Chapter

Novel Formulation of Parzen Data Analysis

David Horn

Abstract

The Parzen analysis associates Gaussian kernels with each data point, thus obtaining a density function which may be viewed as a possible artificial generator of the data. This probability function can be decomposed into the product of two components, weight (W) and shape (S), which represent different aspects of the data. We demonstrate how this naturally leads to a formalism of fields in data space, which are interconnected through relations in one-dimensional scale space, corresponding to the common Gaussian width. We discuss the connection of this formalism to different clustering procedures such as quantum clustering (QC) and mean shift (MS). We demonstrate on various examples the importance of these concepts in the analysis of natural data as well as in image analysis in two or three dimensions.

Keywords: Parzen probability, weight-shape decomposition, quantum clustering, mean shift, image analysis

1. Introduction

Unsupervised machine learning has led to a wealth of clustering methods over the past few decades [1]. One of the important early ideas is that of the Parzen window distribution [2]. It has been introduced in 1962, as a kernel density estimate of a distribution function underlying measured data, and still serves as the basis of clustering algorithms in pattern recognition [1, 3]. Recently, it has been discovered [4] that the Parzen probability function can be decomposed into two components, weight and shape, which represent different aspects of the data. Weight, as its name implies, describes the semi-global strength of the distribution, whereas shape represents local properties which come to light once the bias of the weight is being removed. Moreover, −log(shape) coincides with a potential function V, which has been previously introduced in quantum clustering (QC) [5]. The cluster centers in QC correspond to minima of V. An alternative method, mean shift [6, 7], views the maxima of the probability function as the appropriate candidates of cluster centers. These two different points of view can now be studied and compared within a unified formalism [4].

Here we discuss the novel connections of the Parzen distribution to its potential and show how both can be used for the analysis of data points, leading to alternative clustering possibilities and extracting interesting features from the data. A particularly interesting set of applications appears in image analysis. Scale-space image analysis [8] has developed from the Parzen kernel methodology discussed in [6].

Now it turns out that insights from the potential term, or the shape component, allow for novel applications which are relevant to medical and technical imaging.

2. The Parzen probability distribution and its potential function

Data analysis often involves dimensionality reduction and noise removal, as well as some other tools, which eventually lead to consider a set of preprocessed data points located in a d-dimensional Euclidean space $\mathbf{x}_i \in \mathbb{R}^d$, with possible positive attributes (e.g., intensities) I_i. For this set, we define the non-normalized Parzen window function with Gaussian kernels as

$$\psi_\sigma(\mathbf{x}) = \sum_i I_i e^{-\frac{(\mathbf{x}-\mathbf{x}_i)^2}{2\sigma^2}} \tag{1}$$

Following [4], we introduce a relative probability weight function representing the influence of the kernel at data point $\mathbf{x}_i$ on any arbitrary point $\mathbf{x}$:

$$p_i(\mathbf{x}) = \frac{e^{-\frac{(\mathbf{x}-\mathbf{x}_i)^2}{2\sigma^2}}}{\psi(\mathbf{x})} \tag{2}$$

It obeys $\sum_i I_i p_i(\mathbf{x}) = 1$ and allows for the definition of two new scalar functions over data space $\mathbf{x}$, which are the potential and entropy fields

$$V(\mathbf{x}) = \sum_i I_i \frac{(\mathbf{x}-\mathbf{x}_i)^2}{2\sigma^2} p_i(\mathbf{x}) \tag{3}$$

$$H(\mathbf{x}) = -\sum_i I_i p_i(\boldsymbol{x}) \log p_i(\mathbf{x}) \tag{4}$$

Their difference is related to the Parzen probability function

$$V(\mathbf{x}) = \mathrm{H}(\mathbf{x}) - \log \psi(\mathbf{x}) \tag{5}$$

This can be rewritten as

$$\psi(\mathbf{x}) = W(\mathbf{x}) S(\mathbf{x}) \tag{6}$$

using [4] the concepts of weight and shape: $W(\mathbf{x}) = e^{H(\mathbf{x})}$ and $S(\mathbf{x}) = e^{-V(\mathbf{x})}$. Since $V(\mathbf{x}) \geq 0$, it follows that $S(\mathbf{x}) \leq 1$. Moreover, both W and S are nonnegative. S is integrable over $\mathbf{x}$ and, as such, can also serve as a distribution. From the definitions of (1) and (3), one can derive [4] the Schrödinger equation

$$-\frac{\sigma^2}{2}\nabla^2\psi(\mathbf{x}) + V(\mathbf{x})\psi(\mathbf{x}) = \frac{d}{2}\psi(\mathbf{x}), \tag{7}$$

which has been the cornerstone of the QC algorithm [5].

3. Interplay of scale and data space dependence

All the scalar fields over data space, introduced in the previous section, depend on the parameter σ, the scale of all Gaussian kernels. This dependence leads to further interesting relations between the Parzen probability function and its potential. Thus, from the definitions (1) and (3), it follows that

$$\frac{\sigma}{2}\frac{\partial}{\partial\sigma}\log\psi_\sigma(\mathbf{x}) = V_\sigma(\mathbf{x}) \tag{8}$$

where we keep the index σ which has been suppressed in the previous section. This relation displays a direct connection between the two scalar functions defining the probability and the potential. We proceed now to introduce a vector field $\mathbf{D}_\sigma$ which is defined by

$$-\nabla\log\psi_\sigma(\mathbf{x}) = \mathbf{D}_\sigma \tag{9}$$

and vanishes when the probability reaches its extrema in data space. Interestingly it is also related to the gradient of the potential function, through

$$\frac{-\sigma}{2}\frac{\partial}{\partial\sigma}\mathbf{D}_\sigma = \nabla V_\sigma \tag{10}$$

Hence we conclude that the potential reaches its extrema when $\mathbf{D}_\sigma$ remains stationary with respect to variations of σ.

$\mathbf{D}_\sigma$ may be expressed, in analogy with Eq. (3), as

$$\mathbf{D}(\mathbf{x}) = \sum_i I_i \frac{\mathbf{x}-\mathbf{x}_i}{\sigma^2} p_i(\mathbf{x}). \tag{11}$$

Its square $U = \mathbf{D}^2$ serves as an indicator function whose stationarity

$$\frac{\sigma}{2}\frac{\partial}{\partial\sigma}U_\sigma(\mathbf{x}) = 2\nabla\log\psi_\sigma(\mathbf{x})\cdot\nabla V_\sigma(\mathbf{x}) = 0 \tag{12}$$

implies the existence of extrema of either the probability or the potential. Since $U = \mathbf{D}^2$ is nonnegative, $U = 0$ is a minimum in σ. It corresponds to extrema of ψ which are associated with $\mathbf{D}=0$. Other values of U which obey Eq. (12) are associated with extrema of V which occur whenever $\frac{\partial}{\partial\sigma}\mathbf{D}_\sigma = 0$.Eq. (12) may be viewed as a statement concerning a set of points of interest in the data: all extrema of either the probability or the potential. In analogy with statistics, one may also view this equation as an inference method finding the parameter σ which leads to points of interest at given values of $\mathbf{x}$.

Although all extrema may be regarded as points of interest, some are of more interest than others: extrema that remain fixed in $\mathbf{x}$ for a range of scale values, which is large compared with the range of scales of other points of interest. This criterion, introduced by Roberts [9], allows searching for scales which correspond to natural properties of the data. Thus it subserves the search for good clustering of the data [4, 5, 9].

Finally, we wish to point out that ψ is not a properly normalized distribution function. A proper probability function, whose integral is 1, is defined by

$$P = \frac{1}{N}\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{d}{2}}\psi(\mathbf{x}) \tag{13}$$

where $N = \sum_i I_i$. We note that ψ and V obey a joint integration constraint [10]

$$\frac{1}{N}\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{d}{2}}\int d\mathbf{x}\psi(\mathbf{x})V(\mathbf{x}) = \frac{d}{2}. \tag{14}$$

This may be interpreted as a constraint on the expectation value of the potential function in data space.

Examples of the behavior of $\log\psi$ and of V are demonstrated in **Figure 1** for a data set of 9000 observed galaxies (with redshift in the domain 0.47 ± 0.005) regarded as points in spherical angles θ and φ within some limited range. Whereas for $\sigma = 2$ (in units of angle degrees), the two fields exhibit many extrema; there exist clear differences for larger sigma, for example, σ = 10, where $\log\psi$ has one maximum, while V displays several minima. This figure is taken from [10], a paper which contains a detailed and expanded formulation of the analysis presented in this section.

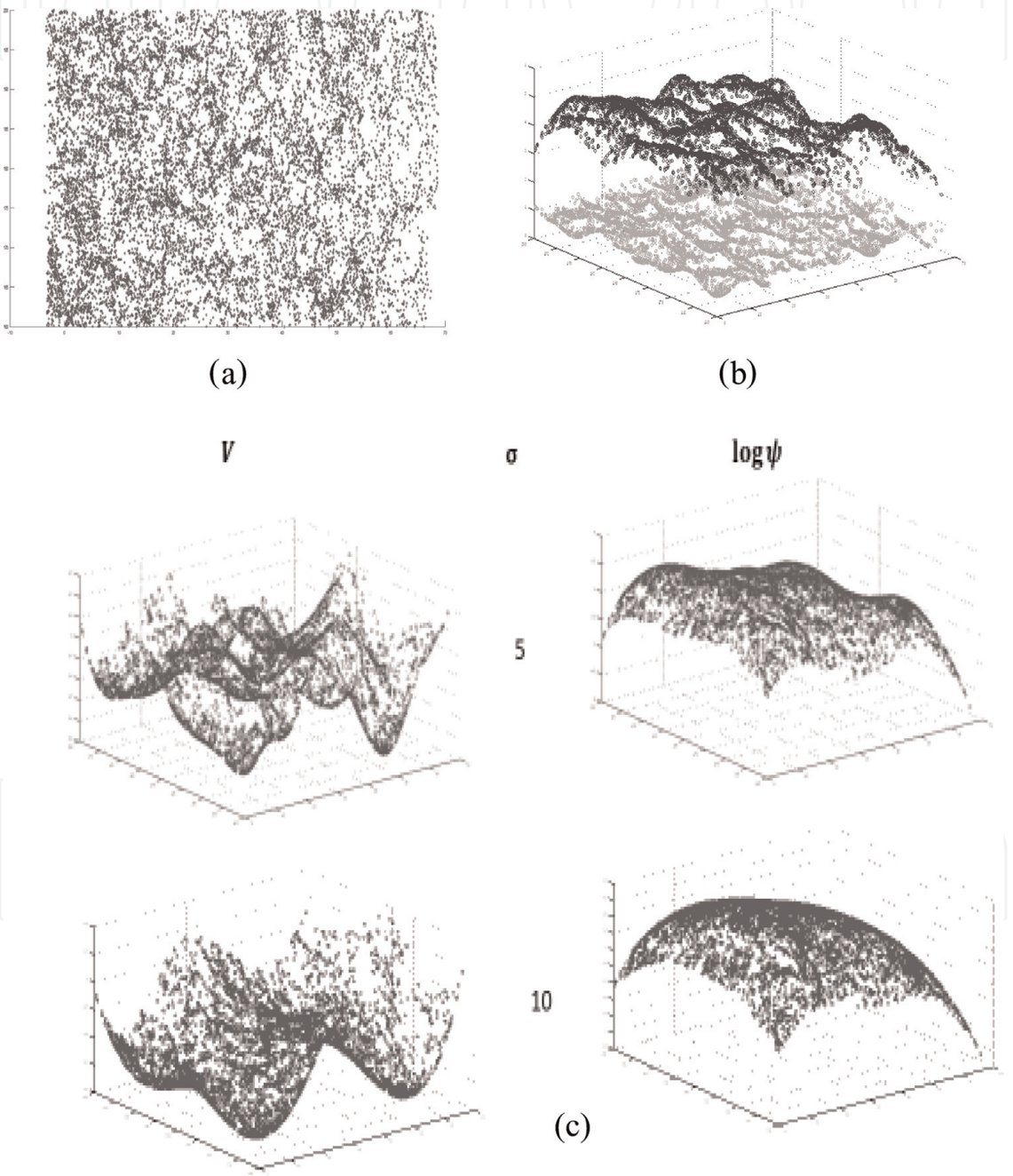

Figure 1.
(a) Loci of 9000 Galaxies, downloaded from the Sloan Digital Sky Server DR12, within some limited range of spherical angles. Reproduced from [10]. (b) log ψ (top) and V (bottom) displayed over the data plane 1a, using σ = 2 in spherical angle units. Reproduced from [10]. (c) Surfaces of V and log ψ for increased values of the Gaussian width. Reproduced from [10].

4. Applications to clustering, image analysis, and more

4.1 Anomalies

Figure 1 serves as an example of different behaviors of $\log\psi$ and of V. We wish to stress that when studying this system with large σ, the probability distribution is smooth; nonetheless underlying structure is observed in V. This means that, for a given σ, representing the large-scale behavior with one Gaussian, as one may be tempted to do after seeing the probability distribution, is wrong as demonstrated by the structures observed in V. On the other hand, the probability function tends to a smooth limit for σ = 10, whereas the fluctuating V changes with σ; hence V may represent random fluctuations in the data. However, comparing with the raw data in **Figure 1a**, we can be convinced that structure of the type discovered by V exists in the data. If these are fluctuations or not, one cannot tell from a single set of data.

A generally important question is if, within changing patterns of V, there exists one (or some) which remains relatively stable as function of σ. Such a structure may be viewed as a possible anomaly in the data. It is therefore advisable to study V when looking for anomalies.

4.2 Clustering

Clustering methodologies based on maximization of the probability and minimization of the potential can be defined by letting replica of data points move in these directions. These methods are known as mean shift (MS) and quantum clustering (QC) correspondingly. A recent review of MS techniques has been presented in [11]. Analyzing the same data with the same width-parameter σ leads to different clustering results for these two different methods, as is expected from **Figure 1**.

For illustration of clustering based on these different methods, we consider the crab data set which is included in Ripley's textbook [12]. It consists of 200 instances belonging to four equally sized classes and is defined in a five-dimensional parameter space. Performing PCA and restricting ourselves to the 2D plane defined by PC2-PC3 lead to a challenging clustering problem which has been discussed by [13], when introducing support vector clustering (SVC), and by [5] when introducing QC. It has been used in other papers employing variations of QC, such as the recent study [14]. Here we will show the results of [4] who applied to these data three clustering methods: Maximal Shape Clustering (MSC) which coincides with QC, Maximal Probability Clustering (MPC) which coincides with MS, and Maximal Entropy Clustering (MEC). The quality of all three methods may be judged by applying the Jaccard score

$$\mathbf{J} = \frac{\mathbf{n_{11}}}{\mathbf{n_{11}} + \mathbf{n_{10}} + \mathbf{n_{01}}}$$

where n_{11} is the number of pairs of points which belong together both in the same class (accepted as "ground truth") and in the same cluster, while $n_{10} + n_{01}$ are numbers of pairs which belong to the same class but different clusters and vice versa. This test is performed in **Figure 2a**, demonstrating that QC wins the competition for a wide range of σ values. The expected asymptotic value is J = 98/398, befitting one cluster and four classes.

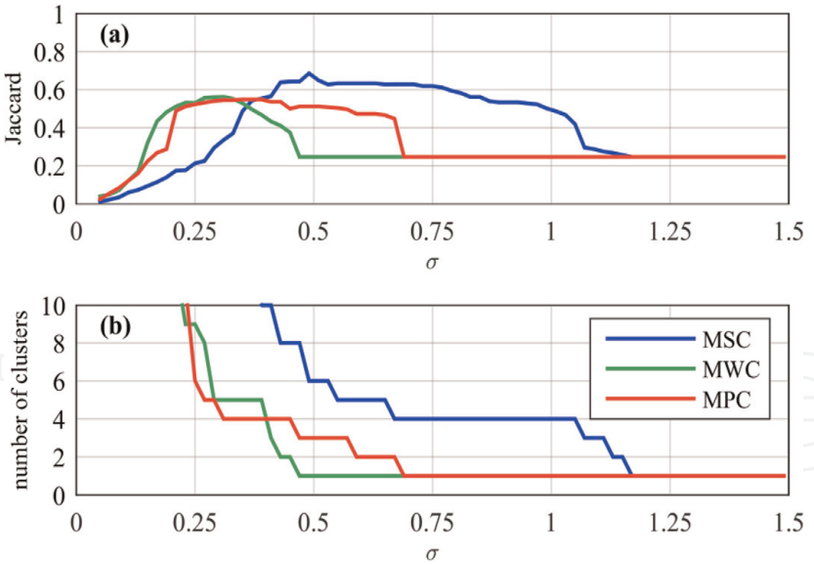

Figure 2.
(a) The Jaccard score, comparing clustering results with expert classification, comparing three clustering methods over a range of σ values. (b) The number of clusters, for each method and value of σ. Reproduced from [4]. MSC, MWC, and MPC stand for maximal shape, weight, and probability clustering accordingly. MSC coincides with quantum clustering and MPC with mean shift.

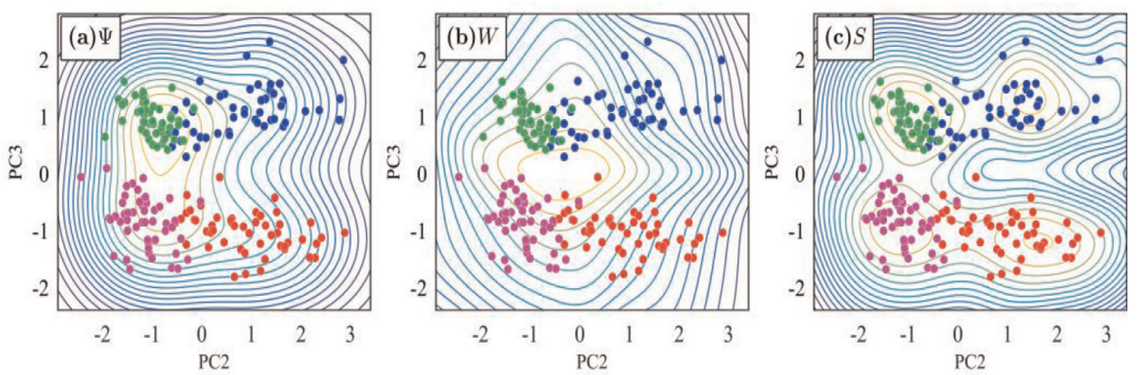

Figure 3.
Topographic maps of probability, weight, and shape, for σ = 0.7. Reproduced from [4].

Another comparison is being made in **Figure 2b**. This follows Roberts' criterion [9] that the preferable clustering method is the one which displays the most stable number of clusters with respect to variation of σ. This criterion is handy when the ground truth is unknown. QC excels also in this test, leading to a stable prediction of four clusters for a wide range of σ. This last figure also serve as a credibility test for Roberts' criterion.

In order to make the clustering results more intuitive, we display in **Figure 3**, also taken from [4], topographic maps of the different fields describing probability, weight, and shape, for σ = 0.7. The points in four different colors represent the four different classes. The topographic maps allow one to understand the clustering results which represent the outcome of gradient ascent applied to replica of data points which climb toward their nearest peak. Comparing the topologies of **Figure 3** with the results for σ = 0.7 in **Figure 2** leads to an understanding of why the three methods differ from each other.

4.3 Image analysis

A gray-scale image may be analyzed as a set of inputs associated with different pixels. In higher dimensional problems, such as 3D MRI data, the pixels are replaced

by voxels. Both may fit well into our analysis which starts with Eq. (1), associating a probability distribution with every image. One may then wonder if the weight-shape decomposition of Eq. (6)

$$\psi(\mathbf{x}) = W(\mathbf{x})S(\mathbf{x})$$

can lead to any novel understanding of image analysis.

In any practical application, non-normalized probability and weight may have very large amplitudes, yet shape will be limited to values ≤1. Nonetheless it carries some important characteristics:

1. Shape may be normalized to become a distribution on its own.
2. Shape serves as a generalized edge detector.
3. Shape is the basis of QC; hence the latter is very relevant to clustering of regions where shape is large. Alternative methods may be relevant when shape is small.

The first claim is trivial since S is limited to the range $0 \leq S \leq 1$, and the Gaussian kernels are integrable. The second property is a result of Eq. (7) which shows that the potential is related to the second derivative of the probability. It has led to an interesting result in [4], demonstrating that line caricatures of images can be produced by thresholded shape drawings.

To demonstrate the third point, we display in **Figure 4** the results of an analysis of a T2 MRI of the brain of a Macaque monkey [15]. Following the general procedure outlined above, and limiting ourselves to large relative values (thresholded distributions) of probability and shape, we find that the latter peaks in cortical regions, whereas the former peaks in internal regions of the brain, as demonstrated in **Figure 4**. Thus, a simple thresholding procedure allows one to easily segment the MR image, for the purpose of further analysis of the cortex by applying QC to the data in the large S domain. In **Figure 5**, we follow these conclusions [15] with a display of QC clusters projected onto the surface of the brain, leading to its

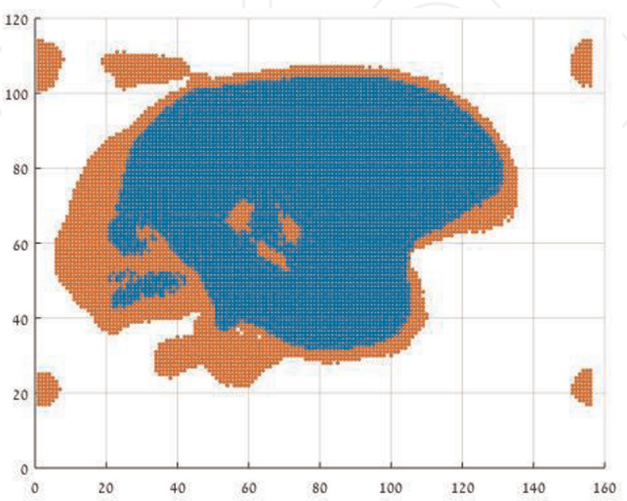

Figure 4.
Thresholded shape (red) and thresholded probability (blue) dominate different regions within the same MR image of a macaque brain, projected on its y-z plane. This analysis used σ = 3 in voxel units. Data outside the brain are due to artifacts and noise in the MR image. These results are due to [15], and they indicate that large shape components dominate cortical regions of the T2 MRI brain image.

parcellation into cortical components which are derived by just computational image analysis.

4.4 Convolutional representation of V

When one analyzes data in a regular underlying structure, such as pixels $\mathbf{m}$ of an image $I(\mathbf{m})$, the translational invariance of the Gaussian kernel allows one to use a convolutional description such as

$$\psi[\mathbf{m}] = \sum_{\mathbf{n}} I[\mathbf{n}]K[\mathbf{m}-\mathbf{n}] = I * K[\mathbf{m}] \tag{15}$$

with K being a discrete representation of the kernel. This leads [4] to the following result for the potential

$$V[\mathbf{m}] = \frac{I * L[\mathbf{m}]}{I * K[\mathbf{m}]} \tag{16}$$

where $L = -K \log K$. Such 3D kernels were applied to brain MR images [15] leading to the results displayed in **Figures 4** and **5**.

Noting that Eq. (15) is reminiscent of a convolutional layer in a deep network [16], we hypothesize that it can be useful to incorporate intermediate layers with nonlinear filters such as Eq. (16), as additional non-trained pooling filters in deep networks.

4.5 Computational remarks

The clustering methodology which has been employed in the different examples shown above is the simplest flavor of gradient descent (or ascent). It calculates the relevant fields on the basis of the data points and continues with straightforward dynamics that have been applied to replica of data points, seeking the extrema of the fields. Various alternatives to this basic application exist. The most important one is hierarchical clustering, which allows for conceptual simplicity and saves computational complexity. Such methodologies were described and discussed in [11] and in [4].

Computational complexity is an obvious issue when working with large data sets. Thus, 3D MRI data may easily comprise 1 M points, whereas their

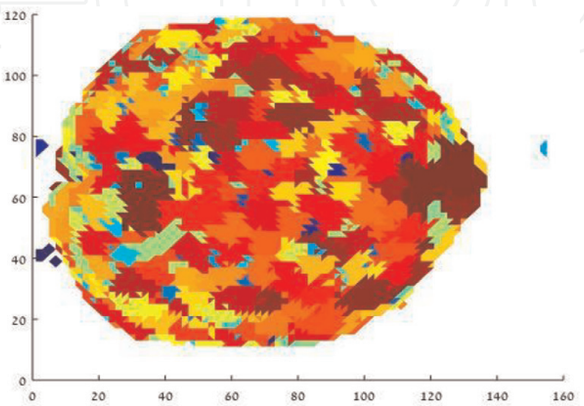

Figure 5.
Characteristic results of QC cortical clusters as mapped onto the surface of the brain and projected onto the x-y plane. This figure displays a map of the largest clusters of shape, each described by a different color. These results are due to Fisher [15].

manipulation within a system like MATLAB may well be limited to handling only 40 K points at a time [15]. One way to overcome such issues is to consider performing the analysis within extended voxels, for example, voxels containing three pixels in each direction in a 3D image problem. Within each new voxel, one may simply sum the intensity of points, leading to a new presentation of the data in the form of Eq. 1 on the smaller extended voxel space. Clearly one has to make sure that such an approximation does not harm interesting features of the data.

When analyzing other big data, no prior dimensional representation may be required. For the sake of noise reduction and computational complexity, it is advantageous to first apply relevant dimensional reduction, as provided, for example, by singular value decomposition (SVD) and principal component analysis (PCA). It is also important to make sure that the different axes are of similar scale, as shown in the example of **Figure 3**. When the data is still large, one may apply the trick of extended voxels described above. For very large data, one may also separate the data into several components, as is customary in supervised learning, to make sure that conclusions are not affected by the random choice of a subset of the data.

5. Conclusions

In the past (see, e.g., [3]), Parzen analysis has not considered the potential field V, which plays an important part in the understanding of different features of the data. In particular, V is sensitive to small changes in the Parzen probability by being related to its second derivative. It is also the basis of quantum clustering whose advantages have been demonstrated here as well as in many other investigations in the literature. The discovery of the weight-shape decomposition of the Parzen probability has led to a focus on shape and on the potential and allows for a meaningful comparative discussion of the different features which may be extracted from data.

Here we have defined a set of fields in data space which hopefully will turn out to serve as useful tools in future data analyses. They seem to be adequately applicable to image analysis, and we expect them to be particularly useful in biomedical and technical image analyses in three dimensions. When analyzing other data, where no visual display constraints exist, noise reduction and computational complexity call for preprocessing by dimensional reduction. Further reduction of computational complexity may be tried by employing extended voxels, which our technique can easily accommodate.

In summary, this extended Parzen method replaces any set of discreet data by a continuous set of fields in data space, with interrelations in scale space. It allows for investigating data properties in terms of these fields and their extrema.

Acknowledgements

This work has been partially supported by the Blavatnik Cyber Center of Tel Aviv University. I thank Itay Fisher for his help with numerical data analysis.

Conflict of interest

The author declares no conflict of interest.

Author details

David Horn
Sackler School of Physics and Astronomy, Tel Aviv University, Tel Aviv, Israel

*Address all correspondence to: horn@tau.ac.il

References

[1] Xu D, Tian Y. A comprehensive survey of clustering algorithms. Annals of Data Science. 2015;**2**(2):165-193

[2] Parzen E. On estimation of a probability density function and mode. The Annals of Mathematical Statistics. 1962;**33**(3):1065-1076

[3] Duda RO, Hart PE, Stork DG. Pattern Classification. 2nd ed. New York: Wiley-Interscience; 2001

[4] Deutsch L, Horn D. The weight-shape decomposition of density estimates: A framework for clustering and image analysis algorithms. Pattern Recognition. 2018;**81**:190-199

[5] Horn D, Gottlieb A. Algorithm for data clustering in pattern recognition problems based on quantum mechanics. Physical Review Letters. 2001;**88**(1): 018702

[6] Cheng Y. Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1995;**17**(8):790-799

[7] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002;**24**(5):603-619

[8] Lindeberg T. Scale-space theory: A basic tool for analyzing structures at different scales. Journal of Applied Statistics. 1994;**21**(1–2):225-270

[9] Roberts SJ. Parametric and non-parametric unsupervised cluster analysis. Pattern Recognition. 1997; **30**(2):261-272

[10] Horn D. Field Formulation of Parzen Data Analysis. arXiv eprint 1808.08776. 2018

[11] Carreira-Perpiñán MA. Clustering methods based on kernel density estimators: Mean-shift algorithms. In: Hennig C, Meila M, Murtagh F, Rocci R, editors. Invited chapter in Handbook of Cluster Analysis. Chapter 18. UK: CRC/ Chapman and Hall (Taylor & Francis Group); 2015. pp. 383-418

[12] Ripley BD. Pattern Recognition and Neural Networks. Cambridge, UK: Cambridge University Press; 1996

[13] Ben-Hur A, Horn D, Siegelmann HT, Vapnik V. Support vector clustering. Journal of Machine Learning Research. 2001;**2**(Dec):125-137

[14] Scott TC, Therani M, Wang XM. Data clustering with quantum mechanics. Mathematics. 2017;**5**:5. DOI: 10.3390/math5010005

[15] Fisher I. Parcellation of brain images using shape analysis [M.Sc. thesis]. Tel-Aviv University; 2018. http://horn.tau.ac.il/publications/FisherThesis.pdf

[16] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;**521**:436-444