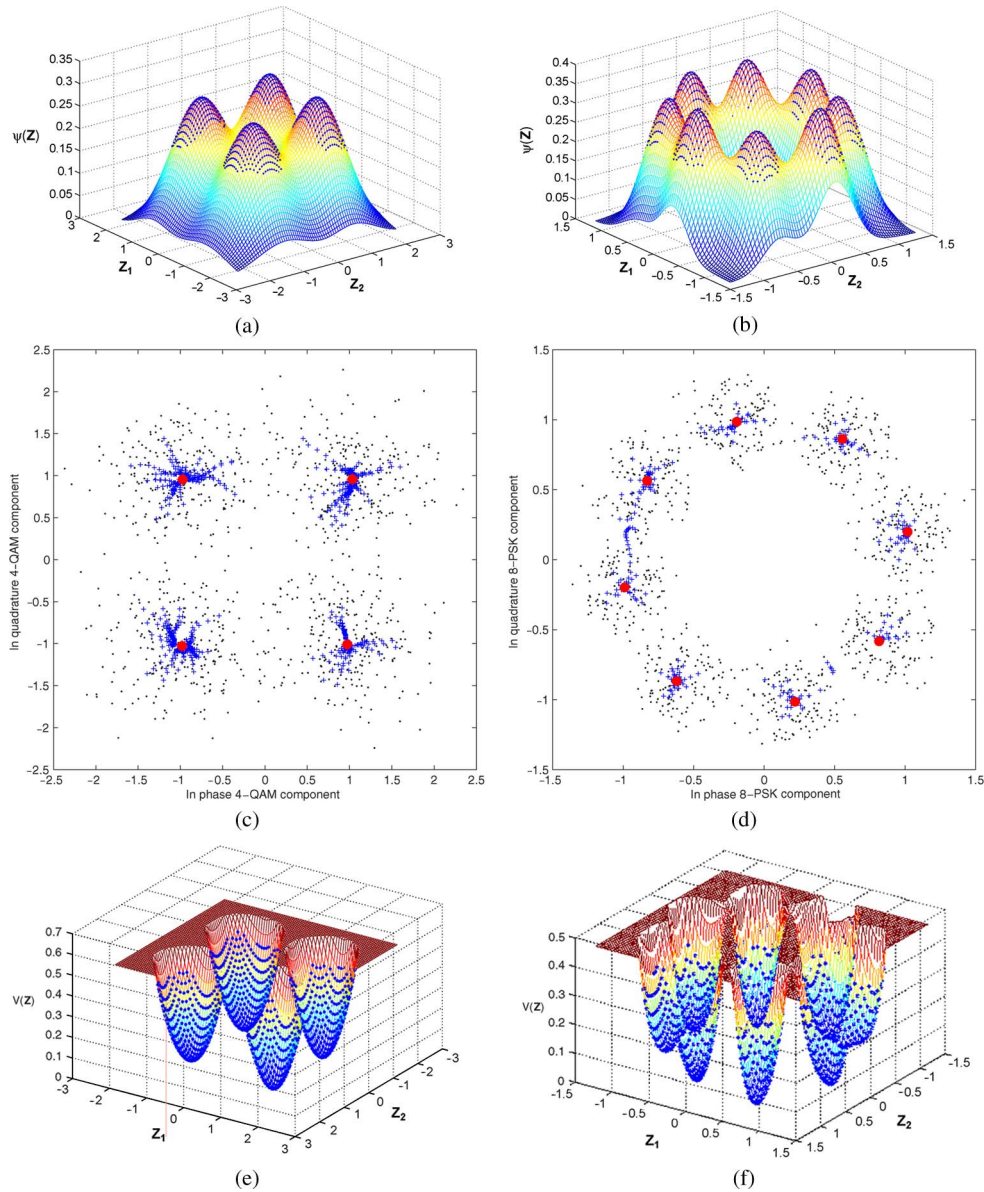


Fig. 6. Defining the influence regions for each symbol in blind detection of modulated signals. In (c) and (d), the intermediate mean shifts are shown by “+”, whereas the mean shifts that were obtained at convergence are indicated with larger circles. (a) $\psi(\mathbf{Z})$ for 4-QAM. (b) $\psi(\mathbf{Z})$ for 8-PSK. (c) Mean shift for 4-QAM. (d) Mean shift for 8-PSK. (e) $V(\mathbf{Z})$ for 4-QAM. (f) $V(\mathbf{Z})$ for 8-PSK.

estimating its parameters $\hat{\alpha}$ and $\hat{\beta}$ according to (29) and (31). We have chosen several values for the bounds K_1 and K_2 , and the results that display the fitting of the Gamma distribution $p(s|\alpha, \beta)$ in (25) are shown in Fig. 2 for 4-QAM and in Fig. 3 for 8-PSK, where the intervals $[K_1, K_2]$ are specified in the figure caption. These plots show a good data fit for the Gamma distribution to the empirical distributions of KNN variances, for $\{K_j|j = 1, \dots, n\}$, and when considering various bounds K_1 and K_2 . The Gamma distribution that models the variance of KNN, for the entire range of neighbors $K \in [K_1, K_2]$, is represented in boldface in all the plots in Figs. 2 and 3. In this paper, we assume a homoscedastic bandwidth $\hat{\sigma}$ that corresponds to the mean of the Gamma distribution, which is indicated by a vertical line in the plots in Figs. 2 and 3, which correspond to 4-QAM and 8-PSK data, respectively.

In the given experimental context, we identify the modes, with each corresponding to a source signal when using all

TABLE I
BANDWIDTH ESTIMATION USING DIFFERENT METHODS

Method	4-QAM	8-PSK
$\hat{\sigma}_{ROT}$	0.2535	0.1729
$\hat{\sigma}_{SROT}$	0.2154	0.1469
$\hat{\sigma}_{TER}$	0.2738	0.1867
$\hat{\sigma}_{SJ}$	0.1220	0.0850
$\hat{\sigma}_{RSW-ROT}$	0.3996	0.0470
$\hat{\sigma}_{RSW-DPI}$	0.3370	0.0696
$\hat{\sigma}_G$	0.46-0.86	0.19-0.36

three machine-learning approaches as described in Section II and when varying the bandwidth σ . The number of detected modes are shown in Fig. 4 for 4-QAM and in Fig. 5 for 8-PSK. The range of $\hat{\sigma}$, which was estimated when choosing various intervals for K , is shown in between two dashed vertical lines

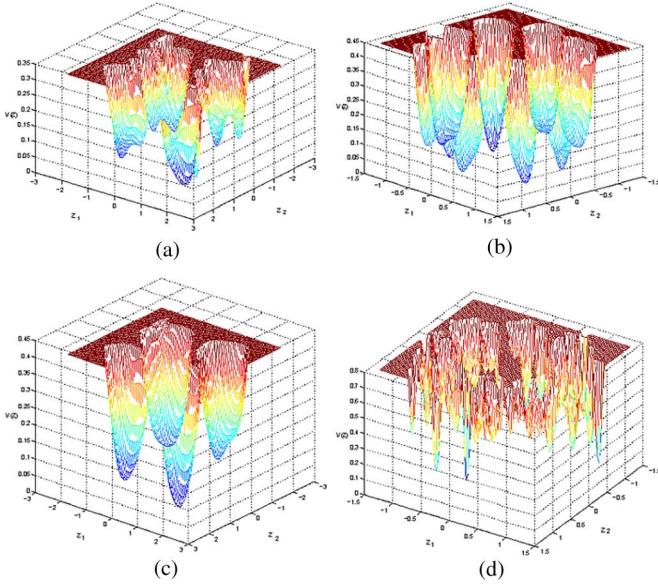


Fig. 7. Quantum potential functions $V(\mathbf{X})$ that embed the bandwidth that was estimated using different estimation methods. (a) Using the $\hat{\sigma}_{ROT}$ bandwidth for 4-QAM. (b) Using the $\hat{\sigma}_{ROT}$ bandwidth for 8-PSK. (c) Using the Ruppert *et al.* direct-plug-in bandwidth for 4-QAM. (d) Using the Ruppert *et al.* direct-plug-in bandwidth for 8-PSK.

in each of the plots in Figs. 4 and 5. The signal constellations for 4-QAM and 8-PSK and the cluster center results that were provided by the mean-shift algorithm are shown in Fig. 6(c) and (d), respectively. The function $\psi(\mathbf{Z})$ for 4-QAM is represented in Fig. 6(a), whereas the quantum potential $V(\mathbf{Z})$ is displayed in Fig. 6(e), where \mathbf{Z} represents a regular grid. The same information for 8-PSK is shown in Fig. 6(b) and (f), respectively. A significant variation of different interval bounds $[K_1, K_2]$ is considered, and as it can be observed, the resulting variation of the bandwidth $\hat{\sigma}$ is limited to a small range of values.

A series of bandwidth estimators are considered for the given data and the resulting estimates are provided in Table I.¹ The bandwidth estimates include $\hat{\sigma}_{ROT}$ based on (17) and a Silverman estimate $\hat{\sigma}_{SROT}$ based on (18). Other comparative bandwidth estimators are the Terrell bandwidth [22] $\hat{\sigma}_{TER}$ based on (19) and Sheather and Jones described in [30], which is denoted as $\hat{\sigma}_{SJ}$. The rule-of-thumb and direct-plug-in methods in [31] are denoted as $\hat{\sigma}_{RSW-ROT}$ and $\hat{\sigma}_{RSW-DPI}$, respectively. The bandwidth that was estimated through the proposed methodology is denoted as $\hat{\sigma}_G$. When looking at the valid mode range of 4 for 4-QAM in the plots in Fig. 4, we can observe that only $\hat{\sigma}_{RSW-ROT}$ provides suitable estimates, whereas $\hat{\sigma}_{RSW-DPI}$ gives appropriate results only for the scale-space method. In the case of 8-PSK, there are three methods that provide suitable bandwidths for the scale-space method but none for the other two KDE-based methods. It can be observed that the results that were provided by the proposed bandwidth estimation method are almost always included in the right range of values for all three KDE methods for 4-QAM and 8-PSK data according to the plots in Figs. 4 and 5, respectively.

We consider the bandwidth values $\hat{\sigma}_{ROT}$ and $\hat{\sigma}_{RSW-DPI}$ that were provided in Table I and use them for modeling the

quantum potentials $V(\mathbf{X})$ of 4-QAM and 8-PSK data. The quantum potentials, when using $\hat{\sigma}_{ROT}$ according to [7] in 4-QAM and 8-PSK, are shown in Fig. 7(a) and (b), respectively. The quantum potentials, when using the direct-plug-in bandwidth of Ruppert *et al.* in 4-QAM and 8-PSK, are shown in Fig. 7(c) and (d), respectively. It can be observed that the modes of 4-QAM are correctly detected only in Fig. 7(c) by using $\hat{\sigma}_{RSW-DPI}$, whereas when using $\hat{\sigma}_{ROT}$ in Fig. 7(b), there are nine modes that were detected instead of eight for the 8-PSK data. The modes are not correctly detected in the quantum potentials in Fig. 7(a) and (d). All the plots in Figs. 4 and 5 indicate that there is a large bandwidth interval of values for which the proper modality can be found when using the methodology in Section II. In particular, the quantum-clustering method provides the right modality for a larger range of values than the scale-space and mean-shift methods. In almost all cases, the proposed bandwidth estimation methodology finds the appropriate number of modes.

For the range of the estimated bandwidth $\hat{\sigma}$, we evaluate the bias in finding the cluster centers, which correspond to the detected symbols. In the case of scale space and quantum clustering, due to the symmetry of the given data, we assimilate the modes with the centers of the clusters, which correspond to the ground-truth signal values. The mean-shift algorithm was applied onto the entire data set, eventually assigning each data sample to the cluster center that was found at convergence. We calculate the mean square error between a detected center by one of three machine-learning algorithms and the closest sample mean (MSE_{SM}) of a cluster. For the scale space and quantum clustering, the bias in estimating the cluster centers is given by

$$MSE_{SM}(\psi(\mathbf{X})) = \frac{1}{h} \sum_{j=1}^h \left\| \frac{\sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i \psi(\mathbf{x}_i)}{\sum_{\mathbf{x}_i \in C_j} \psi(\mathbf{x}_i)} - \frac{\sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i}{N_{C_j}} \right\|^2 \quad (35)$$

$$MSE_{SM}(V(\mathbf{X})) = \frac{1}{h} \sum_{j=1}^h \left\| \frac{\sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i (E - V(\mathbf{x}_i))}{\sum_{\mathbf{x}_i \in C_j} (E - V(\mathbf{x}_i))} - \frac{\sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i}{N_{C_j}} \right\|^2 \quad (36)$$

where N_{C_j} is the number of data assigned to the cluster C_j , $\|\cdot\|$ represents the Euclidean distance, and $h = 4$ is for 4-QAM, whereas $h = 8$ is for 8-PSK. We also calculate MSE_O , i.e., the mean square error between the cluster centers and the noise free centers whose values replace the second fraction in (35) and (36). The numerical results are provided in Table II for both 4-QAM and 8-PSK data. The confidence intervals for the bias estimates, which resulted from several runs, are also indicated in Table II. When considering the bandwidths that were estimated by other methods, as provided in Table I, in the nonparametric potential functions, we obtain much larger bias than the values in Table II, even when the right number of modes is found. As it can be observed in Table II, quantum clustering provides lower bias with respect to the center estimation compared with the other two methods.

In the second set of experiments, we have applied the methodology that was outlined in this paper to segment terrain topography. A set of SAR images of terrain has been processed by considering the signal statistics and radar signal illumination compensation. Eventually, for each block of pixels, we estimate

¹The software for implementing these bandwidth estimators has been adapted from <http://stat-or.unc.edu/webpace/miscellaneous/marron/software/>.

TABLE II
BIAS AND CONFIDENCE INTERVALS IN THE BLIND DETECTION OF MODULATED SIGNALS

Method	4-QAM		8-PSK	
	MSE_{SM}	MSE_O	MSE_{SM}	MSE_O
Scale space	0.0133 ± 0.0065	0.0249 ± 0.0086	0.0024 ± 0.00074	0.0041 ± 0.00073
Mean shift	0.0073 ± 0.0031	0.0110 ± 0.0040	0.0168 ± 0.0097	0.0200 ± 0.0109
Quantum clustering	0.0057 ± 0.0014	0.0111 ± 0.0021	0.00092 ± 0.00026	0.0022 ± 0.0004

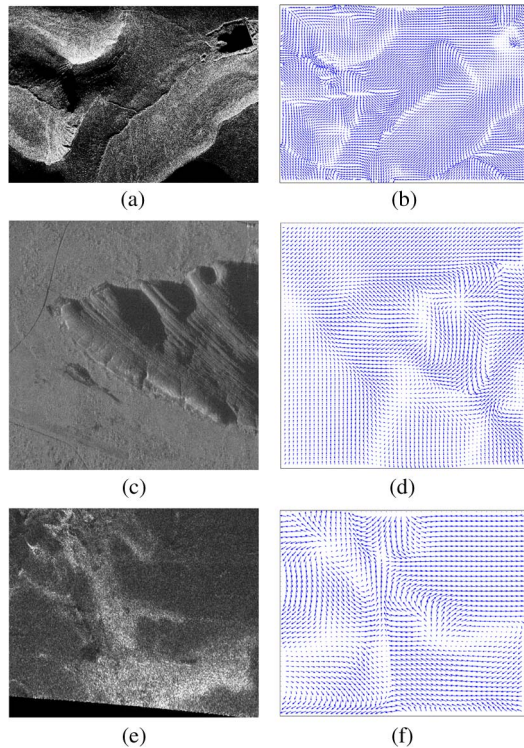


Fig. 8. SAR images that show terrain and their corresponding vector fields of surface normals. (a) SAR image from Wales. (b) Surface normals. (c) SAR image of the Ayers Rock. (d) Surface normals. (e) SAR image of the Titan surface. (f) Surface normals.

the local surface normal by adapting the shape-from-shading methodology to radar images [34]. The resulting vector fields of surface normals have been smoothed using a curvature-consistency approach [38]. Three SAR images of terrain, which display various characteristics, are shown in Fig. 8(a), (c) and (e). These images represent terrain information from Wales, the Ayers Rock from Australia, and an image that was sent by the Cassini orbiter in November 2004, while passing nearby Titan, which is a moon of Saturn, respectively. The smoothed vector fields of surface normals that correspond to these three SAR images are shown in Fig. 8(b), (d) and (f), respectively.

The surface-normal vectors are normalized, $\sqrt{x^2 + y^2 + z^2} = 1$, where (x, y, z) are the components of the surface-normal vector. In the following discussion, we consider only two components of the surface-normal vector along the x - and y -directions, respectively. These data are all contained in the circle of center zero and radius one and possess local correlation due to the smoothing [34], [38]. Given the characteristics of these data distributions and following the acquisition and postprocessing, a parametric method would not provide appropriate results, and a nonparametric approach

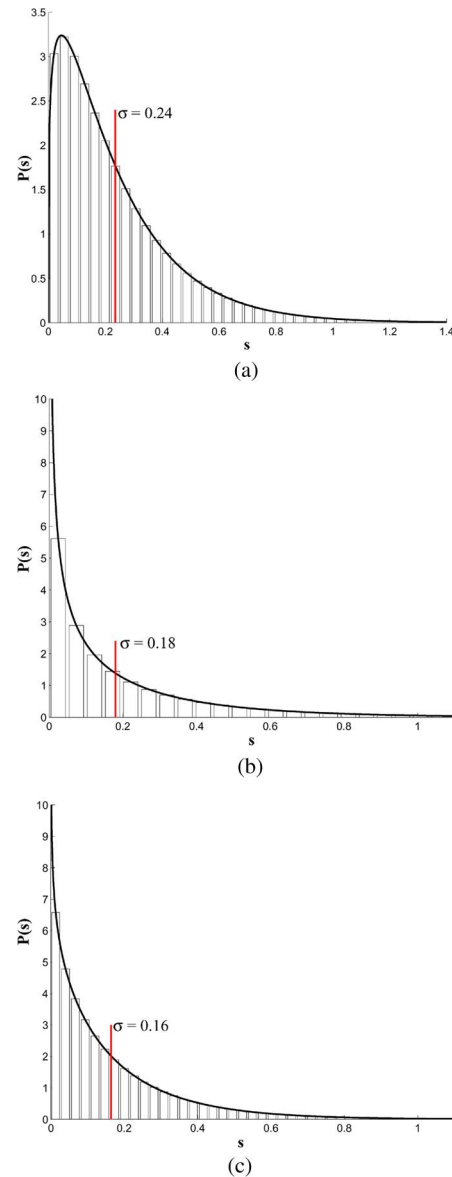


Fig. 9. Histograms of variances of KNN neighborhoods that were sampled from the SAR surface normal data when $K \in [N/20, N/2]$. (a) Bandwidth estimation for the Wales data. (b) Bandwidth estimation for the Ayers Rock data. (c) Bandwidth estimation for the Titan data.

should be adopted. We consider 1518, 1200, and 1050 local vector estimates of surface normals for the Wales, Ayers Rock, and Titan data, respectively. The aim of this application is to identify various topographical regions from SAR images of terrain by using clustering in the given 2-D space.

We apply the methodology in Section III to estimate the appropriate bandwidth. We consider seven different neighborhood

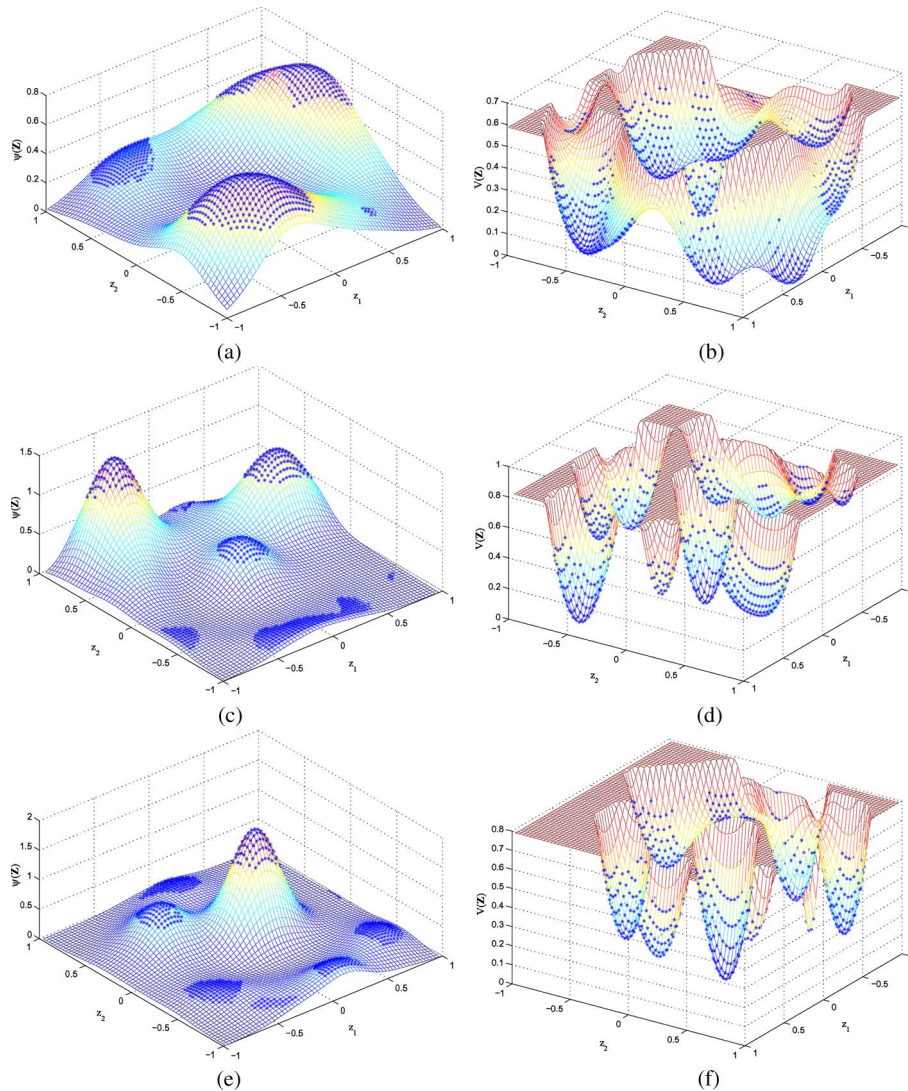


Fig. 10. Kernel density estimation surfaces that correspond to vector fields of surface normals. (a) Using the scale-space representation for the Wales data. (b) Using quantum clustering for the Wales data. (c) Using the scale-space representation for the Ayers Rock data. (d) Using quantum clustering for the Ayers Rock data. (e) Using the scale-space representation for the Titan data. (f) Using quantum clustering for the Titan data.

sizes, which were equally spaced inside the range $K_j \in [N/20, N/2]$, where N is the data size. By applying the proposed statistical method for the bandwidth estimation, we fit the Gamma distribution function, by estimating its parameters, to the histograms that represent empirical distributions of local variances, as illustrated in Fig. 9(a) for the Wales data, in Fig. 9(b) for the Ayers Rock data, and in Fig. 9(c) for the Titan data. Based on these figures, we can observe that, despite the variability in the shape of the resulting histogram for various data sets, we always achieve a good fit to the Gamma distribution. The resulting Gamma distribution means, which were considered bandwidth estimates, are indicated in the plots in Fig. 9.

The same bandwidth is used for all three KDE methods, which were described in Section II: 1) scale space; 2) mean shift; and 3) quantum clustering. A rectangular lattice is considered for each data set, and let us denote by \mathbf{Z} the location of a knot on this lattice. The locations of these knots are used for evaluating the KDE. The pdf surface representations for the given data sets, when using the scale space, i.e., by evaluating $\psi(\mathbf{Z})$ according to (6), are shown in Fig. 10(a) for the Wales data, in Fig. 10(c) for the Ayers Rock data, and in Fig. 10(e) for

the Titan data, respectively. The quantum-potential surfaces, which correspond to $V(\mathbf{Z})$ and were calculated according to (11), are shown in Fig. 10(b) for the Wales data, in Fig. 10(d) for the Ayers Rock data, and in Fig. 10(f) for the Titan data, respectively. The knots that were associated with various segmented regions are marked with “*” onto these surfaces. In the case of the scale space, a series of maxima can be observed. Conversely, data clustering is characterized by the minima of the quantum potential. The modes that were detected by the quantum potential are better defined than those found by the scale-space algorithm as it can be observed from these figures.

The resulting segmentation in topographic regions for the SAR image in Fig. 8(a), according to the surface-normal vector clustering, is shown in Fig. 11. The results, when using the scale-space algorithm according to the evaluation of $\psi(\mathbf{Z})$ in (6), are displayed in Fig. 11(a). The results that were provided by the mean-shift algorithm at convergence, which were calculated according to (8), are shown in Fig. 11(b), whereas the results that were produced by the quantum-clustering algorithm, as defined by the quantum potential $V(\mathbf{Z})$ according to (11), are provided in Fig. 11(c). The number of detected modes is five for

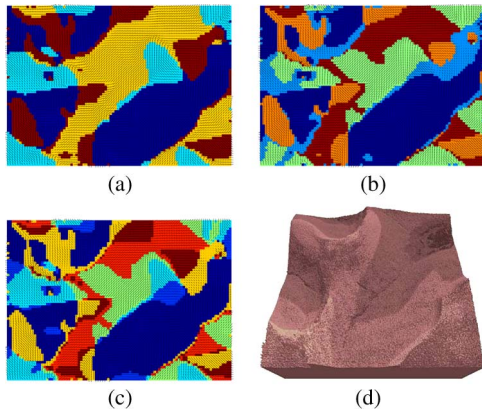


Fig. 11. Topography segmentation of the SAR image from Wales using the kernel density representation of surface normals. (a) Using scale space. (b) Using mean shift. (c) Using quantum clustering. (d) Three-dimensional representation of the terrain.

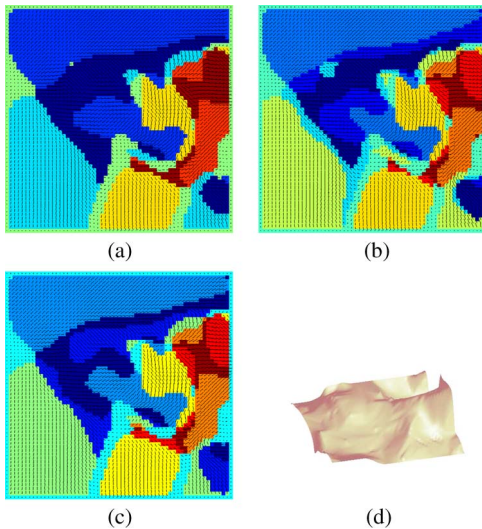


Fig. 12. Topography segmentation of the SAR image from the Ayers Rock using the kernel density representation of surface normals. (a) Using scale space. (b) Using mean shift. (c) Using quantum clustering. (d) Three-dimensional reconstruction of the terrain.

the scale-space algorithm, five for the mean-shift algorithm, and nine for the quantum clustering. The 3-D digital elevation map, which corresponds to the mountainous area in Wales, is shown in Fig. 11(d) and calculated from the field of surface normals according to the algorithm in [40]. At the first glance, all the methods accomplish to roughly capture the dominant regions of terrain in this SAR image, such as the circular lake in the top right corner, the mountainous semicircular area in the left side, and the deep valley that obliquely crosses the image and ends near the lake. The terrain segmentation around the lake is discernible in the quantum clustering, whereas for the other two methods, it is diffused into the area that lies beneath the lake. The same segmentation methodology is applied for the SAR images in Fig. 8(c) and (e). The segmentation results that were obtained for the scale space, mean shift, and quantum clustering are shown in Fig. 12 for the SAR image of the Ayers Rock and in Fig. 13 for the SAR image of Titan. In Figs. 12(d) and 13(d), the reconstructed 3-D heights from their corresponding surface normals are shown after being calculated according to the algorithm in [34]. After segmenting the vector

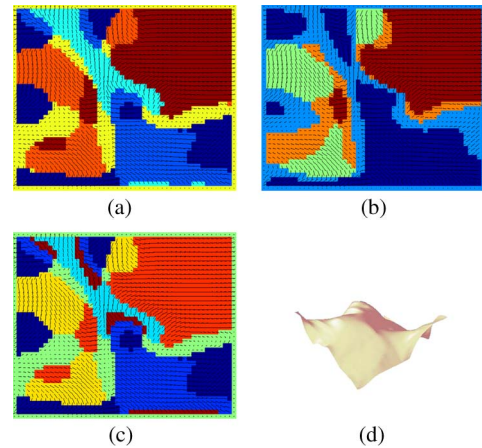


Fig. 13. Topography segmentation of the SAR image from Titan using the kernel density representation of surface normals. (a) Using scale space. (b) Using mean shift. (c) Using quantum clustering. (d) Three-dimensional reconstruction of the terrain.

field of surface normals for the Ayers Rock, a total of seven clusters were found by the scale-space algorithm, ten by the mean-shift, and eleven by the quantum-clustering algorithm. For the SAR image from Titan, the number of nonparametric clusters detected by the scale-space algorithm was eight, whereas five clusters were identified using the mean shift and ten clusters using the quantum-clustering method. Most of the clusters correspond to well-defined and compact segmented regions from the maps. In general, all three methods agree on the mapping of the terrain, particularly with respect to the separation between the homogeneous area that surrounds a clearly mountainous territory. In the case of the Ayers Rock, the quantum-clustering and the mean-shift methods considered the bottom central region as a compact area, whereas the scale-space method failed to perceive its homogeneity. The top side of the mountainous area proved quite challenging for all the techniques due to the shadowing effect, which caused many errors in the estimation of surface normals from that region [34]. The quantum-clustering algorithm, which relies on the second derivative according to (11), consistently produced more regions for the same data set than the other two methods. This result highlights its sensitivity to small variations in the surface-normal orientation. The segmentation results show that the proposed bandwidth estimation methodology was suitable for all the three nonparametric methods that were considered in this paper. The resulting segmentation can be used in a graph-based representation of terrain, where the main graph knots identify relevant terrain components.

V. CONCLUSION

This paper has described a new methodology for kernel bandwidth estimation that is applied to the following KDE methods: 1) scale space; 2) mean shift; and 3) quantum clustering. A Bayesian approach has been proposed to estimate the bandwidth by modeling distributions of variances of localized data subsets. A uniform distribution has been considered to model the data neighborhood size. After sampling K , i.e., the size of the neighborhood, variances of KNN are evaluated. Then, the Gamma distribution function is fitted to the empirical distributions of variances of these neighborhoods. The

Gamma distribution parameters are estimated using a maximum log-likelihood approach. The kernel bandwidth is then inferred from the resulting Gamma distribution function. The proposed methodology is applied to blind detection of modulated signals and to segment radar images of terrain. The bandwidth selection method proposed in this paper can be used to improve the efficiency of other computational-intelligence methods in a large variety of applications.

REFERENCES

- [1] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*. Boston, MA: McGraw-Hill, 2002.
- [2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2000.
- [3] N. Nasios and A. G. Bors, "Variational learning of Gaussian mixture models," *IEEE Trans. Syst., Man, Cybern. B: Cybern.*, vol. 36, no. 4, pp. 849–862, Aug. 2006.
- [4] M. A. Carreira-Perpinan, "Mode finding of Gaussian distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1318–1323, Nov. 2000.
- [5] A. G. Bors and I. Pitas, "Optical flow estimation and moving object segmentation based on median radial basis function network," *IEEE Trans. Image Process.*, vol. 7, no. 5, pp. 693–702, May 1998.
- [6] S. J. Roberts, "Parametric and nonparametric unsupervised cluster analysis," *Pattern Recognit.*, vol. 30, no. 2, pp. 261–272, Feb. 1997.
- [7] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman & Hall, 1986.
- [8] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Stat.*, vol. 27, no. 3, pp. 832–837, 1956.
- [9] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, 1962.
- [10] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Stat.*, vol. 36, no. 3, pp. 1049–1051, 1965.
- [11] H. Wang and D. Suter, "Robust adaptive-scale parametric model estimation for computer vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1459–1474, Nov. 2004.
- [12] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, no. 7, pp. 1151–1163, Jul. 2002.
- [13] D. Comaniciu, "An algorithm for data-driven bandwidth selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 281–288, Feb. 2003.
- [14] M. Girolami and C. He, "Probability density estimation from optimally condensed data samples," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1253–1264, Oct. 2003.
- [15] N. N. Schraudolph, "Gradient-based manipulation of nonparametric entropy estimates," *IEEE Trans. Neural Netw.*, vol. 15, no. 4, pp. 828–837, Jul. 2004.
- [16] S. J. Sheather, "Density estimation," *Stat. Sci.*, vol. 19, no. 4, pp. 588–597, 2004.
- [17] D. F. Specht, "Probabilistic neural networks," *Neural Netw.*, vol. 3, no. 1, pp. 109–118, 1990.
- [18] E. J. Pauwels and G. Frederix, "Finding salient regions in images: Nonparametric clustering for image segmentation and grouping," *Comput. Vis. Image Underst.*, vol. 75, no. 1/2, pp. 73–85, 1999.
- [19] P. Jia, H. Y. Zhang, and X. Z. Shi, "Blind source separation based on nonparametric density estimation," *Circuits Syst. Signal Process.*, vol. 22, no. 1, pp. 57–67, 2003.
- [20] V. Katkovnik, "A new method for varying adaptive bandwidth selection," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2567–2571, Sep. 1999.
- [21] B. W. Silverman, "Choosing the window width when estimating a density," *Biometrika*, vol. 65, no. 1, pp. 1–11, Apr. 1978.
- [22] G. R. Terrell, "The maximal smoothing principle in density estimation," *J. Amer. Stat. Assoc.*, vol. 85, no. 410, pp. 470–477, Jun. 1990.
- [23] C. L. Loader, "Bandwidth selection: Classical or plug-in?" *Ann. Stat.*, vol. 27, no. 2, pp. 415–438, 1999.
- [24] M. C. Jones, J. S. Marron, and S. J. Sheather, "A brief survey of bandwidth selection for density estimation," *J. Amer. Stat. Assoc.*, vol. 91, no. 433, pp. 401–407, 1996.
- [25] K. Fukunaga and L. D. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Inf. Theory*, vol. 21, no. 1, pp. 32–40, Jan. 1975.
- [26] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 790–799, Aug. 1995.
- [27] M. Blatt, S. Wiseman, and E. Domany, "Data clustering using a model granular magnet," *Neural Comput.*, vol. 9, no. 8, pp. 1805–1842, Nov. 1997.
- [28] D. Horn and A. Gottlieb, "Algorithm for data clustering in pattern recognition problems based on quantum mechanics," *Phys. Rev. Lett.*, vol. 88, no. 1, pp. 018 702-1–018 702-4, Jan. 7, 2002.
- [29] N. Nasios and A. G. Bors, "Kernel-based classification using quantum mechanics," *Pattern Recognit.*, vol. 40, no. 3, pp. 875–889, Mar. 2007.
- [30] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *J. R. Stat. Soc. Ser. B*, vol. 53, no. 3, pp. 683–690, 1991.
- [31] D. Ruppert, S. J. Sheather, and M. P. Wand, "An effective bandwidth selector for local least square regression," *J. Amer. Stat. Assoc.*, vol. 90, no. 432, pp. 1257–1270, 1995.
- [32] L. Yang and R. Tschernig, "Multivariate bandwidth selection for local linear regression," *J. R. Stat. Soc. Ser. B*, vol. 61, pt. 4, no. 4, pp. 793–815, 1999.
- [33] D. Comaniciu and P. Meer, "Distribution-free decomposition of multivariate data," *Pattern Anal. Appl.*, vol. 2, no. 1, pp. 22–30, Apr. 1999.
- [34] A. G. Bors, E. R. Hancock, and R. C. Wilson, "Terrain analysis using radar shape from shading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 974–992, Aug. 2003.
- [35] A. C. Cohen, "Estimating parameters of Pearson Type-III populations from truncated samples," *J. Amer. Stat. Assoc.*, vol. 45, no. 251, pp. 411–423, 1950.
- [36] S. C. Choi and R. Wette, "Maximum likelihood estimation of the parameters of the Gamma distribution and their bias," *Technometrics*, vol. 11, no. 4, pp. 683–690, 1969.
- [37] A. G. Bors and N. Nasios, "Kernel bandwidth estimation in methods based on probability density function modeling," in *Proc. Int. Conf. Pattern Recognit.*, Tampa, FL, Dec. 2008, pp. 3354–3357.
- [38] P. L. Worthington and E. R. Hancock, "New constraints on data closeness and needle map consistency for shape from shading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1250–1267, Dec. 1999.
- [39] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, "Separation of nonnegative mixture of nonnegative sources using a Bayesian approach and MCMC sampling," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4133–4145, Nov. 2006.
- [40] G. D. J. Smith and A. G. Bors, "Height estimation from vector fields of surface normals," in *Proc. Int. Conf. Digit. Signal Process.*, Santorini, Greece, 2002, vol. II, pp. 1031–1034.

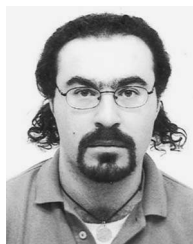


Adrian G. Bors (M'00–SM'04) received the M.S. degree in electronics engineering from the Polytechnic University of Bucharest, Bucharest, Romania, in 1992, and the Ph.D. degree in informatics from the University of Thessaloniki, Thessaloniki, Greece, in 1999.

From 1992 to 1993, he was a Research Scientist with the Signal Processing Laboratory, Tampere University of Technology, Tampere, Finland. From 1993 to 1999, he was a Research Associate with the University of Thessaloniki. In 1999, he joined the

Department of Computer Science, University of York, York, U.K., where he is currently a Lecturer. In 2006, he was a Visiting Scholar with the University of California, San Diego, and an Invited Professor with the University of Montpellier II, Montpellier, France. He is the author or a coauthor of 15 journal papers and more than 60 papers in international conference proceedings. His research interests include computational intelligence, computer vision, image processing, pattern recognition, digital watermarking, and nonlinear digital signal processing.

Dr. Bors has been an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS since 2001. He has also been a member of technical committees for several international conferences.



Nikolaos Nasios was born in Thessaloniki, Greece, in 1976. He received the Diploma degree from the National Technical University of Athens, Athens, Greece, in 2000 and the Ph.D. degree in computer science from the University of York, York, U.K., in 2006.

He is currently with the Department of Computer Science, University of York. His research interests include image processing and pattern recognition.

Dr. Nasios is a member of the Technical Chamber of Greece.